Ambiguity reduction in speaker identification by the relaxation labeling process

Tuan Pham*, Michael Wagner

Faculty of Information Sciences and Engineering, University of Canberra, ACT 2601, Australia

Received 25 November 1997; received in revised form 28 August 1998; accepted 28 August 1998

Abstract

A nonlinear probabilistic model of the relaxation labeling (RL) process is implemented in the speaker identification task in order to disambiguate the labeling of the speech feature vectors. In this proposed algorithm, the deterministic labeling of the vector quantization (VQ)-based speaker identification is relaxed by means of introducing initial probabilistic weights to the labeling process of the speech feature vectors. This process is then iteratively updated until no further significant improvement is found. Experimental results on speaker identification using a commercial speech corpus show that the relaxation labeling outperforms the conventional VQ method. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Relaxation labeling; Nonlinear probabilistic model; Speaker identification; Speaker-based VQ codebook

1. Introduction

Speaker recognition is one of the challenging areas of speech research and has many applications including telecommunications, robotics, security systems, database management, command and control, and others. Speaker recognition is a generic term which refers to the classification of speakers based on their speech characteristics. This general task can be subdivided into two categories: speaker identification and speaker verification. Speaker identification is to assign an unknown or unlabeled voice token of a speaker to one of the reference speakers based on the closet measure of similarity, whereas speaker verification is aimed to either accept or reject the identity of an unlabeled voice token claiming to belonging to a specific reference speaker. The key difference between speaker identification and speaker verification is the number of decision alternatives. The number of decision in speaker identification is equal to the population size, and therefore its performance decreases with increasing population size. Speaker verification, however, is unaffected by the population size, as there are only two possible decisions based on an acceptance/rejection threshold. Both identification and verification categories are also involved in text-dependent and text-independent speech tokens. In a text-dependent speaker recognition, the speech tokens used for both training and testing have the same text; whereas in a text-independent mode, speech tokens in both training and testing do not subject to a specific text. Therefore, speaker recognition in a text-independent mode is more difficult than that in a text-dependent mode since the recognizer must deal with more variability in the unlabeled and the reference tokens. In general, speaker verification task is suitable for most applications due to its tractability. However, speaker identification remains active in research as it is suitable for evaluating the performance of different recognizers [1–4].
There are a number of techniques for speaker identification such as the speaker-based VQ codebook approach, dynamic time warping, second-order statistical method, discrete hidden Markov models, neural networks, and others [2–3,5–7]. Among these, the VQ codebook approach is one of the most popular methods implemented in many speaker recognition systems as it relieves the computational complexity and offers good performance. While the VQ approach has been commonly used for speech and speaker recognition, it is not always effective because the ambiguity inherently existing in the labeling of speech input tokens is treated in an inflexible way by its deterministic rules. Basing our motivation on this reason, we propose an improved algorithm over the speaker-based VQ codebook approach using the relaxation labeling in which the deterministic classification of the VQ-based approach is only an initial process of the probabilistic labeling. Results from this initial labeling will then be updated until convergence is reached. Relaxation labeling was first introduced by Rosenfeld et al. [8] to solve problems in pattern recognition. It is a parallel algorithm that updates the probabilities of labels or classes by using interactive information between unknown objects with respect to the reference labels to reduce uncertainty among labels having interchanging properties. Since being first developed to tackle problems in image analysis, its flexible framework has attracted many researchers in the broad field of pattern recognition [9–13].

In this paper we apply the framework of relaxation labeling to dealing with the VQ-based text-dependent speaker identification. The rest of this paper is organized as follows. In Section 2 we briefly review the standard speaker-based VQ codebook approach, then describe the procedures of relaxation labeling and how the task of speaker identification can be modeled with its algorithm in Sections 3 and 4, respectively. Experiments are given in Section 5 to compare the performance of the proposed approach with the conventional VQ-based method. Finally, conclusions about the present work and suggestion for further investigation are made in Section 6.

