A Comparative Study of Feature and Score Normalization for Speaker Verification

Rong Zheng, Shuwu Zhang, and Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{rzheng, swzhang, xubo}@hitic.ia.ac.cn

Abstract. In speaker verification, it is necessary to reduce the influence of different environmental conditions. In this paper, two stages of normalization techniques, feature normalization and score normalization, are examined for decreasing the mismatch between training and testing acoustic conditions. At the first stage, cepstral mean and variance normalization (CMVN) is modified to normalize the cepstral coefficients with the similar segmental parameter statistics. Next, due to score variability between verification trials, Test-dependent zero-score normalization (TZnorm) and Zero-dependent test-score normalization (ZTnorm) are comparatively presented to transform the output scores entirely and make the speaker-independent decision threshold more robust under adverse conditions. Experiments on NIST2002 SRE corpus show that the normalizations with CMVN in feature stage and ZTnorm in score stage achieved 20.3% relative reduction of EER and 18.1% relative reduction of the minimal DCF compared to the baseline system using CMN and zero normalization.

1 Introduction

For speaker verification tasks, it is necessary to extract information from the speech signal that is speaker specific and robust to noise and various channels. Under adverse conditions, speech features based on cepstral analysis are corrupted. If there is a big mismatch between training and testing speech, the performance of speaker verification system would be deteriorated.

In recent years, some robust feature normalization methods for reducing noise and/or channel effects were proposed, such as CMN, Cepstral Mean and Variance Normalization (CMVN) [1,2], RASTA [3], and feature warping [4]. CMN is proposed to compensate for linear channel variations. When additive noise exists, a natural extension of CMN is CMVN, which normalizes the distribution of cepstral features over some specific window length by subtracting the mean and scaling the standard deviation. RASTA is a kind of modulation spectrum analysis that aims to reduce the effects of convolutional noise in the communication channel. Feature warping algorithm maps the distribution of a cepstral feature stream to a standard normal distribution over a specified time interval.

In output score domain, score normalization has been introduced to deal with score variability and to make speaker-independent threshold more robust and effective. In previous work, the use of score normalization has significantly improved the performance of speaker recognition systems, such as cohort normalization [5], Znorm (Zero normalization) [6], Hnorm (Handset normalization) [7], Tnorm (Test normali-

zation) [8]. Score normalization is the transformation of the distribution of speaker verification scores to enhance the robustness of decision threshold. Hnorm is derived from Znorm. Both Znorm and Hnorm compute scores from a set of impostor speech segments to normalize a speaker model. In Tnorm, each test segment is scored against a set of impostor models at test time. Tnorm parameters are estimated to normalize the score of the test segment against claimed model. Znorm and Tnorm are the mostly used forms of score normalization.

To enhance the robustness of GMM-based text-independent speaker verification systems under adverse conditions, we compare and evaluate a number of feature and score normalization approaches. By appending different types of delta cepstral coefficients, CMVN is improved to be more robust against the effects of linear channel and slowly varying additive noise. After using more robust features, Test-dependent zero-score normalization (TZnorm) and Zero-dependent test-score normalization (ZTnorm) are comparatively presented for score normalization to cope with score variability caused by different speaker models and various test segments.

The remainder of this paper is organized as follows. Section 2 describes the feature normalization. The score normalization is introduced in section 3. The experimental results based on NIST2002 SRE corpus are discussed in section 4. Finally, we will draw a conclusion in section 5.

2 Feature Normalization

In telephony speaker recognition applications, channel variability is the most significant factor decreasing the recognition performance. In order to fit a speaker model to a given speaker, not to background noise, silence and noisy frames are usually removed.

