Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **23** (2023) 015201 (123-137) AKU J. Sci. Eng. **23** (2023) 015201 (123-137)

DOI: 10.35414/akufemubid.1134945

Araştırma Makalesi / Research Article

# Increasing the Robustness of i-vectors with Model Compensated First Order Statistics

**Gökay DİŞKEN** *[1], **Zekeriya TÜFEKCİ**[2]

[1] *Adana Alparslan Turkes Science and Technology University, Department of Electrical-Electronics Engineering, Turkey*

[2] *Cukurova University, Department of Computer Engineering, Turkey*

*Sorumlu yazar e-posta*: gdisken@atu.edu.tr ORCID ID:http://orcid.org/0000-0002-8680-0636*

*e-posta: ztufekci@atu.edu.tr ORCID ID:http://orcid.org/0000-0001-7835-2741*

Geliş Tarihi: 24.06.2022 Kabul Tarihi: 26.01.2023

## Abstract

***Keywords***

Parallel model compensation;Robust speaker recognition;Vector Taylor series; I-vector

Speaker recognition systems achieved significant improvements over the last decade, especially due to the performance of the i-vectors. Despite the achievements, mismatch between training and test data affects the recognition performance considerably. In this paper, a solution is offered to increase robustness against additive noises by inserting model compensation techniques within the i-vector extraction scheme. For stationary noises, the model compensation techniques produce highly robust systems. Parallel Model Compensation and Vector Taylor Series are considered as state-of-the-art model compensation techniques. Applying these methods to the first order statistics, a noisy total variability space training is aimed, which will reduce the mismatch resulted by additive noises. All other parts of the conventional i-vector scheme remain unchanged, such as total variability matrix training, reducing the i-vector dimensionality, scoring the i-vectors. The proposed method was tested with four different noise types with several signal to noise ratios (SNR) from -6 dB to 18 dB with 6 dB steps. High reductions in equal error rates were achieved with both methods, even at the lowest SNR levels. On average, the proposed approach produced more than 50% relative reduction in equal error rate.

# Model Kompanzasyonlu Birinci Derece İstatistikleri ile i-vektörlerin Gürbüzlüğünün Artırılması

## Öz

***Anahtar kelimeler***

Paralel model kompanzasyonu; Gürbüz konuşmacı tanıma;Vektör Taylor serileri; I-vektör

Konuşmacı tanıma sistemleri özellikle i-vektörlerin performansı sebebiyle son on yılda önemli gelişmeler elde etmiştir. Bu gelişmelere rağmen eğitim ve test verileri arasındaki uyumsuzluk tanıma performansını önemli ölçüde etkilemektedir. Bu çalışmada, model kompanzasyon yöntemleri i-vektör çıkarımı şemasına eklenerek toplanabilir gürültülere karşı gürbüzlüğü artıracak bir çözüm sunulmaktadır. Durağan gürültüler için model kompanzasyon teknikleri oldukça gürbüz sistemler üretir. Paralel Model Kompanzasyonu ve Vektör Taylor Serileri en gelişmiş model kompanzasyon tekniklerinden kabul edilmektedir. Bu metotlar birinci dereceden istatistiklere uygulanarak toplanabilir gürültülerden kaynaklanan uyumsuzluğu azaltacak gürültülü tüm değişkenlik uzayı eğitimi amaçlanmıştır. Tüm değişkenlik matrisin eğitimi, i-vektör boyutunun azaltılması, i-vektörlerin puanlanması gibi geleneksel i-vektör şemasının diğer tüm parçaları değişmeden kalmaktadır. Önerilen yöntem, 6 dB'lik adımlarla -6 dB'den 18 dB'ye kadar çeşitli sinyal-gürültü oranlarına (SNR) sahip dört farklı gürültü tipi ile test edilmiştir. Her iki yöntemle de en düşük SNR seviyelerinde bile eşit hata oranlarında yüksek azalmalar elde edilmiştir. Önerilen yaklaşım eşik hata oranında ortalama olarak %50'den fazla göreceli azalma sağlamıştır.

## 1. Introduction

Performance of the text-independent speaker recognition systems have increased considerably with the introduction of i-vectors (Dehak *et al.* 2011). Based on the joint factor analysis (Kenny *et al.* 2007), i-vectors produce a fixed low dimensional vector from variable length utterances. A matrix named total variability space (denoted with T) is trained to extract i-vectors, where the speaker, channel, and session variabilities are modelled. A universal background model (UBM) (Reynolds *et al.* 2000) is also used in the conventional i-vector framework. The low dimensionality of the i-vectors gave the opportunity to develop and use more complex channel compensation techniques (Dehak *et al.*, 2011), and considered as state-of-the-art method for text-independent speaker recognition.