2. Speaker-based VQ codebook approach

For speaker identification based on VQ codebook approach [4], the codebook for each speaker is generated by partitioning a set of training feature vectors \( \{v_1, v_2, \ldots, v_L\} \) into the feature vector space \( \{S_1, S_2, \ldots, S_M\} \), and each partition \( S_j \) is represented by a centroid vector \( b_j \). Thus, there are \( N \) codebooks generated for \( N \) speakers. In the testing phase, the distortion between a set of testing feature vectors \( \{v_1, v_2, \ldots, v_T\} \) and each codebook is to be measured, then an average distortion \( D_i \) to the \( i \)-th codebook is taken, that is

\[
D_i = \frac{1}{L} \sum_{j=1}^{L} d(v_i, b_j), \quad j = 1, 2, \ldots, J,
\]

where \( d(v_i, b_j) \) is the distortion measure (usually a Euclidean distance) between two vector \( v_i \) and \( b_j \).

The recognized speaker \( i^* \) is then decided by taking the minimum of the \( N \) resultant average distortion measures:

\[
i^* = \arg \min_{i} D_i, \quad i = 1, 2, \ldots, N.
\]

3. Relaxation Labeling

The relaxation labeling algorithms consist of four models when first introduced by Rosenfeld et al. [8] – discrete, fuzzy, linear probabilistic, and nonlinear probabilistic models. However, the last model offers the best performance for classification problems and due to this reason, only the nonprobabilistic relaxation model is studied here for the speaker identification. Its algorithm is described as follows.

Let a set of objects \( A = \{a_1, a_2, \ldots, a_n\} \) and a set of labels \( \Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\} \). An initial probability is given to each object \( a_i \) having each label \( \lambda \), which is denoted as \( p_i(\lambda) \). These probabilities satisfy the following condition:

\[
\sum_{\lambda \in \Lambda} p_i(\lambda) = 1, \quad \forall a_i \in A, \quad 0 \leq p_i(\lambda) \leq 1.
\]

The relaxation labeling updates the probabilities \( p_i(\lambda) \) in Eq. (3) using a set of compatibility coefficients \( r_{ij}(\lambda, \lambda') \), where \( r_{ij}(\lambda, \lambda') : \Lambda \times \Lambda \mapsto [-1, 1] \), whose magnitude denotes the strength of compatibility. The meaning of these compatibility coefficients can be interpreted as follows:

\[
r_{ij}(\lambda, \lambda') \begin{cases} < 0: & \lambda, \lambda' \text{ are incompatible for } a_i \text{ and } a_j, \\
0: & \lambda, \lambda' \text{ are independent for } a_i \text{ and } a_j, \\
> 0: & \lambda, \lambda' \text{ are compatible for } a_i \text{ and } a_j,
\end{cases}
\]

The updating factor for the estimate \( p_i(\lambda) \) at \( k \)-th iteration is

\[
q_i^{k+1}(\lambda) = \sum_{j} d_{ij} \left[ \sum_{\lambda'} r_{ij}(\lambda, \lambda') p_j^{k}(\lambda') \right],
\]

where \( d_{ij} \) are the parameters that weight the contributions to \( a_i \) coming from its neighbors \( a_j \), and subject to

\[
\sum_{j} d_{ij} = 1.
\]

The updated probability \( p_i^{k+1}(\lambda) \) for object \( a_i \) is given by

\[
p_i^{k+1}(\lambda) = \frac{p_i^{k}(\lambda) [1 + q_i^{k}(\lambda)]}{\sum_{\lambda'} p_j^{k}(\lambda') [1 + q_j^{k}(\lambda'])}.
\]
Thus, the iterative process given by Eqs. (5) and (7) establish the relaxation labeling, and it is stopped when convergence is achieved. It now becomes clear that for a successful performance of the relaxation process, the initial label probabilities and the compatibility coefficients need to be well determined. Wrong estimates of these parameters will lead to algorithmic instabilities.

