KNOWING THE NON-TARGET SPEAKERS: THE EFFECT OF THE I-VECTOR POPULATION FOR PLDA TRAINING IN SPEAKER RECOGNITION

David A. van Leeuwen and Rahim Saeidi

Centre for Language and Speech Technology CLST/CLS
Radboud University Nijmegen (RUN), The Netherlands
{d.vanleeuwen, r.saeidi}@let.ru.nl

ABSTRACT

Inspired by the NIST SRE-2012 evaluation conditions we train the PLDA classifier in an i-vector speaker recognition system with different speaker populations, either including or excluding the target speakers in the evaluation. Including the target speakers in the PLDA training is always beneficial w.r.t. completely excluding them—which is the normal situation in pre-2012 SRE protocols—even in the $P_{\mathrm{known}} = 0$ evaluation condition. However, adding other speakers than just the targets speakers can slightly increase performance. We also investigated the effect of adding i-vectors extracted from segments with added noise in the PLDA training. This generally makes the system more robust to noise in the test segments, and doesn't hurt performance in the clean condition. The paper further details the 'simple to compound' log-likelihood-ratio conversion necessary for SRE-2012 style calibration.

Index Terms— Speaker recognition, i-vector, PLDA, calibration, noise robustness.

1. INTRODUCTION

Automatic speaker recognition is an area of speech technology that is strongly driven by the series of Speaker Recognition Evaluations (SREs) as organized by the National Institute of Standards and Technology (NIST) [1]. Important aspects of such an evaluation are the task, the data and the evaluation metrics used to assess the performance of the speaker recognition systems. Until 2010, the NIST SREs main task was that of *speaker detection*, where an unknown test segment is compared to a target speaker, for which some training material is available. Traditionally essential in the evaluation was that no knowledge about other target speakers was allowed to be used in trials involving a particular target speaker. This was to ensure a certain application readiness of the system: a comparison score should be given for a given (test segment, target speaker) trial without having to wait for examples of possible non-target speakers that the system may be exposed to at some unknown time in the future.

In the 2012 edition of the NIST SRE, a radical change with respect to this has been made. This time, knowledge of *all* training speakers is allowed to be used for any trial, whereby the cost function reserves a fixed weight for false alarms stemming from *known non-targets* and those stemming from *unknown non-targets*. The latter are speakers used in the test as non-target speaker that are not part of the set of training speakers. Because the unknown non-target trials have the same function as ordinary non-target trials in pre-2012 NIST SREs, the major difference lies in the known non-target trials. In a way, this makes the task more like an (open set) *identification* task, even though the cost function remains specified in a detection framework. There are in fact parallels with the way language recognition is carried out in the NIST Language Recognition Evaluations.

The best performing automatic speaker recognition systems currently are based on subspace modeling using i-vectors [2], with probabilistic linear discriminant analysis (PLDA) modeling [3, 4, 5]. The i-vector approach still is based on the representation of the acoustic feature space using a universal background model (UBM, [6]) and uses subspace modeling techniques developed in the joint factor analysis approach [7]. Originally support vector machines and LDA were used as a back-end classifier [2], but more recently PLDA outperforms other classifiers and relieves the need for score normalization such as s-norm or T-norm, and generally shows relatively good calibration properties [8]. The PLDA model takes care of most of the channel and session compensation, and therefore the choice of training data used for PLDA training is important for the performance of the system.

In pre-2012 SREs, the PLDA training material must be chosen from data not in the evaluation, for reasons explained earlier. Because in SRE12 knowledge of all target speakers is allowed, we can use these speakers for PLDA training as well. This paper studies the effect of training PLDA using speakers from the evaluation, which should lead to more discriminative modeling. The motivation for this work was the NIST SRE-2012, in which the authors participated in both the RUN submission and the I4U collaboration.

Apart from the speaker population, this paper studies the effect of noise addition to PLDA training. Including noisy samples of clean data in the training phase can be carried out in a way to have multiple models for a single speaker (parallel models) [9] or a single overall model (multi condition training) [10]. A GMM-UBM system has also shown quite satisfying identification accuracy on the GRID corpus [11] (speech mixture) when a mixed-UBM and multi-conditioned GMMs are utilized [12]. Multi-condition training for i-vector based representation of utterances and Gaussian PLDA modeling has shown to be an effective way to handle additive noise [10] and reverberation [13].

