Abstract

Gaussian mixture models (GMMs) and ergodic hidden Markov models (HMMs) have been successfully applied to model short-term acoustic vectors for speaker recognition systems. Prosodic features are known to carry information concerning the speaker’s identity and they can be combined with the short-term acoustic vectors in order to increase the performance of the speaker recognition system.

In this paper, a statistical approach using pitch-dependent GMMs for modeling speakers is presented. This new approach is capable of simultaneously modeling the statistical distributions of the short-term acoustic vectors and long-term prosodic features.

1. Introduction

In human decoding of speech, suprasegmental information plays an important role. Suprasegmental features, in particular prosodic features, are known to carry information regarding the identity of the speaker. Several authors have reported on the use of prosodic features such as fundamental frequency which correlates to pitch and on the short-time signal energy which correlates to loudness. However, interest in the use of prosodic features appears to have been diminished in recent years because these features alone could not give the level of performance required for speaker identification and verification in text-dependent systems. In addition, it was also difficult to see how they could be incorporated in a text-independent system. Still, we do not know the best way to combine suprasegmental prosodic measures with segmental acoustic measures extracted from speech signal in the speaker recognition process.

The Gaussian mixture models (GMMs) [1] and ergodic hidden Markov models (HMMs) [2] are widely used statistical models to characterize the short-term spectral envelope and have been successfully applied to text-independent speaker recognition systems. It is well known that the effect of channel distortions and noise on the performance of such systems is a serious concern. Prosodic features are known to be less affected by these impairments than short-term spectral envelope features. Suprasegmental features are therefore worth re-examining for speaker recognition, particularly when used to improve the performance of algorithms based on GMM techniques.

The information carried by the pitch, which is not present in unvoiced regions, is not evident to model. A new statistical modeling approach is necessary to represent, in a unified framework, short-term features and pitch related features, as well as their dependencies.

In this paper, a new approach that takes advantage of the dependencies between the pitch and short-term features is presented. The pitch is modeled in a different manner for voiced and unvoiced regions and the short-term acoustic vector models are dependent on the pitch model.

2. The Proposed Model

Several authors have reported on the usage of pitch in speaker recognition experiments [3] [4]. Works carried out on gender identification lead up to 98% accuracy using the mean pitch alone [5]. These experiments highlight the substantial amount of information about the speaker’s identity carried by the pitch. However, this prosodic feature alone is not discriminative enough for speaker recognition and short-term spectral features can be advantageously used to complete the set of acoustic parameters. The primary problem for this is how to incorporate pitch and short-term features in a unified framework, since the pitch value can be absent (unvoiced speech) or present (voiced speech). Various methods have been investigated for handling the unvoiced region [6] [7] in speech recognition systems. In this paper we propose a new method based on pitch-dependent GMMs for speaker recognition systems.

2.1. The Pitch Model

Although the classical AUTOC method [8], based on the autocorrelation, was used for tracking the pitch, several other robust methods can be used for this task [9] [10] [11]. Following this method, the pitch is set to 0 in unvoiced regions. A technique to model the pitch, as proposed in this paper, is as follows.

Initial model, represented in Figure 1, is a two-state ergodic HMM with one state representing the voiced regions and the other representing the unvoiced regions. The distribution of the pitch values \( \theta \) in the voiced regions (\( \theta > 0 \)) is approximated by only one Gaussian \( \mu_1, \sigma_1 \) distribution, while in the unvoiced regions, the probability of having \( \theta = 0 \) is set to 1. We have already applied a similar ergodic HMM to model the pitch in previous multiresolution experiments [12].

![Ergodic HMM for the pitch](image)

Figure 1: Ergodic HMM for the pitch
The second model aimed at simplifying the first one, the difference being that the transition probabilities \( a_{ij} \) were eliminated and replaced by two weights \( w_1 \) and \( w_2 \). These weights represent respectively the probability of being in a voiced or unvoiced region. One can see that there is a relationship between the transition probabilities \( a_{ij} \) and the weights \( w_i \)

\[
  w_1 = \frac{a_{11}}{a_{21} + a_{12}}, \quad w_2 = \frac{a_{12}}{a_{21} + a_{12}}. \quad (1)
\]

This simplification was made because the experiments demonstrated that the transition probabilities are influenced by the linguistic content when the training set is not large enough. Moreover, identification scores were also highly variable when short test utterances were used (~3ms). The second model was then kept for the experiments that followed.

