2S-Norm: A New Score Normalization for a GMM Based Text-Independent Speaker Identification System: Proceedings of the International Conference, ICERA 2018

Chapter · January 2019

DOI: 10.1007/978-3-030-04792-4_5

1 author:

Huy Van Nguyen
Thainguyen University of Technology
9 PUBLICATIONS **40** CITATIONS

SEE PROFILE

2S-Norm: A New Score Normalization for a GMM Based Text-Independent Speaker Identification System

Van Huy Nguyen(✉)

Thai Nguyen University of Technology, Thai Nguyen, Vietnam
huynguyen@tnut.edu.vn

Abstract. This paper presents a new score normalization method for speaker identification using Gaussian Mixture Model (GMM). The new normalization method has two main advantages: (1) the thresholds are independent to dataset and mapped to the range of $[0\% \div 100\%]$ corresponding to your expected accuracy of the system and (2) better performance comparing to common methods. The experimental results suggest the viability of the proposed approach in terms of shortening the development time and providing regular update for model's parameters.

Keywords: Speaker identification · Score normalization · Decision threshold
2S-Norm · T-Norm · Z-Norm

1 Introduction

Considering a set $\beta = \{\beta_i\}, i = 1, .., N$ in which β_i is a statistical model to represent a known speaker $S_i \in S$, and a group of unknown speakers U represented by β_0, where $U \cap \beta = \varnothing$. With a given acoustic utterance X, in statistical approach, a speaker identification system will decode X as speaker $\hat{S}$ as following [1].

$$\hat{S} = argmax_{\beta} P(X|\beta_i, \beta_0) \tag{1}$$

where $P(X|\beta)$ and $P(X|\beta_0)$ are posteriori probabilities measured on β and β_0 for the given X.

If $\hat{S} = \beta_i$, the system accepts X and identifies X that it was spoken by S_i, otherwise it is rejected as an unknown speaker in U. Assuming that the accuracy of identification for the known speakers is AA (Accurate Accepts), and for unknown speaker is AR (Accurate Rejects). Research efforts are still ongoing in attempts to develop a method which achieves an optimal balance between AA and AR. At present, the common parameterizing method for statistical model (i.e. β_i and β_0) includes Neural Network (NN) and Gaussian Mixture Model (GMM).

Examples of applying NN are (1) the model used to detect any speaker in group S, and (2) larger and balanced training dataset. For both cases, it is always possible to obtain a good quality training database for the group S as the number of known

H. Fujita et al. (Eds.): ICERA 2018, LNNS 63, pp. 9–15, 2019.
https://doi.org/10.1007/978-3-030-04792-4_5

speakers is finite. On the other hand, it is more challenging for group U which generally consists of all speakers who have never been in S.

Comparatively, the GMM model is more suitable to handle the unknown speaker set. By applying GMM to parameterize speaker model with a given input acoustic vector X, the decoder will identify speaker $\hat{S}$ represented by model β_i, if the impostor log-likelihood score meet the condition described in (2) is true:

$$\Lambda(X) = (logP(\mathrm{X}|\beta_i) - logP(\mathrm{X}|\beta_0) > \delta \tag{2}$$

Where: β_i is parameterized by a set $\beta_i = \{w_k, \mu_k, \Sigma_k\}$, $k = 1, \ldots, M$, M is number of mixtures. w_k, μ_k, Σ_k are mixture weight, mean vector and covariance matric. $P(\mathrm{X}|\beta_i)$ is a probability density function described in (3). δ is a threshold,

$$P(\mathrm{X}|\beta_i) = \sum\nolimits_w w_k \mathcal{N}(\mu_k, \Sigma_k) \tag{3}$$

where $\mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} exp\left[-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1}(X-\mu_k)\right]$.

The term δ in Eq. (2) is an experimental and manual parameter. The value range of δ is not the same for every system since the parameters of model β_i depend on the training data and the variants of specific speaker voice. That would be confusing and difficult for developers. The question is that how much the threshold δ should be to stabilize the accuracy in the decoding step for any speaker voice even if their acoustic features at the decoding time have never seen in the training data. To answer that question, score normalization methods have been proposed such as Z-norm [2], T-Norm [3], h-Norm [4], u-Norm [5], and r-Norm [6]. But all these methods could be considered as based on standard deviation normalization theory over a particular training, testing or selected dataset. The basic idea is shown in the Eq. (4). Therefore δ is still a magical and dataset-dependent number.

$$\tilde{\Lambda}(X) = \frac{\Lambda(X) - Mean(\Lambda(X^*))}{Dev((\Lambda(X^*))} \tag{4}$$

where X^* is a specific training, testing or selected acoustic dataset.

