HYBRID TEXT-INDEPENDENT SPEAKER RECOGNITION
USING CHARACTER-BASED BACKGROUND HMMS AND GMMS FOR MANDARIN SPEECH

DENG Hao-Jiang, DU Li-Min, WAN Hong-Jie
Center for Speech Interaction Technology Research,
Institute of Acoustics, Chinese Academy of Sciences, Beijing
denghj, dulm, wanhj@ns.iis.ac.cn

ABSTRACT
In mandarin, the words are composed by the concatenation of Chinese characters. In this paper, we propose a hybrid speaker recognition system based on character-based background HMMS and Gaussian mixture models to combine the advantage of them for text-independent Mandarin speech. Here all characters, spoken by all reference speakers selected to form the background HMMS, is represented by a large HMM, named general-character HMM. The estimating process of background model is much easier and simpler than those word or sub-word based HMMS. The trained character-based HMMS are used to remove the segments only containing silence and noise from utterances, then the speech segments are used to train the GMMs for text-independent speaker recognition and to specify scoring segments for test utterances. Furthermore, it provides speaker-independent background likelihood scores for verification. The normalization effects using the background HMMS with different topological structures are compared. It is shown that score normalization using the background model can improve the verification performance greatly, but the topological structure of general character HMM for Mandarin speech should be defined appropriately.

1. INTRODUCTION
Speaker recognition, which is a subset of pattern recognition area, is to identify a speaker by the inherent differences of his/her voice, such as articulatory organs and the manner of speaking. Statistic approaches can obtain the probabilistic modeling of the speech characteristics so that the mapping from the testing speech to the training speech can be created. Hidden Markov model (HMM) and Gaussian mixture model (GMM) all fall into the category and have been shown to be efficient for speaker recognition task recently [1]-[4].

The HMMs provide a better acoustic model for speech events and a good framework for dealing with noise and channel degradations [5]. Temporal information, which is advantageous for text-dependent tasks, is modeled by HMMs and consecutive frames of data are forced to align to a sequence of sounds of the language being spoken. But it was found the performance didn’t degraded for text-independent task by removing the transition probabilities in HMMS, because the sequences of the sound in training speech didn’t match the testing sound sequences. On the other hand, the GMMs can model arbitrary feature distribution by a nonparametric, multivariate pdf model [6], which are effective for modeling speaker identity. However, both the training and test data for GMMs should be formed of concatenated segments with silence removed [7], so the performance of GMMs can be affected by the front-end feature extracting because of the difficulty to precisely locate endpoints on the input utterance to be free on non-speech regions under the unpredictable background noise.

Based on the result of modern psycho-physical experiments and speech acoustic analysis for Chinese speech, the character-based silence-general-speaker background HMMS are proposed in this paper, the structure and method to estimate the parameters is presented. The background model is employed to remove the silence and noise segments from natural speech, and a hybrid speaker recognition system based on it and GMMs for text-independent Mandarin speech is proposed to combine the advantage of them. A serial of speaker verification experiments are performed using different general-character HMM models with various topological structures and complexity.

2. HYBRID TEXT-INDEPENDENT SPEAKER RECOGNITION
2.1 Character-based background HMMS
In mandarin, the words are composed by the concatenation of Chinese characters. The characters are monosyllabic and can be decomposed into three components, e.g. the initial, medial and final. In the speech recognition, automatic speech labeling, especially the phone-based HMM speaker recognition task for Chinese speech, these three components are confirmed as the phonetic units to design the phonetic conjunction collected in the corpora. Through acoustic analysis for Chinese speech, we propose the character-based silence-general-speaker background HMMS to remove the silence and noise segments from natural speech.
background model named speaker-independent model, which
dependent background model sets; the other is a uniform
claimant, which is called cohort normalization or speaker-
one is using a set of individual reference models close to the
model. Generally two speaker background models are employed,
between the claimed speaker's reference model and background
normalized score is obtained by compute the likelihood ratio
be effective to improve the performance of speaker verification
of model
follow
estimation method. Given a utterance, the average log-likelihood
parameters are usually obtained using maximum likelihood
observation is represented by a Gaussian mixture density [6], the
GMMs for text-independent speaker recognition and to specify
in this paper. These observation vectors can be used to train the
feature vectors, the Mel-frequency cepstral coefficients are used
then the speech segments are parameterized into a sequence of
segments, only containing silence and noise, from utterances,
Model parameters of the background HMMs are estimated
as follows. Every training utterance is represented by a
concatenating sequence of the general-character, sil and sp
HMMs that are generated corresponding to each prompt text, and
then the Forward-Backward algorithm is used to estimate the
parameters of the composite model. The estimating process is
much easier and simpler than those word or sub-word based
HMMs that are usually employed for speaker recognition.