Two possible methods for computing the compatibility coefficients are based on those developed by Peleg and Rosenfeld [11]. The two methods employ the concepts of statistical correlation and mutual information. The correlation-based estimate of the compatibility coefficients is defined by

$$ r_{ij}(\lambda, \lambda') = \frac{\sum [p_i(\lambda) - \bar{p}(\lambda)][p_j(\lambda') - \bar{p}(\lambda')]}{\sigma(\lambda)\sigma(\lambda')}, $$

where $p_j(\lambda')$ is the probability of $a_j$ having label $\lambda'$, and $a_i$ be the neighbors of $a_i$, $\bar{p}(\lambda')$ is the mean of $p_j(\lambda')$ for all $a_j$, and $\sigma(\lambda')$ is the standard deviation of $p_j(\lambda')$. To alleviate the effect of dominance among labels, the modified coefficients are

$$ r_{ij}(\lambda, \lambda') = [1 - \bar{p}(\lambda)][1 - \bar{p}(\lambda')] r_{ij}(\lambda, \lambda'). $$

The mutual-information-based estimate of the compatibility coefficients is

$$ r_{ij}(\lambda, \lambda') = \log \frac{n \sum p_i(\lambda) p_j(\lambda')}{\sum p_i(\lambda) \sum p_j(\lambda')}, $$

where, for the present problem, $n$ is the number of feature vectors of an unknown speaker. The compatibility coefficients in Eq. (10) are divided by 5 in order to take values in the range $[-1, 1]$.

4. Relaxation labeling for speaker identification

For the classification of speech samples from an unknown speaker to the best fit out of a population of speakers, some sets of feature vectors characterizing the variabilities of different speakers are likely to overlap; therefore, in the spirit of relaxation labeling, each feature vector is considered as an object $a_i \in A$ where $A$ is a set of feature vectors of a speaker, and each speaker is considered as a label $\lambda$ in the speaker population $A$. We will discuss how to estimate the initial probabilities, how to effectively implement the relaxation labeling process by Rosenfeld et al. [8] for the problem of speaker identification, and we also outline this proposed algorithm in the following subsections.

4.1. Estimation of initial probabilities

Using the VQ distortion measures, the initial probability that expresses the local measurement of a vector $a_i$ belonging to a speaker $\lambda$ can be estimated as

$$ p_i(\lambda) = \frac{\exp(-D_{i\lambda})}{\sum_a \exp(-D_{i\lambda})} $$

in which $D_{i\lambda}$ is the minimum distortion measure between $a_i$ and the set of codewords of speaker $\lambda$, that is

$$ D_{i\lambda} = \min_k [D(a_i, b_k(\lambda))]. \quad k = 1, 2, \ldots, K, $$

where $K$ is the codebook size.

4.2. Implementation of the RL process for speaker identification

In the case of image analysis, it is important to consider the confidence contributions from pixels lying in the neighborhood of a pixel, as its $m$-connected neighboring pixels may belong to different regions. Therefore, the resulting weight of the pixel is strongly affected by the confidence weights of its neighborhood. However, for speech analysis, it is known that the set of speech feature vectors $\{a_i\}$ (each vector $a_i$ is equivalent to an image pixel) is to belong to a certain speaker $\lambda$. Therefore, there is no need to consider the contributions of its adjacent vectors. On the other hand, we only consider the compatibility of the vector $\{a_i\}$ itself with respect to speaker $\lambda$ and speaker $\lambda'$. With this argument, the original compatibility coefficients as defined in Eq. (4) can be expressed in another form as

$$ r_{ij}(\lambda, \lambda') = r_{ij}(\lambda, \lambda') \begin{cases} <0: \lambda, \lambda' \text{ are incompatible for } a_i, \\ =0: \lambda, \lambda' \text{ are independent for } a_i, \\ >0: \lambda, \lambda' \text{ are compatible for } a_i, \end{cases} $$

Following the above reason, the updating factor for the estimate $p_i(\lambda)$ at $k$th iteration is rewritten as follows:

$$ q_{ik}^{(k)}(\lambda) = \sum_{\lambda'} r_{ik}(\lambda, \lambda') p_i^{(k)}(\lambda'), $$

where the summation of $d_{ij}$ as defined in Eq. (5) is now omitted as the contributions from the adjacent vectors are not considered.