As the majority of speech related systems, the i-vectors' performance degrades with the mismatch between the training and test utterances, caused by environmental noise, channel distortions, etc. (Ming 2007). Although channel variabilities can be compensated effectively within the i-vector space (Dehak *et al.*, 2011), the effects of additive noise can still be harmful for the recognition performance. Various studies can be found in the literature that aim to reduce the noise effects in the feature level, total variability space, i-vector space or even at the late classifying step (i.e. probabilistic linear discriminant analysis (PLDA)). Some of these works are given in the following, and the difference of this work is noted at the end of this section.

The feature extraction is the first step in speaker recognition systems (Tirumala *et al.* 2017). Increasing the robustness of the extracted features will make the classifiers' job easier since the deteriorative effects are minimized, and various studies focused on this step (Dişken *et al.* 2017, Krobba *et al.* 2019). On the other hand, the conventional Mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein 1980) are still preferred in many studies (even in the most recent works reported below), and mismatch effects are dealt in later steps, as proposed in this study.

Many researchers tried to achieve robustness within the i-vector framework, or in the low dimensional i-vector space. In (Ribas and Vincent 2019), uncertainty propagation was employed in both UBM and factor analysis model, and a slight improvement over a speech enhancement algorithm was reported. Clean i-vectors were MAP estimated (called i-MAP) given the noisy i-vectors in (Ben Kheder *et al.* 2015, Ben Kheder *et al.* 2014), assuming the distributions are normal, and noise is additive in the i-vector space. This technique was further improved by applying linear regression based cleaning, where the Gaussian assumption was not required (Baby *et al.* 2017). Computational time of the i-MAP was reduced in (Ben Kheder *et al.* 2017), by including a distribution selection scheme, where a previously observed distribution was selected based on the distance between noisy test i-vector and all noisy i-vectors distributions available in the training data. Jointly modeling the clean and noisy i-vectors, a better performance was achieved (Kheder *et al.* 2018).

(El Ayadi *et al.* 2017) estimated GMM/UBM parameters in a robust manner using robust estimation methods named minimum volume ellipsoid and minimum covariance determinant. Since the UBM takes part in the traditional i-vector scheme, resulted i-vectors were considered as robust against additive noise. Simplified supervised i-vectors were found to be superior than the conventional ones in terms of speed and accuracy (Li and Narayanan 2014), where a look-up table and factor analysis performed on pre-normalized Gaussian Mixture Model (GMM) first order statistics helped to reduce the system's complexity. Frame weighting was taken into account in (Zhang *et al.* 2019), and GMM updating rules were defined which lead to more robust sufficient statistics.

Multicondition training, where clean and noisy versions of data are combined in training, was found to be effective to increase robustness in several studies (Garcia-Romero *et al.* 2012, Lei *et al.* 2012, Li and Mak 2015, Mak 2014, Rajan *et al.* 2013), where robust PLDA classifier was the main concern. Multiple SNR-dependent PLDA models were investigated in (Mak *et al.* 2016). Five back-ends were investigated in (Liu and Hansen 2014), and fusion of them was found to be very effective with a computational burden trade-off. Adaptive boosting was used to combine multiple Support Vector Machine (SVM) classifiers which are trained using noisy i-vectors (Sarkar and Sreenivasa 2014).

Deep neural networks (DNN) have gained popularity in the last decade, thanks to the developments in both software and hardware. They have been successfully applied in many diverse areas. Besides being a recognition system themselves (Snyder *et al.* 2016, Variani *et al.* 2014), DNNs were also included in the i-vector framework at various levels to increase their performances. In (Zhang *et al.*, 2020) DNNs were used for multi-level enhancement; in utterance level, MFCC level, and i-vector level, and frame selection also applied to emphasize noise-invariant frames. Recently, DNN speech enhancement also found to be complementary with PLDA mutlicondition training (Novotný *et al.* 2019). LDA was replaced with a DNN to learn non-linear projection of i-vectors (Wang *et al.* 2018). The sufficient statistics were observed with a DNN (Lei *et al.* 2014), and with a convolutional neural network (CNN) (McLaren *et al.* 2014) instead of the traditional UBM. DNNs were also included in the mixture of PLDA framework to produce posterior probabilities (Li *et al.* 2016, Li *et al.* 2017). DNN based mapping on i-vectors were used to reduce content mismatch between utterances with different lengths (Guo *et al.*, 2018). Two neural networks with noisy versions of the clean i-vectors as inputs were trained to produce denoised i-vectors before applying PLDA classifier (Mahto *et al.* 2017).

