Noise Robust Speaker Verification with Delta Cepstrum Normalization

Naoyuki Kanda, Ryu Takeda and Yasunari Obuchi

Central Research Laboratory, Hitachi Ltd., 1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo, Japan
{naoyuki.kanda.kn, ryu.takeda.qh, yasunari.obuchi.jx}@hitachi.com

Abstract

This paper introduces a delta cepstrum normalization (DCN) technique for speaker verification under noisy conditions. Cepstral feature normalization techniques are widely used to mitigate spectral variations caused by various types of noise; however, little attention has been paid to normalizing delta features. A DCN technique that normalizes not only base features but also delta-features was recently proposed and showed high robustness in speech recognition and language identification. We introduce here DCN for a state-of-the-art speaker verification system that uses iVectors and probabilistic linear discriminant analysis. It is not obvious whether DCN is effective against speaker verification because DCN strongly transforms cepstral features and has the possibility to distort the speaker-specific properties. We compared DCN with cepstral mean normalization (CMN), mean variance normalization (MVN), and histogram equalization (HEQ) using a NIST 2008 SRE dataset with various noise settings, and found that DCN is very effective even for speaker verification. DCN was especially effective under noisy conditions and achieved a maximum 18.5% relative error reduction compared to other competing methods. Combining verification scores from various feature normalization methods further improved the accuracy, and it achieved 9.1% and 16.4% relative error reduction under clean and noisy conditions, respectively.

Index Terms: Speaker verification, histogram equalization, iVector, probabilistic LDA

1. Introduction

This paper describes approaches for speaker verification under noisy conditions. Speaker verification involves determining whether input speech was uttered by a target speaker or another speaker. Many studies on this have been conducted, especially those based on the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) [1]. Elegant super-vector approaches such as iVector extraction [2] and probabilistic linear discriminant analysis (PLDA) [3, 4] have recently been demonstrated to have state-of-the-art performance in this area. These methods work very well under relatively clean conditions and provide a very low error rate in speaker verification. However, real speech in real environments is often deteriorated by various noise sources. It is well known that speaker verification performance severely degrades in noisy environments [5]; therefore, special care is needed to achieve noise robustness of speaker verification.

Cepstral feature normalization techniques are widely used to mitigate spectral variations caused by various types of noise. Many studies on feature normalization techniques have been conducted, especially in the speech recognition area, and well-established techniques such as cepstral mean normalization (CMN) and mean variance normalization (MVN) have been applied for speaker verification [6, 7]. Histogram equalization (HEQ) is also known for its high noise robustness. A sliding window version of HEQ has been widely utilized as feature warping [8]. HEQ enforces the fitting of the distribution of observed cepstral features into the distribution of features of the training corpus. HEQ converts the observed feature vector $y_t$ as,

$$x_t = HEQ(y_t) = C_X^{-1}(C_Y(y_t)).$$

Here, $y_t$ is the observed feature vector at frame $t$, $x_t$ is the normalized feature vector at that frame, and $C_X$ and $C_Y$ respectively indicate the cumulative density function of features in the training corpus and observed features. HEQ and other methods normalize only the cepstral coefficients and do not normalize time-derivative features.

This paper introduces a delta cepstrum normalization (DCN) technique for robust speaker verification. Because delta features are not independent from original features, simply applying HEQ to delta features could deteriorate their properties as differential features. A recently proposed and demonstrated DCN technique that provides a valid extension of HEQ for delta cepstral coefficients showed high robustness in speech recognition [9] and language identification [10]. DCN seems to be promising for speaker verification too, however, a special care is needed in speaker verification because strong cepstral normalization methods often remove not only variations caused by noises but also speaker-specific properties. Experimental studies are needed to evaluate whether the noise robustness obtained by DCN outweighs the expected distortion of speaker-specific properties. We report here our evaluation of DCN using a state-of-the-art speaker verification system with iVector and PLDA. The results indicated that DCN was consistently effective under noisy conditions and it achieved a maximum 18.5% relative error reduction compared to other competing methods. A method that combines various feature normalization methods further improved the accuracy, achieving 9.1% and 16.4% error reduction for clean and noisy conditions, respectively.

2. System description

2.1. Cepstral feature extraction

Input speech is first converted into cepstral features frame-by-frame. In this paper, Mel-frequency cepstrum coefficients (MFCCs) are used as features for speaker verification. After extracting spectral features, we apply the feature normalization method to reduce the statistical difference originating from transmission characteristics and noise. The use of CMN, MVN, HEQ, and DCN is compared in section 3.3. The details of DCN are presented in section 2.4.

