SPEAKER IDENTIFICATION USING RELAXATION LABELING

Tuan Pham, Michael Wagner

University of Canberra
Faculty of Information Sciences and Engineering
ACT 2601, Australia
E-mail: tuanp@ise.canberra.edu.au

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. Identification rates using the RL are higher than those using the conventional VQ (vector quantization) method.

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*. There are a number of techniques for speaker identification such as the speaker-based VQ codebook approach, dynamic time warping, discrete hidden Markov models, neural networks, and others [2,4]. Among these, the VQ codebook approach is one of the most popular methods implemented in many speaker recognition systems as it limits 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 [9] 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.

2. SPEAKER-BASED VQ CODEBOOK APPROACH

For speaker identification based on VQ codebook approach [10], the codebook for each speaker is generated by clustering a set of training feature vectors $\{v_1, v_2, \cdots, v_T\}$ and partitioning the feature vector space $\{S_1, S_2, \cdots, S_J\}$ where each partition S_j is represented by a centroid vector b_j. There are N codebooks generated for N speakers. In the testing, the distortion between a set of testing feature vectors $\{v_1, v_2, \cdots, v_L\}$ and each codebook is to be measured, then an average distortion D_i to the ith codebook is taken:

$$D_i = \frac{1}{L}\sum_{l=1}^{L} \min_j d(v_l, b_j), \quad j = 1, 2, \cdots, J \qquad (1)$$

where $d(v_l, b_j)$ is the distortion measure (usually a Euclidean distance) between two vectors v_l and b_j.

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

$$i^* = \arg\min_i D_i, \quad i = 1, 2, \cdots, N. \qquad (2)$$

3. RELAXATION LABELING

Let a set of objects $A = \{a_1, a_2, \cdots, a_n\}$ and a set of labels $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_m\}$. An initial probability is given to each object a_i, $i = 1, ..., n$, having each label λ, which is denoted as $p_i(\lambda)$. These probabilities satisfy the condition:

$$\sum_{\lambda \in \Lambda} p_i(\lambda) = 1, \quad \forall a_i \in A, \quad 0 \le p_i(\lambda) \le 1. \qquad (3)$$

The relaxation labeling updates the probabilities $p_i(\lambda)$ in (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:

$$r_{ij}(\lambda, \lambda') \begin{cases} < 0 & : \ \lambda, \lambda' \text{ are incompatible for } a_i \text{ and } a_j \\ = 0 & : \ \lambda, \lambda' \text{ are independent for } a_i \text{ and } a_j \\ > 0 & : \ \lambda, \lambda' \text{ are compatible for } a_i \text{ and } a_j \end{cases} \qquad (4)$$

The updating factor for the estimate $p_i(\lambda)$ at kth iteration is

$$q_i^{(k)}(\lambda) = \sum_j d_{ij} \left[\sum_{\lambda'} r_{ij}(\lambda, \lambda') p_j^{(k)}(\lambda') \right] \qquad (5)$$

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

$$\sum_j d_{ij} = 1. \tag{6}$$

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

$$p_i^{(k+1)}(\lambda) = \frac{p_i^{(k)}(\lambda)[1 + q_i^{(k)}(\lambda)]}{\sum_\lambda p_i^{(k)}(\lambda)[1 + q_i^{(k)}(\lambda)]} \tag{7}$$

Thus, the iterative process given by (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 [7]. The two methods employ the concepts of statistical correlation and mutual information. The correlation based estimate of the compatibility coefficients is defined by

$$r_{ij}(\lambda, \lambda') = \frac{\sum_i [p_i(\lambda) - \bar{p}(\lambda)]\,[p_j(\lambda') - \bar{p}(\lambda')]}{\sigma(\lambda)\sigma(\lambda')} \tag{8}$$

where $p_j(\lambda')$ is the probability of a_j having label λ', and a_j 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

$$r_{ij}^*(\lambda, \lambda') = [1 - \bar{p}(\lambda)]\,[1 - \bar{p}(\lambda')]\, r_{ij}(\lambda, \lambda') \tag{9}$$

The mutual-information based estimate of the compatibility coefficients is

$$r_{ij}(\lambda, \lambda') = \log \frac{n \sum_i p_i(\lambda) p_j(\lambda')}{\sum_i p_i(\lambda) \sum_i p_i(\lambda')} \tag{10}$$

where, for the present problem, n is the number of feature vectors of an unknown speaker. The compatibility coefficients in (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 λ in the speaker population Λ. We will discuss how to estimate the initial probabilities, how to effectively implement the relaxation labeling process by Rosenfeld *et al.* [9] 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 λ can be estimated as

$$p_i(\lambda)^{(0)} = \frac{\exp(-D_{i\lambda})}{\sum_\lambda \exp(-D_{i\lambda})} \tag{11}$$

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

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

where K is the codebook size.

