Performance Improvement of Text-Independent Speaker Verification Systems Based on Histogram Enhancement in Noisy Environments

C. H. Kwon 1, J. K. Choi 2, E. Ambikairajah 3

1 Department of Information and Communication Engineering, Daejeon University, Korea
2 BNSWorks, Korea
3 School of EE & T, University of New South Wales, Australia

chkwon@dju.ac.kr, u2u2u2u2u2@nate.com, ambi@ee.unsw.edu.au

Abstract

In this paper a histogram enhancement technique is presented in order to improve the robustness of text-independent speaker verification systems. The technique transforms the features extracted from speech such that the contrast of their histogram is enhanced. Experiments showed significant improvements for this technique compared to standard techniques both in clean testing environments, and in the presence of additive noise.

Index Terms: speaker verification, additive noise, robustness, histogram enhancement

1. Introduction

Speaker recognition is a process that necessarily includes both speaker verification and identification [1]. Speaker verification is defined as a decision whether an unknown speech matches the known speech of a speaker whose identity is claimed. Speaker identification is a decision as to whether a target is a specific person or is among a group of persons. In a text-dependent speaker-recognition system, utterances spoken by a claimant are known to the system and can be fixed or prompted. In a text-independent speaker-recognition system, utterances are unknown to the system and are arbitrary. In this paper we deal with text-independent speaker verification systems.

The main challenges in speaker verification are to find the right feature set and to find an optimal modeling method for recognition [2]. Features such as Mel-frequency Cepstral Coefficients (MFCCs) and statistical approaches based on Gaussian Mixture Models (GMMs) have recently dominated the area of speaker verification. The reason for using MFCCs is that they have worked well in speaker verification systems and in speech recognition systems. It has been shown that GMMs outperform phoneme-based hidden Markov models (HMMs) in text-independent speaker verification [3], and have become the dominant approach for modeling speakers in text-independent speaker-verification systems. The dominant approach to impostor modeling is to pool speech from a number of speakers and train a single model, which is referred to as a universal background model (UBM) [4]. In this paper we use MFCCs for the feature set and GMMs for speaker modeling.

Noise strongly degrades the performance of speaker verification systems, as it introduces distortion of the features of speech. The effect of this distortion depends on the speech representation and the type of noise. Additive noise produces a nonlinear transformation of MFCCs. This distortion causes a mismatch between the training and recognition conditions, such that acoustic models trained with clean speech do not model noisy speech accurately. Compensation methods for robust speaker verification mainly focus on minimizing this mismatch. Cepstral Mean Normalization (CMN) and Mean and Variance Normalization (MVN) have been applied to such systems in order to remove the global shift of the mean of the MFCC, and to normalize the variance [5]. CMN makes the mean of the compensated MFCC zero and so equalizes the first moment of its probability distribution. MVN equalizes the first two moments of the distribution, i.e. the mean and the variance.

This paper proposes a method of compensating for noise affecting speech representation. The proposed method is based on a Histogram Equalization (HEQ) method [5][6]. The HEQ method provides a transformation function, mapping the distribution of each component of the feature vector onto a reference distribution. Given a source speech and a reference distribution, the transformation equalizes all the moments of the source probability distribution to those of the reference probability distribution. This can make the transformation more appropriate than CMN or MVN when dealing with additive noise [5]. In doing so, the statistical mismatch between the training and recognition feature distributions can be reduced. The HEQ method has been applied in image processing as a means of improving brightness and contrast in digital images [5]. It is an effective method for correcting images that are either too bright or too dark or that have a poor contrast. Taking this method of image manipulation and extending an analogy to speech recognition, the proposed method is used to get a ‘clearer picture’ of the source speech in the presence of additive noise.

The remainder of this paper is organized as follows: Section 2 describes the contrast for the distributions of features, and a histogram enhancement method is proposed; Section 3 presents experiments and results for the proposed method; and conclusions and further research areas are given in Section 4.

