IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Phonetic Variability Influence on Short Utterances in Speaker Verification

Ignacio Viñals, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLAB, Aragón Institute for Engineering Research (I3A),
University of Zaragoza, Spain

{ivinalsb,ortega,amiguel,lleida}@unizar.es

Abstract

This work presents an analysis of i-vectors for speaker recognition working with short utterances and methods to alleviate the loss of performance these utterances imply. Our research reveals that this degradation is strongly influenced by the phonetic mismatch between enrollment and test utterances. However, this mismatch is unused in the standard i-vector PLDA framework. It is proposed a metric to measure this phonetic mismatch and a simple yet effective compensation for the standard i-vector PLDA speaker verification system. Our results, carried out in NIST SRE10 coreext-coreext female det. 5, evidence relative improvements up to 6.65% in short utterances, and up to 9.84% in long utterances as well.

Index Terms: Speaker verification, i-vector, short utterance, vocal content.

1. Introduction

Speaker identification is the area of knowledge focused on characterizing a speaker, so any acoustic utterance can be undoubtedly assigned to the active speaker in it. If the range of hypothetical speakers is limited (the target speaker is one out N possible speakers), we refer to the problem as speaker recognition. Otherwise, we normally talk about speaker verification: Given utterances generated by two speakers, enrollment and test, the system must decide whether both speakers are the same or not.

Multiple possibilities have been proposed to best characterize the speaker on an utterance. Some of the first approaches were based on GMMs [1]. These approaches were evolved to subspace techniques such as JFA [2] and the i-vector [3]. Complemented by PLDA [4], i-vectors have become state of-the-art for the last decade, receiving small modifications in its original formula. Some of them imply the use of DNNs in the form of GMM posteriors [5] or bottlenecks [6]. Only in the last years new approaches only based on DNNs are emerging [7] with the intention of substituting the i-vector as utterance embedding.

Originally designed for long utterances with a unique speaker, the referred techniques applied to short acoustic audios have demonstrated a significant loss of performance. This situation is critical when considering other tasks such as diarization, in which long utterances from a single speaker are uncommon. Hence any improvement in the understanding of short utterances can provide a boost in many speaker related techniques. Multiple works have succeed in mitigating the degradation of short utterances. Some works attempt to work on the extraction models [8]. Others handle the situation by compensating the i-vector [9][10]. Some contributions have also worked on the following PLDA [11][12]. Finally, some authors also deal with the problem at the score step [13].

The paper is organized as follows: In Section 2 a review about short utterances in speaker verification is presented. Section 3 contains an analysis of i-vectors with short utterances, including our proposal to mitigate their degradation. The experiments are explained in Section 4. Finally, Section 5 contains our conclusions.

2. Speaker Verification and Short Utterances

Speaker identification is the task focused on the search of patterns to best recognize an individual by means of his voice. Deeply studied for telephone channel, its usefulness has provoked its adaptation to other environments, such as meetings, broadcast, etc.

But, how can we differentiate two speakers by means of their voice, specially when their speech does not have to be the same, i.e. text-independent speaker identification? The answer lies on the different pronunciation of phonemes by two different people. The state-of-the-art, past and present, has been governed by generative models, and specifically the GMM. The GMM-UBM generative model is supposed to represent the average pronunciation for the whole phonetic content. This average pronunciation can assume the role of reference, making possible the estimation of the deviations for each speaker and consequently, speaker identification. This idea has evolved with subspace techniques [2][3], restricting the possible deviations in a limited subspace. Moreover, back-end models such as PLDA [4], have post-processed the obtained deviations, lowering more and more the error rate. Fig 1 illustrates the standard i-vector PLDA framework, which has obtained some of the best results in speaker verification, working with long utterances containing a unique speaker.