Copyright © 2013 ISCA
2.2. iVector extraction
Baum-Welch statistics are first collected from extracted cepstral features with a Gaussian mixture model denoted as a universal background model (UBM). These statistics are represented as a high-dimensional vector called a super-vector. The super-vector is then projected into a fixed-length low-dimensional vector (called an iVector) using an elaborated method based on factor analysis and maximum a posteriori (MAP) estimation [11, 2].

Here, super-vector $M$ and iVector $w$ are correlated as follows,

$$ M = m + Tw. $$

(2)

Variable $m$ represents the mean super-vector between various speech samples. $T$ is a matrix that represents the basis in super-vector space and is called a total variability matrix. In this expression, statistical variations between speech samples are concentrated on low-dimensional vector $w$; therefore, iVector $w$ can be seen as a representative vector of the speech.

2.3. Speaker verification score
Extracted iVectors from two speech samples are then compared based on the PLDA model. The samples are then judged as to whether they were uttered by the same speaker or different speakers. The PLDA model represents iVector $w$ as follows,

$$ P(y) = N(y|y, \Sigma_{ws}), $$

(3)

$$ P(w|y) = N(w|y, \Sigma_{ws}). $$

(4)

Here, $\Sigma_{ws}$ and $\Sigma_{we}$ indicate within-speaker variance and between-speaker variance, respectively. Probabilistic variable $y$ indicates the speaker-inherent feature vector. Based on the above models, a verification score is calculated as follows,

$$ S_{PLDA}(y_1, y_2) = \log \frac{P(w_1|y_1)P(w_2|y_2)P(y_1)dy_1}{P(w_1)P(w_2)} $$

(5)

2.4. Delta cepstral normalization
DCN provides an extension of HEQ for delta (and delta-delta) features. Figure 1 depicts an overview of delta cepstrum normalization. Because delta features are not independent from the original features, simply applying HEQ to delta features could deteriorate their properties as differential features. Various types of delta cepstrum normalizing methods were evaluated in a previous study [9] for their effectiveness in robust speech recognition, and a two-way method called $\Delta$-adjustment achieved the best performance. We introduce this method in this paper. First, HEQ is applied to the input cepstral coefficients $y_t$, which are then converted to normalized cepstral coefficients $z_t$ as

$$ z_t = HEQ(y_t) $$

(6)

Delta-cepstral coefficients are then calculated as follows,

$$ \Delta z_t = \frac{1}{2} (z_{t+1} - z_{t-1}) $$

(7)

DCN is applied to cepstral coefficients $z_t$ so as to reduce the difference between normalized delta features and original delta features, which is called $\Delta$-adjustment, as follows,

$$ z_t = z_t - \alpha (\epsilon_{t+1} - \epsilon_{t-1}). $$

(8)

Here, $\epsilon_t$ indicates the difference between normalized delta features and original delta features at frame $t$.

$$ \epsilon_t = z_t - \alpha (\epsilon_{t+1} - \epsilon_{t-1}). $$

(9)

Finally, delta and delta-delta cepstral coefficients are recalculated by using $z_t$. Variable $\alpha$ in equation (8) is a parameter of DCN and was set to 1 in this paper. If alpha was set to 0, DCN would become equal to HEQ. A method to consider delta-delta features when calculating $\Delta$-adjustment was also proposed in the paper [12], but it showed little improvement; therefore, we used equation (8) for simplicity.

2.5. Combination of verification scores
Different feature normalization methods provide different cepstral features and provides different verification scores described in equation (5). Those scores can be combined to obtain a new verification score. Some studies showed improvements using score combinations based on well-trained logistic regression [13, 14], but we were not able to obtain stable improvements with this method in our preliminary experiment. Instead, simply averaging the verification scores from different feature normalization methods provided very promising results, which are described in section 3.4.

3. Experiments
3.1. Experimental settings
We used the male components of the NIST 2008 SRE database as evaluation data for speaker verification. We selected a task using telephone conversations that were about 5 minutes in duration (telephone-training-and-test set). This dataset comprises 648 male speakers who speak in several languages (i.e., not only in English), and the verification test was conducted 12,922 times. We used the equal error rate (ERR) and detection cost function (DCF) as performance measures. In this experiment, we used classic settings for DCF ($C_{Miss} = 10$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.01$).

