A FAST SEARCH METHOD OF SPEAKER IDENTIFICATION FOR LARGE POPULATION USING PRE-SELECTION AND HIERARCHICAL MATCHING

Zhibin Pan 1)  Koji Kotani 1)  Tadahiro Ohmi 2)

1) Department of electronic engineering, graduate school of engineering, Tohoku University, Sendai, Japan
2) New industry creation hatchery center, Tohoku University, Sendai, Japan
Aza–aoba 05, Aramaki Aoba–ku, Sendai, 980–8579, Japan
E–mail: pzb@sse.ecei.tohoku.ac.jp

ABSTRACT

Performance of search during matching phase in a speaker identification system realized through vector quantization (VQ) is investigated in this paper. Voice of each person is recorded in an office room with personal computers. LPC–cepstrum is selected as feature vector. In order to gain higher success rate of identification, it is necessary to use larger size codebook for each person. Consequently, it is extremely time–consuming to do full matching directly with all of registered larger size codebooks for large population to determine who the current speaker is. To fulfill identification at a tolerable speed, trade–off between success rate and search time is unavoidable usually.

In this paper, a fast search method is proposed that is based on pre–selection and hierarchical matching so as to eliminate a large amount of impossible candidates before doing final fine matching with larger size codebooks meanwhile keep the success rate not degraded. Pre–selection is achieved using divergences between unknown input speech and all registered codebooks. Hierarchical matching is implemented using smaller size codebook and larger size codebook corresponding to each person respectively. With this method, for a population of 290, time required for identification can be reduced to about 13% while memory occupation is increased by just 6% when baselines are taken as those using larger size codebooks in conventional way. Success rate is not degraded as 94% comparing to that acquired in conventional way as well.

Keywords: fast search, large population, pre–selection, hierarchical matching

1. INTRODUCTION

Speaker identification is an approach utilizing person’s voice to identify who you are among a registered population. Speaker identification can be used to access control system and some other applications such as extracting or classifying person’s voice according to speaker’s information from huge amount of recorded sound media. In real applications, it is common to have a rather large population to be identified. For practical purpose, identifying speed and success rate is crucial.

Vector quantization (VQ) is a classical and an effective method for speaker identification [1]. In this method, each person has his own codebook to be registered into database and the amount of speaker–dependent codebooks increases with the persons to be identified linearly. Conventionally, it is a must to use larger size codebook in order to gain higher success rate of identification. For large population, however, it is extremely time–consuming to do full matching directly with all of large size codebooks to determine who the current speaker is.

To reduce computation complexity during identification phase dramatically, this paper proposes a fast search method based on pre–selection and hierarchical matching so as to eliminate a large amount of impossible candidates before doing final fine matching meanwhile keep the success rate of identification not degraded [2]. This method is consisted of 3 steps, namely, pre–selection, coarse matching and fine matching. Because the larger the distance of mean value between unknown input speech and a registered codebook or the smaller the deviation for each of them is the farther these two temples are away from, the ratio of these twoparameters can be adopted as a rough distance measure for pre–selection to determine some possible candidates. The codebooks corresponding to the ratio smaller than a certain threshold value are kept as possible candidates. After that, the distorted speech is used to determine who is this speaker among determined possible candidates, coarse matching with smaller size codebook is carried out. At this step, distortion measures are chosen as the mean, deviation of cepstrum distortion and S/N between unknown input speech and a registered codebook. Most possible candidates are the union set of codebooks appeared at top part of each sorted distortion list [3]. The number of codebooks to be reserved for fine matching should be determined by experiment.

Ultimately, within selected most possible candidates, fine matching with larger size codebook is run. It is realized in the same way as previous step except that decision by majority among 3 minimum distortions is used to determine who the current speaker is uniquely.

