Scientific Research Publishing

Journal of Computer and Communications, 2020, 8, 80-87
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Speaker Verification Based on Log-Likelihood Score Normalization

Wei Cao, Chunyan Liang*, Shuxin Cao

College of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: 846102759@qq.com, *liangchunyan_sdut@163.com, 601933697@qq.com

How to cite this paper: Cao, W., Liang, C.Y. and Cao, S.X. (2020) Speaker Verification Based on Log-Likelihood Score Normalization. *Journal of Computer and Communications*, **8**, 80-87.
https://doi.org/10.4236/jcc.2020.811006

Received: October 26, 2020
Accepted: November 21, 2020
Published: November 24, 2020

Abstract

Due to differences in the distribution of scores for different trials, the performance of a speaker verification system will be seriously diminished if raw scores are directly used for detection with a unified threshold value. As such, the scores must be normalized. To tackle the shortcomings of score normalization methods, we propose a speaker verification system based on log-likelihood normalization (LLN). Without a priori knowledge, LLN increases the separation between scores of target and non-target speaker models, so as to improve score aliasing of "same-speaker" and "different-speaker" trials corresponding to the same test speech, enabling better discrimination and decision capability. The experiment shows that LLN is an effective method of scoring normalization.

Keywords

Speaker Verification, Score Normalization, Log-Likelihood Normalization, Zero Normalization, Test Normalization

1. Introduction

Speaker recognition uses information contained in speech waves to recognize a speaker's identity [1]. Due to advantages such as convenience of speech acquisition, simple collection equipment, and remote recognition through a network, speaker recognition is becoming a major means of biometric recognition [2].

Depending on the purposes of recognition, speaker recognition uses the approaches of speaker identification and speaker verification [3] [4]. Speaker identification finds a speaker whose voice matches a test speech from a given set of speakers. Speaker verification determines whether a test speech belongs to a claimed speaker, *i.e.*, it makes a decision of "true" or "false" for a trial which consists of a test speech and the identification of a claimed speaker.

For speaker verification, there are great differences in the score distributions of trials from the following aspects [5].

1) Inconsistency of a given speaker. Due to factors such as time, health, mental state, recording, and conditions, the scores of different test speeches of the same speaker on the target speaker model are not constant, but follow certain probability distributions.

2) Inconsistency between speakers. Due to the influence of speaking habits, voice, language, and other factors, the scores of trials of different speaker models are inconsistent. The scores of trials associated with some speaker models are generally high, while those of others are relatively low.

3) Inconsistency between test speeches. Under influences such as time duration, environmental noise, and channel conditions, the scores of trials associated with different test speeches show inconsistency. Scores of trials associated with some test speeches may be generally high, and those of other test speeches low. Some test speeches have similar scores on target and non-target speaker models, which makes it difficult to distinguish scores.

For the above reasons, if the scores of all trials are aggregated, then the scores of "same-speaker" and "different-speaker" trials show severe overleap and aliasing; in this case, to use a single threshold to make a true or false decision on trials will seriously affect the performance of a speaker verification system [6]. Therefore, it is necessary to normalize the raw verification scores [7].

The most commonly used score normalization methods are zero normalization (Znorm) [8], test normalization (Tnorm) [9], and their combination, ZTnorm. These methods normalize the scores of "different-speaker" trials to a distribution with a zero mean and standard deviation of 1, thereby eliminating the differences between speaker models and between test speeches, effectively reducing the aliasing part after aggregating the scores of the two types of trials. Score normalization is generally not limited to a speaker model building method. Both the basic Gaussian mixture model-universal background model (GMM-UBM) [10], the joint factor analysis (JFA) and the total variability factor analysis (TVFA) [11] require normalization of raw test scores. Current score normalization methods are also applicable to verification systems based on the above speaker models.

Most normalization methods normalize the "different-speaker" trial score distribution to reduce the overlap of the aggregated scores of the two types of trials. However, they do not effectively increase the separation between scores of the two types of trials associated with the same speaker model or test speech. Also, these methods require the advance selection of a large amount of non-target speaker speech data to estimate the mean and variance of the scores of the "different-speaker" trials. The quality of the non-target speaker's voice data selection will eventually affect the effectiveness of the score normalization.

To tackle the shortcomings of score normalization methods, we propose a log-likelihood normalization (LLN) score normalization algorithm. By increas-

ing the separation between the scores of target and non-target speaker models for the same test speech, this improves the scores of the two types of trials for the same test speech.

The rest of this paper is organized as follows: Section 2 introduces the speaker verification system. Section 3 describes the proposed LLN score normalization algorithm. Section 4 describes our experiment and its results. We relate our conclusions in Section 5.

