# I-VECTORS IN THE CONTEXT OF PHONETICALLY-CONSTRAINED SHORT UTTERANCES FOR SPEAKER VERIFICATION

*Anthony Larcher*[1], *Pierre-Michel Bousquet*[2], *Kong Aik Lee*[1], *Driss Matrouf*[2]
*Haizhou Li*[1], *Jean-Francois Bonastre*[2]
(1) Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
(2) University of Avignon - LIA, France
*alarcher@i2r.a-star.edu.sg*

## ABSTRACT

Short speech duration remains a critical factor of performance degradation when deploying a speaker verification system. To overcome this difficulty, a large number of commercial applications impose the use of fixed pass-phrases. In this context, we show that the performance of the popular *i*-vector approach can be greatly improved by taking advantage of the phonetic information that they convey. Moreover, as *i*-vectors require a conditioning process to reach high accuracy, we show that further improvements are possible by taking advantage of this phonetic information within the normalisation process. We compare two methods, Within Class Covariance Normalization (WCCN) and Eigen Factor Radial (EFR), both relying on parameters estimated on the same development data. Our study suggests that WCCN is more robust to data mismatch but less efficient than EFR when the development data has a better match with the test data.

***Index Terms*—** Speaker verification, Phonetic constraint, *i*-vector , short duration

## 1. INTRODUCTION

Initially introduced for speaker recognition, *i*-vectors [1] have become very popular in the field of speech processing and recent publications show that they are also reliable for language recognition [2] and speaker diarization [3]. Indeed, *i*-vectors extraction can be seen as a compression process aiming at representing speech segments variability in a low-dimensionality space. Hence, *i*-vectors convey the speaker characteristic among other information such as transmission channel, acoustic environment or phonetic content of the speech segment.

In [4], it was shown that for short duration (down to 2s) text-independent speaker verification, *i*-vector systems could reach the same performance as the classical Joint Factor Analysis (JFA) approach but do not provide noticeable improvement. Thus, short duration constraint still poses a serious issue for text-independent speaker verification. One way to improve speaker verification accuracy in the context of short duration is to constrain the lexical content of training and test speech in order to harness the phonetic and temporal structure of the utterances [5, 6]. By nature, the Total Variability framework does not take advantage of the temporal structure of speech and an *i*-vector extracted from a sufficiently long speech segment would have the speaker information characterized uniformly under all the phonetic classes. This is not the case for short utterances, where the *i*-vector will be emphasized toward certain phonetic classes depending on the content of the utterances. This phonetic constraint conveyed by the *i*-vectors could be used to reinforce the speaker characterisation when dealing with short duration utterances. This work focuses on the effect of phonetic-constraint in speech utterances shorter than 3 seconds, on speaker verification performance within the *i*-vector paradigm.

Several normalisation approaches have been proposed for session compensation and *i*-vector conditioning [1, 7, 8]. Two methods that have shown significant improvement for speaker verification are Within Class Covariance Normalisation (WCCN) [1] and Eigen Factor Radial (EFR) [7] which includes also the length normalisation proposed in [8]. Both of these methods are based on dilating the Total Variability space as the mean to reduce the within-class variability. For text-independent speaker verification, the within-class variability corresponds to the speaker inter-session variability. Now that the focus is on phonetically-constrained utterances, we propose to re-define the within-class variability according to both speaker identity and phonetic content of the utterances and to compare its benefits for both WCCN and EFR. Finally, we extend our comparison in order to assess the robustness of WCCN and EFR to data mismatch as such comparison does not exist in the literature according to our knowledge.

Section 2 describes the *i*-vector fundamentals and the session compensation algorithms while Section 3 presents the corpora and experimental protocol used for this study. Section 4 shows the effect of phonetic constraint on speaker verification performance. In Section 5, we show the benefits of including phonetic information in the definition of within-class variability and present a preliminary study of WCCN and EFR robustness to data mismatch. Finally, Section 6 provides conclusions and avenues for future work.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

## 2. TOTAL VARIABILITY PARADIGM

### 2.1. *I*-vector extraction

*I*-vectors are now very popular in the field of speaker recognition and detailed descriptions of the Total Variability paradigm could be found in [1, 2, 4]. The *i*-vector extraction could be seen as a probabilistic compression process that reduces the dimensionality of speech-session super-vectors according to a linear-Gaussian model. The speaker- and channel-dependent super-vector $m_{(s,h)}$ of concatenated Gaussian Mixture Model (GMM) means is projected in a low dimensionality space, named Total Variability space, as follows

$$m_{(s,h)} = m + Tw_{(s,h)} \tag{1}$$

where $m$ is the mean super-vector of a gender-dependent Universal Background Model (UBM), $T$ is called Total Variability matrix and $w_{(s,h)}$ is the resulting *i*-vector .