Model adaptation methods such as parallel model combination (PMC) (Gales and Young 1993, Gales and Young 1996) and Vector Taylor Series (VTS) (Moreno *et al.* 1996) aim to reduce the mismatch between the training and test data by modifying the speech/speaker models' parameters in an efficient manner. Traditional speech recognition systems use Hidden Markov Model (HMM) with Gaussian Mixture model in each state to model the distributions. Model compensation methods modify the model parameters so that the mismatch due to additive noise and/or channel variations is minimized. PMC estimates the noisy models by combining the clean speech and noise models. On the other hand, VTS estimates the noise parameters with (usually) a first-order Taylor Series approximation, then adapts the clean speech model to the noise conditions. One of the advantages of these methods is the requirement of limited adaptation data (Kalinli *et al.* 2010). Due to their state-of-the-art performance, many recognition systems have included these methods to increase robustness (Acero *et al.* 2000, Chung 2016, Gales 1997, Gales and Young 1995, Li *et al.* 2007, Kalinli *et al.* 2009, Kim and Hansen 2009). Modifications on delta parameter estimations for PMC were investigated in (Geng-Xin *et al.* 2006, Sim and Luong 2011). Approximated PMC was proposed in (Sim 2013) to reduce the computational burden of compensating covariance matrices. Mobile (Tao *et al.* 2008) and embedded systems have also benefited from model compensation. VTS preferred in HMM based speech enhancement to modify model parameters for noisy speech (Gao *et al.* 2014). Masking factor was included in VTS before compensation, and a slight improvement compared to the traditional VTS was achieved (Das and Panda 2016). A GMM with a low number of mixtures was used to estimate the noise parameters and another GMM with more mixtures than the first one was used to estimate clean features to reduce the computation load of VTS (Zhou *et al.* 2016). Several PMC approximations were compared in (Gong 2002).

Model compensation was also used in speaker recognition systems (Bellot *et al.* 2000, Ping *et al.* 2001). However, combining the power of the model-based methods with i-vectors have not been investigated, except (Lei *et al.* 2013; Lei *et al.* 2014). In (Lei *et al.* 2013), VTS was used to obtain clean versions of i-vectors. A noisy UBM was constructed for each speech segment with VTS applied to the clean UBM and noise distributions. Noisy models are updated to each utterance with an EM auxiliary function. Expectation maximization (EM) algorithm was also developed to train the total variability matrix. To reduce the computational load of total variability matrix training, a simplified version of this approach was studied, and a minor degradation compared to the original VTS was observed (Lei *et al*. 2014). VTS was replaced with an unscented transform to more accurately estimate the noise-adapted UBM parameters (Martinez *et al.* 2014). The aforementioned methods also include multicondition training.

To the best of authors' knowledge, PMC and i-vector combination have not been tested previously. One of the reasons may be the complexity issues considering that the noise will be injected in UBM, total variability, and scoring models (Ben Kheder *et al.* 2017). On the other hand, previous studies have shown that late steps such as LDA dimensionality reduction and PLDA scoring can remain as in the conventional case with VTS approach (Lei *et al.* 2014). Also, for stationary noise types, noise can be approximated with a single Gaussian. Therefore, the number of mixtures in the UBM will not increase contrary to the non-stationary noise case (Gales and Young 1993). Furthermore, almost every method increases the complexity of the system more or less. For instance, DNN based approaches usually require a high amount of data and a lot of training time. Multicondition training requires noisy data which is not usually available, and producing noisy data inherently increases the training time. Besides, the noise information may not be available a priori. Model based techniques provides the advantage of adapting with a little noise data that can be observed within the test/operating environment. The noise parameters can be estimated by using various methods such as noise tracking, voice activity detecting, speech enhancement, etc. (Chuwatthananurux and Wanvarie 2016, Dişken *et al.* 2017, Ghosh *et al.* 2011, Lin *et al.* 2007, Martin 2001), or even some of the first frames of the incoming utterance can accepted as noise-only frames, which is not always true but still provides a practical solution. The robustness of the recognition system then will be related to the success of the noise estimation methods. In this paper, however, the noise is assumed to be known since the main focus is on the combination of i-vectors and model based compensation. Hence, the performance of the recognition system is going to depend solely on the model based method, and the extracted "noisy" i-vectors. Also, considering the modern devices developed since the first presentation of model based methods, a faster runtime may be anticipated.

