INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Probabilistic approach using joint clean and noisy i-vectors modeling for speaker recognition

Waad Ben Kheder, Driss Matrouf, Moez Ajili and Jean-François Bonastre

LIA, University of Avignon, France

Abstract

Additive noise is one of the main challenges for automatic speaker recognition and several compensation techniques have been proposed to deal with this problem. In this paper, we present a new "data-driven" denoising technique operating in the i-vector space based on a joint modeling of clean and noisy i-vectors. The joint distribution is estimated using a large set of i-vectors pairs (clean i-vectors and their noisy versions generated artificially) then integrated in an MMSE estimator in the test phase to compute a "cleaned-up" version of noisy test i-vectors. We show that this algorithm achieves up to 80% of relative improvement in EER. We also present a version of the proposed algorithm that can be used to compensate multiple "unseen" noises. We test this technique on the recently published SITW database and show a significant gain compared to the baseline system performance.

Index Terms: speaker verification, i-vector, additive noise, joint modeling.

1. Introduction

The performance of i-vector-based SR systems suffers considerably in presence of environment noise and a wide variety of algorithms have been proposed to deal with this problem using different approaches. These techniques can be divided into three groups: The first aims at providing better features either by proposing more robust parameters or by removing noise (such as speech enhancement and stochastic feature compensation techniques). The second aims at improving the quality of i-vectors (by building more robust i-vector extractors or by cleaning-up noisy i-vectors). The third one accounts for the noise effect in the scoring phase by improving the PLDA model.

In the first class of algorithms, spectral and wavelet-based speech enhancement techniques were studied in [1, 2] and were proven to be noise and SNR level-dependent (performance can improve or degrade depending on the noise / SNR). NFM-based speech enhancement algorithms were also presented in [3–5] and showed relatively low improvement compared to other techniques (10% of relative improvement in EER. In the cepstral domain, several stochastic compensation techniques were applied in [6] (such as RATZ [7], SPLICE [8], TRAJMAP [9] and SSM [10]) and were proven to be highly efficient in noisy environments but such algorithms assume prior knowledge about the test noise. Alternatively, a DNN was used in [11] to enhance cepstral features before extracting i-vectors and achieved an improvement rate that varies from 3% up to 26%.

In the second class of algorithms, a set of techniques based on vector Taylor series (VTS) were proposed in [12, 13] then developed using "unscented transforms" [14]. Such algorithms use a non-linear noise model in the cepstral domain and model the relationship between clean and noisy cepstral coefficients. In the recognition phase, the developed noise model is integrated in the i-vector extractor to help estimate a "cleaned-up" version of noisy i-vectors. Despite their efficiency, such models are rigid and can be hard to adapt (adding a normalization step or changing the used parameters could mean to rewrite the whole technique). Alternatively, other algorithms based on uncertainty propagation have also been proposed lately for robust speaker recognition. Based on this idea, a robust i-vector extractor was proposed in [15, 16] in order to make the i-vector extraction system focus on reliable or reliably enhanced features but showed little improvement compared to other methods. Another technique was presented in [17] using a convolutional neural network (CNN) to compute posterior probabilities for speech frames replacing the UBM model. It has been shown to produce more robust i-vector statistics achieving up to 26% of relative EER improvement.

In the third class of algorithms, a robust backend training called "multi-style" was proposed in [18] as a possible solution to account for the effect of noise. This method uses a large set of clean and noisy data (affected with different noises and SNR levels) to build a generic scoring model. The obtained model gives good performance in general (up to 30% of relative improvement) but might not be optimal to for a particular noise because of its lack of generalization (the same system is used for all noises). Recently, an SNR-invariant version of PLDA was proposed in [19]. In this framework, i-vectors extracted from utterances falling within a narrow SNR range are assumed to share similar SNR-specific information and used to develop a more robust version of PLDA which decomposes an i-vector in three components: speaker, SNR, and channel. This model showed an average relative improvement of 25% in EER compared to regular PLDA.