4.2. Implementation of the RL process

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 λ. 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 λ and speaker λ'. With this argument, the original compatibility coefficients as defined in (4) can be expressed in another form as

$$r_{ii}(\lambda, \lambda') = r_i(\lambda, \lambda') \begin{cases} < 0 & : \ \lambda, \lambda' \text{ are incompatible for } a_i \\ = 0 & : \ \lambda, \lambda' \text{ are independent for } a_i \\ > 0 & : \ \lambda, \lambda' \text{ are compatible for } a_i \end{cases} \tag{12}$$

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

$$q_i^{(k)}(\lambda) = \left[\sum_{\lambda'} r_i(\lambda, \lambda') p_i^{(k)}(\lambda')\right] \tag{13}$$

where the summation of d_{ij} as defined in (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 λ, ie. 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 $r_i(\lambda, \lambda')$ tend to be negative as λ and λ' are incompatible for $\{a_i\}$. This also leads to a negative value for the updating factor $q_i^{(k)}(\lambda)$ in (13), which is defined in terms of the compatibility coefficients. From this standpoint, if the equation (7) is used for updating the probability, then the confidence for a distinctive or overlapping vector a_i belonging to the speaker λ will be decreased or increased instead of being increased or decreased, respectively. Therefore, the plus sign in (7) should become a minus sign, that is

$$p_i^{(k+1)}(\lambda) = \frac{p_i^{(k)}(\lambda)[1 - q_i^{(k)}(\lambda)]}{\sum_\lambda p_i^{(k)}(\lambda)[1 - q_i^{(k)}(\lambda)]} \tag{14}$$

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

$$r_i(\lambda, \lambda') = \frac{\sum_i [p_i(\lambda) - \bar{p}(\lambda)]\,[p_i(\lambda') - \bar{p}(\lambda')]}{\sigma(\lambda)\sigma(\lambda')} \tag{15}$$

where $p_i(\lambda')$ is the probability of a_i having label λ', $\bar{p}(\lambda')$ 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

$$r_i^*(\lambda, \lambda') = [1 - \bar{p}(\lambda)]\,[1 - \bar{p}(\lambda')]\,r_i(\lambda, \lambda') \tag{16}$$

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

$$r_i(\lambda, \lambda') = \log \frac{n \sum_i p_i(\lambda) p_i(\lambda')}{\sum_i p_i(\lambda) \sum_i p_i(\lambda')} \tag{17}$$

5. EXPERIMENTS

Both VQ codebook approach and relaxation labeling (RL) were simulated and tested with a set of computer commands from the TI46 speech data corpus. The TI46 corpus contains 46 utterances spoken repeatedly by 8 female and 8 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 8 testing sessions. The corpus is sampled at 12500 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 μ=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 [5], each speaker's 100 training tokens (10 utterances x 1 training session x 10 repetitions) were used to train the speaker-based VQ codebook by clustering the set of all the 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 x 8 testing sessions x 2 repetitions) were tested against all speakers' 10-word models.

For the codebook of 32 entries, the average error rates for speaker identification are shown in Table 1, where for: VQ = 15.02%, RL1 = 8.45 % (RL using correlation-based compatibility coefficients be denoted as RL1), and RL2 = 8.02 % (RL using mutual-information-based compatibility coefficients be denoted as RL2). For the codebook of 64 entries, the average error rates for speaker identification are (Table 2): VQ = 11.00 %, RL1 = 5.97 %, and RL2 = 5.74 %. Finally, for the codebook of 128 entries, the average error rates for speaker identification are (Table 3): VQ = 8.72 %, RL1 = 3.90 %, and RL2 = 3.35 %.

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. Generally, using the RL algorithms the error rates are reduced by half in comparison with those using the VQ approach for all three codebook sizes.