2. A proposed histogram enhancement method

Fig. 1 shows the effect of noise on speech representation. In this case, a speech signal was contaminated with additive white noise at Signal-Noise Ratios (SNRs) ranging from 20 dB to 5 dB. Fig. 1 shows the effect of noise on the first cepstral coefficient. The noise severely affects the probability distributions of the speech, causing a global shift of the mean, and reduction of the variance for the coefficient’s distribution. Similar distortion can be observed for other components. Generally speaking, the effect of the additive noise is to reduce the variance and change the form of the probability distribution and skew the shape of the whole distribution.
Recognition results in speaker verification systems are based on a score or log-likelihood ratio. Given input feature vectors $X = \{x_1, x_2, \ldots, x_T\}$, the likelihood score is described as follows:

$$LLR(X) = \log p(X | \lambda_{\text{target}}) - \log p(X | \lambda_{\text{UBM}})$$

(1)

where $\lambda_{\text{target}}$, $\lambda_{\text{UBM}}$, $\log p(X | \lambda)$, and $LLR(X)$ represent respectively the target model, UBM, log likelihood for target or UBM, and likelihood score. The log likelihood is calculated by the following.

$$\log p(X | \lambda) = \frac{1}{T} \sum_{i=1}^{T} \log p(x_i | \lambda)$$

(2)

where $T$ denotes the length of input feature vectors. The log likelihood of (2) is therefore determined by the frames of values with high frequency near the mean of the feature distributions. From this it follows that when the difference between the means of the target feature distribution and the imposter distribution is large, and the variances of these distributions are small, discriminating scores are given by (1).

It is assumed in the HEQ method that features based on cepstral coefficients generally have a statistical distribution close to a Gaussian distribution [5][6]. Such distributions have values with high frequency near the mean of the distribution, and values with low frequency at the tail ends. We can speak of contrast in this context as the distinction between features, i.e. low contrast means that there is little distinction between features. Contrast is low in regions containing high-frequency values; conversely, contrast is high in low-frequency value regions. Contrast in this paper means the distinction between the features to represent speaker-dependent information. Such contrast is a factor for determining the performance of speaker verification systems.

As mentioned above, additive noise reduces the variance of the feature distribution. Therefore, the features for noisy speech show low contrast and the performance of speaker verification systems is degraded in noisy environments. For this reason the performance of verification systems can be improved if the contrast of speech features is enhanced. In this paper we evaluate the performance of systems with several reference distributions to demonstrate this, and find an optimal reference distribution to enhance low contrast in the region of high-frequency values in the feature distribution.

A histogram enhancement (HEN) technique, proposed in this work, is used to adjust the statistical profile of the features’ dynamic range in order to enhance the contrast of the feature distribution globally. The method is based on the HEQ method, where a Gaussian distribution with zero mean and unity variance is used as the reference distribution for each feature. In the HEN method, we propose a uniform distribution with zero mean and variance equal to three as the reference distribution to adjust contrast of the features. To compare the performance of speaker verification systems with several reference distributions we apply a beta distribution with both alpha and beta equal to 0.5, and a Laplacian distribution with zero mean and variance equal to 0.5 as a contrastive reference distribution. Fig. 2 shows four reference distributions used in this work.

The HEN method transforms the distribution of the original features by matching a cumulative distribution function (CDF) of the original distribution to a CDF of a reference distribution. By comparing the CDFs associated with the distribution of the original features and the reference distributions an adjusted contrast of the features can be seen. Fig. 3 shows the CDFs of four reference distributions. A steep slope of a given CDF means high probability distribution; a gentle slope means low probability distribution. That is, if the slope of the CDF of any reference distribution is larger than that of the Gaussian distribution, that region has values with relatively high frequency and therefore low contrast in speech features. Conversely, a smaller slope than the Gaussian distribution implies high contrast in speech features.
case of the uniform distribution, a lesser slope of the CDF than that of the Gaussian distribution near the mean of the CDF means that the features have high contrast, whilst at both ends the larger slope means that the features have low contrast. In the case of the beta distribution, the features have higher contrast near the mean of the CDF, and lower contrast at both ends of the CDF. The Laplacian distribution is opposite, with lower feature contrast near mean of the CDF, and higher contrast at both ends of the CDF.