We assume that the majority of the speech feature vectors $a_i$ well belong to the speaker $\lambda$, i.e. the amount of the feature vectors having overlapping properties is less than that of the feature vectors having more distinctive properties. If this assumption is true, then the compatibility coefficients $r_{ij}(\lambda, \lambda')$ tend to be negative as $\lambda$ and $\lambda'$ are incompatible for $\{a_i\}$. This also leads to a negative value for the updating factor $q_{ik}^{(k)}(\lambda)$ in Eq. (13), which is defined in terms of the compatibility coefficients. From this standpoint, if Eq. (7) is used for updating the probability, then the confidence for a distinctive or overlapping vector $a_i$ belonging to the speaker $\lambda$ will be decreased or
increased instead of being increased or decreased, respectively. Therefore, the plus sign in Eq. (7) should become a minus sign, that is

$$p_i^{k+1}(\lambda) = \frac{p_i^k(\lambda) [1 - q_i^k(\lambda)]}{\sum_j p_j^k(\lambda) [1 - q_j^k(\lambda)]}.$$  \hfill (14)

We rewrite the computations of the compatibility coefficients according to Eq. (12) as follows. For the correlation-based estimate of the compatibility coefficients:

$$r_i(\lambda, \lambda') = \sum_j [p_i(\lambda) - \hat{p}(\lambda)] [p_j(\lambda') - \hat{p}(\lambda')] \sigma(\lambda) \sigma(\lambda'),$$  \hfill (15)

where $p_i(\lambda')$ is the probability of $a_i$ having label $\lambda'$, $\hat{p}(\lambda')$ is the mean of $p_i(\lambda')$ for all $a_i$, and $\sigma(\lambda')$ is the standard deviation of $p_i(\lambda')$. And the modified coefficients becomes

$$r_i(\lambda, \lambda') = [1 - \hat{p}(\lambda)] [1 - \hat{p}(\lambda')] r_i(\lambda, \lambda').$$  \hfill (16)

Finally, the mutual-information-based estimate of the compatibility coefficients is now rewritten as

$$r_i(\lambda, \lambda') = \log \frac{n \sum_j p_i(\lambda) p_j(\lambda')}{\sum_i p_i(\lambda) \sum_j p_j(\lambda')}.$$  \hfill (17)

### 4.3. Outline of relaxation labeling algorithm for speaker identification problem

The relaxation labeling for speaker identification can be outlined as follows:

1. Estimate the initial probabilities for each unknown feature vector using Eq. (11).
2. Compute the compatibility coefficients using either Eq. (16) or Eq. (17).
3. Calculate the updating factor defined in Eq. (13).
4. Update the probabilities for each unknown vector using Eq. (14).
5. Repeat steps 3 and 4 until the change of the probabilities is less than a prescribed threshold or equal to a prescribed number of iterations.
6. The total probability of the set of vectors for each speaker $\lambda$ is $P(\lambda) = \prod_i p_i(\lambda)$.
7. Deterministic identification for the best guessed speaker $\lambda^*$ can now be carried out using the following classification rule:

$$\lambda^* = \arg \max_{\lambda} P(\lambda).$$

### 5. Experiments on speaker identification

Both VQ codebook approach and relaxation labeling (RL) are simulated and tested with a set of computer commands from the T146 speech data corpus. The T146 corpus contains 46 utterances spoken repeatedly by eight female and eight male speakers, labeled f1–f8 and m1–m8, respectively. The vocabulary contains a set of 10 computer commands: {enter, erase, go, help, no, rubout, repeat, stop, start, yes}. Each speaker repeated the words 10 times in a single training session, and then again twice in each of eight testing sessions. The corpus is sampled at 12,500 samples/s and 12 bits/sample. The data were processed in 20.48 ms frames at a frame rate at 125 frames/s. The frames were Hamming windowed and preemphasized with $\mu = 0.9$. 46 mel-spectral bands of a width of 110 mel and 20 mel-frequency cepstral coefficients (MFCC) were determined for each frame. In the training session, using the LBG algorithm [14], each speaker’s 100 training tokens (10 utterances × 1 training session × 10 repetitions) were used to train the speaker-based VQ codebook by clustering the set of all speakers’ MFCC into codebooks of 32, 64 and 128 codewords.