The proposed method aims to modify the first order statistics with model compensation where there is no requirement for noisy training data or any multicondition training. Since the T matrix is learned from the sufficient statistics, a noisy version of this matrix will be learned due to the injection of the noise to the first order statistics. Also, the model compensation is applied to the UBM since a noisy UBM is needed to train T and to extract sufficient statistics from noisy test data. All other steps of the conventional i-vector extraction scheme remain as is, and no modifications were made in the EM algorithms in any step. Hence, the model based methods fit almost seamlessly with the i-vectors. Both PMC and VTS methods showed very high EER reductions in the experiments realized with different noisy types and various SNR levels.

The rest of the paper is organized as follow. Section 2 reviews the PMC and VTS methods, providing the essential expressions that will be used in the compensation step. Section 3 shows the proposed method to inject noise information into the first order statistics. Section 4 presents the experimental results, along with a discussion part. Section 5 concludes the paper.

## 2. Model Based Compensation

In this section, VTS and PMC methods are reviewed. Various improvements were made after their initial presentations. Therefore, without analyzing the methods detailly or proving the expressions/assumptions, the equations used in this paper were given for completeness.

### 2.1. Vector Taylor Series

VTS is used to characterize the unknown additive noise and channel effects in a computationally efficient manner. The VTS can be applied to the feature vectors, or to the statistics that model them (Moreno *et al.* 1996). In this paper, the latter approach is chosen. As the order of the Taylor series increase, the complexity of the system increases. A first order series usually performs sufficiently. The noisy speech cepstral vector can be expressed as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{h} + g(\boldsymbol{n} - \boldsymbol{x} - \boldsymbol{h}) \tag{1}$$

where x, h, n corresponds to the clean speech, channel, and additive noise cepstral vectors, respectively, with Gaussian distribution assumption (refer to (Acero *et al.* 2000) for the derivation and assumptions). The g(z) is a non-linear function given below,

$$g(\boldsymbol{z}) = \boldsymbol{C}\ln\left(1 + \exp\left(C^{-1}\boldsymbol{z}\right)\right) \tag{2}$$

where C is the discrete cosine transform (DCT) matrix. Since the convolutive channel noise is not considered in this work, it is dropped from the following expressions. Further, channel noise can be compensated in the lower dimensional i-vector space. The additive noise is assumed to be Gaussian, and the noisy speech vector y, and its mean vector (corresponding to a mixture of noisy UBM), $\boldsymbol{\mu}_y$, can be obtained from

$$\boldsymbol{y} \approx \boldsymbol{\mu}_x + g(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x) + \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{\mu}_x) + (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{n} - \boldsymbol{\mu}_n) \tag{3}$$

$$\boldsymbol{\mu}_y \approx \boldsymbol{\mu}_x + g(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x) \tag{4}$$

where $\boldsymbol{\mu}_x$ is the mean vector of clean speech model (a mixture of the clean UBM), $\boldsymbol{\mu}_n$ is the mean vector of additive noise model (single Gaussian in this paper), I is the identity matrix, and A is the Jacobian of Equation 1 with respect to x and can be expressed as

$$\boldsymbol{A} = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^{-1} \tag{5}$$

P is a diagonal matrix whose elements are ($\boldsymbol{\mu} = \boldsymbol{\mu}_n - \boldsymbol{\mu}_x$)

$$\boldsymbol{p}(\boldsymbol{\mu}) = \frac{1}{1+e^{C^{-1}\mu}} \tag{6}$$

The covariance matrix of the noisy UBM can be calculated as

$$\boldsymbol{\Sigma}_y \approx \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^{\boldsymbol{T}} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\Sigma}_{\boldsymbol{n}}(\boldsymbol{I} - \boldsymbol{A})^{\boldsymbol{T}} \tag{7}$$

where $\boldsymbol{\Sigma}_x$ is the covariance matrix of the clean speech (a mixture of the clean UBM), $\boldsymbol{\Sigma}_n$ is the covariance matrix of the additive noise. The noisy covariance matrix is assumed diagonal, although the result of the Equation 7 is not diagonal. The delta parameters can also be estimated by using Equations 8-9.

