A Novel Speaker Verification System Using Samples-Based Acoustical Model

Gongjun Li and Carol Espy-Wilson
Speech Communication Lab, Institute for Systems Research (ISR), Department of Electrical and Computer Engineering
http://www.isr.umd.edu/Labs/SCL

1. INTRODUCTION

A novel text-independent speaker verification system is built up using the sample-based acoustical model (SAM): The training data is viewed as a concatenation of many speech pattern samples and the pattern matching involves a comparison of the pattern samples and the test speech. A tree is generated to index the entries to pattern samples using an expectation and maximization (EM) approach. The leaves in the tree are used to quantize the training data and the obtained leaf number sequences are exploited in pattern matching as a temporal model. We use a DTW technique and a GMM scheme with predefined penalty to extend the search. The results of experiments conducted on NIST'98 speaker recognition evaluation data show that the error rate is lowered and the system can better capture speaker-specific features in the speech patterns.

2. DYNAMIC ACOUSTIC PROPERTY

Text: shaving cream is a popular item on Halloween

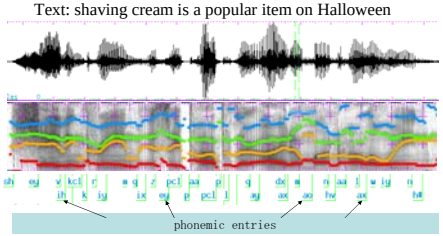


Fig. 1 Dynamic acoustical property and pattern samples in the speech

3. A NOVEL TEXT-INDEPENDENT SPEAKER VERIFICATION SYSTEM

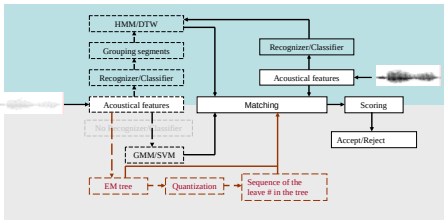


Fig. 2 Supervised approaches and unsupervised approaches are shown in the green and gray blocks, respectively. The dashed frames and solid frames are for training and decoding, respectively.

3. SAM

Like HMM and DTW, SAM is a dynamic model where the training data is viewed as a concatenation of the acoustical pattern samples. The samples are tracked to match with by the test speech

3.1. Comparison of SAM with HMM/DTW and GMM/SVM

Comparison	HMM/DTW	GMM/SVM	SAM
Recognizer	yes	no	no
Training data	more	less	N/A
Clustering	yes	no	no
Flexibility	low	N/A	high
Dynamic	yes	no	yes
Complexity	high	low	medium

3.2. Indexing the entries to the pattern samples using EM tree

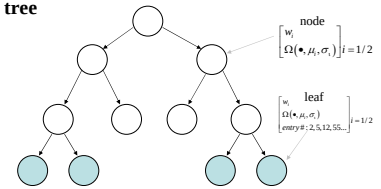


Fig. 3 EM tree used to index the entries to pattern samples

In the EM tree, two nodes generate a 2-mixture GMM and all leaves constitute a *N*-mixture GMM. The parameters of the GMMs are estimated using MLE criterion.

Each leaf contains index information to the entries of the pattern samples in addition to the parameters of a weighted Gaussian density function. Some of them are pseudo-entries and automatically discarded by Viterbi search in the decoding.

3.3. Quantization of the training data

To reduce the computational costs, all feature vectors in the training data are quantized in the following way:

$$index_t = \arg\max_i \left\{ w_i \Omega\left(x_t, \mu_i, \sigma_i\right) \right\}$$

The obtained index sequences are exploited as temporal models in decoding.

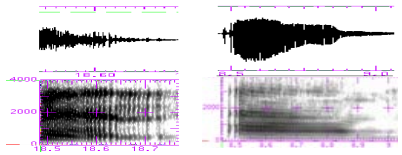
4. DECODING WITH DTW AND GMM

The feature vectors in the pattern samples from the test data are applied to track the similar pattern samples in the training data. Since the boundaries of the pattern samples are unknown, the decoder has to dynamically extend the search inside the pattern samples by DTW and process the transition between the pattern samples by searching EM tree at each time.

5. RESULTS

To verify the viability of SAM, we conducted experiments on the NIST'98 speaker recognition evaluation data. This database is a small portion of the telephony switchboard data and includes speech from 250 male and 250 female speakers. In our experiments, we used speech from 10 female speakers to train the acoustical models under one-session training condition. For testing, we used 500 utterances where 241 of them contained speech from the same speakers in the training data.

Given in Fig. 4 is an example where the syllable /EM/ from the test speech is matched with the syllable /BEN/ in the training data. In the experiments it is noticed that more phonemic pattern samples are tracked. The size of the pattern sample that is tracked is flexible. Since the pattern sample is virtually a segment in a strong correlation, SAM can be applied in identity verification via non-speech.



(a) /BEN/ in the training data (b) /EM/ in the test speech
Fig. 4 Pattern samples in training data tracked by test speech segment

Table 2. Comparison of GMM and SAM at EER point

Speaker #		3030	3207	3317	3346	3349	3369	3384	3528	3740	3753	total
GMM	3s	81.2	88.3	80.0	82.8	86.3	89.1	76.0	82.6	86.8	85.4	83.9
	10s	89.8	89.2	86.4	89.8	90.8	88.8	83.6	88.0	92.3	89.7	88.8
SAM	3s	86.6	90.2	80.9	83.5	87.9	87.7	80.7	84.6	87.9	83.6	85.4
	10s	88.6	89.4	86.9	90.3	90.9	88.8	83.5	88.3	92.8	90.5	89.0