The speaker identification was tested in the text-dependent mode. Each speaker’s 160 test tokens (10 utterances × 8 testing sessions × 2 repetitions) were tested against all speakers’ 10-word models. For the codebook of 32 entries, the average recognition rates for speaker identification are shown in Table 1, where the total average rate for: VQ = 84.98%, RL1 = 91.55% (RL using correlation-based compatibility coefficients be denoted as RL1), and RL2 = 91.98% (RL using mutual-information-based compatibility coefficients be denoted as RL2). For the codebook of 64 entries, the average recognition rates for speaker identification are (Table 2): VQ = 89.00%, RL1 = 94.03%, and RL2 = 94.26%. Finally, for the codebook of 128 entries, the average recognition rates for speaker identification are (Table 3): VQ = 91.28 %, RL1 = 96.02 %, and RL2 = 96.65 %.
Table 2
Speaker identification rates (%) by VQ approach and RL algorithms with codebook size of 64

<table>
<thead>
<tr>
<th>Speaker</th>
<th>VQ</th>
<th>RL1</th>
<th>RL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>95.62</td>
<td>97.50</td>
<td>96.25</td>
</tr>
<tr>
<td>f2</td>
<td>100</td>
<td>99.38</td>
<td>98.75</td>
</tr>
<tr>
<td>f3</td>
<td>91.25</td>
<td>88.12</td>
<td>86.25</td>
</tr>
<tr>
<td>f4</td>
<td>99.38</td>
<td>99.38</td>
<td>99.38</td>
</tr>
<tr>
<td>f5</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>f6</td>
<td>100</td>
<td>100</td>
<td>99.38</td>
</tr>
<tr>
<td>f7</td>
<td>96.25</td>
<td>96.25</td>
<td>97.50</td>
</tr>
<tr>
<td>f8</td>
<td>98.75</td>
<td>100</td>
<td>99.38</td>
</tr>
<tr>
<td>m1</td>
<td>76.97</td>
<td>92.76</td>
<td>92.76</td>
</tr>
<tr>
<td>m2</td>
<td>88.75</td>
<td>96.25</td>
<td>97.50</td>
</tr>
<tr>
<td>m3</td>
<td>99.36</td>
<td>100</td>
<td>99.36</td>
</tr>
<tr>
<td>m4</td>
<td>98.06</td>
<td>98.71</td>
<td>97.42</td>
</tr>
<tr>
<td>m5</td>
<td>98.09</td>
<td>96.18</td>
<td>95.54</td>
</tr>
<tr>
<td>m6</td>
<td>40.88</td>
<td>74.84</td>
<td>84.91</td>
</tr>
<tr>
<td>m7</td>
<td>73.75</td>
<td>86.88</td>
<td>90.62</td>
</tr>
<tr>
<td>m8</td>
<td>66.88</td>
<td>78.12</td>
<td>76.25</td>
</tr>
<tr>
<td>Average</td>
<td>89.00</td>
<td>94.03</td>
<td>94.26</td>
</tr>
</tbody>
</table>

Table 3
Speaker identification rates (%) by VQ approach and RL algorithms with codebook size of 128