Compared to Eigenvoice modeling, which has been shown to capture mainly the speaker characteristics with very short utterances [9], *i*-vectors convey, in addition to the speaker characteristics, other information such as transmission channel, acoustic environment or phonetic content of the speech segments. Session compensation or *i*-vector normalisation should thus be applied in order to isolate the targeted speaker information from other unwanted variability.

### 2.2. *I*-vector normalisation

In order to condition *i*-vectors for a specific task, different normalisation process have been proposed recently [1, 7, 8]. Two of them, WCCN [1] and EFR [7], are especially dealing with session compensation.

WCCN scales the Total Variability space by a matrix $B$ in order to suppress high within-class covariance. For speaker verification, $B$ is obtained by the Cholesky decomposition of the within-class covariance matrix $W_{wccn}$, i.e. $W_{wccn}^{-1} = BB^t$. The matrix $W_{wccn}$ is calculated over a large data set by using:

$$W_{wccn} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s} w_i^s \tag{2}$$

where $S$ is the number of speakers in the data set, and there are $n_s$ number of sessions for each of these speakers. Each utterance is compactly represented as an *i*-vector $w_i^s$. Distance between speech segments could then be computed with a weighted Cosine Similarity Score (CS) given by:

$$CS(w_1, w_2) = \frac{< B^t w_1 | B^t w_2 >}{||B^t w_1||\,||B^t w_2||} \tag{3}$$

EFR has been introduced in [7] to condition *i*-vectors and reduce session variability, as follows

$$w \leftarrow \frac{V^{-\frac{1}{2}}(w - \overline{w})}{\sqrt{(w - \overline{w})V^{-1}(w - \overline{w})}} \tag{4}$$

where $V$ and $\overline{w}$ are respectively the covariance matrix and the mean vector estimated from a large development set of *i*-vectors (note that this normalisation could be iterated to properly condition the data but that does not provide any benefits here). A Mahalanobis-based scoring function could then be used as speaker detection scoring:

$$score(w_1, w_2) = (w_1 - w_2)^t W^{-1}(w_1 - w_2) \tag{5}$$

where $W$ is the within-class covariance matrix computed on the EFR normalized vectors.

The main drawback of these two methods comes from their dependency on the development set that has to be representative of the unseen test material.

## 3. EXPERIMENTAL SET-UP

### 3.1. Corpora

Experiments are performed on the RSR2015[1] database, a new corpus designed to evaluate text-dependent speaker verification engines. This database contains recordings from 100 male speakers using six different cell-phones or tablets. Thirty pass-phrases (each less than 3s) and thirty short commands (each less than 1s) are recorded in nine sessions for each speaker. The pass-phrases and command are the same for all 100 speakers in order to simulate imposture attacks and each speaker records on a minimum of three different devices. A more detailed description of RSR2015 could be found in [10].

Two others corpora were also used in our experiments. We used the entire Switchboard provided by LDC and an in-house corpus which includes the recordings of 118 male speakers recorded in similar condition as RSR2015 but using different portable devices and texts.

### 3.2. Experimental protocol

Our experiments operate on 19 Mel-Frequency Cepstral Coefficients (plus energy) augmented with 19 first ($\Delta$) and 11 second ($\Delta\Delta$) derivatives. The bandwidth is limited to 300-3400Hz. The analysis window is 20ms with 10ms shifting. Lower energy frames are removed and cepstral mean subtraction is applied to the remaining features.

A 512 mixtures UBM and the Total Variability matrix are estimated using 790 speakers and 12,422 sessions taken from Switchboard and the in-house database. The dimensionality of *i*-vectors is 400. The RSR2015 database is divided in two partitions, namely, *RSR2015_norm* and *RSR2015_eval*, each containing 50 speakers. Both WCCN and EFR parameters are estimated for three different development sets:

**PASS-PHRASES,** which is composed of all pass-phrases from the 50 speakers of the *RSR2015_norm* data set (13,500 utterances).

[1] `http://www1.i2r.a-star.edu.sg/~kalee/RSR2015_WEB/RSR2015.html`

**Commands,** which is composed of all short commands from the 50 speakers of the *RSR2015_norm* data set (13,500 utterances).