However, most of these techniques strongly suffer when applied to short utterances. These audios, because of its limited length, do not contain the totality, or at least the majority, of the vocal space. Moreover, the remaining phonemes are poorly represented due to limited information. Unfortunately, state-of-the-art techniques are built to process the totality of the phonetic space assuming the phonetic variability as pronunciation deviations. Consequently, state-of-the-art models invent unreal phonetic content when necessary. Therefore, when comparing short utterances, decisions are made taking into account unreal speech, with no data to back it up. In conclusion, the comparison of short utterances can depend more on the speech rather than the speaker.

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the 2015 FPI fellowship, the project TIN2017-85854-C4-1-R and Gobierno de Aragón /FEDER (research group T36_17R).

10.21437/IberSPEECH.2018-2

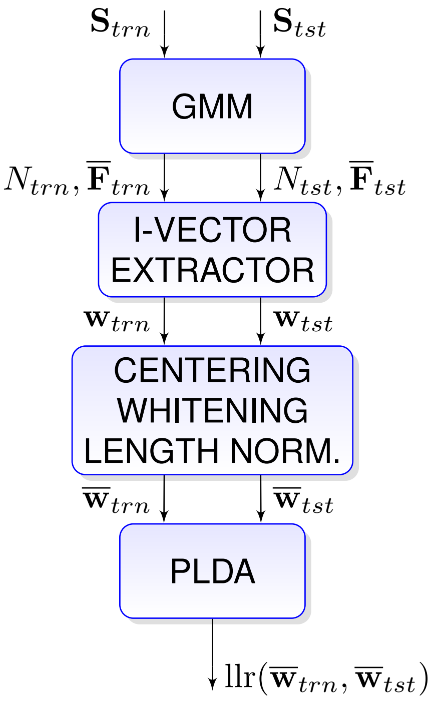
$\mathbf{S}_{trn}$
$\mathbf{S}_{tst}$
GMM
$N_{trn}, \overline{\mathbf{F}}_{trn}$
$N_{tst}, \overline{\mathbf{F}}_{tst}$
I-VECTOR EXTRACTOR
$\mathbf{w}_{trn}$
$\mathbf{w}_{tst}$
CENTERING WHITENING LENGTH NORM.
$\overline{\mathbf{w}}_{trn}$
$\overline{\mathbf{w}}_{tst}$
PLDA
$\mathrm{llr}(\overline{\mathbf{w}}_{trn}, \overline{\mathbf{w}}_{tst})$

Figure 1: *Standard I-vector PLDA framework. Given some input utterances* $\mathbf{S}_{trn}$ *and* $\mathbf{S}_{tst}$*, the system generates* $\mathrm{llr}(\overline{\mathbf{w}}_{trn}\overline{\mathbf{w}}_{tst})$

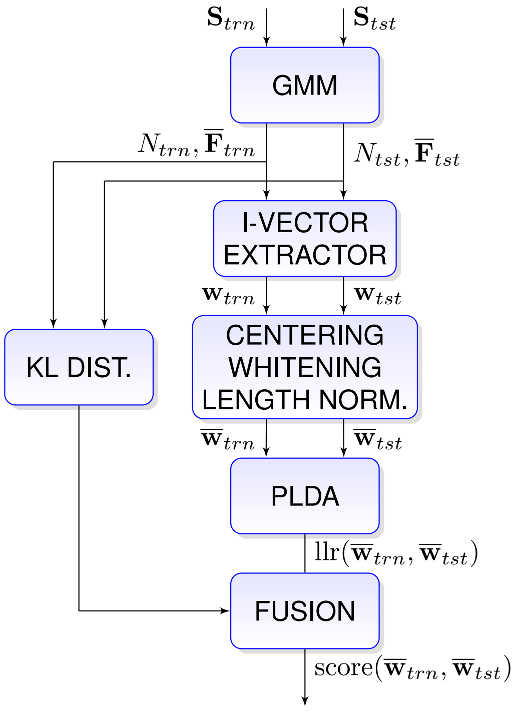
$\mathbf{S}_{trn}$
$\mathbf{S}_{tst}$
GMM
$N_{trn}, \overline{\mathbf{F}}_{trn}$
$N_{tst}, \overline{\mathbf{F}}_{tst}$
I-VECTOR EXTRACTOR
$\mathbf{w}_{trn}$
$\mathbf{w}_{tst}$
KL DIST.
CENTERING WHITENING LENGTH NORM.
$\overline{\mathbf{w}}_{trn}$
$\overline{\mathbf{w}}_{tst}$
PLDA
$\mathrm{llr}(\overline{\mathbf{w}}_{trn}, \overline{\mathbf{w}}_{tst})$
FUSION
$\mathrm{score}(\overline{\mathbf{w}}_{trn}, \overline{\mathbf{w}}_{tst})$