<table>
<thead>
<tr>
<th>Speaker</th>
<th>VQ</th>
<th>RL1</th>
<th>RL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>97.50</td>
<td>98.75</td>
<td>98.75</td>
</tr>
<tr>
<td>f2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>f3</td>
<td>94.38</td>
<td>94.38</td>
<td>94.38</td>
</tr>
<tr>
<td>f4</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>f5</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>f6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>f7</td>
<td>98.12</td>
<td>96.88</td>
<td>98.75</td>
</tr>
<tr>
<td>f8</td>
<td>98.75</td>
<td>99.38</td>
<td>99.38</td>
</tr>
<tr>
<td>m1</td>
<td>78.29</td>
<td>92.76</td>
<td>93.42</td>
</tr>
<tr>
<td>m2</td>
<td>90.62</td>
<td>97.50</td>
<td>97.50</td>
</tr>
<tr>
<td>m3</td>
<td>99.36</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>m4</td>
<td>99.35</td>
<td>99.35</td>
<td>98.71</td>
</tr>
<tr>
<td>m5</td>
<td>99.36</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>m6</td>
<td>51.57</td>
<td>77.36</td>
<td>84.28</td>
</tr>
<tr>
<td>m7</td>
<td>77.50</td>
<td>92.50</td>
<td>93.12</td>
</tr>
<tr>
<td>m8</td>
<td>75.00</td>
<td>88.75</td>
<td>88.12</td>
</tr>
<tr>
<td>Average</td>
<td>91.28</td>
<td>96.10</td>
<td>96.65</td>
</tr>
</tbody>
</table>

It is observed that for the three codebook sizes both VQ and RL methods give similar results when the recognition rates are high as in the case of the female speakers (f1–f8). However, both RL1 and RL2 significantly improve the results when the VQ approach yields the low recognition rates as it can be seen in the case of the male speakers (m1, m2, m6–m8). As the relaxation labeling technique was first proposed to deal with ambiguity and noise in computer vision systems, better results can be obtained by revising the inconsistencies between objects having interchanging properties. For the present problem, possible reasons for the relatively highly improved results in the set of the male speakers are that the male acoustic feature vectors may be subjected to more noise or/and the degree of overlapping between the acoustic properties of the male speakers is more than that of the female speakers. Moreover, when high identification rates for both set of speakers are already given by the VQ then the power of the RL may reach its limit for updating label inconsistencies.

In comparison between the relaxation labeling using the correlation-based (RL1) and the mutual-information-based (RL2) compatibility functions, the total average identification results for the three codebook sizes obtained from RL2 are only slightly higher than those from RL1. Tables 1–3 also show the individual identification rates for the 16 speakers with three different codebook sizes using the VQ and the relaxation labeling (RL1 and RL2) algorithms. The convergence of both RL1 and RL2 are obtained about 25 iterative steps, therefore the computer running time is not a problem for implementation. Based on the present experimental results, it can be concluded that speaker identification using the relaxation labeling outperforms the VQ approach.

6. Conclusions

A relaxation labeling algorithm has been presented for solving classification problem in the speaker identification task. The flexibility embedded in the framework of relaxation labeling as well as the improved experimental results appear to be promising as a new approach for speech research including the task of speaker verification as the convergence property of the relaxation processes will certainly ease the verification decision based on an accept/reject threshold, i.e. improves the speaker separability. This is also under our present investigation. Even such promising results have been presented, what has been discussed here is an early step of applying the relaxation algorithms to speaker recognition, therefore further study with other proposed relaxation methods [9,15–17] should be encouraged in order to fully explore the power of the relaxation labeling that can offer to the field of speech and speaker recognition.

7. Summary

A nonlinear probabilistic relaxation labeling for speaker identification is presented in this paper. This relaxation scheme, which is an iterative and parallel process, offers a flexible and effective framework for dealing with uncertainty inherently existing in the labeling of the speech feature vectors. Basic concepts and formulations of the relaxation algorithms are outlined. We then discuss how to model the relaxation scheme to the labeling of the speech feature vectors for the speaker
identification task. The results using several codebook sizes obtained from the proposed approach are more favorable than those from the conventional VQ (vector quantization)-based method.

Acknowledgements

The authors would like to thank Dat Tran for his assistance in computer programming, and the referees for their constructive comments.

References