Alternatively, robust speech recognition techniques have been introduced to reduce the effect of linear channel and slowly variable additive noise. Feature warping and CMVN algorithms were proposed to normalize distribution parameters of cepstral coefficients over specified time interval using sliding windows. In feature warping method [4], mapped features of short-time distribution of acoustic features are quickly found by looking up in a standard normal cumulative distribution function table. It has been proven to outperform CMN. A natural extension of CMN, CMVN, improves normalization of the mean and variance parameters. CMVN yields a better compensation of the mismatch caused by additive noise [4,9].

The use of time derivative cepstral parameters improved recognition performance in previous studies. In original CMVN algorithm, it was reported to only append corresponding delta coefficients after static coefficients have been normalized [1]. If we use NMFCC to represent MFCC coefficients after CMVN performed, then static and delta cepstral coefficients through original CMVN can be represented as $\text{NMFCC} + \Delta \text{NMFCC}$. All of the feature normalizations mentioned above assumes that the mean and/or the variance, even all the moments of the probability distribution of the cepstral coefficients [10], are irrelevant information for speaker recognition. Since the separation of relevant and irrelevant information contained in the speech is not clear, it is hard to theoretically judge the proper normalization strategies for delta

coefficients. We further evaluate two types of additional derivation of delta coefficients compared with the original CMVN method. In figure 1 (a), delta coefficients derived from NMFCC are further processed by CMVN, which is denoted as NMFCC+N Δ NMFCC. In figure 1 (b), CMVN is applied to static and delta coefficients, respectively. The cepstral coefficients are expressed as NMFCC+N Δ MFCC.

Fig. 1. (a) Delta coefficients derived from NMFCC are further processed by CMVN. (b) CMVN is applied to static and delta coefficients, respectively.

3 Score Normalization

Decision-making process in speaker verification based on GMM-UBM (Gaussian Mixture Models-Universal Background Model) is to compare the likelihood ratio obtained from the claimed speaker model and the UBM model with a decision threshold [7]. Due to the score variability between verification trials, the tuning of decision thresholds is an important and troublesome problem. Score variability mainly consists of two different sources. The first one is the different quality of speaker modeling caused by enrollment data varying. The second one is the possible mismatches and environment changes among test utterances.

Znorm [6] calculates the scores of the target speaker model k against a set of impostor speech utterances. Then the mean μ_k^Z and standard deviation σ_k^Z of these scores are estimated to normalize the target speaker score $S(x, k)$ computed from each test segment x against this target model.

$$S_{znorm}(x, k) = \frac{S(x, k) - \mu_k^Z}{\sigma_k^Z}. \quad (1)$$

Tnorm [8] parameters are estimated from scores of each test segment x against a set of impostor speaker models at test time. Then the mean μ_x^T and standard deviation σ_x^T of the impostor scores are used to adjust the target speaker score $S_{tnorm}(x, k)$.

$$S_{inorm}(x, k) = \frac{S(x, k) - \mu_x^T}{\sigma_x^T}. \quad (2)$$

Znorm is mostly utilized to scale various output scores caused by different speaker models, and Tnorm is to transform output scores caused by various test utterances. These two normalization methods have proven to be effective. Some comparative experiments are shown in [9].

In order to enhance the robustness of decision threshold and normalize the uncertainty of score variability between trials entirely, two kinds of combination mode, Test-dependent zero-score normalization (TZnorm) (Figure 2) and Zero-dependent test-score normalization (ZTnorm) (Figure 3), are presented as follows.

Fig. 2. Schematic diagram of Test-dependent zero-score normalization

Figure 2 shows a block diagram of Test-dependent zero-score normalization (TZnorm). In the application of TZnorm, we first compute the log-likelihood ratio scores for verification trial (i) against a target speaker model (k) and against impostor model set (1,2,...,M). We get the score of target speaker S_{ki} and a set of impostor scores $S_{li}, S_{2i}, \dots, S_{Mi}$, respectively. Then, S_{ki} and $S_{li}, S_{2i}, \dots, S_{Mi}$ perform zero normalization processing separately using impostor speech set (1,2,...,N). We assume $S_{li}^Z, S_{2i}^Z, \dots, S_{Mi}^Z$ have a Gaussian distribution and estimate the mean μ_i^Z and standard deviation σ_i^Z . Finally, TZnorm normalized output score is obtained as follows.