Figure 2: *Proposed system. The zeroth order Baum Welch statistic from both enrollment and test utterances are compared by means of KL distance, which is then fused to the original final score.*

3. Phonetic Mismatch Compensation

Short utterances are an already known problem in speaker verification, with several contributions [8][9][10][11][12][13]. Most of these solutions assume a sort of uncertainty term because of the missing information, which must be compensated. This uncertainty term summarizes about how limited is the information in the utterances, but do not pay attention to the detailed missing phonemes. Therefore, this term is used as a sort of quality measure of the utterance representations. Consequently, scores are only compensated by these representation quality approximations, without any concern about the conditional dependencies when comparing enrollment and test. It is not as harmful comparing utterances with similar limited information as with totally mismatched phonetic content.

According to our understanding, the detailed phonetic information is an impressive side information to pay no attention to. Besides its quality is increasing as long as ASR systems evolve. This sort of knowledge allows the identification of the missing acoustic content, making possible some sort of compensation for the missing acoustic content and a fair comparison of short utterances with only the available audio.

Therefore, our proposal is a proof of concept, as a first attempt to include the phonetic information in the evaluation of the trial. In this work we work on the phonetic mismatch between enrollment and test utterances in a speaker verification system. For this reason we have defined a distance between the enrollment and the test utterance for a trial. This distance is defined to measure how different is the acoustic content of enrollment and test utterances, hence measuring how fair the trials are in terms of acoustic similarity. The higher the number of matching phonemes, the more reliable the score is. Similarly, the lower is the phoneme similarity, the less restrictive should the score be, in order to gain robustness against mismatches.

Considering the i-vector PLDA standard framework, we consider the KL distance as the metric between enrollment and test utterances. This metric, is formulated as follows:

$$KL_{dist}(p,q) = KL(p||q) + KL(q||p) \tag{1}$$

$$KL(p||q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x \tag{2}$$

where KL represents the Kullback Leibler divergence between distributions p and q. KL divergence is not symmetric, hence not a distance, so we make use of the symmetric version instead. This distance will compare our phonetic information, in this work the zeroth order Baum-Welch statistics, extracted from the GMM-UBM step in the pipeline. This information, related with the acoustic content in the utterance, has strong relationships with the desired phonetic content. Nevertheless, no side information is required for its extraction.

The obtained distance can be taken into account in any posterior point of the speaker verification system (i-vector extractor, PLDA, etc). In this work, a fusion of the PLDA score with the distance is considered, made by means of logistic regression. The schematic for the tested framework is illustrated in Fig 2

With this fusion, the new score is able to compensate the phonetic mismatch in the trials, providing at the same time some sort of quality measure. However, this distance just analyzes the

Table 1: *EER(%) and minDCF metrics for the original long utterances, the chopped short utterance and the phonetically balanced short utterance*

Utterance	EER(%)	MinDCF
Baseline Long Utterance	3.25	0.16
Chopped Short Utterance	8.57	0.40
Phoneme Balanced Short Utterance	4.11	0.20

interaction between enrollment and test utterances in the scoring process, but does not analyze the utterance representation itself.

