CSLT TECHNICAL REPORT-20160020 [Saturday 7th January, 2017]

Feature Transformation For Speaker Verification Under Speaking Rate Mismatch Condition

Askar Rozi[1,2]
, Lantian Li[1,2]
, Dong Wang[1,3]
and Thomas Fang Zheng[1,3*]

*Correspondence: fzheng@tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Speaker verification suffers from serious performance degradation under speaking rate mismatch condition. This degradation can be largely attributed to the spectrum distortion caused by different speaking rates. This paper proposes a feature transform approach which projects speech features in slow speaking rates to features in normal speaking rates. The feature space maximum likelihood linear regression (fMLLR) is adopted to conduct the transform, under the well-known GMM-UBM framework.

The proposed approach has been evaluated on the CSLT-SPRateDGT2016 corpus which consists of normal and slow speech. The experiments show that with the transform, the equal error rate (EER) of the GMM-UBM system was reduced by 19.04% relatively. More interestingly, the transform learned based on the GMM-UBM system can improve i-vector systems as well. Our experiments show a 10.16% relative EER reduction after the transform was applied to the i-vector/PLDA system.

Keywords: speaker verification; speaking rate; fMLLR

1 Introduction

Speaker verification authenticates the claimed identity of a person by speech input. After a decade of research, current speaker recognition (also known as voice print recognition, or VPR) systems have attained rather satisfactory performance, given that the enrollment and test speech are sufficient and the quality is high [1, 2]. However, when there is mismatch between the enrollment and test speech data, the performance is often seriously degraded.

Speaking rate is one particular mismatch that causes the degradation. If a speaker enrolled itself in a normal speaking rate, but test with a slower or faster speech utterance, mismatch occurs. Unfortunately, abnormal speaking rate is often observed in practical systems. For instance, people tend to speak faster if he/she is in a rush state. Conversely, one may speaks slowly due to the exhaustion or illness. Little difference in speaking rate between enrollment and test speech will not be a problem, but substantial mismatch may lead to serious performance degradation. Considering that robustness is a major concern for most applications, it is necessary to study the impact of speaking rate to speaker verification systems.

For fast speech, the speaking rate challenge can be largely attributed to the reduced speaker information caused by shortened speech signals. A possible solution is to change the frame rate so that more data can be accumulated [3]. In our experiments, we found that fast speech caused just minor performance reduction, and so do not deal with it in this study. In contrast, we found slow speech caused clear performance reduction. A thorough analysis shows that when people speak slowly, the spectrum of the speech signals are often damaged seriously. On one hand, this may be attributed to the abnormal articulatory movement caused by the lengthened pronunciation, and on the other hand, it may be caused by the abnormal behavior (e.g., special emotion expression) when a speaker intentionally speaks slowly. Put in another way, the spectrum damage may be not *caused* by slow speaking, but *associated* with slow speaking. The spectrum damage was also found by Zeng et al. [4] in speech recognition.

The research on the speaking rate is still preliminary in speaker recognition. Early studies mainly focused on the impact of speaking rate to speaker recognition systems. For example, performance degradation was confirmed by [5] when mismatch on speaking rate (fast and slow) exists. A similar study was proposed in [6], where mismatch in both channel and speech style (including speaking rate) was studied.

A normalization approach was proposed to mitigate the impact of speaking rate mismatch [7]. In this approach, phoneme duration was used as an extra feature and was augmented to the conventional Mel frequency cepstral coefficients (MFCCs). Their experiments on the YOHO corpus confirmed that with this normalization, both robustness and accuracy of their speaker verification was improved.

This paper proposes a feature transform approach to deal with the speaking rate mismatch problem. The basic idea is to learn a linear transform that maps acoustic features of slow speech to features of normal speech. With this transform studied, test utterances in slow speaking rate can be transformed to utterances in normal speaking rate, therefore mitigating the mismatch between enrollment and test. In this study, the feature space maximum likelihood linear regression (fMLLR) [8] is adopted to train the transform, due to its simplicity and effectiveness. Although the fMLLR learning is based on the Gaussian mixture model-universal background model (GMM-UBM) architecture, the learned transform is model independent and can be applied to speaker verification systems based on any model, for instance the popular i-vector/PLDA architecture.

The rest of the paper is organized as follows: Section 2 introduces the fMLLR approach, and Section 3 describes the experimental settings and results. Section 4 concludes the paper and discusses some future work.

