Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Multi-channel i-vector combination for robust speaker verification in multi-room domestic environments

Alessio Brutti[1], *Alberto Abad*[2,3]

[1] Fondazione Bruno Kessler, CIT, Trento
[2]L[2]F - Spoken Language Systems Lab, INESC-ID Lisboa
[3] IST - Instituto Superior Técnico, University of Lisbon
brutti@fbk.eu, alberto.abad@l2f.inesc-id.pt

Abstract

In this work we address the speaker verification task in domestic environments where multiple rooms are monitored by a set of distributed microphones. In particular, we focus on the mismatch between the training of the total variability feature extraction hyper-parameters, the enrolment stage, which occurs at a fixed position in the home, and the test phase which could happen in any location of the apartment. Building upon a previous work, where a position independent multi-channel verification system was introduced, we investigate different i-vector combination strategies to attenuate the effects of the above mentioned mismatch sources. The proposed methods implicitly select the microphones in the room where the speaker is, without any knowledge about the speaker position. An experimental analysis on a simulated multi-channel multi-room reverberant dataset shows that the proposed solutions are robust against changes in the speaker position and orientation, achieving performance close to an upper-bound based on knowledge about the speaker location.

1. Introduction

One of the most typical problems of domestic scenarios regarding automatic speech processing methods, and particularly speaker recognition systems, is the fact that users can give commands from any position of any room. Consequently, a mismatch is often present between the enrolment and test phases, resulting in a noticeable decrease in the verification performance [1].

In practice, creating suitable speaker models able to cope with any spatial condition is a fundamental challenge for speaker recognition in real-world home automation applications. One possibility to partially tackle this problem would consist of adopting a network of multiple distributed microphones that continuously monitors the target environment. Previous works have shown the advantages of exploiting the information from multiple channels. The most common strategy implements speech enhancement techniques based on multi-channel processing to reduce the amount of reverberation and noise in the speech signals before the identification step [2, 3, 4, 5, 6, 7]. The majority of these techniques (e.g. beamforming, spectral subtraction, etc.) requires the availability of a compact microphone array. The latter constraint is relaxed in [8] where a post-processing fusion of independent classifiers, applied to each channel, is presented. On the other hand, the investigation of multi-channel solutions in the Total Variability (TV) framework [9, 10] is still in a preliminary stage. In [11, 12] solutions to improve robustness against condition mismatch (i.e. telephone and distant microphones) are presented, without considering multiple simultaneous recording of the same utterance. Recently, multi-condition training [13, 14] and an extended version of the Probabilistic Linear Discriminant Analysis (PLDA) [15, 16] have been adopted in multi-channel recordings to improve the verification performance. The recent widespread interest in deep learning methods has also reached the TV framework for speaker recognition [17, 18]. In [19] methods to address channel mismatch in DNN/i-vector based systems are explored, and the effect of artificial noise and reverberation on speaker verification performance is analysed.

Given a domestic environment equipped with multiple distributed microphones which capture the same acoustic scene, we investigate suitable i-vector combination strategies to achieve a position independent speaker verification system able to accomplish the verification task whatever the speaker location is in the home. This paper builds upon our previous work [20] where a variety of multi-channel implementations of TV were investigated, based on two strong assumptions:

- training material for TV hyper-parameter estimation is available in the same environmental conditions as enrolment;
- the room where the speaker is located during test is known and only the microphones inside this room are used.

This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

doi: 10.21437/Odyssey.2016-36

In the present work, we attempt to relax those highly impractical constraints. This is attained by selecting or combining i-vectors through confidence measures derived from the Universal Background Model (UBM) Log-likelihoods (UL).

The paper is organised as follows. Section 2 introduces the domestic scenario addressed in this work. The multi-channel TV framework proposed in [20] is described in Section 3. Section 4 presents the experimental validation, including data description, experimental framework and attained results. Finally, Section 5 concludes the paper with final remarks and future work.

2. Domestic application scenario

In this paper we address a speaker verification application in a domestic scenario where the multiple rooms of an apartment are monitored by a network of distributed microphones. Realistic applications in such a scenario can envisage that the speaker enrolment stage, that is controlled by the system, takes place in a specific spot in the apartment. On the other hand, the speaker should be free to interact with the system from any position in the home and, consequently, the system must be able to verify the identity of the speaker anywhere.

