Robust Speaker Verification over Narrowband and Wideband Communication Channels

Jozef Polacky

Audiolab, Dept. of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Univerzitna 8215/1, 010 26 Zilina, Slovakia

jozef.polacky@fel.uniza.sk

Abstract. Modern speaker recognition applications involve the authentication of users by their voices. A wide range of systems requires reliable personal recognition techniques to either determine or confirm the identity of a person requesting some type of their services. The main purpose of these techniques is to ensure that provided services are accessed only by a legitimate user and no one else. Voice biometrics for user authentication is a task in which the goal is to perform convenient and secure authentication of speakers. In this work we investigate the use of voiced segments of speech utterances and features normalization techniques on the text-independent speaker verification system based on GMM-UBM approach. A variety of narrowband and wideband codecs were used for simulation real communication channel. Result shows that feature enhancement technique are better for narrowband verification especially.

Keywords

Biometrics, speaker verification, GMM-UBM, codec, voice

1. Introduction

The science of biometrics has been developing approaches that can be used to automatically identify individuals by personal characteristics. Humans have used body characteristics such as face, gait or voice for hundreds of years to recognize each other. Biometric recognition or, simply, biometrics refers to the automatic recognition of persons based on their physiological and/or behavioral characteristics. These characteristics should meet the following requirements:

- each person should have this characteristic
- any two persons should be sufficiently different in terms of the characteristic.
- the characteristics should be enough stable during a longer period of time

However, in systems which use biometrics for personal recognition, there are lots of other issues that should be taken into account, including:

- performance – measurable achievable recognition accuracy and speed
- acceptability - willingness of people to accept the use of a particular biometric characteristic in their daily lives
- robustness - it shows the system’s ability to resist fraudulent methods

All practical biometric systems should meet these specified conditions.

During the previous few years, voice biometrics technology has overcome many obstacles that prevented its wide, trusted use. The number of commercial applications of biometric systems significantly increase nowadays. The current trend of using mobile devices and the internet requires new methods of strong authentication of users to eliminate security gaps. Due to higher users comfort and a risk of the device getting in the hands of unauthorized users classic user-id/password methodology is not very suitable. New progresses in voice biometrics offer great potential for strong authentication using voice. This is of particular interest in the banking area, where financial institutes are looking for ways to offer mobile customers easy authentication while respecting security and reducing fraudulent usage.

2. Biometric systems

A biometric system can be regarded as a pattern recognition system which obtains biometric data from an individual. From these obtained data it is then extracted feature set and subsequently compared against the template set in the database. Automated biometric systems have only become available over the last few decades, due to significant advances in the field of computer processing. Depending on the use of application, a biometric system may work either in verification or identification mode:

- Verification mode – system makes a one-to-one comparison to determine whether the claim is true or not
- Identification mode - The system conducts a one-to-many comparison to establish an individual’s identity
In real applications, the core speaker recognition task is usually defined as a text-independent verification problem. Verification is typically used as positive recognition, for purposes, where the aim is to prevent multiple people from using the same identity [1].

2.1 Voice as biometric

Voice verification techniques use the different characteristics of a person’s voice for differentiation of speakers. These characteristics are based on physiological and behavioral components. The physical shape of the vocal tract forms the primary physiological component. The behavioral component is made up of movement, manner and pronunciation. Physiological characteristics of human speech are invariant for an individual, but the behavioral part of the speech of a person changes over time due to age, medical conditions and emotional state, etc. The combination of these two unique aspects of speaking enable verification of the person who is speaking. Voice verification systems can be used to verify a person’s claimed identity. The downside to the technology is that it is not well suited in a noisy environment. Speaker recognition is most appropriate in phone-based applications but the voice signal over phone is typically degraded in quality by the microphone and the communication channel.

Voice authentication has a number of advantages. The cost of implementation is low because there is no special hardware required. A simple telephone or microphone is all that a user needs to authenticate using her voice. Other methods of biometric authentication like fingerprinting and retinal scans require special devices. Voice authentication is easy to use and easily accepted by users. The concept of identifying people by voices is also quite natural and it is the only biometric that allows users to authenticate remotely [2].