3. Experiments and results

In this section, we present our experimental method and results for the proposed method. Experiments were conducted on the ETRI speaker recognition corpora, compiled by a Korean national institution. The ETRI corpora were collected by a Sennheiser MD425 microphone in quiet office environments, and consist of variable length utterances (0.5 – 3 sec) of number sequences and short sentences.

The UBM is a large GMM trained to represent the speaker-independent distribution of the features. Parameters for the UBM were trained using the iterative expectation-maximization (EM) algorithm. One hour of 50 male speech segments, and one hour of 50 female speech segments, with 72 seconds for each speaker, were used for training the UBM, creating a gender-independent UBM. A single UBM with 512 mixtures was trained by pooling all the training data together.

A GMM-based system employing Bayesian adaptation of speaker models from the UBM is referred to as a GMM-UBM speaker verification system [4]. In the GMM-UBM system, we derived the speaker model by adapting the parameters of the UBM using the target speaker’s training speech. We made 50 male and 50 female models adapted from the gender-independent UBM. Each model was adapted with one minute of data, and only means were adapted. Recognition was performed with 15 short sentences for each target speaker. The duration of each test sentence varied from 2 seconds to 3 seconds. The ratio between target and impostor trials is 1:10.

The original speech data recorded in quiet conditions were then contaminated by the addition of several noise types at various SNRs based on NOISEX-92 [7]. Noise types used at this point were white noise, car noise (Volvo), and babble noise, at SNR levels of 5, 10, 15, and 20dB. This task therefore consisted of two kinds of verification experiments: (1) using a verification system trained and tested with clean speech; and (2) using a verification system trained with clean speech and tested with speech contaminated by 3 kinds and 4 levels of noise.

The front-end analysis included several processing steps. First, the speech was segmented into frames by a 25 msec Hamming window progressing at a 10 msec frame rate. A speech activity detector was then used to discard silent frames. Next, 13 MFCCs including a log-spectral energy were extracted from the speech frames. We used the Hidden Markov Model Toolkit (HTK) ver. 3.3 for feature extraction and GMM-UBM modeling.

Score distributions produced for target speakers and impostors by the verification system trained and tested with clean speech are shown in Fig. 4. A basic system and four systems with various reference distributions are shown. Overlapping regions in score distributions means recognition error, half the area of the overlapping region is the equal error rate (EER). The difference between the means of the two score distributions and the sum of variances of those are an index for the performance of the speaker verification system.

Fig. 4 shows that applying a uniform or beta distribution increases the difference between the means and the sum of the variances. Applying a Laplacian distribution, on the other hand, decreases this performance index. Table 1 lists the difference between the means and the sum of the variances for Laplacian, uniform and beta systems, compared to those for a Gaussian system. The Laplacian system deteriorates contrast of the features and is expected to result in worse performance of the speaker verification system.
As outlined above, feature contrast can be observed from the slope of the CDF, as in Fig. 3. The performance of the speaker verification system is improved if the difference between the means becomes large and the sum of the variances becomes small, however the performance of the systems can deteriorate if contrast goes beyond a suitable range. Near the mean of the CDF, the order of magnitude for contrast is highest for the beta distribution, followed by the uniform, Gaussian, and Laplacian distributions. At both ends of the CDF the Laplacian distribution has the highest contrast, and the beta distribution the lowest. According to the difference between the means and the sum of variances in Table 1, the uniform distribution has the greatest magnitude contrast, followed by the beta, Gaussian, and Laplacian distributions.