Figure 1 shows the layout of the real apartment we used in the experiments. Microphone positions and the grid of possible source positions and orientations are also shown in the figure. The "fixed" position/orientation for speaker enrolment is in the Livingroom (black square). Conversely, in the verification stage the speaker positions and orientations are randomly selected among those available in 3 rooms: Kitchen (red), Livingroom (blue) and Bedroom (yellow), choosing a different position/orientation pair for each test utterance. The number of microphones considered in each room has been restricted to one representative channel per wall/ceiling, i.e. 4 microphones in the Kitchen, 5 in the Livingroom and 3 in the Bedroom.

3. Multi-channel Total Variability

3.1. Baseline System

The baseline speaker verification system is the multi-channel TV framework presented in [20]. Let us denote with $x_m(t)$ the signal acquired by the m-th microphone ($m = 1, \ldots, M$). We consider the channel-independent single system $\Lambda = \{\mathbf{U}, \mathbf{T}\}$, where $\mathbf{U}$ is the UBM and $\mathbf{T}$ is the TV matrix. We consider two systems:

- $\Lambda_{\text{ct}} = \{\mathbf{U}_{\text{ct}}, \mathbf{T}_{\text{ct}}\}$: trained on clean close-talking signals;
- $\Lambda_{\text{mt}} = \{\mathbf{U}_{\text{mt}}, \mathbf{T}_{\text{mt}}\}$: trained on reverberant material matching the conditions of the fixed-position enrolment.

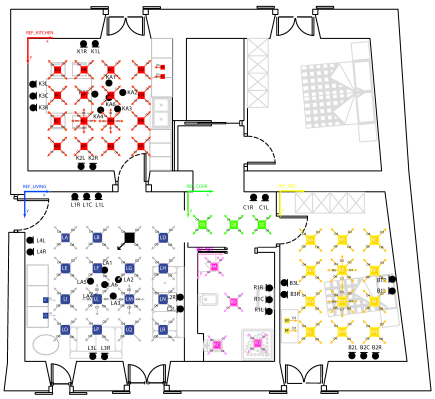

Figure 1: Layout of the apartment used in the experiments. Squares and arrows indicate the possible speaker positions and orientations. The black square and arrow represent the enrolment position.

During enrolment, the signals $x_m(t)$ are used together with the system Λ to obtain M_e speaker model vectors $w_m(s)$ ($m = 1, \ldots, M_e$) where M_e is the number of microphones available in enrolment, for each enrolment speaker s. Given the multiple enrolment observations, a single i-vector model is derived by simple averaging all the vectors of a given speaker s:

$$\overline{w}(s) = \frac{1}{M_e} \sum_{m=1}^{M_e} w_m(s). \tag{1}$$

Likewise, during test, an i-vector $w_m(u)$ of the test utterance u is obtained for each channel m ($m = 1, \ldots M_t$), where M_t is the number of microphones in the room and typically $M_t \neq M_e$. A single test i-vector is then derived averaging all the channels:

$$\overline{w}(u) = \frac{1}{M_t} \sum_{m=1}^{M_t} w_m(u). \tag{2}$$

At this point a single verification score can be obtained applying the cosine scoring on the average enrolment speaker and test utterance i-vectors:

$$\overline{C}(s,u) = < \overline{w}(s), \overline{w}(u) > . \tag{3}$$

Figure 2 shows a summary diagram of the multi-channel TV framework described in this section.

We also adopt two i-vector processing stages that are expected to contribute to a partial reduction of the channel variability: i-vector centering and whitening [21]. Both mean and decorrelation matrix parameters are estimated using training data from unknown speakers (not belonging to the set of enrolment speakers) from all the available channels in the case of the Λ_{mt} system and from close-talking speech in the case of Λ_{ct}.

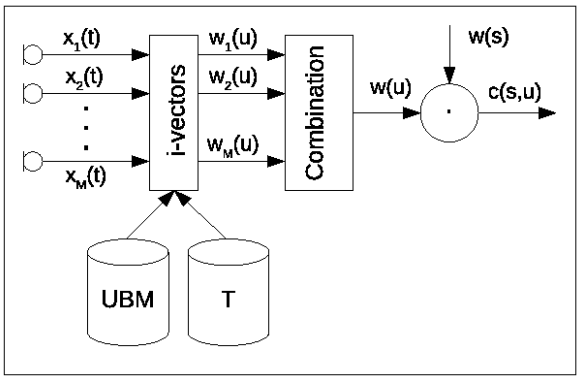

Figure 2: Block diagram of the multi-channel speaker verification system.