2 Feature space linear transform

In this section, we first analyze the impact of speaking rate on speech features. Based on this analysis, the fMLLR-based feature transform is proposed to address the speaking rate mismatch problem.

2.1 Feature analysis

As has been discussed in Section 1, we argue that a slow speaking rate damage speech spectrum, hence reducing performance of speaker verification systems. To verify this argument, we trained two UBMs (UBM_{norm} and UBM_{slow}) that represent the normal and slow speech, respectively. These two UBMs were adapted from a global UBM via MAP using normal and slow speech respectively. In this MAP adaptation, only the mean vectors were updated. By this configuration, the mean vectors of the two UBMs can be regarded as representations of normal and slow speech. We draw these mean vectors in a two-dimensional space using t-SNE [9], and the results are shown in Fig. 1, where each point represents a mean vector of a particular Gaussian component. It can be seen that the Gaussian components of UBM_{norm} and UBM_{slow} clearly deviate from each other in a systematic way. Particularly, several most significant deviations are aligned to a similar direction. Although this can not be regarded as an evidence that the corresponding Gaussian components are changed in a linear way (due to the nonlinear nature of t-SNE), we can still assume a linear transform that can map speech features of slow speech to those of normal speech.

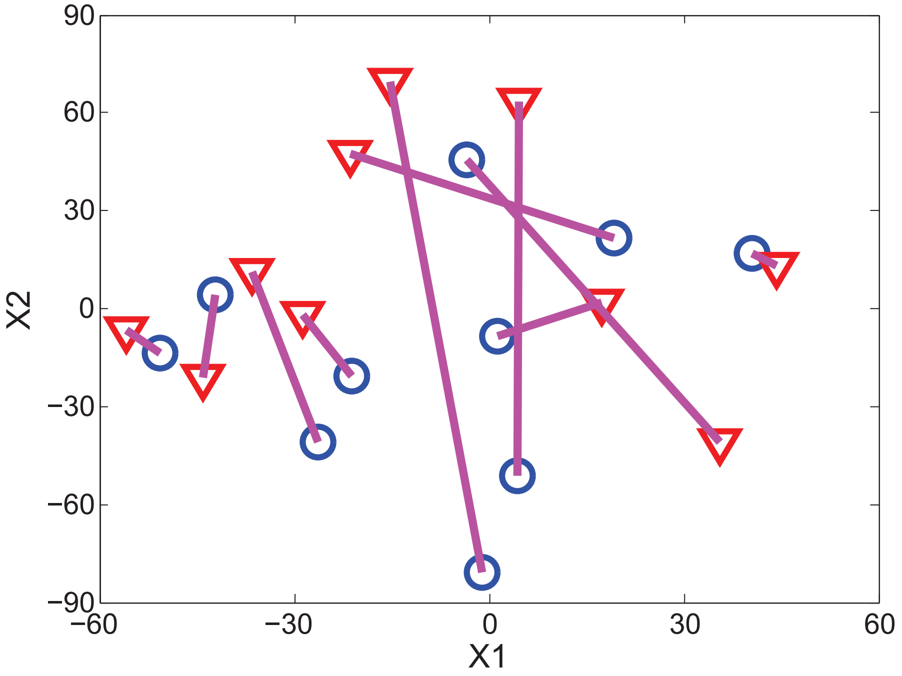

Figure 1 The mean vectors of the Gaussian components of UBM_{norm} and UBM_{slow}, visualized by t-SNE. The circles represent Gaussian mean vectors of UBM_{norm}, and the triangles represent Gaussian mean vectors of UBM_{slow}.

2.2 Feature space linear transform

We design a linear transform that can project speech features of slow speech to those of normal speech. Ideally, different speaking rates should have different transforms as they may cause different damage on speech spectrum. However, training speaking-rate-dependent transforms suffers from data sparsity. We therefore train and apply a single projection for all the slow speech, in spite of its exact speaking rate.