2. Speaker Verification System

2.1. Basic Framework

The speaker verification system, as shown in **Figure 1**, is divided into the three parts of feature extraction, model construction, and score decision [12].

2.2. Evaluation Measures

Each test in the speaker verification system makes true and false decisions on a set of trials. To evaluate a "different-speaker" trial as true (non-target speaker is accepted) is referred to as a false alarm; and to declare a "same-speaker" trial to be false (target speaker is rejected) is called a miss. The probabilities of these two kinds of errors are respectively referred to as false-alarm and miss rates.

1) Equal error rate (EER) [13]

In practical applications, the false-alarm and miss rates should be reduced simultaneously. However, the two error probabilities are mutually constrained, as they change in accordance with opposite trends with changing decision threshold value. When the false-alarm and miss rates are roughly the same, the performance of the system is considered to have reached its maximum, and the error rate at this time is referred to as the equal error rate (EER).

2) Minimum Value of Detection Cost Function (minDCF) [14]

Application scenarios have different requirements for false-alarm and miss rates. Thus, the setting of the threshold needs to be adjusted accordingly. To better describe the system performance in different situations, we introduce the

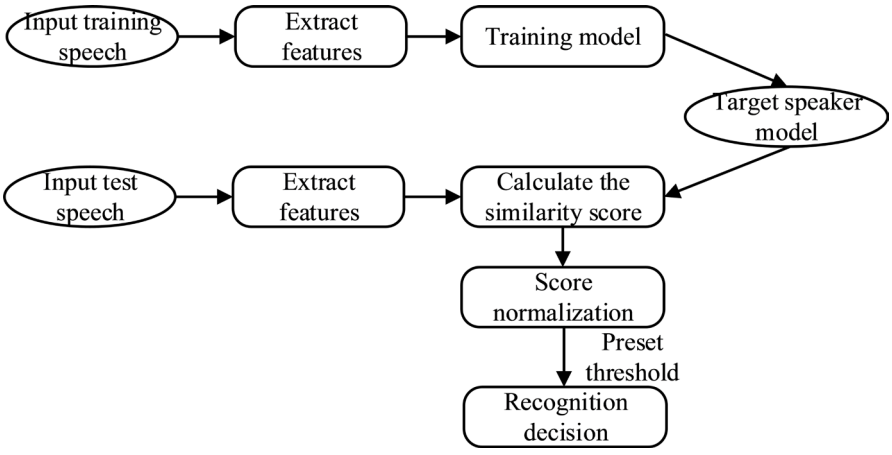

Figure 1. Speaker verification system.

detection cost function (DCF),

$$C_{\text{det}} = C_{\text{M}} \times P_{\text{M|T}} \times P_{\text{T}} + C_{\text{FA}} \times P_{\text{FA|NT}} \times \left(1 - P_{\text{T}}\right) \tag{1}$$

where C_{M} and C_{FA} are the costs associated with the miss rate $P_{\text{M|T}}$ and false-alarm rate $P_{\text{FA|NT}}$, respectively, and P_{T} and $\left(1 - P_{\text{T}}\right)$ are the probabilities that a trial should be classified as true and false, respectively. The detection cost function describes a loss after a recognition error occurs and can represent the performance of a system well. The DCF value corresponding to the system threshold can be obtained, and this can be iterated to obtain a minimum detection cost function (minDCF), which is the most important metric in the National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE).

2.3. Zero Normalization (Znorm) and Test Normalization (Tnorm)

Znorm uses a large number of non-target speaker speeches to score the target speaker model, calculates the mean μ_λ and standard deviation σ_λ of the auxiliary parameters corresponding to the target speaker model λ, and uses the calculated values to normalize the difference in the score distribution. The score normalization is

$$S_\lambda^* = \frac{S_\lambda - \mu_\lambda}{\sigma_\lambda} \tag{2}$$

where S_λ is the raw score of the test speech against the model λ, and S_λ is the normalized score.

Tnorm uses a test speech to calculate scores on a large number of non-target speaker models, and obtains the auxiliary parameters corresponding to the test speech, which are also the mean and standard deviation, and uses the obtained values to reduce the effects of test speech environments on the score distribution. The equation for score calculation is the same as (2).

For the speaker verification system, Znorm parameter calculation is completed in the model training stage, and Tnorm parameter calculation in the test stage. ZTnorm combines the training model and test speech information in the score domain, *i.e.*, the score normalization combines Znorm and Tnorm. These three methods have the shortcoming that they do not effectively increase the separation between the scores of the two types of trials associated with the same speaker model or the same test speech. In addition, a priori knowledge must be introduced. A small part of the training data must be set aside as a development dataset to estimate the parameters required for score normalization. The quality of the selection of the development dataset will affect the final score normalization performance.