In addition, T-norm score normalisation (only mean) is applied [22], in which the mean off-set is estimated on all enrolment speaker model scores S as follows:

$$\overline{C}'(s,u) = \overline{C}(s,u) - \frac{1}{S}\sum_{i=1}^{S}\overline{C}(i,u). \quad (4)$$

3.2. UL-based i-vector combination

A reasonable hypothesis regarding the uniform combination of speaker vectors in Eq. 1 and Eq. 2 is that it should be more effective if all the channels considered present a similar degree of match (or mismatch) with respect to the training data characteristics. In particular, in our previous work in [20], we considered that this similarity assumption holds only for the subset of microphones located in the room where the speaker is, and consequently, only those microphones were used. Thus, speaker position knowledge was implicitly assumed. However, recent studies addressing the problem of speaker localisation in multi-room environments based on microphone networks have shown that the room detection error rates are not insignificant, unless very articulated and computational expensive algorithms are employed [23, 24].

In this work, we consider that no knowledge about the speaker position is available during verification. Consequently, considering the former hypothesis, a suitable i-vector combination (or selection) strategy is needed to avoid speaker verification performance deterioration due to the inclusion of less informative channels in the uniform vector combination. To this end, we propose to replace the averaging of Eqs. 1 and 2 with the following weighted combination:

$$\overline{w}(u) = \frac{1}{M_t}\sum_{m=1}^{M_t} l_m(u)w_m(u); \quad \sum_{m=1}^{M_t} l_m(u) = 1, \quad (5)$$

where, following the method presented in [23], the weights $l_m(u)$ are proportional to the log-likelihood obtained evaluating the UL of the TV system (Λ_{mt} or Λ_{ct}) against the incoming utterance u. The basic underlying goal is to emphasise more those channels that better match the training conditions. The same process is applied both to the enrolment and verification utterances.

3.3. UL-based N-best i-vector selection

As an alternative to the weighted combination approach, an N-best combination strategy can be considered. Microphone channels are ranked according to the UL attained and only the N-best i-vectors are used either for uniform or weighted combination. Without loss of generality, let us assume that:

$$l_1(u) \geq l_2(u) \geq \ldots l_{M_t},$$

Eq. 5 is modified as:

$$\overline{w}(u) = \frac{1}{N}\sum_{m=1}^{N} l_m(u)w_m(u); \quad \sum_{m=1}^{N} l_m(u) = 1, \quad (6)$$

where $l_m(u) = \frac{1}{N}$ in the case of uniform averaging.

4. Experimental evaluation

4.1. Data description

We simulated a multi-channel reverberant data-set through convolution of clean speech segments with a set of Room Impulse Response (RIR)s, measured in the target apartment [25]. Figure 1 shows the floor plan of the apartment and the possible positions and orientations of the speaker. Note that in this work, we are considering the effects of reverberation only and no additive noise has been incorporated into the data generation process. The clean speech segments used for simulation are obtained from the Portuguese corpus *Base de Dados em Português eUropeu, vocaBulário Largo, Independente do orador e fala COntínua* (BD-PUBLICO) [26] that contains data from 60 female and 60 male speakers. Speakers were divided into two gender-balanced disjoint sets: enrolment and unknown speakers. The unknown speakers' data were used to train the UBM, TV matrices and centering and whitening transformation parameters. For the enrolment of the speakers, we limited the total duration of speech to approximately 60s per speaker.

4.2. Evaluation protocol

The test set consisted of the remaining data, not used in the enrolment sets, totalling 3107 test utterances: 1045, 1000 and 1062 in the Kitchen, Livingroom and Bedroom respectively. Considering that in typical domestic applications only the home owners are registered into the system and have access to the services, an open-set speaker verification task better resembles the problem.

To this end, and given the relatively small number of enrolment speakers, we apply a sort of 2-fold open-set validation: enrolment speaker models are randomly split in two halves, one is kept as actual enrolment speakers and the others are disregarded (and the corresponding test utterances are considered to be from "unknown" speakers). This evaluation process is repeated using 10 different random partitions and the mean performance is computed. Thus, the total number of gender-independent trials at each random partition is 93210. Note that in enrolment only the 5 microphones of the Livingroom are used ($M_e = 5$) and best-N i-vector selection is not applied.