$$\Delta\boldsymbol{\mu}_y \approx \boldsymbol{A}\Delta\boldsymbol{\mu}_x \tag{8}$$

$$\Delta\boldsymbol{\Sigma}_y \approx \boldsymbol{A}\Delta\boldsymbol{\Sigma}_x\boldsymbol{A}^{\boldsymbol{T}} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\Delta\Sigma}_{\boldsymbol{n}}(\boldsymbol{I} - \boldsymbol{A})^{\boldsymbol{T}} \tag{9}$$

### *2.2. Parallel Model combination*

The basic idea behind PMC is to obtain modified models of the acoustic environment (such as HMM, GMM), so that the mismatch between training and test data are minimized (Gales and Young 1993, Gales and Young 1996). The modification is simply done by combining a clean speech model with a noise model. The combination of the parameters is performed in the linear spectral domain. Therefore, parameters of each model must be mapped from cepstral domain. Once the models are combined, the observed noisy model parameters are mapped back to the cepstral domain. One of the advantages of the PMC is that no change is required in the further process such as scoring. Some assumptions made for the PMC are as follows (Gales and Young 1996);

- The speech and noise are independent.
- They are additive in the time and power spectrum domains.
- A single Gaussian or a GMM well presents the distribution of the observation vectors in the cepstrum or log filter-bank energy domain.
- The frame alignment used to generate the speech models from clean data is not changed by the addition of noise.

Additional assumptions to use log normal approximation (Gales and Young 1993, Tufekci *et al.* 2006) are given below.

- The sum of two log normal distributed random variables results in a log normal distributed random variable.
- The variances of $(\frac{S_i}{S_i+N_i})$ and $(\frac{N_i}{S_i+N_i})$ are negligible. Si and Ni are the ith components of the clean speech observation vector and noise observation vector, respectively, in the mel-scaled filter-bank energy domain.
- $\mathrm{E}(\frac{S_i}{S_i+N_i}) \approx \frac{\mu_i}{\mu_i+\tilde{\mu}_i}) = \gamma_i$, $\mathrm{E}(\frac{N_i}{S_i+N_i}) \approx \frac{\tilde{\mu}_i}{\mu_i+\tilde{\mu}_i}) = \eta_i$, where E is expectation operator, $\mu_i$ and $\tilde{\mu}_i$ are the ith components of the clean speech and noise mean vectors in the mel-scaled filter-bank energy domain.

For the rest of the equations, superscripts are used to denote the domain, i.e., c indicates the cepstral domain, l indicates log domain. Absence of a superscript indicates linear domain. The symbols ~ and ^ are used to depict noise and estimated noisy speech parameters.

The model parameters (μ: mean vector, Σ: covariance matrix) are mapped to the log energy domain as follows:

$$\boldsymbol{\mu}^l = \boldsymbol{C}^{-1}\boldsymbol{\mu}^c \tag{10}$$

$$\Delta\boldsymbol{\mu}^l = \boldsymbol{C}^{-1}\Delta\boldsymbol{\mu}^c \tag{11}$$

$$\boldsymbol{\Sigma}^l = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}^c(\boldsymbol{C}^{-1})^{\mathrm{T}} \tag{12}$$

$$\Delta\boldsymbol{\Sigma}^l = \boldsymbol{C}^{-1}\Delta\boldsymbol{\Sigma}^c(\boldsymbol{C}^{-1})^{\mathrm{T}} \tag{13}$$

Then, exponential function is applied to transform into linear domain:

$$\mu_i = exp(\mu_i^l + \frac{\Sigma_{ii}^l}{2}) \tag{14}$$

$$\Delta\mu_i = exp(\Delta\mu_i^l + \frac{\Delta\Sigma_{ii}^l}{2}) \tag{15}$$

$$\Sigma_{ij} = \mu_i\mu_j[exp(\Sigma_{ii}^l) - 1] \tag{16}$$

$$\Delta\Sigma_{ij} = \Delta\mu_i\Delta\mu_j[exp(\Delta\Sigma_{ii}^l) - 1] \tag{17}$$

Then, the noisy model parameters are estimated by using Equation 18 and Equation 19,

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu} + g\widetilde{\boldsymbol{\mu}} \tag{18}$$