In our experiment for a population of 290, LPC–cepstrum [4] is selected as feature vector and Kohonen’s LVQ [5] is used to generate codebook. Smaller and larger codebook size corresponding to each speaker for hierarchical matching is 8 and 128, respectively.
At pre-selection step, computation complexity is quite low. It is just necessary to have about 30% codebooks of all to be kept averagely for next step. By experiment, it is guaranteed to have the real codebook in possible candidates. At coarse matching step in full search way, only 30 codebooks at top part of each sorted distortion list should be chosen as most possible candidates. Because computation cost is proportional to the codebook size, coarse matching is also rather quick. Then, for just a very limited part of most possible candidates extracted from all registered codebooks, fine matching is performed in conventional way [6]. Totally, in the case of 290 speakers, identification can be fulfilled in about 13% of time required by conventional way meanwhile success rate is not degraded as 94%.

2. IMPLEMENTATION METHOD OF PRE-SELECTION AND HIERARCHICAL MATCHING USING VQ

For the purpose of speaker identification, voice of each person is recorded in a office room with personal computers and digitized with 8 bit at 11 kHz. Background noise is reduced by exponential average in time domain. LPC-cepstrum is adopted as feature vector. By front end processing, each speech is converted into a sequence of cepstrum vector for the purpose of further registration or identification. During registration phase, Kohonen’s LVQ is used for a whole speech learning continuously to generate two codebooks of different size for each person.

Based on our previous work in conventional full matching method of VQ, for a population of 290, relation of success rate versus codebook size is shown in Figure 1. In this case, it is very time-consuming for larger size codebooks.

Figure 1: Success rate versus codebook size in conventional full matching method of VQ when population is 290

According to this result, it is clear that the larger the codebook size is, the higher the success rate of identification will be. Success rate of identification has a saturated trend. In order to gain a little improvement of success rate by using larger size codebook, it is usual to have to pay too much computation cost to direct matching. Because it is the ultimate purpose for speaker identification system to acquire a success rate as high as possible, it is necessary to use large size codebook to guarantee the success rate while using other methods to speed up the matching process. Speaker identification is essentially a process to find a nearest template among all registered persons by minimum distance. For large population, it implies that there does exit a certain amount of templates that are not near to the current unknown speech. If it is possible to find and abandon these templates by some simple and fast method before doing final matching to determine who the current speaker is uniquely, the identification process can be made fast and success rate is not degraded at all.

The architecture of speaker identification system using pre-selection and hierarchical matching method is demonstrated in Figure 2. It is consisted of 3 steps.

Figure 2: Block diagram of speaker identification system for large population using pre-selection and hierarchical matching. Refers to pre-selection, stands for coarse matching and represents fine matching

The current unknown input speech and each of codebooks registered in database can be considered as templates in vector space. To find the nearest template to the current unknown input speech, it is necessary to measure distance between them. The distance between centers of them (mean value) can be used as a measure but the widespread of each template (deviation) is not included in this way. Therefore, the effectiveness of this assessment is not so good.

In this paper, these two factors are combined into one distance measure as a rough classification criteria between templates. It is called divergence as defined below,

\[
D_{iv} = \frac{\sum_{i=1}^{k} (m_{i2} - m_{i1})^2}{\sum_{i=1}^{k} \sigma_{i1}^2 + \sum_{i=1}^{k} \sigma_{i2}^2}
\]  

(1)

where \(m_{i1}, m_{i2}\) and \(\sigma_{i1}, \sigma_{i2}\) are elements of mean vector and deviation vector of template 1 and template 2 respectively. \(k\) is the dimension of each vector.

In order to obtain mean vector and deviation vector of current unknown input speech, they are computed twice to eliminate outliers. Any input vector beyond mean vector ± 2\(s\) deviation vector is discarded as an outlier. Because the number of vectors in a codebook is small, it is necessary to compute unbiased mean vector and deviation vector for each codebook and store in a database.

To identify an unknown person, firstly, divergence between this current unknown input speech and each of codebooks registered in database is computed. Since it is computed in a scalar way, the computation complexity is extremely low. Then all divergences are sorted in an ascending order and codebooks corresponding to divergences less than a threshold value is kept...
as possible candidates for succeed coarse matching. The number of possible candidates kept should be determined by experiment. Suppose d terms is sufficient under the condition of all real codebooks are preserved at this step.