We recently presented a new efficient Bayesian cleaning technique operating in the i-vector domain named I-MAP [20–22]. It is a "data-driven" i-vector cleaning method based on an additive noise model in the i-vector space. It estimates a clean i-vector given its noisy version and the noise distribution using MAP approach. It uses a full-covariance Gaussian modeling of the clean i-vectors and noise distributions in the i-vector space. Even though the noise is known to be non-additive in this space, using such model with a MAP estimator makes the derivations very simple while giving up to 60% of relative improvement compared to the baseline system performance.

In this paper, we present a new model-free i-vector denoising technique which operates in the i-vector space inspired from the "Stereo Stochastic Mapping" (SSM) algorithm [10,23]. Unlike I-MAP which supposes statistical independence between clean and noisy i-vectors distributions, this technique takes advantage of the joint information between clean and noisy i-vectors. This algorithm starts by building a joint distribution between clean i-vectors and their noisy versions (generated artificially). Then, this distribution is integrated in an MMSE

http://dx.doi.org/10.21437/Interspeech.2016-1292

(Minimum Mean Square Error) estimator in the test phase to compute a "cleaned-up" version of noisy test i-vectors. This technique provides more flexibility than I-MAP and offers three advantages: First, this technique takes advantage of the joint information between clean and noisy i-vectors distributions. Second, the estimated i-vectors are supposed to be noise-free, so it is possible to use a clean backend for scoring. Finally, it is possible to use this model for multiple noises while achieving high recognition rates. We show that this algorithm achieves up to 80% of relative improvement in EER when used to compensate a specific noise. Then, we develop a fast version that can be used to compensate multiple "unseen" noises and show that it yields up to 70% of relative EER improvement. We also test this technique on the recently published SITW database and show a significant gain compared to the baseline system performance.

This paper is structured as follows: Section 2 presents the proposed technique based on a joint probability model of clean and noisy i-vectors. Section 3 presents the experimental protocol and Section 4 presents the results and analysis.

2. A joint probability model of clean and noisy i-vectors for i-vectors denoising

This technique is inspired from the "Stereo Stochastic Mapping" algorithm (SSM) which was first introduced for robust speech recognition [10, 23] then adapted to speaker verification [6]. In speaker recognition context, SSM was used to create a mapping between noisy and clean cepstral features based on their joint distribution. Unlike other denoising techniques operating in the i-vector space such as I-MAP [21] which relies on statistical independence between clean and noisy i-vectors distributions, we aim at developing a model-free denoising technique that takes advantage of both clean and noisy i-vectors distributions along with the joint information between the two.

Let's define two random variables x and y representing respectively clean i-vectors and their noisy versions (affected by a certain noise at a certain SNR level) and let M be the dimension of the i-vector space. We define a third random variable called z as the concatenation of x and y:

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \tag{1}$$

Such variable lies in a $2M$-dimensional space and can be modeled using a mixture of Gaussians:

$$p(z) = \sum_{k=1}^{K} c_k \mathcal{N}(z; \mu_{z,k}, \Sigma_{z,k}) \tag{2}$$

where:

- K is the number of GMM components.
- c_k is the weight of the k^{th} Gaussian.
- $\mu_{z,k}$ corresponds to the mean vector of the k^{th} component.
- $\Sigma_{z,k}$ corresponds to the covariance matrix of the k^{th} component.

This GMM represents the joint distribution between clean and noisy i-vectors (affected by a certain noise/SNR level) for each Gaussian k. For each component, it is possible to decompose the mean and covariance matrix as:

$$\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix} \tag{3}$$

$$\Sigma_{z,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix} \tag{4}$$

Where $\Sigma_{yx,k} = \Sigma_{xy,k}^T$ and $\Sigma_{xy,k}$ models the joint covariance between the two representations (clean and noisy i-vectors).