To evaluate robustness under noisy conditions, three types of noise were added to the evaluation data: white noise, computer room noise, and exhibition hall noise. The computer room and exhibition hall noise were selected from the JEIDA (Japan Electronic Industry Development Association) noise database [15]. Each type of noise was added at five different SNR levels: 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. As a result, 16 types of conditions (including a clean condition) were prepared and evaluated. Note that we added noise only to the test speaker’s speech. The target speaker’s speech remained clean.

$$ S_{PLDA}(w_1, w_2) = \log \frac{P(w_1|y_1)P(w_2|y_2)P(y_1)dy_1}{P(w_1)P(w_2)} $$

(5)
Table 1: EER and DCF under white noise.

<table>
<thead>
<tr>
<th>Method</th>
<th>EER(%) / DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
</tr>
<tr>
<td>CMN</td>
<td>4.36 / 0.022</td>
</tr>
<tr>
<td>MVN</td>
<td>4.51 / 0.024</td>
</tr>
<tr>
<td>HEQ</td>
<td>4.95 / 0.025</td>
</tr>
<tr>
<td>DCN</td>
<td>4.88 / 0.024</td>
</tr>
</tbody>
</table>

CMN + MVN + HEQ  | 4.08 / 0.019          | 4.59 / 0.025 | 5.75 / 0.029 | 7.78 / 0.035 | 10.43 / 0.047 | 15.49 / 0.058 |
| ALL             | 3.96 / 0.019          | 4.36 / 0.024 | 5.34 / 0.028 | 7.00 / 0.035 | 9.57 / 0.045 | 13.31 / 0.057 |

Table 2: EER and DCF under computer room noise.

<table>
<thead>
<tr>
<th>Method</th>
<th>EER(%) / DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
</tr>
<tr>
<td>CMN</td>
<td>4.36 / 0.022</td>
</tr>
<tr>
<td>MVN</td>
<td>4.51 / 0.024</td>
</tr>
<tr>
<td>HEQ</td>
<td>4.95 / 0.025</td>
</tr>
<tr>
<td>DCN</td>
<td>4.88 / 0.024</td>
</tr>
</tbody>
</table>

CMN + MVN + HEQ  | 4.08 / 0.019          | 4.47 / 0.022 | 4.98 / 0.026 | 6.61 / 0.031 | 9.49 / 0.044 | 14.16 / 0.056 |
| ALL             | 3.96 / 0.019          | 4.27 / 0.021 | 4.82 / 0.025 | 6.44 / 0.031 | 8.95 / 0.042 | 12.45 / 0.054 |

3.2. System settings

We used 250 hours of speech from the NIST 2005 & 2006 SRE databases as development data for the UBM and PLDA models. Any noise was not added to the development data. We used 60 dimensional MFCCs (20 base coefficients and their first and second time-derivatives). Those coefficients were normalized by CMN, MVN, HEQ, or DCN. Then, 400-dimensional iVectors were extracted by using gender-dependent diagonal covariance UBM with 2048 mixtures. The extracted iVectors were normalized by using within-class covariance normalization (WCCN) and length-normalization methods. Finally, a verification score was calculated based on the PLDA model [3]. We used maximum likelihood trained PLDA based on a Gaussian prior.

3.3. Comparison of cepstrum normalization techniques

We compared DCN to CMN, MVN, and HEQ under various noise conditions. The results are presented in Tables 1, 2 and 3. The bottom two rows of each table (under the double line) indicates the results obtained by combining feature normalization techniques, which is discussed in the next subsection.

Table 1 depicts the results of various feature normalization methods under white noise conditions. The best performance in clean conditions was obtained with CMN, which had an EER of 4.36% and a DCF of 0.022. The second best was MVN, which obtained a 4.51% EER and 0.024 DCF. HEQ and DCN showed the worst performance in clean conditions. In contrast, in highly noisy conditions under 10 dB, DCN had the best performance of all methods. Under 0-4 dB conditions, DCN obtained an EER of 13.11%, which corresponded to a 17.0% reduction in relative error over the result obtained using HEQ (second best at 0 dB). CMN had the worst results in noisy conditions under 10 dB.

In summary, the best results in clean conditions were obtained with CMN, and DCN was best in noisy conditions. CMN translates features moderately (only translates the mean of features) compared to other methods; therefore, most of the spectral properties of speakers would be maintained. However, CMN could not mitigate the spectral variance caused by noise, whereas methods that translate features heavily such as DCN, obtained improved results under noisy conditions.

