Chapter 6

Person Authentication using Speech

In the previous chapter, we have presented two methods for person authentication using static nature of the visual speech. A method of extracting the features from the speech signal is described in Section 6.2. The modeling of text independent and text dependent person authentication using acoustic features is described in Sections 6.3.1 and 6.3.2, respectively. The experimental results of text independent and text dependent person authentication are described in Section 6.4.

6.1 Introduction

The process of accepting or rejecting identity claim of a person from his/her voice is termed as speaker verification. Identity of a speaker may exist in the physiological and behavioral characteristics. The physiological characteristics correspond to the characteristics of the vocal tract system and that of the voice source. The behavioral characteristics are due to the manner in which speakers have learnt to use their speech production apparatus. Automatic speaker identification or verification systems rely mainly on features derived from the physiological characteristics of the speaker. For speaker identification or verification, speech data are collected from a speaker, and are used to develop a model for capturing the speaker-specific information. Once the model is developed, verification of the identity claim involves determining the confidence or probability that a given test utterance belongs to the claimant model.
The most important vocal tract related features are those derived from the short-time spectral analysis aiming at capturing the spectral envelope and thus the formant structure. There are two major branches in the short-time spectral analysis: linear predictive analysis and short-time Fourier transform (STFT). Correspondingly, there are two prevalent vocal tract features: linear predictive cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC). In this work, we use only MFCC features which are extracted from speech signal as described below.

6.2 Acoustic Feature Extraction

The speaker specific information present in the speech signal can be used for person authentication. The speech signal is segmented in successive frames, overlapping with each other. Each frame is multiplied by a Hamming window to reduce the effects of spectral leakage [139], [140]. Acoustic features representing the speaker information can be extracted from each windowed frame. These features represent the short-time spectrum of the speech signal. The short-time spectrum envelope of the speech signal is attributed primarily to the shape of the vocal tract. The spectral information of the same sound uttered by two persons may differ due to change in the shape of the individual’s vocal tract system and the manner of speech production. An outline of MFCC feature extraction process is shown in Fig. 6.1.

- **Mel-frequency wrapping**: The Mel-frequency Cepstral Coefficient (MFCC) [66] is able to represent the dynamic features of a signal as they extract both linear and non-linear properties. The MFCC can be a useful tool of feature extraction in vibration signals as vibrations contain both linear and non-linear features. Normal speech waveform may vary from time to time depending on
Fig. 6.1: MFCC feature extraction process.

the physical condition of speakers’ vocal cord. Rather than the speech waveforms themselves, MFCCs are less susceptible to the said variations. The speech signal consists of tones with different frequencies. For each tone with an actual frequency $f$, measured in Hz, a subjective pitch is measured on the ‘Mel’ scale. A ‘Mel’ is a unit of special measure or scale of perceived pitch of a tone. The MFCC is a type of wavelet in which frequency scales are placed on a linear scale for frequencies less than 1 kHz and on a log scale for frequencies above 1 kHz. As a reference point, the pitch of a 1 kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore the following formula is used to compute the mels for a given frequency $f$ in Hz:

$$mel(f) = 2595 \log_{10}(1 + f/700)$$  \hspace{1cm} (6.1)

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.
• **Cepstrum:** In the final step, the log mel spectrum has to be converted back to time. The result is called the mel-frequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT). In practice the last step of taking inverse DFT is replaced by taking discrete cosine transform (DCT) for computational efficiency. The cepstrum is computed by taking the inverse DFT of the logarithm of the magnitude spectrum of the frame. This is represented in the following equation:

\[
\text{cepstrum}(\text{frame}) = \text{IDFT}(\log(|\text{DFT}(\text{frame})|))
\]  

(6.2)

The final MFCC feature vector is obtained by retaining about first 12 lowest DCT coefficients. For acoustic feature extraction, the differenced speech signal is divided into frames of 20 ms, with a shift of 10 ms. The 12 MFCC coefficients \((a_1, a_2, a_3, \cdots, a_{12})\) for a frame are extracted from the speech signal at the segmental level. The 'null' MFCC coefficient \(a_0\) is excluded from the DCT, since it represents the mean value of the input signal which carries little speaker specific information.