It is shown in Table 1 that the means of score distributions are increased after the application of the HEN method except in the case of the Laplacian distribution. This leads to a larger separation between target and impostor score distributions. The HEN method also remaps the distribution to improve its shape and scale and to result in improved contrast.

Table 2 shows EERs for various verification systems in clean training and testing environments. The beta system reduces recognition error by 10.38%, and the uniform system 7.64%, when compared to the basic system. The HEN method in this work was applied on a sentence-by-sentence basis and the length of a sentence is very short. It is for this reason that the Gaussian system increases recognition error in clean recognition environments.

Table 2. Comparison of EERs for various verification systems in clean training and testing environments (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>Male</th>
<th>Female</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>6.41</td>
<td>7.47</td>
<td>6.94</td>
</tr>
<tr>
<td>Gaussian</td>
<td>7.76</td>
<td>7.63</td>
<td>7.70</td>
</tr>
<tr>
<td>Laplacian</td>
<td>9.08</td>
<td>8.40</td>
<td>8.74</td>
</tr>
<tr>
<td>Uniform</td>
<td>6.27</td>
<td>6.55</td>
<td>6.41</td>
</tr>
<tr>
<td>Beta</td>
<td>6.0</td>
<td>6.43</td>
<td>6.22</td>
</tr>
</tbody>
</table>

Table 3. Comparison of EERs for various verification systems in clean training and noisy testing environments (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>White noise</th>
<th>Volvo noise</th>
<th>Babble noise</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>39.43</td>
<td>19.34</td>
<td>22.16</td>
<td>26.98</td>
</tr>
<tr>
<td>Gaussian</td>
<td>28.61</td>
<td>8.43</td>
<td>10.79</td>
<td>15.95</td>
</tr>
<tr>
<td>Laplacian</td>
<td>30.08</td>
<td>9.83</td>
<td>12.46</td>
<td>17.46</td>
</tr>
<tr>
<td>Uniform</td>
<td>27.38</td>
<td>8.25</td>
<td>10.32</td>
<td>15.32</td>
</tr>
<tr>
<td>Beta</td>
<td>28.87</td>
<td>7.86</td>
<td>10.34</td>
<td>15.69</td>
</tr>
</tbody>
</table>

Experimental results for clean training and noisy testing environments are shown in Table 3, which shows that there is approximately a 40% improvement in the performance of the speaker verification systems after the application of the HEQ or HEN method. Furthermore, there is a marginal improvement in the EER obtained for the HEN method (the beta and uniform systems) over that obtained for the HEQ method (the Gaussian system).

The beta system increases recognition error by 0.91% in the case of the additive white noise, but reduces it by 6.77% in Volvo noise, 4.17% in babble noise and 1.63% on the average compared to the Gaussian system. The uniform system reduces recognition error by 4.30% in the case of the additive white noise, 2.14% in Volvo noise, 4.36% in babble noise and 3.95% on the average compared to the Gaussian system.

The beta system therefore has the best performance in a clean testing environment, and the uniform system performs best in a noisy testing environment. Additive noise modifies the means and variances for the feature distributions. As the HEN method transforms the distributions in a non-linear fashion it can result in performance degradation. We can see from Table 3 that the uniform system is more flexible to distortion of the feature distributions by the additive noise than the beta system.

4. Conclusions

In this paper we proposed the HEN method to improve the robustness of a speaker verification system in noisy environments. The technique is applied to enhance contrast for the feature distributions and is shown to improve system performance above that obtained for the HEQ method by 3.95%. The results show that the compensation provided by the HEN method is a crucial step in improving the robustness of speaker verification systems in adverse environments.

We did not perform experiments for the HEN method in channel mismatch environments. The HEQ method showed good performance in such environments in the literature, and we expect the HEN method to perform similarly.

5. Acknowledgements

This research was supported by Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation, and also by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-D00741(100101))

6. References