Feature space maximum likelihood linear regression (fMLLR), also known as Constrained Maximum Likelihood Linear Regression (CMLLR), is adopted in this work. The effectiveness of fMLLR has been demonstrated in [10] in which speech features in one language is projected to another language. For a clear presentation, we give a brief introduction to the fMLLR approach. Define a transformation matrix $W = [b \quad A]$ that projects an input speech signal x_i as follows:

$$\hat{x}_i = Ax_i + b = W\xi_i \tag{1}$$

where A is a rotation matrix and b is a bias term. $\xi_i = [1 \quad x_i]^T$ is the extended observation vector. The optimal W can be attained by maximizing the following likelihood function

$$Q(W; X, M) = \sum_i log(p(W\xi_i; M)) \tag{2}$$

with respect to W, where $M = \{\mu_c, \sigma_c\}$ represents the GMM based on which the fMLLR is conducted, and $p(x; M)$ is the probability of signal x given by the GMM parameterized by M. The optimization process can be found in [8].

When applying fMLLR to transform acoustic features, a projection matrix W_{SN} should be learned, where S is the features in slow speaking rate and N is the target features. First of all, a UBM is trained using all the enrollment speech in the development set, via MAP adaptation from a global UBM trained with all the development data. Since all the enrollment speech are in a normal speaking rate, the UBM represents normal speech, and is denoted by UBM_{norm}. Assume X_{slow} denotes the speech features of all slow speech in the development set, the projection matrix W_{SN} is trained by maximizing the objective function $Q(W_{SN}; X_{slow}, UBM_{norm})$ in (2).

It should be highlighted that although the fMLLR transform matrix is trained under the GMM-UBM architecture, the learned transform can be applied to verification systems based on any models, e.g., the i-vector architecture and the recent DNN-based system [11], so long as they use the same features.

3 Experiments

The proposed fMLLR approach was evaluated on a speech database collected by ourselves. We first present the data, and then report the results with the fMLLR transform, on a GMM-UBM baseline and an i-vector baseline.

3.1 CSLT-SPRateDGT2016 database

The goal of this research is to study the impact of speaking rate on speaker verification. Therefore, the variations caused by other factors such as channel and linguistic content should be excluded.

We had several choices at the beginning. For example, the CHAINS corpus [12] that consists of SOLO (in a comfortable speaking rate) and FAST speech recordings. However, there is clear channel variation: they recorded the speech signals with different microphones (Neumann U87 condenser microphone and AKG C420 headset condenser microphone). Furthermore, the SOLO and FAST recordings were recorded within a 4-months interval, which may causes temporal variation. In short, existing databases could not meet our requirements for the speaking rate research,

Table 1 EER results of the GMM-UBM baseline

Condition	Enroll	Test	EER%
Matched	norm.	norm.	2.55
Mismatched	norm.	slow.	7.64

so we decided to record a new database by ourselves. The resultant database was named as CSLT-SPRateDGT2016.

The speech signals in CSLT-SPRateDGT2016 were recorded by a smart phone, with the sampling rate set to 16 KHz and the sample size set to 16 bits. As described above, we want to exclude variations caused by any factors besides speaking rate, so the same phone was used in the entire recording. For the same reason, the transcripts of the enrollment and test speech were identical for all the speakers.

The recordings involve standard Chinese digital strings, and there are 26 speakers in total in the database. Each speaker recorded 5 digital strings for enrollment and 25 digital strings for test. Each digital string of these 25 utterances was recorded in both normal and slow speaking rate. We chose data of 15 speakers from the database as the development set, and data of the rest 11 speakers as the evaluation set. The development set was used to train the feature transform and develop the speaker verification systems, and the evaluation set was used to test system performance.

3.2 Baseline System

The baseline system in this study was based on the GMM-UBM framework. We used 60-dimensional MFCCs as the acoustic feature, which includes 20-dimensional static components and their first and second derivatives. Cepstral mean and variance normalization(CMVN) was used to remove the channel effect.

The UBM was trained using a large volume of data in standard Chinese, using the expectation-maximization (EM) algorithm. The speaker GMMs were adapted from the UBM via MAP. We report the test results in two conditions, the matched condition involves test speech in a normal speaking rate, and the mismatched condition involves test speech in a slow speaking rate. There were 3,025 trials in each condition. The Kaldi toolkit [13] was used to build the system.

Table 1 shows the performance in terms of equal error rate (EER). It can be seen that the EER value in the speaking rate mismatched condition is much higher than that in the matched condition. This means that speaking rate mismatch between enrollment and test indeed causes significant performance reduction in speaker verification.