The dynamic parameters derived from 13th order static cepstral coefficients \((a_0, a_1, a_2, a_3, \cdots, a_{12})\) have been suggested and shown to improve the performance in speaker recognition. These dynamic features include the delta-cepstrum (the first-order difference of the short-time static cepstrum), the delta-delta-cepstrum (the second-order difference of the static cepstrum), and delta- and delta-delta-energy. Especially, the dynamic features are verified to be more robust than the static features in noisy conditions. A 39th order MFCC is used to capture the static and dynamic features of a
speech signal spectrum which contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration (delta-delta) coefficients. Altogether, a 39 dimensional MFCC for each frame is used as a feature vector.

6.3 Person Authentication using Acoustic Features

6.3.1 Modeling of Text Independent Person Authentication

6.3.1.1 Autoassociative Neural Network Models

The five layer autoassociative neural network model as shown in Fig. 4.6 is used to capture the distribution of the feature vectors as explained in Section 4.3.1. The AANN model with the structure \(39L \ 78N \ 8N \ 78N \ 39L\) is used for capturing the distribution of the acoustic features of a subject. For testing the identity claim, each acoustic feature vector extracted from the test utterance is given as input to the claimant speaker model. The output of the model is compared with its input to compute the normalized error. The normalized error \(e_i\) for the \(i^{th}\) feature vector \((y_i)\) is given by, \(e_i = \frac{\|y_i - o_i\|^2}{\|y_i\|^2}\), where \(o_i\) is the output vector given by the model. The error \((e_i)\) is transformed into a confidence score using \(c_i = \exp(-e_i)\). The average confidence score \(c = \frac{1}{n_f} \sum_{i=1}^{n_f} c_i\) is used to accept or reject the identity claim of the subject, where \(n_f\) is number of acoustic feature vectors in the test utterance.

6.3.1.2 Support Vector Machines

Support Vector Machines as shown in Fig. 4.8 are used to construct the optimal separating hyperplane for acoustic features as explained in Section 4.3.2. The classification ability of the acoustic features of a subject is analyzed using SVM. SVM is trained to distinguish between the acoustic features (class label '+1') of a subject and all other acoustic features (class label '-1') in the training set. For testing the identity claim of a subject, acoustic feature vectors are extracted and are given as input to the claimant
SVM model and the distance between each of the feature vectors and the SVM hyper-plane is obtained. The claim is accepted if the number of positive matches (distance greater than zero) is greater than a threshold; otherwise it is rejected.

### 6.3.1.3 Radial Basis Function Neural Network Model

Radial Basis Function Neural Networks as shown in Fig. 4.10 are used to construct RBF centers (hidden layer) using k-means algorithm for acoustic features as explained in Section 4.3.3. The classification ability of the acoustic features of a subject is analyzed using RBFNN. In the training phase, 39 dimensional feature vectors are extracted from television broadcast news audio and are given as input to the RBFNN model. The weights are determined using least squares algorithm. For testing, acoustic feature vectors are extracted from television broadcast news audio and are given as input to the RBFNN acoustic model and the average output is calculated for the respective neuron. The claim is accepted if the average output of the respective neuron is greater than a threshold; otherwise the claim is rejected.

### 6.3.2 Modeling of Text Dependent Person Authentication

#### 6.3.2.1 Dynamic Time Warping

Dynamic time warping [82], [141] uses the principle of dynamic programming [principle of optimality], in order to compute the overall distortion between the two speech templates. Comparing the template with incoming speech might be achieved via a pairwise comparison of the feature vectors in each. The problem with this approach is that if constant window spacing is used, the length of the input and stored sequences is unlikely to be the same. Moreover, within a word, there will be variation in the length of individual phonemes. The matching process needs to compensate for length differences and takes account of the non-linear nature of the length differences within
Fig. 6.2: An example warping path.