2.2 Biometric system errors

Metrics that reflect accuracy are related to a typical hypothesis testing test (i.e., based on false positives (referred to as false alarms) and false negatives (misses)). We can express the error in the verification system mathematically as follows. Let us stored biometric template of the user I represented by \( X_I \) and the actual input for recognition is represented by \( X \). Subsequently the null and alternate hypotheses are:

- \( H_0 \) - input \( X \) and template \( X_I \) do not come from the same person
- \( H_1 \) - input \( X \) comes from the same person as template \( X_I \)

The associated decisions are:

- \( D_0 \) - person is not who she claims to be
- \( D_1 \) - person is who she claims to be

The decision rule is based on the following arguments. If the matching score \( S(\bar{x}, x_I) \) is greater than the system threshold, then decide \( D_1 \), else decide \( D_0 \). This hypothesis testing formulation contains two types of errors. The first type of error is false match. This means that \( D_1 \) is decided when \( H_0 \) is true. The second type false nonmatch which means that \( D_0 \) is decided when \( H_I \) is true.

False positive rate (FPR) is the probability of the first type error and False negative rate (FNR) is the probability of the second type error as follows:

\[
FPR = P(D_1 | H_0) \\
FNR = P(D_0 | H_I)
\]  

Demands the accuracy of biometric systems are different depending on their use in a particular application. For example, in forensic applications such as criminal identification is FNR very important for the design of the system. On the other hand, FPR is one of the most important factors high security applications. The main aim here is to discourage impostors.

3. Proposed approach

The developed ASV (Automatic Speaker Verification) system is based on conventional Gaussian Mixture Model – Universal Background Model (GMM-UBM) [3] of speech represented by Mel-Frequency Cepstral Coefficients (MFCCs) [4].

3.1 GMM-UBM classification

The GMM-UBM approach is the dominant one in text-independent speaker recognition. The universal background model (UBM) is a large GMM trained to represent the speaker-independent distribution of features. This approach is based on a statistical modeling paradigm, where a hypothesis is modeled by a GMM model:

\[
p(x | \theta) = \sum_{k=1}^{K} p_k \mathcal{N}(x | \mu_k, \Sigma_k)
\]  

where \( K \) is the number of Gaussian components, \( p_k \) is the prior probability (mixture weight), \( \mu_k \) is mean vector and \( \Sigma_k \) are covariance matrices.
is covariance matrix. UBM is used as an initial model for training speaker-specific GMM during speaker enrollment. Adaptation of the parameters is usually done using MAP (a maximum a posteriori approach) technique.

During a test, the system has to determine whether the recording \( X \) was pronounced by a given speaker \( I \). This question is modeled by the log-likelihood ratio:

\[
\log \frac{p(X|\lambda_I)}{p(X|\lambda_{UBM})} \geq t \tag{3}
\]

where \( X \) is the test speech recording, \( \lambda_I \) is the model of the hypothesis that \( I \) pronounced \( X' \), \( \lambda_{UBM} \) corresponds to the model of the negated hypothesis (i.e. \( I \) did not pronounce \( X' \)) and \( t \) is the decision threshold. The model \( \lambda_{UBM} \) is a generic background model.

### 3.2 Codec-degraded speech problem

Speech as a complex signal can be influenced by various external variability. This variability is largely detrimental to the accuracy of speaker recognition. The recorded speech can be changed depending on various factors which are independent of the identity of the speaker, including:

- Channel
- Audio degradation through compression
- Acoustic environment
- Speaker’s physical condition

In our study, we focused on audio degradation through compression. Codec-degraded speech is commonplace in modern communications, such as mobile communication, Voice over Internet Protocol (VoIP), voicemail and gaming communication. In all of the above mentioned cases, lossy speech codecs are deployed. Moreover, it is worth noting here that codec degradations are currently considered as one of the most prominent degradations encountered in current telecommunication networks. Therefore, an analysis of codec-induced degradations in the context of speaker recognition, and development of the ASV techniques that are robust against this type of degradations are of great interest to researchers around the world.

The codecs selected for the experiment represent ones commonly used in current telecommunication networks and also cover all range of degradations currently introduced by narrowband (NB) and wideband (WB) codecs. The selected bit rates cover the most popular ones.