The problem of i-vector denoising can be formulated using an MMSE (Minimum Mean Square Error) estimator. For a given noisy i-vector y_0, the corresponding clean version can be computed as:

$$\begin{aligned} \hat{x} = E[x|y_0] &= \int_x p(x|y_0) x dx = \sum_k \int_x p(x,k|y_0) x dx \\ &= \sum_k p(k|y_0) \int_x p(x|k,y_0) x dx = \sum_k p(k|y_0) E[x|k,y_0] \end{aligned} \tag{5}$$

For each component k, the Schur complement [24] can be used to compute $E[x|k, y_0]$ as:

$$E[x|k,y_0] = \mu_{x,k} + \Sigma_{xy,k}\Sigma_{yy,k}^{-1}(y_0 - \mu_{y,k}) \tag{6}$$

The final solution can then be written as:

$$\begin{aligned} \hat{x} &= \sum_k p(k|y_0) E[x|k,y_0] \\ &= \sum_k p(k|y_0)(\mu_{x,k} + \Sigma_{xy,k}\Sigma_{yy,k}^{-1}(y_0 - \mu_{y,k})) \end{aligned} \tag{7}$$

Equation 7 can be re-written as:

$$\hat{x} = \sum_{k=1}^{K} p(k|y_0)(F_k y_0 + g_k) \tag{8}$$

with:

$$F_k = \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \tag{9}$$

$$g_k = \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k} \tag{10}$$

In other words, this MMSE-based mapping is a weighted sum of linear functions contributed by each Gaussian component k from the joint GMM distribution $p(x, y)$. The weight is the posterior probability $p(k|y)$ and the linear function is built using the hyper-parameters of each component.

3. Experimental protocol

Our experiments operate on 19 Mel-Frequency Cepstral Coefficients (plus energy) augmented with 19 first (Δ) and 11 second ($\Delta\Delta$) derivatives. A mean and variance normalization (MVN) technique is applied on the MFCC features estimated using the speech portion of the audio file. The low-energy frames (corresponding mainly to silence) are removed.
A gender-dependent 512 diagonal component UBM (male model) and a total variability matrix of low rank 400 are estimated using 15660 utterances corresponding to 1147 speakers (using NIST SRE 2004, 2005, 2006 and Switchboard data). The LIA_SpkDet package of the LIA_RAL/ALIZE toolkit is used for the estimation of the total variability matrix and the i-vectors extraction. The used algorithms are described in [25]. Finally a PLDA-based scoring [26] is applied. The eigenvoice rank the PLDA models is equal to 100 and the eigenchannel matrix is kept full-rank (400). PLDA is preceded by 2 iterations of LW-normalization (spherical nuisance normalization [27]). The equal-error rate (EER) over the NIST SRE 2008 male test data on the "short2/short3" task under the "det7" conditions [28] will be used as a reference to monitor the performance improvements compared to the baseline system.

Table 1: Comparison of the 3 systems using clean enrollment and noisy test data.

Test condition		EER				
		Baseline	I-MAP	JointModel		
				1 Gauss.	2 Gauss.	3 Gauss.
Car driving noise	0dB	11.15	4.99	3.87	8.03	9.11
	5dB	5.90	2.96	2.28	4.36	4.96
	10dB	3.64	2.28	1.82	2.49	2.80
	15dB	2.54	2.03	1.80	1.89	2.12
Air-cooling noise	0dB	11.84	4.78	3.47	7.59	8.81
	5dB	6.80	3.64	2.74	5.01	5.67
	10dB	4.11	2.93	2.31	3.02	3.38
	15dB	3.18	2.23	1.82	2.22	2.53

Table 2: Comparison of the 3 systems using noisy enrollment and test data (enrollment segments are affected by air-cooling noise and test segments are affected by car-driving noise).