The paper is organized in the following way. After specifying the data (Section 2), experimental setup (Section 3) and system (Sec-

This research was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

tion 4) we present in Section 5 the calibration strategy for SRE12 before the results are presented in Section 6.

2. DATA PREPARATION

In the preparation for SRE12 we worked with two trial sets, coined 'dev' and 'eval' for development and evaluation. Because SRE12 deals with known non-targets, we chose our development set to contain all target speakers in SRE12 known previously from SRE08 and SRE10 training and evaluation data. In SRE12, there are 100 additional target speakers defined, for which one conversation side training is available. We used these 'singleton' speakers in two ways: we used them in the development set as unknown non-target test segments, and in the evaluation set as target model (for which we would not have target test segments, unfortunately). The available speech data from SRE08 and SRE10 were distributed over train, dev-test and eval-test data sets according to a number of criteria. Firstly, we made sure that all segments with the same LDC session identification would appear within the same data set. Secondly, we attempted to populate the training segments for every target speaker with both microphone and telephone segments. The training material was used for both development test models and evaluation test models, where the latter included all segments per speaker of the former and had about twice as many in total. The evaluation test further had additional models for the 100 new singleton speakers. The dev and eval test segments were augmented with many segments from SRE06, in order to have abundant unknown non-targets in the test sets. The dev and eval test segment sets were completely disjoint.[1]

It was mentioned in the SRE12 evaluation plan [14] that some of the test segments contained noise, so therefore we added noisy versions of all segments used in training or testing, which were obtained by adding HVAC[2]-noise and speaker babble at 6 dB and 12 dB signal-to-noise ratio (SNR). The noise added using FaNT,[3] which takes into account the spectral properties of speech and noise in order to have a meaningful SNR. The babble noise was generated by mixing the noise of 100 speakers from one of our speech databases. The HVAC noise was obtained from public resources. The noise levels and types were inspired by information from NIST about these parameters, disclosed to the participants of SRE12.

3. EXPERIMENTAL SETUP

In the first experiment we want to study the effect of the speaker population for PLDA training, and specifically, whether target speakers from the evaluation are preferred or not. In order to do this, three PLDA training conditions were defined:

target All training i-vectors are from target speakers

non-target All training i-vectors are from other speaker than the target speakers

combined The PLDA training i-vectors are from both sources

Of course, it will make a difference if there are any speakers and channels in the test that are different from what is observed in PLDA training, because PLDA is the main classifier in the i-vector system that tries to separate speakers and reduces the influence of channels and sessions. In SRE12, three different analysis conditions are defined, where the weight of the unknown non-target test segments to the false alarm rates varies. This weight is indicated as P_{known}, and has the discrete values 0, $\frac{1}{2}$, and 1 for the three analysis conditions. We will therefore study the effect of PLDA training using these three evaluation conditions.

In a second study, we continue with the most optimal configuration from the first experiment, and this time concentrate on the effect of having noisy segments in the PLDA training. Again, we have several PLDA training conditions. The first three are 'clean', '15 dB' and '6 dB' corresponding to i-vectors conditioned on corresponding noise addition level in the speech segments, and a fourth condition is 'combined' which pools all three sets of i-vectors—this corresponds to one of the conditions in the first experiment. Again, the effect of training will depend on the expected test data, so here we will analyze results separately for test segments without additional noise, and with 15 dB and 6 dB added noise.