At hierarchical matching steps, suppose memory occupation and time consumption in conventional full matching VQ method be the baselines for comparison when codebook size is the larger one. With this novel approach, it is necessary to let a speech that is used to generate a codebook learn twice. One is for smaller size codebook (cb1) and the other is for larger size codebook (cb2). Then the increase of memory occupation for registration is

\[ \text{ratio}_m = \frac{\text{cb1}}{\text{cb2}} \]  

where ratio\(_m\) is relative increment of memory.

The mean value of VQ distortion over all frames for a whole piece of speech is considered as the measure of similarity between an unknown input speech and a registered codebook as shown in (3).

\[ \mu = \frac{1}{L} \sum_{m=1}^{L} \sum_{n=1}^{k} (c_{x,n} - c_{cb,n})^2 \]  

where \((c_{x,n}), (c_{cb,n})\) is the \(n\)th element of cepstrum vector for input speech and winner codevector, respectively. \(k\) is the number of elements in a vector. \(L\) is the number of total frames in an unknown input speech.

In this paper, not only mean value of VQ distortion but also standard deviation of VQ distortion centered at \(\mu\) and S/N are used simultaneously to identify a person as given in (4).

\[ \sigma = \mu + \kappa \times \sigma \]

\[ = \mu + \kappa \times \left( \frac{1}{L} \sum_{m=1}^{L} \sum_{n=1}^{k} (c_{x,n} - c_{cb,n})^4 - \mu^2 \right) \]  

\[ S/N = \frac{1}{L} \sum_{m=1}^{L} \log_{10} \left( \frac{\sum_{n=1}^{k} (c_{x,n})^2 / \sum_{n=1}^{k} (c_{cb,n})^2}{\mu} \right) \]

means how far the centroid of an unknown input speech is from that of each codebook and \(\mu\) to what extent the distribution of an unknown input speech is mismatched with that of each codebook. S/N describes relative distortion between an unknown input speech and each codebook in power.

At coarse matching step, mean and deviation of distortions are sorted in ascending order while S/N is sorted in descending order. Most possible candidates are selected as the union set of codebooks appeared at top part of each sorted distortion list. The number of codebooks to be reserved at this step for succeed fine matching should be determined by experiment as well.

Assume population registered in database be \(p\) and totally \(k\) terms at top of sorted list of distortions be kept as most possible candidates after coarse matching in order to do fine matching furthermore. Because at fine matching step, all \(k\) terms chosen from previous coarse matching must match with larger size codebook in conventional way to decide which speaker the current piece of speech belongs to, decision by majority method is utilized among minimum value of \(\mu\), \(\sigma\), and maximum value of S/N.

Then the decrease of time consumption for whole identification is

\[ \text{ratio} = 1 - \left( \frac{d \times \text{cb1}}{p \times \text{cb2}} - \frac{k}{p} \right) \]  

where ratio\(_t\) is relative decline of time consumption.

The aim is to make ratio\(_m\) as small and ratio\(_t\) as large as possible so that speaker recognition system can be realized at an acceptable speed with least memory cost.

### 3. EXPERIMENT

The experiment is implemented under conditions summarized in Table1 for the same 290 speeches as those used in conventional full matching method.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling</td>
<td>11.025 kHz, 8 bit</td>
</tr>
<tr>
<td>Low pass filtering</td>
<td>3 kHz</td>
</tr>
<tr>
<td>Pre-emphasis</td>
<td>0.96</td>
</tr>
<tr>
<td>Length of frame</td>
<td>30 ms</td>
</tr>
<tr>
<td>Shift of frame</td>
<td>10 ms</td>
</tr>
<tr>
<td>Window function</td>
<td>Hamming</td>
</tr>
<tr>
<td>LPC order</td>
<td>16</td>
</tr>
<tr>
<td>Cepstrum number</td>
<td>21</td>
</tr>
<tr>
<td>Size of codebook for coarse</td>
<td>8</td>
</tr>
<tr>
<td>Size of codebook for fine</td>
<td>128</td>
</tr>
</tbody>
</table>

Table 1: Conditions for speaker identification experiment with pre-selection and hierarchical matching method.