Suppose we have two time series $Q$ and $S$, of length $n$ and $m$ respectively, where:

\[ Q = q_1, q_2, \ldots, q_i, \ldots, q_n \]  
\[ S = s_1, s_2, \ldots, s_j, \ldots, s_m \]

To align two sequences using DTW, we construct an $n$-by-$m$ matrix where the $(i^{th}, j^{th})$ element of the matrix contains the distance $d(q_i, s_j)$ between the two points $q_i$ and $s_j$. Typically the Euclidean distance is used, so $d(q_i, s_j) = (q_i - s_j)^2$. Each matrix element $(i, j)$ corresponds to the alignment between the points $q_i$ and $s_j$. This is illustrated in Fig. 6.2. A warping path $W$, is a contiguous (in the sense stated below) set of matrix elements that defines a mapping between $Q$ and $S$. The $k^{th}$ element of $W$ is defined as $w_k = (i, j)_k$ so we have:

\[ W = w_1, w_2, \ldots, w_k, \ldots, w_K \quad \text{max}(m, n) \leq K < m + n - 1 \]
• **Boundary conditions**: \( w_1 = (1, 1) \) and \( w_K = (m, n) \), simply stated, this requires
the warping path to start and finish in diagonally opposite corner cells of the
matrix.

• **Continuity**: Given \( w_k = (a, b) \) then \( w_{k-1} = (a', b') \) where \( a - a' \leq 1 \) and \( b - b' \leq 1 \).
This restricts the allowable steps in the warping path to adjacent cells (including
diagonally adjacent cells).

• **Monotonicity**: Given \( w_k = (a, b) \) then \( w_{k-1} = (a', b') \) where \( a - a' \geq 0 \) and
\( b - b' \geq 0 \). This forces the points in \( W \) to be monotonically spaced in time.

There are exponentially many warping paths that satisfy the above conditions.
However, we are interested only in the path which minimizes the warping cost:

\[
DTW(Q, S) = \min \left\{ \sqrt{\sum_{k=1}^{K} \frac{w_k}{K}} \right\}
\]  

(6.6)

The \( K \) in the denominator is used to compensate for the fact that warping paths
may have different lengths. This path can be found very efficiently using dynamic
programming to evaluate the following recurrence which defines the cumulative dis-
tance \( \gamma(i, j) \) as the distance \( d(i, j) \) found in the current cell and the minimum of the
cumulative distances of the adjacent elements:

\[
\gamma(i, j) = d(q_i, c_j) + \min \{ \gamma(i - 1, j - 1), \gamma(i - 1, j), \gamma(i, j - 1) \}
\]

(6.7)

**Backtracking** (B) and **Termination**: The optimal (minimum) distance is \( D(Q,S) \)
and the optimal path is \((w_1, w_2, \cdots, w_K)\) and \( w_i = B(i+1, w_{i+1}, i = K - 1, K - 2, \ldots, 1) \).

### 6.4 Experimental Results

Performance of the text independent person authentication system is evaluated using
Indian TV broadcast news video (Sun Network: Sun TV and Sun News) for 50 subjects,
32 females and 18 males as described in Section 6.4.1. For enrolling (training) a subject, an AVI file of 60 sec (4 clips, each of 15 sec) duration at 12 fps is recorded with a resolution of $320 \times 240$ using a TV tuner card. The speech signal is recorded at 8000 samples per second. Performance of the text dependent person authentication system is also evaluated in real time in the laboratory environment for 50 subjects, 32 females and 18 males, using a camera as described in Section 6.4.2.