3.1. Evaluation metric

This research has been conducted in preparation to SRE12, with the SRE12 in mind, and therefore we will report results in terms of the official evaluation metric, C_{primary}. This metric measures both discrimination and calibration capabilities of the speaker recognition systems, and more specifically, at $N_\omega = 2$ operating points,

$$C_{\text{primary}} = \frac{1}{N_\omega} \sum_\omega C_{\text{norm}}(\omega). \tag{1}$$

Here ω represents the prior log odds used as operating points, for SRE12 equal to $\{\log \frac{1}{99}, \log \frac{1}{999}\}$, and $C_{\text{norm}}(\omega)$ is the normalized Bayes' error rate

$$C_{\text{norm}}(\omega) = (1 + e^{|\omega|}) C_{\text{det}}(\omega), \tag{2}$$

where $C_{\text{det}}(\omega)$ is the traditional detection cost for the prior log odds ω. Because SRE12 contains known non-target trials, the contribution of the false alarms to C_{det} is split in two parts, weighted by P_{known} and $1 - P_{\text{known}}$, respectively:

$$P_{\text{FA}}(\omega) = P_{\text{known}} P_{\text{FA}}^{\text{kn}}(\omega) + (1 - P_{\text{known}}) P_{\text{FA}}^{\text{unk}}(\omega) \tag{3}$$

where

$$P_{\text{FA}}^{\text{kn}}(\omega) = \frac{\sum_{t \in \mathcal{K}} u(s_t + \omega)}{||\mathcal{K}||} \tag{4}$$

$$P_{\text{FA}}^{\text{unk}}(\omega) = \frac{\sum_{t \in \mathcal{U}} u(s_t + \omega)}{||\mathcal{U}||}, \tag{5}$$

where $\mathcal{K}, \mathcal{U}$ are the sets of known an unknown non-targets trials, respectively, and $||\dots||$ is the cardinality operator that counts trials. The unit step function u counts recognition scores s_t that are higher than $-\omega$, the decision threshold that leads to a Bayes' decision. With the traditional definition for the miss probability, using the set of target trials $\mathcal{T}$, C_{det} can be completed

$$P_{\text{miss}}(\omega) = \frac{\sum_{t \in \mathcal{T}} u(-s_t - \omega)}{||\mathcal{T}||} \tag{6}$$

$$C_{\text{det}}(\omega) = S(\omega) P_{\text{miss}}(\omega) + S(-\omega) P_{\text{FA}}(\omega), \tag{7}$$

where we have used the logistic function $S(x) = 1/(1 + e^{-x})$ to express the prior for targets and non-targets in terms of the prior log odds ω.

[1] We have made the development and evaluation sets available via `http://lands.let.ru.nl/~saeidi/`.

[2] Heating, Ventilation and Air Conditioning

[3] `http://dnt.kr.hsnr.de/download.html`

4. THE SPEAKER RECOGNITION SYSTEM

The speaker recognition system at RUN consists of a standard i-vector configuration with PLDA modeling. We use 19 MFCC's plus log energy computed every 10 ms over a 30 ms window, and augment these features with first and second order derivatives computed over 9 consecutive frames, followed by short time Gaussianization [15]. After speech activity detection, the system is gender-dependent. For each gender, a 2048-component UBMs has been trained, using segments from the data sets NIST SRE04–06, Switchboard cellular phase 1 and 2, and Fisher English. Using the UBMs, from each relevant utterance (background, train, or test segments) 0th, 1st and 2nd order Baum-Welch statistics are computed w.r.t. the UBM. An i-vector extractor matrix $\mathbf{T}$ of rank 400 has been trained using the statistics from the same utterances used for training the UBM. Next, for each relevant utterance, an i-vector is extracted using the statistics and $\mathbf{T}$. After applying LDA to map the i-vectors on a space with 200 dimensions, i-vectors are centered, whitened and length-normalized [16]. The speaker and session dependent i-vector distribution is modeled using PLDA [4]. Finally a score for a trial is based on the log likelihood ratio expression of the likelihoods of the pair of i-vectors originating from the same speaker versus different speakers. Scores are then converted to calibrated log-likelihood-ratios using a linear calibration transformation. Finally, the calibrated log-likelihood-ratios are converted to 'compound LLRs' suitable for submission to NIST SRE-2012.