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + g^2\widetilde{\boldsymbol{\Sigma}} \tag{19}$$

where g is a gain matching term calculated with averages of noisy speech signal energy ($E_{ns}$), noise energy ($E_n$), and clean speech energy ($E_s$) as

$$g = \frac{E_{ns} - E_n}{E_s} \tag{20}$$

Similarly, delta parameters of the noisy model are estimated by using the following equations.

$$\Delta\hat{\mu}_i = \gamma_i\Delta\mu_i + \mathrm{g}\eta_i\Delta\tilde{\mu}_i \tag{21}$$

$$\Delta\hat{\Sigma}_{ij} = \gamma_i\gamma_j\Delta\Sigma_{ij} + \mathrm{g}^2\eta_i\eta_j\Delta\tilde{\Sigma}_{ij} \tag{22}$$

Once the noisy model is constructed, its parameters must be transformed back to the cepstral domain by first taking the logarithm (Equations 23-26), then applying the DCT (Equations (27-30).

$$\hat{\mu}_i^l = ln(\hat{\mu}_i) - \frac{1}{2}ln(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1) \tag{23}$$

$$\Delta\hat{\mu}_i^l = ln(\Delta\hat{\mu}_i) - \frac{1}{2}ln(\frac{\Delta\hat{\Sigma}_{ii}}{\Delta\hat{\mu}_i^2} + 1) \tag{24}$$

$$\hat{\Sigma}_i^l = ln(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i\hat{\mu}_j} + 1) \tag{25}$$

$$\Delta\hat{\Sigma}_i^l = ln(\frac{\Delta\hat{\Sigma}_{ii}}{\Delta\hat{\mu}_i\Delta\hat{\mu}_j} + 1) \tag{26}$$

$$\widehat{\boldsymbol{\mu}}^c = \boldsymbol{C}\widehat{\boldsymbol{\mu}}^l \tag{27}$$

$$\Delta\widehat{\boldsymbol{\mu}}^c = \boldsymbol{C}\Delta\widehat{\boldsymbol{\mu}}^l \tag{28}$$

$$\widehat{\boldsymbol{\Sigma}}^c = \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}^l\boldsymbol{C}^T \tag{29}$$

$$\Delta\widehat{\boldsymbol{\Sigma}}^c = \boldsymbol{C}\Delta\widehat{\boldsymbol{\Sigma}}^l\boldsymbol{C}^T \tag{30}$$

In the experiments, four preceding and four succeeding frames were considered to obtain delta features as given in Equation 31, where y is the static feature vector, $\Delta\boldsymbol{y}$ is the delta feature vector, k and i are the frame indexes, and N=4.

$$\Delta\boldsymbol{y}^{\boldsymbol{l}}(k) = \frac{\sum_{i=-N}^{N} i\boldsymbol{y}^{\boldsymbol{l}}(k-i)}{\sum_{i=1}^{N} i} \tag{31}$$

However, in the PMC method, the equations for deltas were derived considering that the deltas were calculated using the present frame and its succeeding one. To apply the same formulas, it is assumed that the delta features can be expressed as,

$$\Delta\boldsymbol{y}^{\boldsymbol{l}}(k) \cong \Delta\boldsymbol{y}^{\boldsymbol{l}}(k-\tau) - \Delta\boldsymbol{y}^{\boldsymbol{l}}(k+\tau) \tag{32}$$

Considering that the noisy speech is the sum of the speech (***x***) and noise (***n***) signals in the linear domain, Equation 32 can be transformed into Equation 33.

$$\Delta\boldsymbol{y}^{\boldsymbol{l}}(k) \cong \log(\frac{e^{\boldsymbol{x}(k-\tau)} + e^{\boldsymbol{n}(k-\tau)}}{e^{\boldsymbol{x}(k+\tau)} + e^{\boldsymbol{n}(k+\tau)}})$$

$$= \log\,(e^{\boldsymbol{x}(k-\tau)-\boldsymbol{x}(k+\tau)}\,\frac{e^{\boldsymbol{x}(k+\tau)}}{e^{\boldsymbol{x}(k+\tau)} + e^{\boldsymbol{n}(k+\tau)}} +$$

$$e^{\boldsymbol{n}(k-\tau)-\boldsymbol{n}(k+\tau)}\,\frac{e^{\boldsymbol{n}(k+\tau)}}{e^{\boldsymbol{x}(k+\tau)} + e^{\boldsymbol{n}(k+\tau)}}) \tag{33}$$

The expressions within the log operation follows Equation 21, where the ratios correspond to $\gamma$ and $\eta$, and the exponentials correspond to the speech signal and noise signal, respectively. Therefore, we do not need to modify the original PMC equations for the deltas.