Enrol./test SNR	EER				
	Baseline	I-MAP	JointModel		
			1 Gauss.	2 Gauss.	3 Gauss.
0dB	26.22	10.48	7.28	15.13	17.16
5dB	21.17	8.89	4.04	8.39	9.52
10dB	14.80	6.51	2.52	5.23	5.94
15dB	10.27	4.21	2.05	4.26	4.83

We use 5 noise samples from the free sound repository FreeSound.org [29] as background noises: {air-cooling, car-driving, crowd-noise, nature noise and rain noise}. The open-source toolkit FaNT [30] was used to add these noises to the full waveforms generating new noisy audio files for each noise / SNR level.

4. Experiments and results

In this section, we start by testing the proposed denoising technique in clean enrollment / noisy test and noisy enrollment / noisy test setups, then we analyze the amount of train data needed for optimal results. Finally, we present a real implementation of this version by using it to compensate multiple "unseen" noises.

4.1. Used systems and preliminary results

In this section, we test the proposed denoising technique in two configurations:

- Clean enrollment / noisy test: In this configuration, test segments are affected by car-driving noise at 0dB, 5dB, 10dB and 15dB.
- Noisy enrollment / noisy test: In this configuration, enrollment segments are affected by air-cooling noise at 0dB, 5dB, 10dB and 15dB and test segments are affected by car-driving noise at 0dB, 5dB, 10dB and 15dB.

We compare the performance of 3 systems:

- **Baseline:** A clean PLDA backend is used (no noise compensation is performed).
- **I-MAP:** Noisy enrollment/test i-vectors are denoised using I-MAP as described in [21] (500 train i-vectors are generated to estimate the noise distribution for each enrollment/test noise), then scored using a clean PLDA backend.
- **JointModel:** For each noise $\mathcal{N}$ and SNR level $\mathcal{S}$:
  1. The clean train set (15660 sessions) is affected by noise $\mathcal{N}$ at $\mathcal{S}$dB.
  2. The i-vectors are extracted for the clean sessions $\{x_i\}$ and their noisy versions $\{y_i\}$.
  3. The distribution $p(z)$ is estimated using 20 iterations of the EM algorithm (different number of components are tested for each experiment; $K \in \{1, 2, 3\}$).
  4. Each noisy enrollment/test i-vector affected by $\mathcal{N}$ at $\mathcal{S}$dB is denoised using Equation 8.

  Then, the resultant i-vectors are scored using a clean PLDA backend.

Tables 1 and 2 show respectively the performance of these 3 systems for clean enrollment/noisy test and noisy enrollment/noisy test conditions.

It is easy to see that the proposed joint model denoising technique achieves up to 80% of relative improvement compared to the baseline system performance. This system also outperforms I-MAP (60% of relative improvement in EER). We observe that training $p(z)$ using more than one Gaussian deteriorates the results (compared to the 1-Gaussian version).

A large set of train i-vectors was used in this experiment to estimate $p(z)$. This constraint is unpractical in real applications. In the next subsection, we try to find the minimum amount of data needed to have an optimal performance.

4.2. Effect of the amount of train data used to train the joint model on performance

In this experiment, we use the 1-Gaussian modeling since it gave the best results and investigate the amount of data needed by the joint model to achieve good results. Figure 1 presents the performance of the JointModel system compared to the baseline when used on clean enrollment/noisy test configuration (affected by Car Driving noise at 0dB, 5dB, 10dB and 15dB). It is