5. CALIBRATION

Calibration is based on the log-likelihood-ratio (LLR) representation that we are familiar with in SRE10 and earlier, i.e., under the assumption that the speakers in a non-target test segment have not been observed before in training. This is equivalent to the testing condition $P_{\text{known}} = 0$ in SRE12. Our calibration transformation of these so-called 'simple[4] LLRs' $s(x, y)$ for a trial involving training speech segments x and test segment y is

$$\lambda(x, y) = w_0 + w_1 s(x, y) \tag{8}$$

The offset w_0 and scaling w_1 are found by minimizing the multi-class cross entropy H_{mc} [17] over the development set. H_{mc} is defined in terms of the posterior probability of the true class, by

$$H_{\text{mc}} = \sum_{i=0}^{N} \frac{\pi_i}{N_i} \sum_{j=1}^{N_i} -\log P(i \mid x, y_j). \tag{9}$$

Here i indexes the N target speakers, using $i = 0$ for an unknown speaker, and j runs over all N_i test segments for which i is the the speaker. For the priors π_i we were inspired by the NIST SRE core conditions, setting $\pi_0 = 1 - P_{\text{known}}$ and $\pi_{i>0} = P_{\text{known}}/2N$. The posterior in (9) is computed using

$$P(i \mid x, y_j) = \frac{\pi_i e^{\lambda_i(x,y_j)}}{\pi_0 + \sum_{k=1}^{N} \pi_k e^{\lambda_k(x,y_j)}}. \tag{10}$$

Note that we use the notation $\lambda_i(x, y_j)$ to indicate the simple likelihood ratio for test segment y_j with speaker i in the target hypothesis using all available training material x. We used a standard general numerical optimizer `nlm` from the R software package for finding the calibration parameters.

[4]Term coined by Niko Brümmer

5.1. Compound LLRs

The denominator in the 'simple LLRs' λ_i is the likelihood of the test segment given the fact it is not any of the known target speakers. This is different from the log likelihood ratio required for NIST SRE12, for which the denominators condition indicates 'not the target speaker,' i.e., including any known non-target speaker. For the conversion from our simple LLRs in (8) it is easiest to start with the posterior defined in (10), and compute the posterior odds using the fact that $P(\neg i \mid x, y_j) = 1 - P(i \mid x, y_j)$

$$\frac{P(i \mid x, y_j)}{P(\neg i \mid x, y_j)} = \frac{\pi_i e^{\lambda_i(x,y_j)}}{\pi_0 + \sum_{k=1}^{N} \pi_k e^{\lambda_k(x,y_j)} - \pi_i e^{\lambda_i(x,y_j)}} \tag{11}$$

$$= \frac{\pi_i e^{\lambda_i(x,y_j)}}{\pi_0 + \sum_{k \neq i} \pi_k e^{\lambda_k(x,y_j)}}. \tag{12}$$

Here the last summation runs over all known speakers excluding the target speaker. The prior odds, in SRE12 sense, are just

$$\frac{P(i)}{P(\neg i)} = \frac{\pi_i}{\pi_0 + \sum_{k \neq i} \pi_k}. \tag{13}$$

The required log-likelihood-ratio now simply is the posterior odds (12) divided by the prior odds (13):

$$\lambda_i^{\text{comp}}(x, y_j) = \log \frac{\big(\pi_0 + \sum_{k \neq i} \pi_k\big) e^{\lambda_i(x,y_j)}}{\pi_0 + \sum_{k \neq i} \pi_k e^{\lambda_k(x,y_j)}} \tag{14}$$

This expression, while nicely not explicitly dependent on the prior π_i—which is one of the reasons to use a likelihood ratio—shows a dependency on all other priors *and* all other 'simple LLRs' $\lambda_k(x, y_j)$. This is the reason why the form (14) is also known as the 'compound[5] LLR.' The expression can be simplified in appearance by including the unknown non-target speakers $i = 0$ in the summations, and using $\lambda_0(x, y_j) = 0$.

6. EXPERIMENTAL RESULTS

In the first experiment, we trained three different PLDA models, for both male and female, and computed 'simple LLR's for both the dev and eval data. We then trained a global calibration transformation using each of the six development set scores, and applied these to the eval set scores using (8). Next, the simple→compound transformation (14) was carried out using the values $P_{\text{known}} = \{0, \frac{1}{2}, 1\}$, as according to the SRE12 evaluation protocol, leading to 18 calibrated 'compound LLR' sets. Each of these sets was further analyzed separately for the noise conditions 'clean,' '15 dB' and '6 dB,' and a value of C_{primary}, using corresponding values of P_{known} was computed for each of these conditions, leading to 54 values of C_{primary} for the factors gender (2), PLDA training (3), P_{known} (3), and noise (3). In Table 1 the values for C_{primary} averaged over noise condition are shown for the evaluation set, and in Figure 1 a 'box plot' is shown to indicate the interaction between the factors 'PLDA training' and 'P_{known}'.