We consider two test conditions:

- in the *matched* test set the speaker is always located in the enrolment position;
- the *3-Rooms* test set simulates the target application scenario with random speaker positions and orientations in 3 different rooms of the apartment.

4.3. TV implementation

In the following speaker verification experiments, i-vectors are based on Mel-Frequency Cepstral Coefficients (MFCCs) features extracted in 20 ms frames, with 10 ms overlap. Each feature vector is composed of 15 static MFCCs with its derivatives, totalling 30 dimensions. The TV matrix is estimated according to [27], starting from a gender independent UBM modelled by a mixture of 1024 Gaussian distributions. The dimension of the total variability subspace is fixed to 400. Zero and first-order sufficient statistics of the training sets are used for training the TV matrix $\mathbf{T}$: 10 EM iterations are applied for both ML and minimum divergence update. The covariance matrix is not updated in any of the EM iterations. The estimated $\mathbf{T}$ matrix is in turn used for extraction of the total variability factors from the speech segments as described in [27]. Finally, the resulting factor vectors are normalised to be of unit length, which we will refer to as i-vectors.

4.4. Experimental results

Figure 3 shows the Equal Error Rate (EER) obtained with the reverberant system Λ_{mt} as a function of the number of microphones in both the *3-Rooms* and the *matched* scenarios. Note that, overall, verification performance of the systems addressed in this work are in general quite good in absolute terms (approx. $2.5 - 3\%$ in the worst case). This is mainly related to the absence of additional environmental noise in the experimental data.

Focusing on the matched scenario, the best performance is achieved using the 5-best i-vectors, which is the number of channels available in the Livingroom, where the enrolment position is defined. This means that the UL is a good metrics to identify the most reliable channels. This claim is supported by the fact that the weighted

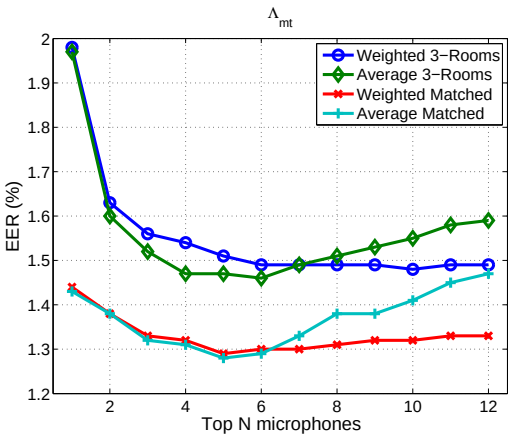

Figure 3: EER as a function of the number of microphones used, with the system Λ_{mt}. The speaker during verification is in the enrolment position (light blue and red) or randomly located in the apartment (dark blue and green).

combination saturates for $N > 5$, implying that bad microphones correctly receive low weights. For $N = 5$, i-vectors averaging delivers slightly better results than the weighted combination, implying that the UL provides a good ranking of the i-vectors but it is not so reliable in providing a fine weight definition between similar channels. This ability of UL to coarsely classify different channels is consistent with the results reported in [23] for multi-microphone ASR in home environments. Note however that the performance of the averaging deteriorates when more microphones are included, in contrast to the saturation observed in the weighted combination. Therefore, the weighted combination seems to be preferable when there is no prior knowledge about the optimal number of N-best microphones. This is more evident when considering the *3-Rooms* scenario in Figure 3: in this case the best N value is 6, which does not seem possible to be determined a-priori since it does not correspond to the number of microphones of the enrolment room (or any other room) and because it can be strongly conditioned by the specific characteristics of the test set (note that this is the best value on average over all the tests). Conversely, at the cost of a minor performance deterioration, the weighted combination allows using all the channels without taking hard decisions about the number of microphones.