The speech samples were coded by the following codecs at the specified bit rates to introduce codec specific degradations induced when speech is transmitted over telecommunication network:

- G.711 speech codec [5] (a typical PCM (Pulse-Code Modulation) speech codec) operating at 64 kbps
- G.729 speech codec [6] (a very popular parametric codec dominantly deployed in fixed networks) operating at 8 kbps
- AMR-NB [7] speech codec (typically deployed in 3G mobile networks) operating at 5.9 kbps, 7.95 kbps and 12.2 kbps
- EVS speech codec [8] (a brand new 3GPP codec recently standardized by 3GPP and designed to be deployed in 4G (LTE) networks) operating at 5.9 kbps, 8 kbps and 13.2 kbps
- G.711.1 [9] speech codec operating at 96 kbps
- G.729.1 [10] speech codec operating at 32 kbps
- AMR-WB [11] speech codec operating at 6.6 kbps, 8.85 kbps and 12.65 kbps
- EVS-WB [7] speech codec operating at 5.9 kbps, 8 kbps and 13.2 kbps
- Speex [12] speech codec (a codec optimized for a low-latency speech communication over an unreliable packet networks) operating at 27.8 kbps

### 3.3 Feature normalization techniques

#### 3.3.1 Feature warping

For speech, the true distribution of a feature is speaker dependent and multi-modal in nature. However, various channel and additive noise influences can corrupt this distribution. We aim to perform a mapping that will condition the feature distribution.

The main role of feature warping is to create a more robust representation of the cepstral feature distribution. This is obtained by adjusting the individual cepstral feature streams such that they follow a specific target distribution over a window of speech frames. The first step in the warping process is extraction of set of cepstral coefficients from speech utterances. Each coefficient is subsequently analyzed as a freestanding feature stream over warping time. A window of features is processed in the warping algorithm to achieve a mapped feature for the initial cepstral feature in the middle of the window. The sliding window is shifted by a single frame each time and the analysis is repeated [13].

#### 3.3.2 Short - time mean and variance normalization

In the short-time mean and variance normalization (STMVN) technique each feature vector is normalized by the following transformation:

\[
x_{norm}(i) = \frac{x(i) - \mu(t)}{\sigma(t)} \tag{4}
\]

where \( x(i) \) is the \( i \)-th component of the original feature vector at time \( t \). The mean \( \mu(t) \) and standard deviation \( \sigma(t) \) of each feature vector component \( i \) are calculated over a sliding finite length window of length \( N \) as follows:

\[
\mu(t) = \frac{1}{N} \sum_{n=t-N/2}^{t+N/2} x_n(i) \tag{5}
\]
\[ \sigma^2_T(i) = \frac{1}{i+N/2-1} \sum_{n=i-N/2}^{N} (x_n(i) - \mu_T(i))^2 \]  \hspace{1cm} (6)

Details of the calculation of transformation parameters at the beginning and end of recognition are given in [14].

4. Experiment

4.1 Description of the experiment

Four different scenarios formed our experiment with aim to examine an effect of using feature normalization techniques and voiced segments of speech for the speaker verification from codec-based speech. We continue in our experiments published in [15, 16] where we didn’t use any feature enhancement techniques and we have worked with the entire recordings. This scenario is denoted as baseline here. In the scenario 2 and 3 feature normalization techniques mentioned above were used to improve the robustness of cepstral features to channel effects. In last one scenario voiced segments were extracted from the speech wave forms and then parameterized and used for verification process. The voiced segments were detected using combination of the autocorrelation method, subharmonic-to-harmonic ratio, short time energy and zero crossing rate.

Each scenario consists of two versions. In the first version speaker enrollment (training) is performed on clean speech while testing is done on degraded speech affected by a codec (named as fully mismatched case hereinafter). In the fully matched case, both training and testing phases are carried out on speech data coded by the same codec.

In all scenarios, speaker models were obtained by MAP based adaptation of the UBM model, which consists of 512 gaussians. For UBM training, EM algorithm was used, a relevance factor was set to 10, and K-means algorithm with 100 iterations was applied. Note that all of the GMM parameters (i.e. weights, means and covariance matrices) were modified during the adaptation process.