From Figure 1 we can observe that knowing the target speakers in PLDA always helps, even for $P_{\text{known}} = 0$. In that case, apparently the target trials are helped by having target speaker data explicitly modeled by the PLDA. This is a condition that was not sensible to investigate with pre-2012 SRE evaluation protocols. Inspecting Table 1, there appears to be a slight preference for adding other speaker data in the PLDA training—if anything, it doesn't seem to hurt.

[5]Again, a term coined by Niko Brümmer

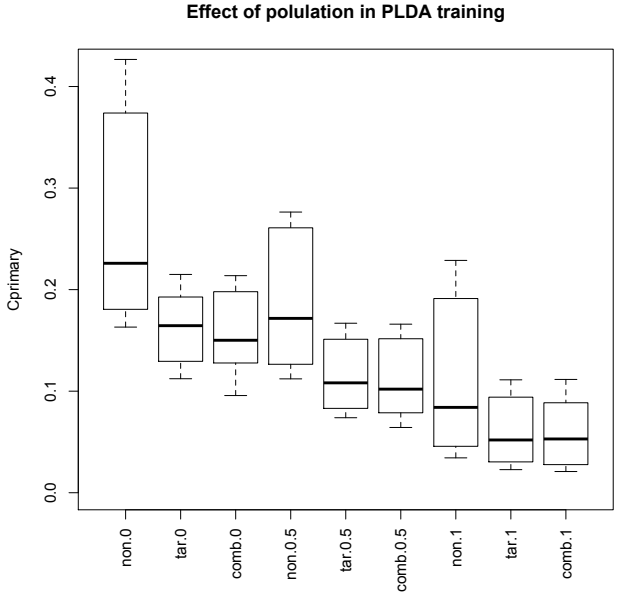

Fig. 1. Box plot showing the interaction between PLDA training condition (non, tar, comb) and P_{known} (0, 0.5, 1). Each box represents 6 values of C_{primary} (gender, noise).

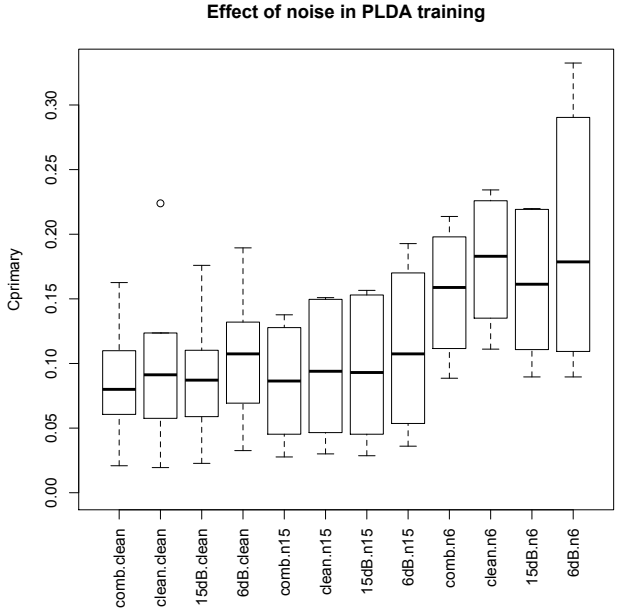

Fig. 2. Effect of the factors *PLDA training* (combined, clean, 15 dB, 6 dB) and *test noise level* (clean, n15, n6) for the eval test. Each box represents 6 values

Table 1. The values of C_{primary} of the evaluation set, averaged over the three noise conditions.