**SWB,** which is composed of 672 speakers from Switchboard databases recorded through telephone channel (6,522 utterances).

Note that the Pass-Phrases development set has the closest match to the test data (derived from the *RSR2015_eval*) in terms of channel, duration and phonetic content are similar. The Commands set match on the channel of the test set since the duration and phonetic content are different. Finally, the SWB set is strongly mismatched with the test data in terms of channel and duration, thus it could be considered as the most different one. The average duration of utterances for the three development sets (Pass-Phrases, Commands,SWB) are $0.75s$, $0.43s$ and $79.66s$, respectively. The test segments, as described below, has an average duration of $0.93s$.

The test set is derived from the *RSR2015_eval*. For all the 50 speakers, the three first recordings are used for training and the remaining six sessions are used as test segments. A trial would simply involves comparison of *i*-vector extracted for a training utterance of a speaker with the *i*-vector extracted from test segment. We use all the cross-pairs between training and test segments made available in the *RSR2015_eval* partition. As we consider phonetically-constrained speaker verification task, we separate the trials into four categories according to the condition whether the user is the target client or an impostor and whether the phonetic content is the same for training and test segments. Table 1 shows the number of each type of trials resulting from our protocol. Notice that the *Same phonetic content* condition of our protocol only considers the case where the full utterances are the same (i.e, text-dependent speaker recognition). Future work has to include cases where the phonetic content still the same when sequences differ.

| | Same phonetic contain | Different phonetic contain |
|---|---|---|
| Target User | CLIENT-same<br>(26,913) | CLIENT-diff<br>(390,185) |
| Impostor | IMP-same<br>(659,286) | IMP-diff<br>(19,119,255) |

**Table 1**. Different types and numbers of trials in phonetically-constrained speaker verification

## 4. INFLUENCE OF THE PHONETIC CONTENT ON *I*-VECTOR SPEAKER VERIFICATION

The first experiment is performed in order to assess the contribution of phonetic constraint for short duration speaker verification. Figure 1 shows the performance of the *i*-vector system using the Cosine Scoring without any normalisation process depending on the nature of the target and impostor trials. The first configuration, similar to text-independent condition, is provided as a baseline. The phonetic content used during test for both target speakers and impostor is different from the one used for training (*CLIENT-diff / IMP-diff*). The Equal Error Rate in this case, $43.06\%$, drops by $74\%$ to $15.38\%$ in a second configuration where both target and impostor users pronounce the same phonetic-content that was used for training (*CLIENT-same / IMP-same*). A third configuration shows that EER falls to $8.02\%$ in the optimal case where only target users know the proper phonetic content and impostors pronounce a different one (*CLIENT-same / IMP-diff*). This experiment shows that the phonetic information conveyed by *i*-vectors could be used to improve accuracy of speaker verification in short duration context. All experiments in the rest of this paper consider the case where both target users and impostors pronounce the same phonetic content as the one used for training (*CLIENT-same / IMP-same*). Indeed, this configuration is closer to the realistic case of a phonetically-constrained application.

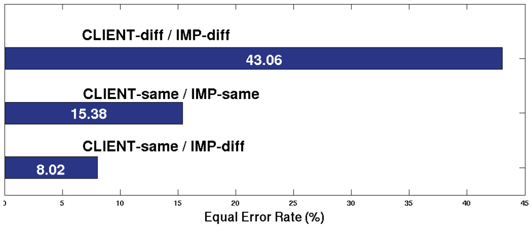

**Fig. 1**. EER for different trials configuration.

## 5. INFLUENCE OF THE NORMALISATION TRAINING SET

The second experiment is designed to compare the influence of development dataset and class definition on both WCCN and EFR.

Rows 1, 2 and 3 of Table 2, shows the performance of the *i*-vector system using these two normalisation methods when moving the development dataset closer to the test set. Within-class variability is defined according to speaker identity only. Performance of *i*-vectors using Cosine Scoring without normalisation is reported as a baseline and is enhanced by all normalisations. As expected, the performance improves when the development set get more similar to the test data for both WCCN and EFR. EER reduces from $13.85\%$ to $10.54\%$ and $9.37\%$ for EFR and from $13.36\%$ to $10.42\%$ and $10.01\%$ for WCCN when moving from SWB to Commands and then Pass-Phrases respectively. These results highlight the importance of development set for *i*-vector normalisation as the reduction of Equal Error Rate observed when moving from SWB to Pass-Phrases is more than $33\%$ relative for EFR conditioning.