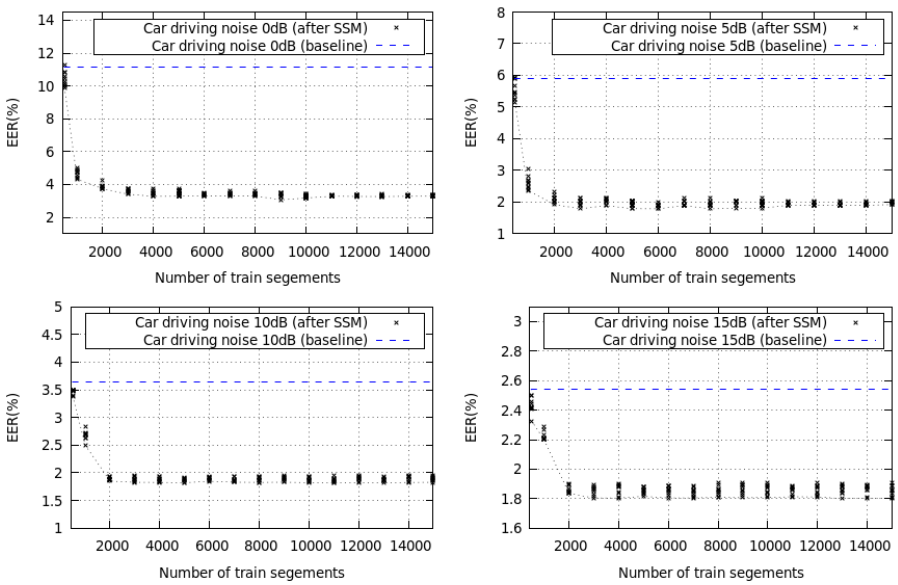

Figure 1: EER variation with the amount of i-vectors used to train $p(z)$ for the "car-driving noise" at 0dB, 5dB, 10dB and 15dB (10 measures for each number of segments).

clear from Figure 1 that more than 3000 pairs of i-vectors are needed to train $p(z)$ in order to achieve good results using the joint model. Using more training data will not yield better results. It is important to mention that the I-MAP algorithm [21] requires only 500 sessions for the training.

4.3. Using the joint model on unseen noises:

In real applications, time is a very important factor and generating 3000 i-vector for each test noise is not feasible. This motivates us to use a single generic model to compensate multiple noises. The three noises {crowd-noise, nature noise and rain noise} are used to affect randomly clean train segments at different SNR levels between 0dB and 15dB (one noise and SNR level are used for each segment) and the resultant i-vectors are used to train a generic version of the JointModel system.

Table 3 compares the performance of the baseline system and the JointModel when trained using test noise and the generic version.

Table 3: Comparison of joint model denoising using test noise and a mixture of noises (not observed in test conditions).

		EER		
Test and enrol. condition		**Baseline**	**JointModel (test noise)**	**JointModel (generic)**
Car-driving noise	**0dB**	21.18	4.78	6.37
	5dB	15.67	3.84	3.91
	10dB	11.64	2.73	2.50
	15dB	8.46	2.28	2.05
Air-cooling noise	**0dB**	18.39	5.01	6.51
	5dB	16.17	3.18	3.82
	10dB	13.21	2.72	2.95
	15dB	10.47	2.50	2.61

4.4. Performance on SITW:

In order to test our technique in real conditions, we will apply the joint model denoising technique on the recently published SITW database [31] (Speakers in the Wild). In this experiment we compared the baseline system performance to the generic version of the JointModel system; the three noises {crowd-noise, nature noise and rain noise} are used to affect randomly clean train segments at different SNR levels between 0dB and 15dB (one noise and SNR level are used for each segment) and the resultant i-vectors are used to estimate $p(z)$. In this experiment, only long and noisy sessions ($> 30s$ of speech and $SNR < 10dB$) corresponding to male speakers are used. Table 4 shows the results of the proposed technique compared to the baseline system performance.

Table 4: Performance of the JointModel system used on long male noisy test utterances.

EER	
Baseline	**JointModel (generic)**
12.69	4.24

The generic system achieves 66% of relative EER improvement compared to the baseline system. These results confirm the efficiency of the proposed technique on real conditions.

5. Conclusion

In this paper, we presented a new i-vector denoising technique based on a joint model of clean and noisy i-vectors. The joint distribution is estimated using a large set of i-vectors pairs (clean i-vectors and their noisy versions generated artificially) then integrated in an MMSE estimator in the test phase to compute a "cleaned-up" version of noisy test i-vectors. We tested this algorithm is various configurations and showed it can achieve up to 80% of relative improvement in EER. Then, we developed a version that can be used to efficiently compensate multiple "unseen" noises and tested it on the SITW database and showed a significant gain compared to the baseline system performance.