4. Experiments

Our experiments try to analyze the relevance of the phonetic mismatch in short utterances with limited acoustic information. We have opted for a speaker verification task with NIST SRE datasets, with available long utterances (around 5 minutes of speech with a unique speaker). Cohorts from SRE04, SRE05, SRE06 and SRE08 are used to construct the speaker verification system, the i-vector PLDA standard framework. This system consists of a 2048-Gaussian GMM-UBM and a 400-dimension i-vector extractor. I-vectors are centered, whitened and length-normalized [14] before being evaluated in a 400-dimension PLDA. Gaussianized MFCCs with first and second derivatives are the input for the system.

The described system has evaluated three subsets based on NIST SRE10 det. 5 core-extended core-extended female experiment: The original utterances constitute the reference system with long utterances. Short utterances are created by chopping random segments from the original ones, only reassuring that the short utterance contains between 3 and 60 seconds of speech. Another short utterance subset is also extracted, selecting the frames so that the original and the extracted utterances share the same phoneme distribution. This subset is referred as phonetically balanced in the paper. This latest subset is impossible to find in real life, but allows the analysis of short utterances with (short utterances) and without (phonetically balanced) phonetic variability, making comparisons possible.

The first analysis compares the performance of the three subsets, evaluated with the reference speaker verification system. Both sorts of short utterances are expected to yield degraded performance with regard to the original long ones due to the limited information. The relevance of the phonetic variability is checked by direct comparison between these two results. In Table 1 we present the obtained performances for the long original utterances with respect to their shorter versions, either chopped or phonetically balanced.

The results indicate that both types of variability imply a loss of performance, as expected. However, the level of degradation is far from being the same. Whilst the standard chopped short segments obtain 163.69% relative degradation in terms of EER, the phonetically balanced short utterances only gets degraded a relative 26.46%. Therefore, the estimation variability, i.e. how robust is our i-vector due to limited information, is not nearly as influential as the the phonetic variability, present in real life short utterances. It is important to bear in mind that the amount of data evaluated per short utterance is significantly smaller than the original long utterance, sometimes ruling out up to 95% of the original audio. The obtained results for long and short utterances are considered the baseline results for the experiments onwards.

Table 2: *KL distance and Error (%) for both target and non-target trials depending on the trial length: long utterances (Long), chopped short utterances (Short) and phonetically balanced short utterances (Phon. Balanced). Error estimated at NIST operation point.*

Utterance	Long	Short	Phon. Balanced
Distance			
Target	1.06	3.62	2.55
Non-target	1.74	4.61	3.47
Error (%)			
Target	28.43	80.40	40.79
Non-target	0.06	0.01	0.03

The previous results show a significant impact of the phonetic variability on the utterance modeling capabilities. However, it is still unclear how this variability affects the performance of our system. Hence we have performed a study comparing acoustic mismatch between enrollment and test utterances with the error score. As a first approach, the acoustic mismatch is measured by means of the KL distance between the distributions of the zeroth Baum Welch statistics for both the enrollment and the test utterances. Thus we are comparing which components of the GMM contribute to the i-vector extraction for enrollment and test. The results are exposed in Table 2, comparing the newly proposed distance with the error at the evaluation operating point ($C_{MISS} = 10$, $C_{FA} = 1$, $P_{tgt} = 0.01$). The results are differentiated between target and non-target populations for a better understanding.

The results indicate that short utterances suffer from the acoustic mismatch between enrollment and test, being much more significant than in long utterances. This extra mismatch occurs with both target and non-target trial populations. However, this extra mismatch does not have the same effect in the error term. Whereas non-target populations are not affected in terms of error, target trials do, explaining the degradation of short utterances. Trials with short utterances fail because the speaker verification system considers the phonetic variability as speaker variability, not differentiating between them.

The proposed solution is the compensation of the original scores by means of the phonetic distance between enrollment and test utterances. As a first approach, we propose a simple yet effective linear regression fusing two systems, the i-vector PLDA and the KL distance. This first approach helps the speaker verification system to notice whether the acoustic mismatch can be degrading the score or not. The results with this score are shown in Table 3.