## 3. Model Compensated First Order Statistics

In this section, the conventional i-vector extraction scheme is reviewed first, then the combination of the model based methods with i-vectors is explained. The main idea is observing noisy first order statistics. Since the total variability matrix (T) is learned from sufficient statistics, a noisy version of T will be estimated. Further, as T is also called as i-vector extractor, noisy i-vectors will be observed at the final stage. The model compensation methods are applied to the UBM and first order statistics. All of the other steps and training conditions remain the same (i.e. training of T, applying channel normalization and/or dimensionality reduction methods, scoring the i-vectors). Hence, there is no requirement to develop new EM algorithms, apply multicondition training, or modify the scoring process.

### *3.1. Extraction of i-vectors*

The conventional i-vector scheme (Dehak *et al.* 2011) is reviewed for convenience. A speaker and channel dependent GMM supervector can be defined as

$$M = m + T\omega \tag{34}$$

where m is the mean supervector taken from the UBM, T is the i-vector extractor, and $\omega$ is a random vector with a normal distribution. For each utterance, an i-vector is obtained by the maximum a posterior (MAP) estimate of $\omega$. Sufficient statistics (Baum-Welch), which are used in the training of T and in the extraction of i-vectors, are collected using the UBM (Ω) as follows:

$$N_c = \sum_{t=1}^{L} P(c|f_t, \Omega) \tag{35}$$

$$F_c = \sum_{t=1}^{L} P(c|f_t, \Omega) f_t \tag{36}$$

*N* and *F* called as the zero and first order statistics, respectively, calculated for a sequence of *L* frames. $P(c|f_t, \Omega)$ is the posterior probability of mixture component $c = 1, \dots, C$ generating the observation vector $f_t$. Centralized first order statistics can be calculated by substituting the UBM mean supervector.

$$\overline{F_c} = \sum_{t=1}^{L} P(c|f_t, \Omega)(f_t - m_c) \tag{37}$$

Given an utterance, the posterior estimation of i-vector is obtained by using Equation 38.

$$\omega = (I + T^t \Sigma^{-1} N(u) T)^{-1} T^t \Sigma^{-1} \bar{F}(u) \tag{38}$$

$N(u)$ is a diagonal matrix whose diagonal block are $N_c I$, $\bar{F}(u)$ is a supervector constructed by concatenating all $\overline{F_c}$ for a given utterance, and $I$ is the identity matrix. Σ is the covariance matrix that can be copied from the UBM (Kenny 2012). Assuming the first and second order moments $\langle\omega(s)\rangle$ and $\langle\omega(s)\omega(s)^T\rangle$ have been calculated, T can be updated using the formula below (Kenny, 2012).

$$T_c = (\textstyle\sum_s \bar{F}_c(s)\langle\omega^T(s)\rangle)(\sum_s N_c(s)\langle\omega(s)\rangle\langle w^T(s)\rangle)^{-1} \tag{39}$$

As seen in Equation 39, T matrix depends on the sufficient statistics. It is assumed that if the statistics are noisy, the resulted T will be noisy, Hence, robustness of the system will be increased due to the fact that noisy i-vectors can be extracted from the clean training data, and the mismatch between noisy test data will be reduced.

### *3.2. Extraction of noisy i-vectors with model compensation*

The proposed method modifies the UBM parameters and the first order statistics. The UBM has a critical role in the i-vector scheme. It is used to estimate sufficient statistics, centering the first order statistics, and the covariance matrices used in the MAP estimation of the i-vectors. Considering these facts, it is clear that a noisy UBM is needed to before further processes. The noise is assumed to be stationary and modeled as a single Gaussian. Hence, using the PMC and VTS methods described in the previous section, a noisy UBM is obtained with the number of mixtures equal to the clean UBM. It should be noted that both methods applied independently at the exact same stages. Hence, the noisy model, or compensated model means that the model parameters are modified either PMC or VTS.