3. Log-Likelihood Normalization (LLN)

We propose LLN-based score normalization, which can increase the separation

between the scores of "same-speaker" and "different-speaker" trials for the same test speech to effectively mitigate the score aliasing problem of the two types of trials. The test score can be directly adjusted without a priori knowledge, so there is no need to allocate training data.

Let $\vec{\boldsymbol{S}} = [S_1\ S_2 \cdots S_L]^t$ denote the score of a test speech on all L speaker models. Let S_t be the score of the test speech and its target speaker model, *i.e.*, the "same-speaker" trial score; the remaining $L-1$ scores S_n ($n \neq t$) are the scores of the test speech and the non-target speaker models, *i.e.*, the "different-speaker" trial scores. The score of the test speech on the target speaker model will normally be higher than the scores on the non-target speaker models, *i.e.*, $S_t > S_n$ ($n \neq t$). Each score is normalized as

$$S_i' = S_i - \ln\left(\frac{1}{L-1}\sum_{j\neq i} e^{s_j}\right) \tag{3}$$

where S_i is the raw score of the test speech on the i-th speaker model, and S_i' is the normalized score. $\ln\left(\frac{1}{L-1}\sum_{j\neq i} e^{s_j}\right)$ is the normalization quantity for the score S_i. Let $N_i = \ln\left(\frac{1}{L-1}\sum_{j\neq i} e^{s_j}\right)$. Then N_i is calculated by the remaining $L-1$ scores.

1) If $i = t$, the value of S_i which represents the "same-speaker" trial score is big, and the value of the normalization quantity N_i is small.

2) If $I \neq t$, the value of S_i which represents the "different-speaker" trial score is small, and the value of the normalization quantity N_i is big.

Thus, for a given test speech, the score normalization with Equation (3) further increases the separation between the scores of the test speech for the target and non-target speaker models so as to better discriminate the scores of "same-speaker" and "different-speaker" trials. This makes it easier to set thresholds and improves system verification performance.

4. Experiment

4.1. Experimental Setup

Experiments were conducted in telephone training (short2-short3) and testing (tel-tel) on the 2008 NIST SRE core test set, and conducted on male and female voice test sets respectively. The male voice test set consists of 12,857 trials, involving 894 test speeches and 648 target speaker models. In the LLN score normalization stage, the score of a trial was obtained using Equation (3) based on the matching scores between the test speech data and all 648 speaker models. The female test set consists of 23,385 trials, involving 1674 test speeches and 1140 target speaker models.

The experiment used the 36-dimensional Mel frequency cepstral coefficient (MFCC) feature [15]. Each frame feature consisted of an 18-dimensional basic cepstral coefficient and its first-order difference (delta). The silent part of the

data was removed using a phoneme decoder to perform voice activity detection (VAD) on the voice data, and 36-dimensional MFCC features were extracted by sliding a 25-ms window by 10 ms. Score normalization methods are universal and are not limited to speaker modeling methods used by a system. Hence the speaker recognition system based on total variability factor analysis-Probabilistic Linear Discriminant Analysis (TVFA-PLDA) [16], which is basic and mainstream, was selected to verify LLN score normalization method. Both the UBM [17] (gender related, mixed number is 1024) and the Total Variability Matrix T [18] (dimension 100, iteration number is 5) used in the system were trained by the telephone voice data of NIST SRE 2004, 2005 and 2006. The female voice contains 755 speaker data, a total of 9854 voice files, and the male voice contains 477 speaker data, a total of 6824 voice files. In addition, 367 male data and 340 female data were selected from NIST SRE 2006 data for Tnorm score normalization, and 280 male data and 340 female data were selected for Znorm score normalization. This essentially ensured only one piece of voice data for each speaker in these two small datasets.

4.2. Experimental Results

Table 1 compares the experimental results of the Znorm, Tnorm and LLN normalization methods on the TVFA-PLDA system. Table 1 shows that LLN has good normalization performance, without the need of development dataset. Compared with the TVFA-PLDA system without normalization, LLN leads to a relative improvement of 19.11% in EER and 15.84% in minDCF for male testing dataset, and 15.68% in EER and 17.87% in minDCF for female testing dataset. Znorm and Tnorm have no significant effect on the performance of the TVFA-PLDA system.

5. Conclusion

Log-likelihood normalization (LLN) was proposed to address the shortcomings of score normalization in speaker verification systems. LLN can further increase the separation between the scores of the test speech for the target and non-target speaker models so as to better discriminate the scores of "same-speaker" and "different-speaker" trials. Therefore, LLN makes it easier to set thresholds to make the decision of "true" or "false" for trials and improves the system performance of

Table 1. Performance comparison of Znorm, Tnorm and LLN on NIST SRE 2008 test set.