According to the results, consistent improvements have been obtained, reaching up to 10% relative improvements. Significantly enough, not only short utterances get improved but so do long utterances.

Finally, it is possible to analyze the benefits of the phonetic compensation and its impact with the different populations (target and non-target) in our trial subsets. The comparison between our baseline system and our compensated version is included in Table 4

The results indicate a significant reduction of the target trials error (False Negative cases) with both short and long utter-

Table 3: *EER(%) and MinDCF metrics for trials with original long utterances and short utterances, evaluating with the standard i-vector PLDA system (Baseline) and our proposed compensated version (Compensated)*

Utterance	EER (%)	MinDCF
Long Utterance		
Baseline	3.25	0.15
Compensated	**2.93**	0.15
Short Utterance		
Baseline	8.57	0.39
Compensated	**8.00**	0.39

Table 4: *Error (%) in NIST2010 evaluation point estimated for the baseline and the compensated score with both target and non-target trials. The error is expressed for both long utterances (Long) and chopped short utterances (Short)*

Utterance	Long	Short
Baseline score		
Target	28.43	80.40
Non-target	0.06	0.01
Compensated score		
Target	8.18	32.18
Non-target	0.76	0.80

ances. Nevertheless, this compensation generates a small increase of False Positive decisions, which were almost null in the baseline situations.

5. Conclusions

In this work we have successfully analyzed the impact of the different variability factors that short utterances have: phonetic and estimation, providing a simple solution to start dealing with them. This solution is able to generate up to 10% relative improvements in terms of error ratios.

We have identified two main sources of variability in short utterances to degrade the performance in speaker verification systems: phonetic variability (what is said) and estimation variability (how reliable is our representation). According to the experiments, both sources of variability imply a decrease of performance, being phonetic variability significantly much more harmful than the estimation one. This is due to the fact that state-of-the-art technologies invent acoustic content when it is missing, a common situation in short utterances.

Besides, our proposed distance metric between utterances has revealed that the loss in performance in short utterances is due to the mismatch in target trials. Phonetic variability is not differentiated from pronunciation variability, sustain of the speaker variability. Therefore, the speaker verification system does not differentiate between speech and speaker mismatch, hence significantly increasing the amount of False Negative evaluated trials.

Finally, our proposal to take advantage of this mismatch distance has obtained limited but consistent improvements. The fusion of the original PLDA log-likelihood ratio score with the KL distance has obtained improvements up to 10% for short and long utterances. This result is specially satisfactory thanks to its simplicity, leaving the remaining framework (i-vector extractor, PLDA, etc.) unaltered. Further work should be done in order to determine best and more efficient ways to make use of this phonetic information.

6. References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions On Speech And Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, pp. 1–17, 2005.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[5] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A Novel Scheme for Speaker Recognition Using a Phonetically-aware Deep Neural Network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014)*, pp. 1714–1718, 2014.

[6] M. McLaren, Y. Lei, and L. Ferrer, "Advances in Deep Neural Network Approaches to Speaker Recognition," *International Conference on Acoustics, Speech and Signal Processing*, pp. 4814–4818, 2015.

[7] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep Neural Network-based Speaker Embeddings for End-to-end Speaker Verification," *iEEE Spoken Language Technology Workshop (SLT)*, pp. 165–170, 2016.

[8] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 853–856, 2008.

[9] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez, "Improving short utterance based I-vector speaker recognition using source and utterance-duration normalization techniques," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. August, pp. 2465–2469, 2013.

[10] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, pp. 69–82, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2014.01.004

[11] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 7644–7648, 2013.

[12] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration," *Journal of Chemical Information and Modeling*, vol. 53, no. 9, pp. 1689–1699, 2013.

[13] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. a. Van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 238803, pp. 7663–7667, 2013.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 249–252.