When comparing the methods, results suggest that WCCN tends to be slightly more robust than EFR to the mismatch between development and test data. When training the normalisation parameters on SWB, EER obtained by using WCCN (13.36%) is 3.6% less than when using EFR (13.85%) relatively. For normalisation parameters trained on COMMANDS (closer to test data), the gap between normalisation methods is less but still in favour of WCCN with EER of 10.42% and 10.54% for EFR. However, when development share similar phonetic content, channel and duration with test data (PASS-PHRASES), EFR outperforms WCCN. The EER obtained with WCCN is 10.01% when it is 9.37% for EFR ($-6.4\%$ relative). This result suggests that EFR is more effective than WCCN but less robust to data mismatch.

| Development Set | *i*-vector scoring | | |
|---|---|---|---|
| | EFR | CS + WCCN | CS |
| SWB | 13.85 | 13.36 | 15.38 |
| COMMANDS | 10.54 | 10.42 | |
| PASS-PHRASES | 9.37 | 10.01 | |
| PASS-PHRASES<br>speaker + phonetic | 7.88 | 9.67 | |

**Table 2**. Performances of Eigen Factor Radial (EFR) and Cosine Scoring (CS) with and without WCCN in terms of EER (%) for different development datasets and classes definitions.

In the context of short duration, speaker identity and phonetic content could be expected to be the main sources of variability. For applications where the text pronounced during training and test is fixed, it is possible to use this knowledge in order to improve the *i*-vector normalisation. Indeed, both WCCN and EFR are conditioning the *i*-vectors in order to minimize the within-class variability. In this work, we propose to define those classes according to both speaker and phonetic content instead of grouping all sessions from a same speaker. Rows 3 and 4 of Table 2 respectively show the performance of the two normalisation methods when using the classical definition of within-class variability or when including the phonetic information. Defining the normalisation classes by using speaker and phonetic information improves the accuracy for both WCCN and Eigen Factor Radial. As observed above, when the classes defined for normalisation training exactly match the test classes, i.e. one speaker and one phonetic content per class, we observe that the gain for EFR is more important than in the case of WCCN, respectively 15, 9% and 3, 4% of relative improvement.

Further experiments has to be performed in future works in order to distinguish between the effect of duration, channel and phonetic mismatch and to confirm the benefit of adding phonetic information in the within-class definition when dealing with channel mismatch.

## 6. CONCLUSION

In this paper we focused on the influence of phonetic constraint in short utterances for speaker verification. We showed that using the phonetic information conveyed by *i*-vectors could lead to substantial improvement, up to 74% in terms of EER. We underlined the importance of an adequate development dataset on WCCN and Eigen Factor Radial methods. Our analysis suggests that WCCN is more robust to data mismatch when Eigen Factor Radial performs better for similar conditions. This preliminary work needs to be continued as several questions remain regarding the importance of individual factors such as duration, channel or phonetic content on the robustness of the different normalisations. Finally we showed that it is possible to take advantage of a phonetic constraint for *i*-vector normalisation by using a phonetic classification of the development data. This adaptation of WCCN and Eigen Factor Radial has led to relative reduction of EER of 3.4% and 5.9% respectively. In the future, we intend to continue exploring the impact of phonetic information on *i*-vector normalisation by considering the correlation between the existing speaker discrimination scoring and different between -utterances phonetic distances for very short durations.

## 7. REFERENCES

[1] 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. 99, pp. 1, 2009.

[2] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Interspeech*, 2011.

[3] J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez, "ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation," in *FALA "VI Jornadas en Tecnologa del Habla" and II Iberian SLTech Workshop*, 2010, pp. 415–418.

[4] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector Based Speaker Recognition on Short Utterances," in *Interspeech*, 2011.

[5] M. Hébert, *Text-dependent speaker recognition*, Springer-Verlag, Heidelberg, 2008.

[6] A. Larcher, J.F. Bonastre, and J.S.D. Mason, "Reinforced temporal structure information for embedded utterance-based speaker recognition," in *Interspeech*, 2008, pp. 371–374.

[7] P.M. Bousquet, D. Matrouf, and J.F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Interspeech*, 2011.

[8] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249 – 252.

[9] R. Vogt, C. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*. 2008, IEEE.

[10] K.A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home," in *Interspeech*, 2011.