Table 2 lists the results under computer room noise conditions. This noise was almost stationary over time and large portion of the power was concentrated at a low frequency. We first observed that CMN worked best in upper 15 dB SNR conditions. In noisy conditions under 10 dB, DCN again outperformed other methods. DCN achieved 11.98% EER and 0.051 DCF in 0-dB SNR conditions, which corresponds to an 18.0% and 8.2% error reduction compared to CMN (second best in this condition). Note that, unexpectedly, MVN and HEQ could not outperform CMN even in heavy noise conditions such as 0-dB SNR conditions. It indicated that the accuracy gain from noise robustness obtained by those methods could not outweigh the loss caused by the distortion of speaker-specific properties. Such trade-off would not be considered in speech recognition or language identification, in which spectral differences between speakers should be removed by feature normalization.

Finally, Table 3 gives the results under exhibition hall noise. This noise has rapid time-varying properties, and the power spreads from low to high frequencies. In this condition, DCN again showed superior results under low SNR conditions. CMN was the best in relatively clean conditions. The other methods almost always performed somewhere between CMN and DCN. DCN achieved 9.31% EER and 0.048 DCF in 0-dB SNR conditions, which corresponds to an 18.5% and 6.3% error reduction compared to CMN (second best in this condition).

The results of these experiments indicate that DCN is consistently effective for speaker verification under noisy conditions. Other feature normalization methods such as MVN and HEQ sometimes could not show their noise robustness. Note that DCN was better than HEQ in almost all conditions¹. These results suggested that the deterioration of delta features by using HEQ was critical for speaker verification and a Δ-adjustment in DCN could appropriately restore the properties of delta features.

¹The only exception was 20-dB SNR condition under exhibition hall noise.
### 3.4. Evaluation of combined method

The score-combination method described in section 2.5 was evaluated in three noise conditions. The results are given in the bottom two rows of Tables 1, 2, and 3. The combination of CMN, MVN, and HEQ was first evaluated (shown as CMN+MVN+HEQ). This method showed consistent improvement compared to single methods; however, the results was still worse than DCN under 5-dB conditions. These results again indicated the superiority of DCN in terms of noise robustness. Adding DCN to the combination (shown as ALL) provided further improvements in all conditions. In clean conditions, the ALL score-combination method achieved 3.96% EER and 0.024 DCF, which corresponds to 9.1% and 10.3% relative error reductions from CMN (best among single methods). This method also achieved a large improvement in moderate noise conditions. For example, in 20-dB conditions under white noise, the score-combination method achieved 4.36% EER and 0.024 DCF, which corresponds to 16.4% and 11.6% relative error reductions from CMN (single best in this condition). Under heavy noise conditions such as 0-dB SNR conditions, the combination method provided almost the same (sometimes, little bit worse) results as those for DCN. This suggests that DCN provided much better results than the other methods, and the results from the other methods did not make a noticeable contribution.

### 4. Related work

The work presented here is focused on the robustness of speaker verification under noisy conditions. Approaches to improve noise robustness can be categorized into two main approaches, which extract cepstral features, and back-end approaches, which calculate verification scores (including i-vector extraction). Front-end approaches include noise reduction [5], reverberation mitigation [17], and robust feature extraction [18, 19]. Back-end approaches include multicondition PLDA [7, 20]. The work reported in this paper belongs in the front-end category and is focused especially on the cepstral feature normalization method. The contributions of this paper are summarized as follows: (1) it introduces delta cepstrum normalization (DCN) for state-of-the-art speaker verification and describes its superiority under various noise conditions, and (2) it presents an experimental analysis of a combination of various feature normalization methods.

DCN provides a valid extension of HEQ for delta features. The RASTA method [21] and other filtering approaches make use of inter-frame information, but they do not use the entire distribution of delta parameters. The vector Taylor series (VTS) [22] is known for its high noise robustness in speech recognition. An experimental study on speech recognition showed that DCN could be used in combination with VTS [9]. In the future, we plan to do an experimental study on speaker verification technique.

### 5. Conclusion

This paper introduced delta cepstrum normalization (DCN) for speaker verification under noisy conditions. We compared DCN with CMN, MVN, and HEQ using the NIST 2008 SRE dataset and found that DCN achieved a maximum 18.5% relative error reduction on EER compared to other competing methods. Combining verification scores from various feature normalization methods further improved the accuracy; it achieved 9.1% and 16.4% relative error reduction under clean and noisy conditions, respectively.

### 6. References