At pre-selection step, divergences between unknown input speech and each of codebooks are computed and sorted in ascending order. For each of all 290 speeches, the threshold value of divergence varies dramatically. There does not exist a unique threshold value for all of unknown input speech. In this experiment, within the sorted list of divergences for each of unknown input speech, the mean value and standard deviation of divergences are computed and then the threshold value of divergence is determined as the mean value minus 0.5 times of standard deviation by experiment. That is to say codebooks corresponding to divergence less than the threshold value determined above are kept for succeed coarse matching. For all 290 speeches, Experiment result is illustrated in Figure2.

Figure 3: Occurrence frequency for divergence less than the threshold value that is corresponding to real codebook versus its sequence in sorted divergence list.
Under these conditions, it is clear that to guarantee real codebook be kept for succeed coarse matching, the maximum number of codebooks to be selected is 101 at which occurrence frequency is 98.5%. In this case, possible candidates are about 30% of total registered persons. The codebooks corresponding to possible candidates are labeled with a flag 1 and the others with a flag 0.

At coarse matching step, identification experiment using candidate codebooks selected previously with size 8 is implemented. Because each person has low different size codebooks, the increase of memory occupation for registration is 8/128=6% after Eq.(2). Now, VQ is run for unknown input speech over all selected codebooks to find the distortions between them using 1 or 0 flag list obtained in pre–selection. When coarse matching is over, all three distortions (i.e., and S/N) due to dissimilarity between the unknown input speech and all registered codebooks are sorted in ascending order for , and descending order for S/N. At this step, unique decision can not be made to determine who the current speaker is because success rate is too low. But some ones as most possible candidates can be determined. Therefore within these three sorted distortion lists, top several or several tens codebooks are extracted as most possible candidates for further fine matching.

The codebooks corresponding to most possible candidates are also labeled with a flag 1 and the others with a flag 0.

At fine matching step, using 1 or 0 flag list obtained in coarse matching, VQ is progressed for unknown input speech over all selected codebooks with size 128. At this time, only a very limited part of registered codebooks labeled by 1 is used for fine matching so that computation cost or time consumption can be reduced depending on the number of selected most possible candidates. After fine matching, distortions based on larger codebook size are sorted once again. Among these sorted distortion lists, by minimum three distortions at position top 1, decision by majority can be made uniquely. The current speaker is recognized as the person at position top 1 in sorted distortion lists.

It is clear that if top terms kept during coarse matching phase for possible candidates are too few, some of possible candidates would be lost and it is impossible to pick them up during fine matching phase so as to make correct identification afterwards. On the other hand, if top terms kept are too much, computation cost will be too high.

If possible candidates have not been discarded during coarse matching phase, correct identification certainly can be made during fine matching phase. In this way, success rate corresponding to codebook size being 128 can be realized while computation cost is just that of matching with possible candidates of codebook size 8 and that of matching with most possible candidates of codebook size 128.

The experiment result for hierarchical matching is demonstrated in Figure 4 when top terms kept during coarse matching phase varies.

Figure 4: Success rate versus top terms kept during coarse matching for fine matching

From Figure 4, it can be seen that when top terms kept reaches 30, success rate saturates. Under the conditions depicted above, in pre–selection and hierarchical matching method, the decrease of time consumption for identification is 1−(101/290*8/128 + 30/290)=87.5% after Eq.(5) while the final success rate is 94%, the same as that achieved in conventional full matching method.

4. CONCLUSIONS

A fast search method is investigated using pre–selection and hierarchical matching. In this method, each person is registered into 3 different databases simultaneously with various representations. That is to say the mean vector, deviation vector, smaller size codebook and larger size codebook corresponding to him. Identification is progressed step by step and at each step a rather large amount of impossible candidates are eliminated so as to reduce the burden of final matching with large size codebook greatly while keeping success rate not degraded. This method is evaluated using a collection of 290 speeches recorded in ordinary office room. With the cost of only 6% increase of memory occupation for registration, time needed for identification can be cut down dramatically to 13% while the success rate keeps the same as 94% comparing to performances obtained in conventional method.

5. REFERENCES