All the experiments were carried out on the TIMIT speech database [17] containing recordings of phonetically-balanced English speech of 630 speakers of eight major dialects regions of the United States (each reading ten phonetically rich sentences resulting in 6300 sentences. TIMIT database, in spite of its original design intention for a speech recognition, is currently also widely used for a speaker recognition. In addition, it suits our intention to built speaker’s model on data of constraint duration (TIMIT provides only several seconds of speech per speaker). In order to maintain sampling frequency of NB speech communication we have down-sampled all of recordings from 16 kHz to 8 kHz via an anti-aliasing low-pass FIR filter with no further processing. Furthermore, the NB and WB speech samples were transcoded (i.e. encoded-decoded) by the codecs listed in section 3.2, at the specified bit rates (to introduce codec specific degradations induced when speech is transmitted over telecommunication network).

For the UBM model training, the utterances from all speakers were used - i.e. 10 clean (uncoded) recordings and 8 transcoded speech recordings per speaker. Note that each of these eight transcoded recordings was transcoded either by different codecs or by the same codec operating at different bit rates, as listed above.

In each scenario, 190 speakers from the first three dialect regions of the “training part” of the TIMIT database were used during speaker enrollment (i.e. for the speaker specific GMM development) and in testing phase. A set of 10 utterances of each speakers was divided into two non-overlapping parts - one half of the utterances was utilized during enrollment and the rest for the testing, and vice versa.

This means that for each scenario, verification trials consist of all possible model-test combinations, resulting in a total of 180500 trials (950 target versus 179550 impostor trials) for each test condition. For each session, AVS performance was evaluated by calculating an EER in percentage.

As a front-end, speech analysis was performed frame-by-frame with 30 ms frame duration and 15 ms overlap, 12 MFCC features (excl. log-energy and 0-th coefficient) were extracted from each speech frame.

4.2 Results

The real verification applications usually don’t have information about the codec that is used for the voice transmission. For this reason it is quite difficult to select a suitable codec for the system training, offering a good performance over wide range of NB or WB codecs currently used in telecommunication networks. For this reason, we cover the two versions mentioned above and we tried to improve baseline systems performance. Table 1 shows the verification performance of the individual narrowband scenarios. It is clearly seen that feature normalization techniques significantly improve the performance of verification system especially in the case of fully matched version. The most significant improvement was achieved in the case of G.729 and AMR-NB codecs. On the other hand when only voiced segments of speech utterance were used in the testing phase worse results were obtained compared to the baseline system. It means that this approach isn’t very suitable for this application.

Results for wideband scenarios are in table 2. We can see that none of the used approaches (scenarios) improve the performance of baseline system in fully mismatch version of experiments. Verification performance increases for AMR – WB and EVS –WB codecs when we used only voiced segments for verification in the case of fully matched version. Feature enhancement methods are benefit mainly in the case of AMR – WB codec similar to the narrowband version.
We have also found that EVS codec achieved very good results in all scenarios. It is very promising result as this codec is going to be widely deployed in voice communication over LTE (a successor of 3G mobile communication system). Therefore a huge amount of the voice communication is going to be coded by this codec in a near future as mobile networks generate a dominant portion of voice communication nowadays.

The overall results are summarized in table 1 and 2. For better clarity we attach graphical representation (Fig.1-2) of the results too.

![Fig. 1. EERs for narrowband/wideband fully mismatched version of individual scenarios](image1)

![Fig. 2. EERs for narrowband/wideband fully mismatched version of individual scenarios](image2)

![Fig. 3. EERs for narrowband/wideband fully matched version of individual scenarios](image3)