2.2 Hybrid system and score normalization

The trained character-based HMMs are used to remove the
segments, only containing silence and noise, from utterances,
then the speech segments are parameterized into a sequence of
feature vectors, the Mel-frequency cepstral coefficients are used
in this paper. These observation vectors can be used to train the
GMMs for text-independent speaker recognition and to specify
scoring segments for test utterances.

In the GMM model of speaker $i$, the distribution of acoustic
observation is represented by a Gaussian mixture density [6], the
parameters are usually obtained using maximum likelihood
estimation method. Given an utterance, the average log-likelihood
model $\lambda_i$ used for speaker recognition can be computed as follow

$$S_i = \frac{1}{T} \log \left[ p(X | \lambda_i) \right] = \frac{1}{T} \sum_{t=1}^{T} \log \left[ p(x_t | \lambda_i) \right]$$  \hspace{1cm} (1)

Score normalization or likelihood ratio has been verified to be
effective to improve the performance of speaker verification
systems [8]-[10] and open speaker identification systems. The
normalized score is obtained by compute the likelihood ratio
between the claimed speaker’s reference model and background
model. Generally two speaker background models are employed,
one is using a set of individual reference models close to the
claimant, which is called cohort normalization or speaker-
dependent background model sets; the other is a uniform
background model named speaker-independent model, which
doesn’t require additional step to train and select background sets
just like the former.

Using a speech database of a large number of speakers, the
character-based HMMs can be trained to establish a speaker-
identifiable model, and it provides a speaker-independent
background likelihood scores for normalized verification. The
log likelihood score of background model is calculated using
Viterbi decoding to find the optimal state segmentation for the
corresponding utterance. Then the accumulated score over all
speech segments is normalized by the total length of the speech
segments. The average score is used for score normalization.

In term of log likelihood, the normalized likelihood score of
the hybrid system is computed as the difference of the log
likelihood between the claimant reference GMM model $\lambda_i$, and
background HMM model $\lambda^{(b)}$,

$$S_i^{(n)} = \frac{1}{T} \log \left[ p(X | \lambda_i) \right] - \log \left[ p(X | \lambda^{(b)}) \right]$$  \hspace{1cm} (2)

3. EXPERIMENTS

3.1 Database

The database used in the following experiments is derived from
the standard Mandarin microphone speech database 863Bag. The
number of speakers used to train the character-based background
HMMs should be large enough to contain the general acoustic
features of potential imposters. So a set of 76 speakers (38 males
and 38 females) are selected from 863 database, and 5 utterances
per speaker selected randomly from every speakers’ speech data
are used for background modeling. The speaker verification
system based on GMM model consists of 90 speakers (45 males
and 45 females) separate from the speakers of background
HMMs. 15 utterances per speaker are used to construct the
reference speaker GMMs, and the training speech duration
(silence removed) for each speaker is limited to 45s. The other 30
utterances per speaker are used to test, and the test duration of
each test utterance ranges from 2s to 4s.

3.2 Feature extracting

The utterances are pre-emphasized by coefficient 0.97, then
the 26-dimension feature vectors containing 12 Mel-frequency
cepstral coefficients (including 0’ cepstral coef.) and their delta
coefficient (including delta 0’ cepstral coef.) are obtained using
Hamming window (width 25s and shift 10s), the filterbank has
26 channels.

The feature vectors training the background HMM models
are extracted over the whole utterance duration, and the feature
vectors training the reference GMM models are extracted from
the speech segments labeling using the background HMM models.
3.3 Experiment results

In the speaker verification experiments, the character-based background HMM models are initialized by the global speech means and variances and trained using Baum-Welch algorithm. If the number of mixtures of each states in general character HMM is more than one, each state is initialized by a mono Gaussian mixture component at first, after several epochs of re-estimation using standard Baum-Welch algorithm, the number of mixtures of each states in general character HMM is increased gradually by splitting the mixture component with the largest weight in each state. The two new components split is constructed as follows, the weights of the two are half of the old one, 0.1 standard deviation is added and subtracted to the old mean vector to build the new ones, then the re-estimations and mixtures increasing are repeated until the needed number of mixture components of each states is reached.