PLDA	female			male		
P_{known}	0	0.5	1	0	0.5	1
non	0.273	0.207	0.1323	0.259	0.1665	0.0904
tar	0.181	0.128	0.0717	0.145	0.1027	0.0491
comb	0.171	0.123	0.0725	0.140	0.0982	0.0457

In the second experiment, we chose the 'combined' training condition for PLDA in order to further inspect the effect of having included noisy versions of the segments in the PLDA training. The experimental setup was very similar to the first experiment, but this time we conditioned the PLDA training on noise in the segments from which the i-vectors were extracted, choosing the i-vectors from the combined target speakers and external non-target speakers. Again, we applied gender and PLDA-training dependent calibration, applied simple→compound LLR transform for three values of P_{known}, and analyzed C_{primary} in corresponding values of P_{known} for three different noise level subsets. For the eval set this gave again 54 values of C_{primary}, and we have shown a cross section of these for the interacting factors PLDA training and test set noise level in Figure 2. In this figure, we included the 'combined' data from experiment 1 for comparison.

From Figure 2 we see that the noise condition in PLDA training should not deviate too much from what is expected in the test, but again, it doesn't seem to hurt to simply add all noise conditions in the PLDA training, which is computationally not a limiting factor.

7. DISCUSSION AND CONCLUSION

The biggest change in paradigm in NIST SRE-2012 probably is the concept of the known non-target trial. The counterpart, the unknown non-target trial, is what in pre-2012 SREs was known as a ordinary non-target trial, i.e., the evaluation protocol forbade any knowledge of other target speakers for any trial. The knowledge of non-target speakers can be utilized in, e.g., discriminative modeling. In the first experiment we saw that knowledge of the non-target speakers in PLDA modeling was always beneficial w.r.t. ignoring the information, even when only evaluating with unknown non-targets speakers (i.e., $P_{\text{known}} = 0$). Even though there is no overlap in non-targets in the evaluation and those in the PLDA model fore the $P_{\text{known}} = 0$ case, the fact that the target trials have speakers observed by the PLDA model helps for a better performance. Further, the inclusion of *other* speakers than targets in the PLDA does not seem to hurt performance for any P_{known} condition. This is in line with the second experiment, where multi-condition noise level training (the same segments occurring multiple times in the PLDA training set with different amounts of added noise) seems always at least as accurate as training with noise levels specifically tailored to the test condition. Since it appears we can simply add more 'versions' of i-vectors extracted from the same speech segment in different conditions to make the model more robust, we suggest that also *artificial* perturbation of the i-vectors can have the same effect. Indeed, in experiments conducted on similar data where the effect of reduced duration was studied, such synthesized i-vectors in PLDA traing has led to an improved performance with variable duration test segments, which we have reported in another submission to this conference [18].

8. REFERENCES

[1] Craig S. Greenberg, Alvin F. Martin, Bradford N. Barr, and George R. Doddington, "Report on performance results in the NIST 2010 speaker recognition evaluation," in *Proc. Interpseech*. 2011, pp. 261–264, ISCA.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2009.

[3] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[4] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. ICASSP*. May 2011, pp. 4832–4835, IEEE.

[5] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Proc. of Odyssey Speaker and Language Recognition Workshop*, 2012.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[8] Mitchell McLaren and David A. van Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Proc. Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, 2012, ISCA, pp. 55–61.

[9] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. ICASSP*. April 1987, vol. 12, pp. 705–708, IEEE.

[10] Yun Lei, Lukáš Burget, Luciana Ferrer, Martin Graciarena, and Nicolas Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. ICASSP*. 2012, pp. 4253–4256, IEEE.

[11] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421, 2006.

[12] R. Saeidi, P. Mowlaee, T. Kinnunen, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR 2010)*, 2010, pp. 4545–4548.

[13] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multi-condition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. ICASSP*. IEEE, 2012.

[14] Craig S. Greenberg, "The NIST year 2012 speaker recognition evaluation plan," 2012.

[15] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*. 2001, pp. 213–218, Crete, Greece.

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

[17] Luis Javier Rodríguez-Fuentes, Niko Brümmer, Mikel Penagarikano, Amparo Varona, Mireia Diez, and Germán Bordel, "The Albayzin 2012 language recognition evaluation plan," Nov. 2012.

[18] Taufiq Hasan, Rahim Saeidi, John H. L. Hanson, and David A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. ICASSP*. IEEE, 2013.