### Tab. 1. EER value (in %) for narrowband scenarios

<table>
<thead>
<tr>
<th></th>
<th>G.711</th>
<th>G.729</th>
<th>AMR_6.6</th>
<th>AMR_8.65</th>
<th>AMR_12.65</th>
<th>EVS_5.9</th>
<th>EVS_8</th>
<th>EVS_13.2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>ver.1</td>
<td>4.11</td>
<td>3.16</td>
<td>3.68</td>
<td>3.25</td>
<td>3.21</td>
<td>3.48</td>
<td>3.26</td>
</tr>
<tr>
<td></td>
<td>ver.2</td>
<td>4.37</td>
<td>4.11</td>
<td>10.74</td>
<td>8.22</td>
<td>6.85</td>
<td>8.95</td>
<td>7.01</td>
</tr>
<tr>
<td><strong>CMVN</strong></td>
<td>ver.1</td>
<td>4.95</td>
<td>4.56</td>
<td>5.32</td>
<td>5.37</td>
<td>5.36</td>
<td>6.32</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td>ver.2</td>
<td>4.64</td>
<td>4.84</td>
<td>6.74</td>
<td>5.89</td>
<td>5.68</td>
<td>8.00</td>
<td>7.42</td>
</tr>
<tr>
<td><strong>WARP</strong></td>
<td>ver.1</td>
<td>6.17</td>
<td>5.47</td>
<td>6.63</td>
<td>6.42</td>
<td>6.06</td>
<td>7.53</td>
<td>7.26</td>
</tr>
<tr>
<td></td>
<td>ver.2</td>
<td>6.11</td>
<td>5.70</td>
<td>7.62</td>
<td>7.04</td>
<td>6.99</td>
<td>8.78</td>
<td>8.42</td>
</tr>
<tr>
<td><strong>Voiced</strong></td>
<td>ver.1</td>
<td>5.70</td>
<td>5.26</td>
<td>11.32</td>
<td>9.26</td>
<td>7.54</td>
<td>10.10</td>
<td>7.89</td>
</tr>
<tr>
<td><strong>segments</strong></td>
<td>ver.2</td>
<td>4.80</td>
<td>3.77</td>
<td>4.81</td>
<td>4.42</td>
<td>4.32</td>
<td>4.84</td>
<td>4.51</td>
</tr>
</tbody>
</table>

### Tab. 2. EER value (in %) for wideband scenarios

<table>
<thead>
<tr>
<th></th>
<th>G.711</th>
<th>G.729</th>
<th>AMR_5.9</th>
<th>AMR_7.95</th>
<th>AMR_12.2</th>
<th>EVS_5.9</th>
<th>EVS_8</th>
<th>EVS_13.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ver.2</td>
<td>5.15</td>
<td>29.00</td>
<td>37.74</td>
<td>31.55</td>
<td>45.34</td>
<td>6.53</td>
<td>11.37</td>
</tr>
<tr>
<td><strong>CMVN</strong></td>
<td>ver.1</td>
<td>6.67</td>
<td>7.43</td>
<td>9.16</td>
<td>8.11</td>
<td>7.05</td>
<td>6.54</td>
<td>6.20</td>
</tr>
<tr>
<td></td>
<td>ver.2</td>
<td>7.26</td>
<td>9.30</td>
<td>10.70</td>
<td>9.68</td>
<td>8.33</td>
<td>7.66</td>
<td>7.00</td>
</tr>
<tr>
<td><strong>WARP</strong></td>
<td>ver.1</td>
<td>8.35</td>
<td>9.34</td>
<td>10.42</td>
<td>10.36</td>
<td>8.74</td>
<td>8.62</td>
<td>8.13</td>
</tr>
<tr>
<td></td>
<td>ver.2</td>
<td>9.05</td>
<td>11.05</td>
<td>12.42</td>
<td>11.37</td>
<td>10.11</td>
<td>9.28</td>
<td>9.15</td>
</tr>
<tr>
<td><strong>Voiced</strong></td>
<td>ver.1</td>
<td>10.56</td>
<td>12.63</td>
<td>12.53</td>
<td>12.11</td>
<td>10.59</td>
<td>10.88</td>
<td>10.63</td>
</tr>
<tr>
<td><strong>segments</strong></td>
<td>ver.2</td>
<td>7.76</td>
<td>29.81</td>
<td>37.92</td>
<td>31.66</td>
<td>45.29</td>
<td>7.42</td>
<td>12.67</td>
</tr>
</tbody>
</table>

### 5. Conclusion and future work

According to our results we can conclude that features normalization techniques have a positive impact on the performance of the speaker verification system in the case of narrowband codecs. The best improvement is obtained when G.729 and AMR codecs are used for training and testing. On the other hand these techniques impair performance in wideband fully mismatched version of scenarios. The best results for wideband fully matched version were obtained if voiced segments of speech were used for testing phase. As a future work we plan to extend our study for another features and new classification techniques.
References


About Authors

Jozef POLACKY was born in Povazska Bystrica, Slovakia in the 1988. In 2013 he finished MSc at University of Zilina, Faculty of Electrical Engineering, Department of Telecommunications and Multimedia. Currently he studies doctor degree. His research is focused on speaker recognition for biometric purposes.