6.4.1 Text Independent Person Authentication

6.4.1.1 Autoassociative Neural Network Models

The performance of the text-independent speaker authentication system is evaluated for television broadcast news audio considered earlier. The speech signal is recorded for 60 sec at 8000 samples per second. For enrolling a speaker, the differenced speech signal is analyzed by dividing it into frames of 20 msec, with a shift of 10 msec. A 39th order MFCC is used to capture the static and dynamic features of a speech signal spectrum as described in Section 6.2 which contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration coefficients. Altogether, a 39 dimensional MFCC for each frame is used as a feature vector. The distribution of the 39 dimensional MFCC feature vectors in the feature space is captured using an AANN model. Separate AANN models are used to capture the distribution of feature vectors of each speaker. For testing the identity claim of a subject, the speech signal is recorded for 10 sec, one month after collecting the training data.

For identification, acoustic feature vectors extracted from the test utterance are given as input to the model. The output of the model is compared with the input to compute the normalized squared error. The normalized squared error ($e$) for the feature vector $y$ is given by, $e = \frac{||y - o||^2}{||y||^2}$, where $o$ is the output vector given by the model. The confidence score [0 to 1] is same as confidence level which indicates the
Table 6.1: Performance of speaker authentication in terms of number of units in the compression layer.

<table>
<thead>
<tr>
<th>$N_c$</th>
<th>EER (%)</th>
<th>$N_c$</th>
<th>EER (%)</th>
<th>$N_c$</th>
<th>EER (%)</th>
<th>$N_c$</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12.1</td>
<td>4</td>
<td>10.2</td>
<td>8</td>
<td>8.6</td>
<td>10</td>
<td>9.6</td>
</tr>
</tbody>
</table>

correctness of matching (1 for best match and 0 for worst match). The error ($e$) is transformed into a confidence score ($c$) using $c = \exp(-e)$. The average confidence score is calculated from AANN model. The average confidence score gives better performance than using confidence score for each frame. The identity of the subject is decided based on the highest confidence score. The confidence scores from the models can be used to decide the similarity of a subject to other subjects.

For authentication, acoustic feature vectors extracted from the test utterance are given as input to the claimant model, and the confidence score is calculated. The confidence scores for all the authentic and impostor claims are calculated and they are used to measure the performance of the system. If the confidence score is greater than a threshold, then the claim is accepted, otherwise the claim is rejected. For each threshold, the FAR and FRR are computed using the confidence scores. The EER is calculated using person-independent thresholds. The selection of the acoustic frame is an important issue to be addressed in order to discriminate the authentic and impostor claims at the frame level. In the database of 50 subjects, there are 50 authentic claims and $49 \times 50$ impostor claims. The structure of AANN model plays an important role in capturing the distribution of the feature vectors.

Speaker verification experiments were conducted by varying the number of units in the compression layer ($N_c$). The results shown in Table 6.1 indicate that the system gives an optimal performance in terms of EER at $N_c = 8$. The performance is also evaluated by varying the number of units in the second layer (expansion layer), keeping
Table 6.2: Performance of speaker authentication in terms of number of units in the expansion layer.

<table>
<thead>
<tr>
<th>$N_e$</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>9.5</td>
</tr>
<tr>
<td>74</td>
<td>9.0</td>
</tr>
<tr>
<td>78</td>
<td>8.6</td>
</tr>
<tr>
<td>82</td>
<td>8.6</td>
</tr>
<tr>
<td>86</td>
<td>8.6</td>
</tr>
</tbody>
</table>

Table 6.3: Performance of text independent speaker identification and authentication using AANN.

<table>
<thead>
<tr>
<th>Television broadcast news audio</th>
<th>RR (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>91.0</td>
<td>8.60</td>
</tr>
</tbody>
</table>

the number of units in the compression layer to 8. The performance remains unchanged if the number of units in the expansion layer ($N_e$) is increased to 86, and there is a slight decrease in the performance (EER=9.0%) if the number of units is 74. The results are shown in Table 6.2. The experimental results show that the network structure $39L\ 78N\ 8N\ 78N\ 39L$ gives optimal performance in terms of computation time and EER. Performance of text independent speaker identification and authentication using AANN model is given in Table 6.3.