In this paper, we propose a new normalization method named 2S-Norm (2 Scores Normalization), consisting of two scores: the identification score (IS) and the confidence score (CS). IS is used to accept an utterance if it was spoken by a known speaker in S or reject otherwise. IS is estimated over a training dataset, similar to the Z-Norm method. CS represents the confidence of a decoding decision, avoiding mistakes where a known speaker was correctly accepted but incorrectly identified as a different individual. IS and CS are first normalized and then mapped to the range of [0% ÷ 100%]. Thresholds will be simple as the choices according to your expected accuracy are in range from 0% to 100%.

This paper is structured as follows. In Sect. 2, a basic idea of speaker identification based on Gaussian Mixture Model is introduced, which is followed by the description of score normalization and common methods in Sect. 3. The detail of the proposed method including definitions and decision algorithm are presented in Sect. 4. Section 5

provides the evaluation results for the new method compared to T-Norm and Z-Norm. Finally, Sect. 6 describes the salient pointer derived from this study.

2 Speaker Identification Based on GMM

A speaker identification system based on GMM model is usually designed in two stages [1]. In the first stage, a speaker-independent background GMM model called universal background model (UBM) is trained using all the speakers' acoustic data. This model is parameterized by the set $GMM_{ubm} = \{w_k, \mu_k, \Sigma_k\}$. It is used to describe common acoustic features of human voices.

Once GMM_{ubm} is trained, in the second stage, speaker-dependent GMM models are trained by adapting GMM_{ubm} to speaker-dependent acoustic data for each target speaker. These model are parameterized by $\beta_j = \{w_k, \mu_k, \Sigma_k\}$ to represent a target speaker S_j, where $j = 1, \ldots, N$.

Within the scope of this study, the toolkit scikit-learn [9] will be applied to estimate and adapt parameters for these models, while our main contribution is on the score normalization phrase, which is expected to make a better decision based on likelihood values which are obtained from the trained GMM_{ubm} and β_j models.

3 Common Normalization Techniques

The most widely used normalization techniques are Z-Norm [2] and T-Norm [3]. The basic idea was described in Eq. (4). The difference between Z-Norm and T-Norm is the way to estimate $Mean(\Lambda(X^*)$ and $Dev(\Lambda(X^*)$. In Z-Norm, after GMM models are trained, all the available impostor utterances are scored against each potential claimant model. The resulting impostor score distribution is used to estimate the parameters. In T-Norm, during the test stage, the test utterance is scored against a pre-selected set of cohort models based on the claimant model. But no matter how $Mean(\Lambda(X^*)$ and $Dev(\Lambda(X^*)$ parameters are estimated, the value of the $\tilde{\Lambda}(X)$ in (4) and the variants of X are still considered in the training dataset during real operations. This results that to set up a threshold in (2) is respected into a specific dataset, and could be different for the same expected accuracy.

4 2S-Norm: More Accurate and Simple Decision Method

4.1 Average Likelihood Distance and Confidence Definitions

Giving a set of acoustic vectors $X_i = \{x_t^i\}, t = 1, \ldots, T$ spoken by speaker $S_i, i = 1, \ldots, N$. Let $P_i^{UBM} = P(X_i|GMM_{UBM})$, and $P_i = P(x_t^i|\beta_i)$ which are computed using the Eq. (3), where GMM_{UBM} and β_i are the trained universal background model and potential claimant model of speaker S_i. An average likelihood distance (ALD) for each speaker S_i is defined in (5). It represents the average likelihood distance from acoustic features of each speaker

S_i to common acoustic features of human voices, and is used as a standard reference at the decoding stage.

$$ALD_i = \frac{1}{T}\sum_{t=1}^{T}(\log P_i - logP_i^{UBM}) \tag{5}$$

An average confidence (AC) for each speaker S_i is defined in (6). This score represents the average likelihood distance from speaker S_i to its most similar speaker S_j.

$$AC_i = \frac{1}{T}\sum_{t=1}^{T}\left(logP_i - \max(\{\log P_j\} : j = \{1,\ldots,N\}\backslash\{i\})\right) \tag{6}$$

where $P_j = P(x_t^i|\beta_j)$.