$$S_{ki}^{TZ} = \frac{S_{ki}^Z - \mu_i^Z}{\sigma_i^Z}. \quad (3)$$

Figure 3 shows a schematic diagram of Zero-dependent test-score normalization (ZTnorm). When considering ZTnorm, we first compute the log-likelihood ratio scores for speaker model (k) with verification trial (i) and impostor speech set (1,2,...,N). S_{ki} and $S_{k1}, S_{k2}, \dots, S_{kN}$ are obtained respectively. Then, S_{ki} and

$S_{k1}, S_{k2}, \dots, S_{kN}$ perform test normalization processing using impostor model set $(1, 2, \dots, M)$ separately. Supposing $S_{k1}^T, S_{k2}^T, \dots, S_{kN}^T$ has a Gaussian distribution, the mean μ_k^T and standard deviation σ_k^T are estimated. The output score of ZTnorm is derived from the following equation 4.

$$S_{ki}^{ZT} = \frac{S_{ki}^T - \mu_k^T}{\sigma_k^T}. \quad (4)$$

Fig. 3. Schematic diagram of Zero-dependent test-score normalization

4 Experimental Results

The described speaker verification system has been developed and evaluated using data from NIST2001 and NIST2002 speaker recognition evaluations (SRE) [11]. In this section, we will describe the corpus, the baseline system and the performance of speaker verification based on feature normalization and score normalization, respectively.

4.1 Corpus

The NIST2002 corpus includes cellular data extracted from the Switchboard Cellular part 2. This corpus consists of 139 male and 191 female speakers with 2 minutes of training speech from a single cellular phone call. There are 3570 test segments (1442 males and 2128 females) which range from few seconds to a minute (with a primary focus on segments with 15s~45s). All of the verification utterances have to be scored against gender-matching impostors and true speaker. A detailed description of the evaluation corpus can be found in [11].

NIST2001 SRE is used to train universal background model and to collect the needed speech for detection score normalization. This corpus includes 60 development speakers (38 males and 22 females), 174 test speakers (74 males and 100

females) and a total of 2038 verification speech segments (850 males and 1188 females). To avoid bimodal distribution, the impostor data for score normalization is of the same gender as the target speaker [6].

4.2 Baseline System

Speech utterances are divided into 24ms frames, shifted by 12ms, ignoring about 10%~15% low energy frames which do not contain much speaker information. 16 Mel-Frequency Cepstral Coefficients (MFCCs) and 16 delta coefficients are calculated. CMN is applied to mitigate linear channel effects. Znorm is performed to transform output scores. Two gender-dependent Universal Background Models are trained on the 60-speaker development set of the NIST2001 SRE. Each gaussian mixture in the UBM has a diagonal covariance matrix. All speaker models are obtained by Bayesian adaptation of the UBM. Details of the adapted GMM-UBM system can be found in [7].

Performance measure is evaluated using Detection Error Trade-off (DET) curves [12] and Detection Cost Function (DCF) [11]. The DCF is defined to weight two types of errors, that is, miss detections and false alarms. For all results, we report the Equal error rate (EER) and the minimal DCF obtained in a posteriori way.

4.3 Comparisons of Feature Normalization and Score Normalization

For CMVN processing, we have evaluated three types of combinations of static and delta coefficients mentioned in section 2. For short,

- (1) CMVN1: $\text{NMFCC} + \Delta \text{NMFCC}$; (2) CMVN2: $\text{NMFCC} + \text{N} \Delta \text{NMFCC}$; (3) CMVN3: $\text{NMFCC} + \text{N} \Delta \text{MFCC}$.