6.4.1.2 Support Vector Machines

The performance of the text-independent speaker verification system is evaluated for television broadcast news audio. The speech signal is recorded for 60 sec at 8000 samples per second. For enrolling a speaker, the differenced speech signal is analyzed by dividing it into frames of 20 msec, with a shift of 10 msec. A 39th order MFCC is used to capture the static and dynamic features of a speech signal spectrum as described in Section 6.2 which contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration coefficients. Altogether, a 39 dimensional MFCC for each
Table 6.4: Performance of text independent speaker identification and authentication using different kernel functions.

<table>
<thead>
<tr>
<th>SVM Kernel Function</th>
<th>Television broadcast news audio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR (%)</td>
</tr>
<tr>
<td>Polynomial</td>
<td>90.0</td>
</tr>
<tr>
<td>Gaussian</td>
<td>94.0</td>
</tr>
<tr>
<td>Sigmoidal</td>
<td>92.0</td>
</tr>
</tbody>
</table>

frame is used as a feature vector. SVM is trained to distinguish acoustic features of a subject (+1) and all other acoustic features (-1) in the training set. One SVM is created for each subject. For evaluating the performance of the system, an AVI file of 10 sec duration at 12 fps is recorded, one month after collecting the training data. For testing the identity claim of a subject, the speech signal is recorded for 10 sec, one month after collecting the training data.

For identification, acoustic feature vectors extracted from the test utterance are given as input to the SVM model and the distance between each of the feature vectors and the SVM hyperplane is obtained (the number of positive matches). The average distance (average the number of positive matches) is calculated for each model. The average distance gives better performance than using distance for each frame. The identity of the subject is decided based on the maximum distance.

For authentication, acoustic feature vectors extracted from the test utterance are given as input to the claimant SVM model and the distance between each of the feature vectors and the SVM hyperplane is obtained. The claim is accepted if the number of positive matches (average distance greater than zero) is greater than a threshold. Performance of text independent speaker identification and authentication using different kernel functions is given in Table 6.4. SVM with gaussian kernel gives
better performance when compared to all other kernel functions as shown in Table 6.4. The system achieves a recognition rate (RR) of 94% an equal error rate (EER) of about 7.42%.

6.4.1.3 Radial Basis Function Neural Network Model

The performance of the text-independent speaker verification system is evaluated for television broadcast news audio. The speech signal is recorded for 60 sec at 8000 samples per second. For enrolling a speaker, the differenced speech signal is analyzed by dividing it into frames of 20 msec, with a shift of 10 msec. A 39th order MFCC is used to capture the static and dynamic features of a speech signal spectrum as described in Section 6.2 which contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration coefficients. Altogether, a 39 dimensional MFCC for each frame is used as a feature vector. A 39 dimensional acoustic feature vector is extracted from each speaker and is given as input to the RBFNN model. The RBF centers are located using k-means algorithm. The weights are determined using least squares algorithm. The value of k varies from 1 to 10 in our studies for each speaker. The system gives optimal performance for k=5 as shown in Fig. 6.3. The weight matrix of size $251 \times 50$ is calculated using the least squares algorithm discussed in Chapter 4.3.3 for 50 speakers. For evaluating the performance of the system, an AVI file of 10 sec duration at 12 fps is recorded, one month after collecting the training data. For testing the identity claim of a subject, the speech signal is recorded for 10 sec, one month after collecting the training data.

For identification, acoustic feature vectors extracted from the test utterance are given as input to the RBFNN speaker model. The average output is calculated for each of the output neurons. The identity of the subject is decided based on the highest output. The identification performance is measured in terms of recognition rate (RR).
Table 6.5: Performance of text independent speaker identification and authentication using RBFNN.