4.2 2S-Norm Normalization Method

The idea of 2S-Norm normalization is that two scores, the identification score (IS) and the confidence score (CS), are used to make decision. These score are estimated in the decoding stage. For any given acoustic vector X^* spoken by speaker S^*, and trained models GMM_{UBM}, $\{\beta_i\}, i = 1,\ldots,N$, IS and CS are defined in (7) and (8).

$$IS = \frac{logP(X^*|\beta_{i^*}) - \log P(X^*|GMM_{UBM})}{ALD_{i^*}} \times 100\% \tag{7}$$

$$CS = \frac{logP(X^*|\beta_{i^*}) - \max(\{\log P(X^*|\beta_j\} : j = \{1,\ldots,N\}\backslash\{i^*\})}{AC_{i^*}} \times 100\% \tag{8}$$

where: $i^* = argmax_{i\in N}(logP(X^*|\beta_i))$, ALD_{i^*} and AC_{i^*} are estimated in (5) and (6). If IS and CS greater than 100% they will be set to 100%.

The algorithm using IS and CS for speaker identification is given below.

Algorithm 1: Speaker identification using 2S-Norm	
1:	Input: Trained GMM_{UBM}, $\{\beta_i\}$, ALD_i, AC_i, X*.
2:	Output: "Unkown" or $S_{i^*} \in S$.
3:	Determining thresholds δ_{IS}, and δ_{CS} in range of [0% ÷ 100%].
4:	Calculating IS, CS and determining i^* for the input X^* based on equation (7) and (8).
5:	if $(IS > \delta_{IS})$ and $(CS > \delta_{CS})$ then $output = S_{i^*}$ else $output =$ "$Unkown$".

5 Experiment

5.1 Speech Corpus

The IOIT2013 speech corpus [10] was used to test the 2S-Norm normalization method. IOIT2013 was developed by Institute of Information and Technology (IOIT) – Vietnam Academic of Science. It is a reading speech with sentences chosen from 5 M sentences of daily news websites. There are 206 speakers (103 male and 103 female). Four datasets were produced. The first one is a common voice dataset (C-data) used to train UBM model. It was created by randomly selecting 10 unique utterances for each speaker from all speakers in IOIT2013 corpus. IOIT2013 was further divided into two parts. One part consists of 106 speakers, the so-called unknown dataset (U-data), is used to evaluate the rejects accuracy of the proposed method. The remaining part consisting of 100 speakers is also divided into two sub parts with a proportion of 8:2. The bigger subset is used as the training dataset (T-data) for impostor models, and the smaller subset is used as the testing dataset (E-data) for the trained impostor models. All data is stored in wave format with sample rate of 16 kHz and analog/digital conversion precision of 16 bits.

5.2 System Setup

In order to train speaker identification models, a GMM_{UBM} model with 1024 mixtures is firstly trained using the C-data and Scikit-learn framework [9]. Once UBM is trained, each target speaker model is estimated by adapting UBM model to each particular speaker data in the T-data. In this step, only the parameters of mean vectors of impostor models are updated. The impostor speaker models have the same structure as the UBM model. The input feature is Mel-frequency cepstral coefficients (MFCC) extracted with a Hamming window of 32 ms that was shifted at the interval of 16 ms. Each MFCC vector consists of 39 coefficients which are 13 MFCCs, the first and the second order derivatives. All models are trained with 150 iterations of Expectation-Maximization (EM) [11].

5.3 Experimental Results

To test the new normalization method, three testing tasks were performed. The first test is to evaluate the effect of training data duration. In this test, audio files in E-data and U-data are divided to 3-second files. These files are then used as inputs to test the performance of the system. Subsets of T-data were created which were 15s-data, 30s-data, 45s-data and 60s-data according to the total duration of training data for each particular speaker (15, 30, 45 and 60 s respectively). The results of this test are shown in Table 1.

The second test is to evaluate the effect of length for an utterance to get a good accuracy. In this test, the 60s-data was used to train the impostor models. Audio files in the T-data are divided into 3, 5, 7 and 10 s files to produce datasets named as 3s-T, 5s-T, 7s-T and 10s-T. In which, for instance, the 3s-T dataset means that it consists of divided audio files in length of 3 s. The results of this experiment are given in Table 2.