Since \( w_2 = 1 - w_1 \), the pitch model \( \lambda_\phi \) is then completely defined by

\[
  \lambda_\phi = \{ w_1, \mu_1, \sigma_1 \}. \quad (2)
\]

### 2.2. Pitch-Dependent GMMs

Short-term acoustic features that belong to voiced regions present common characteristics (presence of the response to the glottal excitation, specific distribution of voiced features, etc.). These characteristics are different than those of features from unvoiced regions. In order to better capture the variations and to better model the distributions of voiced and unvoiced features, a pitch-dependent (voiced/unvoiced) model is proposed.

In this work, the spectral envelope is represented by the Mel-frequency cepstral coefficients (MFCCs). These coefficients are modeled with two GMMs with diagonal covariance matrices. The first GMM is trained to fit the distribution of the MFCCs in the voiced regions, while as the second GMM models the unvoiced regions. Splitting the short-term features in this way, we reduce the variability of each group of features and GMMs can better fit their distributions. This model is represented in Fig. 2.

The spectral content of the utterance can then be represented by two GMMs

\[
  \lambda_i = \{ c_{im}, \mu_m, \sigma_m \} \quad i = 1, 2 \quad ; \quad m = 1, \ldots, M, \quad (3)
\]

where \( i \) represents the pitch state, \( c_{im} \) the weight associated to the \( m \)th mixture and \( M \) the number of mixtures per state.

The speaker can now be represented by the following model:

\[
  \lambda = \lambda_\phi \cup \lambda_1 \cup \lambda_2. \quad (4)
\]

### 2.3. Training and Testing

#### 2.3.1. Training

Given the short-term spectral features \( \{ \vec{z}_i \} \) and the pitch contour \( \{ \phi_t \} \) training sets, voiced and unvoiced regions are trained separately following the Expectation-Maximization algorithm (E-M) described in [1]. The parameters of the Gaussian distribution for the pitch can be directly obtained from the data.

#### 2.3.2. Likelihood Estimation

Let now \( \vec{z}_i \) be a short-term feature vector that belongs to a test utterance and \( \phi_t \) the corresponding value of the pitch. Given the model, the likelihood for the couple \( \eta_t = (\vec{z}_i, \phi_t) \) is

\[
  p(\eta_t | \lambda) = p(\vec{z}_i | \phi_t | \lambda_i) \cdot p(\phi_t | s_{\phi t} = i), \quad (5)
\]

where \( s_{\phi t} \) is the state of the pitch at time \( t \).

Since the likelihood \( p(\phi_t | s_{\phi t} = 0) \) in unvoiced regions is zero and the likelihood \( p(\phi_t | s_{\phi t} = 1) \) in voiced regions can also be considered zero, Eq. 5 can therefore be simplified and the likelihood for voiced and unvoiced regions can be written as

\[
  p(\eta_t | V_i | \lambda) = \prod_{t | \eta_t \in V_i} w_i \cdot p(\vec{z}_i | \lambda_i) \cdot p(\phi_t | s_{\phi t} = i), \quad (6)
\]

where \( V_i = \{ \eta_t | s_{\phi t} = i \} \).

In order to make the likelihood independent of the number of vectors in the set \( V_i \), we can express the log-likelihood as

\[
  LLH(V_i | \lambda) = \frac{1}{T_{V_i}} \sum_{t | \eta_t \in V_i} log(w_i \cdot p(\vec{z}_i | \lambda_i) \cdot p(\phi_t | s_{\phi t} = i)), \quad (7)
\]

where \( T_{V_i} \) is the number of vectors in the set \( V_i \).

The likelihood of the complete sequence \( O = \{ \eta_1, \ldots, \eta_T \} \) can then be expressed as:

\[
  LLH(O | \lambda) = \sum_{i} LLH(V_i | \lambda). \quad (8)
\]

### 3. Experiments

Six speaker identification experiments (E1,...,E6) were performed for 38 speakers of the TIMIT database downsampled to 8 kHz. Thirteen MFCCs obtained from windows of 256 samples (31.25ms) and overlapped by 128 samples were extracted from speech signals. The models were trained with ~24ms of speech, and two utterances of ~3ms were used for the tests, giving a total of 2812 inter-tests and 76 intra-tests.
The goal of the following experiments was to compare different statistical models in order to show the advantages of a pitch-dependent modeling.