The reference speakers are modeled by a set of 75 component GMMs with a nodal, diagonal covariance matrix trained using the expectation-maximization (EM) algorithm. The reference models are initialized using the LBG vector quantization algorithm [11], and the minimum variance value 0.01 is applied to the variance estimates to avoid singularities in the final model.

During speaker verification test, each test utterance is used to play a true claimant to match its own reference model, and impostors to all the other reference models. So the number of true speaker and impostor attempts for each reference model is 30 and 2670 respectively. Totally it is 2700 and 80100 respectively. The average log-likelihood of the reference model for a test utterance is calculated using formula 1 and then the normalized score is calculated using formula 2. The normalized score is compared with a speaker independent threshold given in advance, if it is higher than the threshold, accept the claim to the reference speaker; otherwise refuse.

The general-character HMM model with various topological structures and complexity is employed and the verification

![Graphs showing EER versus number of Gaussian mixtures for different states and models](image)

Figure 2. Speaker verification results using different background models for score normalization
performance is tested in the paper. The posterior EER (Equal Error Rate) is calculated to measure the performance of speaker verification system. EER means that when an appropriate posterior threshold is defined, the error rate of false rejection is equal to that of false acceptance, i.e. equal error rate.

Figure 2 shows a plot of the test results using different background models for speech segmentation and score normalization. The horizontal axis indicates the number of Gaussian mixtures in each state, and the vertical axis indicates the EER obtained by averaging all the evaluation speaker scores for all test utterances. In each of the panels, the general character HMMs using different topological structures are used for background modeling and the results are shown. In the figures, the symbols ‘rmLn’ indicate the different topological structures; ‘r’ means each state has non-zero transition probabilities forward to the next m states, and ‘Ln’ means each state has non-zero transition probabilities backward to the front n states. For example, the left-to-right topology without skip usually employed is denoted by ‘R’ The number of states for general character HMM is 10, 15, 20. 30 in the panels of Figure 2, and keep the number of states of general character HMM immutably the best performance is listed in Table 1.

It is shown in Figure 2 and Table 1 that score normalization using character-based background HMM model can improve the verification performance greatly (EER from 5.57% to 1.15%); the topological structure of general character HMM and complexity can influence the normalization effect: multi-mixture no more than 4 mixture components in each state is better than mono-mixture always, the number of non-zero transition probabilities per state shouldn’t be too big, the backward transition is not needed, i.e. the Ergodic and the left-to-right without skip general-character HMM topology is not appropriate for the hybrid speaker recognition system. So the best topological structure of general character HMM for Mandarin speech should be defined as follows, the number of states is about 15, the product of number of states and number of Gaussian mixture per state is about 40, the number of non-zero transition of each state is 2 or 3.

<table>
<thead>
<tr>
<th>Number of states</th>
<th>Number of mixtures per state</th>
<th>Topological structure</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>R2</td>
<td>0.0117</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>R2</td>
<td>0.0115</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>R3</td>
<td>0.0116</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>R3</td>
<td>0.0122</td>
</tr>
<tr>
<td>No score normalization</td>
<td></td>
<td></td>
<td>0.0557</td>
</tr>
</tbody>
</table>

5. CONCLUSION

The hybrid speaker recognition system based on character-based background HMMs and Gaussian mixture models for text-independent Mandarin speech is proposed in this paper. The estimating process is much easier and simpler than those word or sub-word based HMMs The trained character-based HMMs are used to remove the segments containing silence and noise from utterances, then the speech segments are used to train the reference speaker GMMs for text-independent speaker recognition and to specify scoring segments for test utterances. Using a speech database of a large number of speakers, the character-based HMMs can be trained to establish a speaker-independent model, and it provides a speaker-independent background likelihood scores for normalized verification. It is shown that score normalization using the background model can improve the verification performance greatly. The topological structure and complexity of general character HMM can influence the normalization effect, the Ergodic and the left-to-right without skip general-character HMM topology is not appropriate for the hybrid speaker recognition system. The best topological structure of general character HMM for Mandarin speech should be defined as, the number of states is about 15, the product of number of states and number of Gaussian mixture per state is about 40, the number of non-zero transition of each state is 2 or 3.

5. REFERENCE