Table 1. Experimental results on duration of training data

	15s-data	30s-data	45s-data	60s-data
U-data	83.11	83.40	84.12	84.12
E-data	75.20	76.01	80.10	82.40
Average	79.16	79.71	82.11	83.26

Table 2. Experimental results on duration of testing data

	3s-T	5s-T	7s-T	10s-T
U-data	84.12	87.91	88.02	88.72
E-data	82.4	85.03	89.71	90.06
Average	83.26	86.47	88.87	89.39

The final test is to compare to T-Norm and Z-Norm methods. This implementation is the same as the second test, except that only the 7s-T was used for decoding, and the normalization method to make decision is replaced by T-Norm, and Z-Norm. Values of thresholds in range of $[5 \div 30]$ with the step of 0.5 were all tested. The thresholds that gave the best results as shown in Table 3 were 10.5 and 18.5 for T-Norm and Z-Norm.

Table 3. Comparison to T-Norm and Z-Nom

	2S-Norm	T-Norm	Z-Norm
U-data	88.02	85.91	85.38
E-data	89.71	87.01	86.17
Average	88.87	86.46	85.78

The performance is evaluated using an accuracy score that is calculated in (9). The thresholds δ_{IS} and δ_{CS} for all experiments are 75%.

$$AC = \frac{Total\ correct\ identifications}{Total\ utterrances} \tag{9}$$

6 Discussion and Conclusion

The results in the Table 3 shown that the 2S-Norm normalization performance is significantly better than T-Norm and Z-Norm methods (about 2–3% in absolute). In the viewpoint of a developer, it is simple to choose thresholds at the decoding stage, because they are mapped to the range of $[0\% \div 100\%]$. They can be considered as the expected accuracy for identification. Another advantage of 2S-Norm is that the same chosen thresholds can be applied for any application even if the training and testing data is different. That is useful when developing real applications for the real world.

The experimental results also showed that the minimum duration of the training data for each speaker should be 45 s to get a good accuracy. For decoding, the length of an input utterance should be more than 3 s that normally can contain 4 Vietnamese syllables. This condition is possible to reach compare to the requirements for speech recognition or neural network based identification systems. Therefore, at the beginning, when developing a speaker identification system, only 45 s of voice recording for each speaker are needed. Furthermore, we can regularly update model parameters without changing the thresholds.

Acknowledgements. This work was supported by Thai Nguyen university of Technology, Vietnam.

References

1. Roberts, W.J.J., Willmore, J.P.: Automatic speaker recognition using Gaussian mixture models. In: Information, Decision and Control on Proceedings, pp. 465–470. IEEE, Australia (1999)
2. Li, K.P., Porter, J.: Normalizations and selection of speech segments for speaker recognition scoring. In: International Conference on Acoustics, Speech, and Signal Processing on Proceedings ICASSP, pp. 595–598. IEEE, New York (1988)
3. Auckenthalera, R., Michael, C., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Sig. Process. **10**, 42–54 (2000)
4. Reynolds, D.A.: Comparison of background normalization methods for text-independent speaker verification. In: EUROSPEECH on Proceedings, Rhodes, Greece (1997)
5. Garcia-Romero, D., Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ortega-Garcia, J.: U-norm likelihood normalization in pin-based speaker verification systems. In: AVBPA on Proceedings, pp. 208–213. Springer, Heidelberg (2003)
6. Vandyke, D, Wagner, G.M.: R-Norm: improving inter-speaker variability modelling at the score level via regression score normalization. In: International Speech Communication Association INTERSPEECH, pp. 1–5. ISCA, New York (2013)
7. Reynolds, D.: A Gaussian mixture modeling approach to text independent speaker identification, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, USA (1992)
8. Reynolds, D., Douglas, A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Sig. Process. **10**, 19–41 (2000)
9. Scikit-learn: Machine Learning in Python. http://scikit-learn.org. Accessed 21 June 2018
10. Luong, C.M.: Vietnam country report 2016: Updated activities on resources development for Vietnamese Speech and NLP. In: Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) on Proceedings. IEEE, Bali, Indonesia (2016)
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. J. Royal Stat. Soc. Series B **39**(1), 1–38 (1977)