In our experiments, feature warping is also applied to the cepstral coefficients to obtain corresponding normalized feature parameters. We only report the best experimental results. The length of sliding windows is the same as in [4].

Figure 4 (a) shows the DET curves of different types of feature normalization for speaker verification: CMN, Feature warping, and three kinds of cepstral coefficients derived from CMVN. Table 1 gives the corresponding EER and the minimal DCF. It has shown that CMVN3+Znorm brings 12.9% relative reduction of EER and 4.6% relative reduction of the minimal DCF compared to the baseline system. So this type of robust cepstral parameters is used in the following comparative experiments on score normalization. It is also notable that three kinds of feature normalization strategies for delta coefficients caused different recognition results. CMVN3 strategy produced modest improvement.

Speaker verification performance using various score normalizations based on CMVN3 features is shown in Figure 4 (b). Corresponding EER and the minimal DCF are reported in Table 2. The results showed that recognition performance is significantly improved by employing the combined modes of score normalization. CMVN3+ZTnorm yields the best performance, which achieves 20.3% relative reduction of EER and 18.1% relative reduction of the minimal DCF compared to the baseline system using CMN and zero normalization.

Fig. 4. (a) Speaker verification performance for different types of feature normalization. The minimal DCF operating point is indicated with a circle. (b) Speaker verification performance for different types of score normalization.

Table 1. EER and minimal DCF for different types of feature normalization

Feature normalization	EER (%)	DCF
CMN+Znorm (Baseline)	10.8	0.0457
FeatureWarping+Znorm	9.7	0.0472
CMVN1+Znorm	9.9	0.0439
CMVN2 +Znorm	9.7	0.0437
CMVN3 +Znorm	9.4	0.0436

Table 2. EER and minimal DCF for different types of score normalization

Feature normalization	EER (%)	DCF
CMN+Znorm (Baseline)	10.8	0.0457
CMVN3+Znorm	9.4	0.0436
CMVN3+Tnorm	12.0	0.0410
CMVN3 +TZnorm	9.4	0.0391
CMVN3 +ZTnorm	8.6	0.0374

5 Conclusions

In this paper, we have compared two different transformation methods, feature normalization and score normalization respectively, for speaker verification tasks over cellular data. Experimental results showed that fine-tuning derivation of delta coefficients when normalizing local distributions of cepstral parameters improves the recognition performance. TZnorm and ZTnorm detection score normalization have been

presented to cope with the variability of output score distributions entirely. The performance of speaker verification system has been significantly improved by combining modified CMVN with ZTnorm.

References

1. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, Vol.25 (1998) 133-147
2. Segura, J.C., Benítez, C. et al.: Cepstral domain segmental nonlinear feature transformations for robust speech recognition. *IEEE Signal Processing Letters*, Vol.11 (2004) 517-520
3. Hermansky, H., Morgan, N.: RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, Vol.2 (1994) 578-589
4. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. *Proc. Speaker Odyssey Conf.* (2001) 213-218
5. Rosenberg, A.E., DeLong, J. et al.: The use of cohort normalized scores for speaker verification. *Proc. ICSLP Vol.2* (1992) 599-602
6. Reynolds, D.A.: Comparison of background normalization methods for text-independent speaker verification. *Proc. EuroSpeech* (1997) 963-966
7. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, Vol.10 (2000) 19-41
8. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, Vol.10 (2000) 42-54
9. Barras, C., Gauvain, J.: Feature and score normalization for speaker verification of cellular data. *Proc. ICASSP Vol.2* (2003) 49-52
10. Molau, S., Pitz, M., Ney, H.: Histogram based normalization in the acoustic feature space. *Proc. ASRU* (2001) 21-24
11. NIST SRE plans. <http://www.nist.gov/speech/tests/spk/index.htm>
12. Martin, A., Doddington, G. et al.: The DET curve in assessment of detection task performance. *Proc. EuroSpeech* (1997) 1895-1898