Similarly, Figure 4 shows the EER obtained with the system Λ_{ct} trained on close-talking data. In this case the gap for the best case ($N = 5$) between the averaging and the weighted combination is slightly larger (approximately 0.1% compared to the very small difference observed for Λ_{mt}). This can be explained by the fact that the huge mismatch between the clean data used to train the

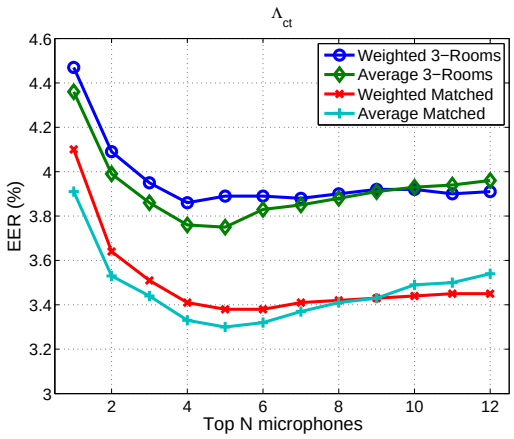

Figure 4: EER as a function of the number of microphones used, with the system Λ_{ct}. The speaker during verification is in the enrolment position (light blue and red) or randomly located in the apartment (dark blue and green).

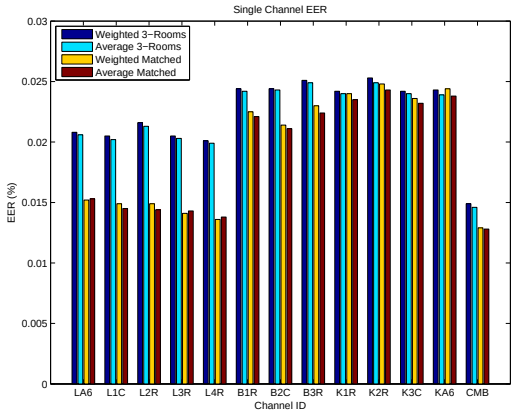

Figure 5: Speaker verification performance of the Λ_{mt} system for each single channel in comparison with the proposed combinations (for the i-vector average, the optimum number of microphones is used). The first letter of the channel ID represents the room where the microphone is located: L=livingroom, B=bedroom, K=kitchen. The remaining two characters identify the channel in the room (see Figure 1). The final bar (CMB) refers to either average or weighted combination (see legend).

system and the enrolment and test signals makes the UL less reliable. Nevertheless, also in this case the weighted combination allows considering all the channels without knowledge about the microphone and speaker positions.

Table 1 summarises some of these results in comparison with the upper-bound system described in [20]. This system assumes knowledge about the speaker position and consequently uses only the microphones of the room where the speaker is. Note that here numbers are slightly different with respect to [20] because we are considering an open-set verification task. The proposed methods achieve very close performance to the upper bound. As observed above, the gap between the weighted combination and the N-best averaging is larger when Λ_{ct} is used. This behaviour makes sense because the large mismatch between training and test signals leads to less reliable UL scores. The gap is particularly wide in the *3-Rooms*-Λ_{ct} combination: this case is particularly challenging because both mismatches (training and position) are present and the weighted combination results less effective.

Finally, Figure 5 shows the speaker verification performance of the reverberant Λ_{mt} system obtained with each single channel in both test conditions: *matched* and *3-Rooms*. The first letter of the channel ID represents the room where the microphone is located. Note that performance of individual channel for uniform and weighted combination may differ due to different processing of the enrolment data. As discussed in [20], the figure shows that combining test i-vectors provides performance improvements with respect to single channel verification. Note that in the *matched* test conditions, the microphones in the Livingroom outperform all the other channels, as expected, but still both i-vector combination strategies are able to bring additional improvements. In the *3-Rooms* case, the difference between the channels is smaller and the combination allows for considerable performance improvements. Interested readers can find a more detailed comparison between single channel baselines and different multi-channel TV systems in [20].

5. Conclusions

In this paper we presented two methods for combining and selecting i-vectors extracted from the signals observed by multiple distributed microphones in a domestic scenario. The proposed weighted average based on UL is shown as an effective combination strategy of i-vectors, even when the position of the speaker is not known. The same method is also effective when the TV system is trained on mismatched close-talking material.

Moving towards a more realistic application scenario, we plan to remove the constraint on the fixed and known enrolment position, allowing the speaker to enrol from anywhere in the home, eventually in an unsupervised manner. A further issue to address towards a real application is robustness against environmental noise.

Future work will address the development of more informative combination weights (or more articulated combination functions). In fact, the UL provides a reliable ranking of the channels but the i-vectors resulting from the weighted combinations are less accurate than those obtained with a simple uniform combination of the N-best microphones.