<table>
<thead>
<tr>
<th>Television broadcast news audio</th>
<th>RR (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92.0</td>
<td>8.50</td>
</tr>
</tbody>
</table>

For authentication, acoustic feature vectors extracted from the test utterance are given as input to the claimant RBFNN model, and the average output is calculated for the respective neuron. The claim is accepted if the average output of the respective neuron is greater than a threshold; otherwise the claim is rejected. Performance of speaker identification and authentication using RBFNN model is given in Table 6.6. The FAR and FRR curves for the average output of neuron score for RBFNN acoustic model are given in Fig. 6.4. The EER is about 8.50%.

The comparative recognition rate chart for performance of different models (AANN, SVM and RBFNN) of text independent speaker identification is shown in Fig. 6.5. The comparative equal error rate chart for performance of different models (AANN, SVM
Fig. 6.4: FAR and FRR curves for speaker authentication using RBFNN.

Fig. 6.5: Comparative RR chart for performance of text independent speaker identification.
The performance of text dependent speaker authentication is evaluated in the laboratory environment for 50 subjects, 32 females and 18 males, using a camera with a resolution of $160 \times 120$. The speech signal is recorded at 8000 samples per second, using 16-bits per sample. For enrolling a speaker, the differenced speech signal is analyzed by dividing it into frames of 20 msec, with a shift of 10 msec. The speech signal for the sentence “kindly allow me to transact with my account” is used for training and testing. A 39th order MFCC is used to capture the static and dynamic features of a speech signal spectrum as described in Section 6.2 which contains 13th order static coefficients, 13th order delta coefficients and 13th order acceleration coefficients. Altogether, a 39 dimensional MFCC for each frame is used as a feature vector.

Dynamic time warping (DTW) algorithm [82] is used for identification and authentication. For identification, dynamic time warping template is first created for
Table 6.6: Performance of text dependent speaker identification and authentication using DTW.

<table>
<thead>
<tr>
<th>Model</th>
<th>RR (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTW</td>
<td>98.0</td>
<td>3.10</td>
</tr>
</tbody>
</table>

Fig. 6.7: FAR and FRR curves for text dependent speaker authentication using DTW.

For each enrolled speaker \((S_1, S_2, \cdots, S_{50})\) and then test speaker sample \(Q\) is matched against all templates. The Euclidean distance \(D(.)\) measure is used for matching and the speaker with minimum distance is selected as the identified speaker \((Id)\).

\[
Id = \arg \min_{i=1,\cdots,50} \{D(Q, S_i)\} \tag{6.8}
\]

For authentication, dynamic time warping template \(Q_i\) is unknown (test) sample which is matched with enrolled speaker template \((S_1, S_2, \cdots, S_{50})\). The speaker is accepted if Euclidean distance \(D(.)\) is less than a threshold; otherwise the speaker is rejected. The value of threshold can be experimentally determined. The performance of text dependent speaker identification and authentication system using DTW is
evaluated in the laboratory environment and the result is given in Table 6.6. The Euclidean distance for text dependent speaker authentication is normalized from 0 to 1 (1 for minimum distance and 0 for maximum distance). These normalized scores are used in combining the modalities for text dependent audio-video based person authentication system as described in the next chapter. The FAR and FRR curves for text dependent speaker authentication using DTW are given in Fig. 6.7.

6.5 Summary

This chapter describes a method for text independent and text dependent speaker verification. In text independent speaker verification, the acoustic feature vectors were extracted from the speech signal, and its distribution was captured using an AANN model. Separate AANN model was used to capture the distribution of feature vectors of each speaker. The significance of number of units in the compression layer and expansion layer of the AANN model for speaker verification were studied. The classification ability of SVM and RBFNN models were analyzed for text independent speaker verification. The performance of text independent speaker verification system was evaluated for 50 newsreaders. In text dependent speaker verification, the acoustic feature vectors were extracted from the speech signal. Dynamic time warping template was unknown (test) sample which was matched with enrolled speaker template. The speaker was accepted if Euclidean distance was less than a threshold; otherwise the speaker is rejected. The value of threshold could be experimentally determined. The performance of text dependent speaker verification system was evaluated for 50 subjects (real time recorded audio).