3.3 fMLLR results for GMM-UBM system

The fMLLR transform was trained based on the CSLT-SPRateDGT2016 database. A normal (slow) UBM was firstly trained using the normal (slow) speech in the development set via MAP adaptation (adapted from the global UBM used in the GMM-UBM baseline). The fMLLR was then trained to maximize the objective function (2), where the speech signal X was the slow speech in the development set. The EER results with the fMLLR transform are shown in Table 2. It can be seen that the fMLLR transform improved the system performance when condition mismatch exists, and the relative EER reduction is 19.11 %.

Table 2 EER results with FMLLR-based transform

Condition	Enroll	Test	EER%	Transform
Matched	norm.	norm.	2.55	no
Mismatched	norm.	slow.	7.64	no
Mismatched	norm.	slow.	6.18	yes

Table 3 EER results with the FMLLR-based transform under i-vecoter framework.

Condition	Enroll	Test	EER(%)		Transform
			Cosine	PLDA	
Matched	norm.	norm.	4.00	1.82	no
Mismatched	norm.	slow.	13.09	3.64	no
Mismatched	norm.	slow.	12.36	3.27	yes

3.4 fMLLR for i-vector system

Since the fMLLR transform operates in feature space and is model-independent, it can be applied to any system, so long as the system use the same feature as the GMM-UBM system. To verify this argument, we constructed an i-vector system and evaluated the contribution of the fMLLR transform learned from the GMM-UBM system. The UBM was the same as the GMM-UBM system, and the i-vector dimensionality was set to 100. The i-vector model was trained using the same data as in the GMM-UBM framework. But the PLDA was trained using a larger database that consists of $4,329$ utterances from 231 speakers. We used a larger database for PLDA is because the number of speakers of CSLT-SPRateDGT2016 is too small to train a reasonable PLDA model.

To apply the fMLLR transform to the i-vector system, we first extracted MFCC features for all the test utterances in slow speaking rate, and then transformed these features using the fMLLR learned under the GMM-UBM system. Note that the enrollment data were not transformed as they are normal speech already. After extracting i-vectors, we evaluated the system performance, using both cosine scoring and PLDA scoring.

Table 3 shows the EER results of the i-vector system. we can observe that in the mismatched condition, the system performance was improved with the fMLLR transform, with both cosine scoring and PLDA scoring. The relative EER reduction are 5.58% and 10.16% with cosine scoring and PLDA scoring, respectively. This confirms our conjecture that fMLLR transform is global and can be used freely for various systems based on different model.

4 Conclusions

In this paper, we study an fMLLR-based feature transformation method to mitigate the impact of speaking rate mismatch on speaker verification systems. Our preliminary experiments show that slow speech tends to cause substantial performance reduction, and the proposed fMLLR-based approach can diminish this reduction when there is a mismatch in the speaking rate between enrollment and test. Despite the promising results, our study is still limited. We conjecture that training different transformation for different speaking rates may offer more improvement. Future work involve collecting more data and studying more complicated transforms.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No.61271389 and NO.61371136 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 210–229, 2006.
2. Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.
3. Stephen M. Chu and D. Povey, "Speaking rate adaptation using continuous frame rate normalization," in *ICASSP'10*, 2010.
4. Xiangyu Zeng, Shi Yin, and Dong Wang, "Learning speech rate in speech recognition," in *Interspeech'15*, 2015.
5. L. Zhang M. Xu and L. Wang, "Database collection for study on speech variation robust speaker recognition," in *Proc. O-COCOSDA*, 2008.
6. Cummins Fred Grimaldi Marco, "Speech style and speaker recognition: a case study," in *Interspeech'09*, 2009.
7. Van Heerden E. et al Van Heerden C.J., Barnard E., "Speech rate normalization used to improve speaker verification," *SAIEE Africa Research Journal*, vol. 98, pp. 129–135, 2007.
8. M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.
9. Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.
10. Dong Wang Rozi Askar and Fanhu Bie et al, "Cross-lingual speaker verification based on linear transform," in *ChinaSIP 2015*, 2015.
11. R. Dehak N. Dehak and P. Kenny et al, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Interspeech'09*, 2009.
12. Marco Grimaldi Fred Cummins and Thomas Leonard et al, "The chains corpus: Characterizing individual speakers," in *International Conference on Speech and Computer SPECOM-2006*, 2006.
13. Ghoshal A Povey D and Boulianne G et al, "The kaldi speech recognition toolkit," in *Proc of ASRU*, 2011.