Scenario	System	Upper Bound[20]	Avg 12ch	Avg N-best [5-6]	Weighted 12ch	Weighted N-best [5-6]
matched	Λ_{mt}	1.20	1.47	1.28	1.33	1.29
	Λ_{ct}	3.29	3.54	3.30	3.45	3.38
3-Rooms	Λ_{mt}	1.42	1.59	1.46	1.49	1.49
	Λ_{ct}	3.57	3.96	3.83	3.91	3.89

Table 1: EER(%) of the proposed methods in comparison with an upper-bound based on knowledge of the room where the speaker is located. 5-best channels are considered in the *matched* scenario, while 6-best is the optimal solution in the *3-Rooms* case.

A further interesting research direction is the use of PLDA [28] to handle the multiple observations. Finally, the solutions for channel mismatch robustness proposed by [11, 12] can also be investigated for comparison or to further improve the performance of our multi-channel solutions.

6. References

[1] P.J. Castellano, S. Sradharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.

[2] Qin Jin, Runxin Li, Qian Yang, Kornel Laskowski, and Tanja Schultz, "Speaker identification with distant microphone speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[3] Zhaofeng Zhang, Longbiao Wang, and Atsuhiko Kai, "Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2014, pp. 15, 2014.

[4] Qiguang Lin, Ea-Ee Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 622–629, Oct 1994.

[5] Iain McCowan, Jason Pelecanos, and Sridha Sridharan, "Robust speaker recognition using microphone arrays," in *Speaker Odissey*, 2001.

[6] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Providing single and multi-channel acoustical robustness to speaker identification systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[7] Jack W. Stokes, John C. Platt, and Sumit Basu, "Speaker identification using a microphone array and a joint HMM with speech spectrum and angle of arrival," July 2006, Institute of Electrical and Electronics Engineers, Inc.

[8] Jordi Luque and Javier Hernando, "Robust speaker identification for meetings: UPC CLEAR'07 meeting room evaluation system," in *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR and RT 2007, Revised Selected Papers*, 2007, pp. 266–275.

[9] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Interspeech*, 2009.

[10] 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, May 2011.

[11] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Speaker Odyssey*, 2010.

[12] N. Dehak, Z.N. Karam, D.A. Reynolds, R. Dehak, W.M. Campbell, and J.R. Glass, "A channel-blind system for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[13] P. Rajan, T. Kinnunen, and V. Hautamaki, "Effect of multicondition training on i-vector PLDA configurations for speaker recogntion," in *Interspeech*, 2013.

[14] A. R. Avila, M. Sarria-Paja, F. J. Fraga, D. O'Shaughnessy, and T. Falk, "Improving the performance of far-field speaker verification using multi-condition training: The case of GMM-UBM and i-vector systems," in *Interspeech*, 2014.

[15] J. Villalba and E. Lleida, "Handling i-vectors from different recording conditions using multi-channel simplified PLDA in speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[16] J. Villalba, M. Diez, A. Varona, and et al., "Handling recordings acquired simultaneously over multiple channels with PLDA," in *Interspeech*, 2013.

[17] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[18] Yun Lei, Luciana Ferrer, Mitchell McLaren, and Nicolas Scheffer, "A deep neural network speaker verification system targeting microphone speech," in *Interspeech*, 2014.

[19] Mitchell McLaren, Yun Lei, and Luciana Ferrer, "Advances in deep neural network approaches to speaker recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4814–4818.

[20] M.J. Correia, A. Abad, and A. Brutti, "Multi-channel speaker verification based on total variability modelling," in *Interspeech*, 2015.

[21] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, 2011.

[22] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, 2000.

[23] Miguel Matos, Alberto Abad, Ramón Fernandez Astudillo, and Isabel Trancoso, "Recognition of distant voice commands for home applications in portuguese," in *Advances in Speech and Language Technologies for Iberian Languages - Second International Conference, IberSPEECH*, 2014, pp. 178–188.

[24] G. Panagiotis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *EUSIPCO*, 2015.

[25] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *LREC*, 2014.

[26] J. P. Neto, C. A. Martins, H. Meinedo, and L. B. Almeida, "The design of a large vocabulary speech corpus for Portuguese," in *Eurospeech*, 1997.

[27] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[28] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision*, 2007.