6. References

[1] A. El-Solh, A. Cuhadar, and R. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on*. IEEE, 2007, pp. 235–239.

[2] S. O. Sadjadi and J. H. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions." in *INTERSPEECH*, 2010, pp. 2138–2141.

[3] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors." in *ICASSP*, 2008, pp. 4029–4032.

[4] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising." in *Interspeech*, 2008, pp. 411–414.

[5] S. Liu, Y. Zou, and H. Ning, "Nonnegative matrix factorization based noise robust speaker verification," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 2015, pp. 35–39.

[6] S. Sarkar and K. Sreenivasa Rao, "Stochastic feature compensation methods for speaker verification in noisy environments," *Applied Soft Computing*, vol. 19, pp. 198–214, 2014.

[7] P. J. Moreno, B. Raj, E. Gouvea, and R. M. Stern, "Multivariate-gaussian-based cepstral normalization for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 137–140.

[8] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments." in *INTERSPEECH*, 2000, pp. 806–809.

[9] H. Zen, Y. Nankaku, and K. Tokuda, "Stereo-based stochastic noise compensation based on trajectory gmms," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4577–4580.

[10] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1325–1334, 2009.

[11] S. Du, X. Xiao, and E. S. Chng, "Dnn feature compensation for noise robust speaker verification," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 2015, pp. 871–875.

[12] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6788–6791.

[13] Y. Lei, M. McLaren, L. Ferrer, and N. Scheffer, "Simplified vts-based i-vector extraction in noise-robust speaker recognition," *ICASSP, Florence, Italy*, 2014.

[14] D. Martınez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," *ICASSP, Florence, Italy*, 2014.

[15] C. Yu, G. Liu, S. Hahm, and J. H. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4017–4021.

[16] D. Ribas, E. Vincent, and J. R. Calvo, "Uncertainty propagation for noise robust speaker recognition: the case of nist-sre," in *Interspeech 2015*, 2015.

[17] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions." in *INTERSPEECH*, 2014, pp. 686–690.

[18] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4253–4256.

[19] N. Li and M.-W. Mak, "Snr-invariant plda modeling for robust speaker verification," *Proc. Interspeech'15*, 2015.

[20] W. B. Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, and M. Ajili, "Robust speaker recognition using map estimation of additive noise in i-vectors space," in *Statistical Language and Speech Processing*. Springer, 2014, pp. 97–107.

[21] W. Ben Kheder, D. Matrouf, J.-F. Bonastre, M. Ajili, and P.-M. Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4190–4194.

[22] D. Matrouf, W. Ben Kheder, P. Bousquet, M. Ajili, and J. Bonastre, "Dealing with additive noise in speaker recognition systems based on i-vector approach," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 2092–2096.

[23] J. Du and Q. Huo, "Synthesized stereo-based stochastic mapping with data selection for robust speech recognition," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 122–125.

[24] D. V. Ouellette, "Schur complements and statistics," *Linear Algebra and its Applications*, vol. 36, pp. 187–295, 1981.

[25] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." in *INTERSPEECH*, 2007, pp. 1242–1245.

[26] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[27] 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 *Odyssey*, 2012, pp. 157–164.

[28] "The NIST year 2008 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig//tests/sre/2008/, 2008, [Online; accessed 15-May-2014].

[29] "Freesound.org," http://www.freesound.org.

[30] H. G. Hirsch, "FaNT - Filtering and Noise Adding Tool," http://dnt.kr.hsnr.de/download.html, [Online; accessed 15-May-2014].

[31] D. C. A. L. Mitchell McLaren, Luciana Ferrer, "The speakers in the wild (sitw) speaker recognition database," in *Submitted to Interspeech 2016*.