System	male		female	
	EER (%)	minDCF	EER (%)	minDCF
TVFA-PLDA	6.54	3.22	8.61	4.14
TVFA-PLDA + Znorm	6.78	3.21	9.20	4.19
TVFA-PLDA + Tnorm	7.09	3.21	8.82	4.19
TVFA-PLDA + LLN	5.29	2.71	7.26	3.40

speaker verification.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 11704229).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Atal, B.S. (2005) Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America*, **52**, 1687-1697. https://doi.org/10.1121/1.1913303

[2] Reynolds, D.A. (1995) Speaker Identification and Verification Using Gaussian Mixture Speaker Models. *Speech Communication*, **17**, 91-108. https://doi.org/10.1016/0167-6393(95)00009-D

[3] Larcher, A., Bonastre, J.F. and Mason, J.S.D. (2013) Constrained Temporal Structure for Text-Dependent Speaker Verification. *Digital Signal Processing*, **23**, 1910-1917. https://doi.org/10.1016/j.dsp.2013.07.007

[4] Yue, X.C. and Ye, D.T. (2001) A Survey: Text-Independent Speaker Identification. *Pattern Recognition and Artificial Intelligence*, **14**, 194-200.

[5] Singh, S. (2018) Support Vector Machine Based Approaches for Real Time Automatic Speaker Recognition System. *International Journal of Applied Engineering Research*, **13**, 8561-8567.

[6] Chen, J.X., Liu, M.H., Dai, B.Q. and Li, H. (2006) A New Method of Score Normalization for Text-Independent Speaker Verification. *Signal Processing*, **22**, 545-549.

[7] Pisani, P.H., Poh, N., de Carvalho, A.C.P.L.F. and Lorena, A.C. (2017) Score Normalization Applied to Adaptive Biometric Systems. *Computers & Security*, **70**, 565-580. https://doi.org/10.1016/j.cose.2017.07.014

[8] Kenny, P., Boulianne, G. and Dumouchel, P. (2005) Eigenvoice Modeling with Sparse Training Data. *IEEE Transactions on Speech and Audio Processing*, **13**, 345-354. https://doi.org/10.1109/TSA.2004.840940

[9] Ramos-Castro, D., Fierrez-Aguilar, J. and Gonzalez-Rodriguez, J. (2007) Speaker Verification Using Speaker- and Test-Dependent Fast Score Normalization. *Pattern Recognition Letters*, **28**, 90-98. https://doi.org/10.1016/j.patrec.2006.06.008

[10] Zhang, D. (2016) A Robust Speaker Recognition Technique Based on GMM-UBM. Thesis, Huazhong University of Science and Technology, Wuhan.

[11] Stafylakis, T., Alam, M.J. and Kenny, P. (2016) Text-Dependent Speaker Recognition with Random Digit Strings. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, **24**, 1194-1203. https://doi.org/10.1109/TASLP.2016.2546458

[12] Kinnunen, T. and Li, H. (2010) An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Communication*, **52**, 12-40. https://doi.org/10.1016/j.specom.2009.08.009

[13] Li, Y.P., Ding, H. and Tang, Z.M. (2010) A New Score Normalizaion Algorithm Based on EMD-Tnorm for Speaker Verification. *Engineering Science*, **12**, 95-100.

[14] Zheng, R., Zhang, S.W. and Xu, B. (2006) Research on Feature and Score Normalization for Speaker Verification. *Journal of Chinese Information Processing*, **20**, 77-84.

[15] Li, X.H., Dai, B.Q. and Fang, S.W. (2001) The Recognition Performance and Robustness of High Order MFCC for Speaker Recognition. *Signal Processing*, **17**, 124-129.

[16] Wang, W.C., Xu, J. and Yan, Y.H. (2019) Identity Vector Extraction Using Shared Mixture of PLDA for Short-Time Speaker Recognition. *Chinese Journal of Electronics*, **28**, 357-363. https://doi.org/10.1049/cje.2018.06.005

[17] Yessad, D. and Amrouche, A. (2014) Robust Regression Fusion of GMM-UBM and GMM-SVM Normalized Scores Using G729 Bit-Stream for Speaker Recognition over IP. *International Journal of Speech Technology*, **17**, 43-51. https://doi.org/10.1007/s10772-013-9204-6

[18] Laskar, M.A. and Laskar, R.H. (2020) A Fuzzy-Clustering-Based Hierarchical I-Vector/Probabilistic Linear Discriminant Analysis System for Text-Dependent Speaker Verification. *Expert Systems*, **37**, e12524. https://doi.org/10.1111/exsy.12496