Table 1. Identification rates (%) and average errors (%) using VQ and RL with codebook size of 32

Speaker	VQ	RL1	RL2
f1	95.62	95.62	93.12
f2	98.75	98.75	98.75
f3	84.38	87.50	83.12
f4	98.75	98.12	98.75
f5	100	99.38	99.38
f6	98.75	98.12	97.50
f7	95.62	91.88	90.00
f8	96.25	97.50	96.88
m1	79.61	90.79	91.45
m2	78.12	94.38	96.88
m3	99.36	100	99.36
m4	93.55	94.84	91.61
m5	96.18	94.90	95.54
m6	40.88	69.18	83.65
m7	48.12	76.25	83.12
m8	55.62	77.50	72.50
Average	84.98	91.55	91.98
Av. Error	**15.02**	**8.45**	**8.02**

Table 2. Identification rates (%) and average errors (%) using VQ and RL with codebook size of 64

Speaker	VQ	RL1	RL2
f1	95.62	97.50	96.25
f2	100	99.38	98.75
f3	91.25	88.12	86.25
f4	99.38	99.38	99.38
f5	100	100	100
f6	100	100	99.38
f7	96.25	96.25	94.38
f8	98.75	100	99.38
m1	76.97	92.76	92.76
m2	88.75	96.25	97.50
m3	99.36	100	99.36
m4	98.06	98.71	97.42
m5	98.09	96.18	95.54
m6	40.88	74.84	84.91
m7	73.75	86.88	90.62
m8	66.88	78.12	76.25
Average	89.00	94.03	94.26
Av. Error	**11.00**	**5.97**	**5.74**

Table 3. Identification rates (%) and average errors (%) using VQ and RL with codebook size of 128

Speaker	VQ	RL1	RL2
f1	97.50	98.75	98.75
f2	100	100	100
f3	94.38	94.38	94.38
f4	100	100	100
f5	100	100	100
f6	100	100	100
f7	98.12	96.88	98.75
f8	98.75	99.38	99.38
m1	78.29	92.76	93.42
m2	90.62	97.50	97.50
m3	99.36	100	100
m4	99.35	99.35	98.71
m5	99.36	100	100
m6	51.57	77.36	84.28
m7	77.50	92.50	93.12
m8	75.00	88.75	88.12
Average	91.28	96.10	96.65
Av. Error	**8.72**	**3.90**	**3.35**

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. In fact we have also reported a successful application of this relaxation labeling to the task of speaker verification [8]. 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 [1, 3, 6] 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.

Acknowledgement – The authors thank Dat Tran for his assistance in computer programming.

7. REFERENCES

1. Q. Chen and J.Y.S. Luh, Relaxation labeling algorithm for information integration and its convergence, *Pattern Recognition*, **28**, 1705-1722 (1995).

2. G.R. Doddington, Speaker recognition evaluation methodology – An overview and perspective, *Proceedings of Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, Avignon (France), 1998, pages 60-66.

3. A.M.N. Fu and H. Yan, A new probabilistic relaxation method based on probability space partition, *Patern Recognition*, **30**, 1905-1917, 1997.

4. S. Furui, An overview of speaker recognition technology, *Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, pp. 1-9, Martigny, Switzerland (1994).

5. Y. Linde, A. Buzo and R.M. Gray, An algorithm for vector quantization, *IEEE Trans. Comm.*, **28**, 84-95, 1980.

6. H. Ogawa, A fuzzy relaxation technique for partial shape matching, *Pattern Recognition Letters*, **15**, 349-355, 1994.

7. S. Peleg and A. Rosenfeld, Determining compatibility coefficients for curve enhancement relaxation processes, *IEEE Trans. Systems, Man, and Cybernetics*, **8**, 548-555, 1978.

8. T.D. Pham, D. Tran and M. Wagner, Speaker verification using relaxation labeling, *Workshop RLA2C (Speaker Recognition and its Commercial and Forensic Applications)* (Avignon, France, 1998) 29-32.

9. A. Rosenfeld, R.A. Hummel and S.W. Zucker, Scene labeling by relaxation operations, *IEEE Trans. Systems, Man, and Cybernetics*, **6**, 420-433, 1976.

10. F.K. Soong, E. Rosenberg, B.H. Juang and L.R. Rabiner, A vector quantization approach to speaker recognition, *AT & T Technical Journal*, **66**, 14-26, 1987.