- E1: Only the pitch model (two weighted states) was used to identify the speakers.
- E2: Only the unvoiced speech regions were used to train a GMM. This experiment was carried out in order to point out the contribution of these features in the total identification score.
- E3: Only the voiced speech regions were used to train a GMM. Just as for E2, the goal was to determine the relative amount of information carried by the voiced speech.
- E4: MFCCs are modeled with a GMM, this experiment was carried out for comparison purposes only.
- E5: MFCCs from voiced and unvoiced regions were modeled with different GMMs; however, the pitch model was not included.
- E6: Pitch and MFCCs were combined using the proposed model. Here, GMMs for voiced and unvoiced regions depend on the pitch model (two weighted states).

Table 1 displays the identification scores reached for each experiment.

Table 1: Identification scores for each experiment.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Score Test1</th>
<th>Score Test2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1: Pitch</td>
<td>42%</td>
<td>29%</td>
</tr>
<tr>
<td>E2: Unvoiced</td>
<td>84%</td>
<td>84%</td>
</tr>
<tr>
<td>E3: Voiced</td>
<td>81%</td>
<td>89%</td>
</tr>
<tr>
<td>E4: MFCC-GMM</td>
<td>92%</td>
<td>87%</td>
</tr>
<tr>
<td>E5: Voiced-Unvoiced</td>
<td>92%</td>
<td>94%</td>
</tr>
<tr>
<td>E6: Pitch-Dependent GMMs</td>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>

3.1. Measure of the Discrimination Level

Let \( \psi = \{ l_h \} \) be the set of speakers, \( h = 1, \ldots, N \), and let \( \lambda_{ij} \) be the model of the \( j^{th} \) speaker. The likelihood for a test sequence \( O_h \) produced by \( l_h \) given the model \( \lambda_{ij} \) is \( LLH(O_h | \lambda_{ij}) \).

The identified speaker in each case corresponds to the model \( \lambda_{ij} \) for which the likelihood score is maximum. We define the normalized likelihood \( L(h, j) \) as

\[
L(h, j) = \frac{LLH(O_h | \lambda_{ij})}{LLH(O_h | \lambda_{ij*j})}.
\]

Figure 4 shows the matrices \( L(h, j) \) for the experiments E1,...,E6. The choice of \( j^* \) is more or less sensitive and prone to errors depending on how close the likelihoods are to the values on the diagonal.

In order to measure the level of discrimination of the features associated to a model, the following function is proposed

\[
D(\psi, \lambda) = \frac{\sum_h \delta(h) \cdot \sum_{k=1}^{K} L(h, k) - L(h, \xi(k))}{K \cdot N},
\]

where \( \delta(h) \) is 1 if \( l_h \) has been correctly identified and 0 otherwise. \( K \) is the number of closest speakers for which the measure is calculated and \( \xi(k) \) is the \( k^{th} \) closest speaker.

This function is not intended to replace the identification score, but to rather provide a measure of its quality.

The first test (Test 1) shows that the classical MFCC-GMM E4 experiment obtains a better identification score compared to E3, where only the voiced regions are used. Nevertheless, the \( L(h, j) \) matrix for E3 presents a better discrimination measure compared to E4. The second test shows that the identification score of E4 is lower than the E3 score. These results were already predicted by the discrimination measure \( D \). Also, the second test with
the pitch-dependent GMMs (E6) gave a score of 100%. This method had already obtained the higher discrimination measure D when the first test was achieved.

4. Conclusions
Although only 16 mixtures were used to model voiced and unvoiced regions, the identification scores reached 100%. The pitch contour is proven to help the identification score and the experiments show that the proposed model is capable of capturing and advantageously exploiting the speaker information carried by the pitch. The measure of discrimination introduced gives us an idea regarding the quality of the identification score, which can present variations when performing several tests. Since the pitch contour varies slowly in comparison to short-term spectral features, it can be downsampled in order to reduce redundant data (Figure 5). In this case the pitch-dependent model does not change and training and testing algorithms remain the same. In addition, the computational load compared to classical GMM training and testing remains nearly the same, the main difference being the estimation of the pitch model.

When MFCCs were modeled with a GMM (E4) with doubled number of mixtures (32), the identification score reached 96%, which is still lower than the score obtained by the proposed pitch-dependent method. The coefficients distribution characteristics are then better modeled by two pitch-dependent groups of 16 mixtures than by only one group of 32 mixtures trained by the E-M algorithm.

5. References