The next step of the proposed method is to extract the sufficient statistics, as in the conventional i-vector framework. In this case however, using the noisy UBM will produce erroneous results since the training data is clean. Also, as mentioned in Section 2.2, The frame alignment used to generate the speech models from clean data is not changed by the addition of noise. Therefore, there is no need to modify the zero order statistics. The first order statistics, on the other hand, are multiplied by the observation vectors which will be noisy in the test data. Model compensation is applied to the first order statistics so that noisy i-vectors can be extracted from the clean training data. The zero and first order statistics can be thought as weights and mean vectors of the UBM, respectively. In fact, dividing the first order statistics to the zero order statistics yields to updated mean values as in the M-step of the GMM/UBM training. Hence, after reshaping and dividing, the first order statistics are available for model compensation. The covariance matrix of the clean UBM is used in conjunction with the first order statistics to apply PMC method. For the VTS, the first order statistics can be directly modified without considering the covariance matrices. The modified equations are given below for convenience.

Let $\breve{\boldsymbol{F}}$ denote the reshaped and divided first order statistics (mean vector of a mixture). For the VTS method,$\boldsymbol{\mu}_x = \breve{\boldsymbol{F}}$, and $\boldsymbol{\mu}_{\Delta x} = \Delta\breve{\boldsymbol{F}}$ so the noisy first order statistics $(\widehat{\boldsymbol{F}})$ and its deltas $(\Delta\widehat{\boldsymbol{F}})$ can be estimated as in Equation 4 and Equation 8,

$$\widehat{\boldsymbol{F}} \approx \breve{\boldsymbol{F}} + g\big(\boldsymbol{\mu}_n - \breve{\boldsymbol{F}}\big) \qquad (40)$$

$$\Delta\widehat{\boldsymbol{F}} \approx \boldsymbol{A}\Delta\breve{\boldsymbol{F}} \qquad (41)$$

For the PMC method, $\boldsymbol{\mu}^c = \breve{\boldsymbol{F}}$ and$\Delta\boldsymbol{\mu}^c = \Delta\breve{\boldsymbol{F}}$, and the related equations in Section 2.2 should be handled accordingly, with the covariance matrix taken from the clean UBM. Equation 27 and Equation 28 give the $\widehat{\boldsymbol{F}}$ and its deltas $\Delta\widehat{\boldsymbol{F}}$, respectively.

Note that model compensation is applied before calculating the centralized statistics. In order to follow the i-vector framework, the noisy first order statistics must be multiplied with the zero order statistics, then concatenated (to reverse the reshape and divide operations). Equation 37 can be rewritten as

$$\overline{F_c} = F_c - N_c m_c \qquad (42)$$

Using the noisy first order statistics and the mean vectors from the noisy UBM $(\widehat{\boldsymbol{m}})$ , the centered statistics can be expressed as

$$\widehat{\boldsymbol{F}} \approx \breve{\boldsymbol{F}} + g\big(\boldsymbol{\mu}_n - \breve{\boldsymbol{F}}\big) \qquad (43)$$

The remaining process is completely the same as the standard i-vector process. T matrix is trained using the noisy UBM, clean zero order statistics, and noisy first order statistics. The observed matrix can be considered as noisy total variability space. Finally, the noisy i-vectors can be extracted using the noisy T matrix.

In the test stage, contrary to the training, sufficient statistics are extracted using the noisy UBM. The reason for this approach is that test data contains noise, and we do not have any frame alignment information as in the UBM training with clean data. Using the noisy UBM, noisy statistics will be observed inherently so there is no need to add any compensation in this stage. Experimental results given in the next section confirmed that this approach effectively reduce the mismatch due to the additive noises, even in severely degraded situations such as -6 dB and 0 dB SNR levels.

To summarize the proposed approach, a block diagram is given in Figure 1. The leftmost blocks show processes for the conventional i-vector system. The dashed blocks are the proposed modifications. Note that clean sufficient statistics are extracted using the clean UBM. In the proposed method, noisy statistics are obtained directly with the model compensation in the training stage. As mentioned in the previous paragraph, in the test stage, the noisy UBM is used to obtain the sufficient statistics.

## 4. Speaker Verification Experiments

### *4.1. Experimental setup*

250 male speakers from the NIST SRE 1998 database were used in the verification experiments. For each speaker, approximately 5 minutes of training data were available. The durations of test data were 30

seconds. 1308 test utterances were used to measure the performance of the proposed method. A simple energy-based VAD was used to remove the silence parts found in the utterances (Kinnunen and Li 2010). Four different noises (F16, factory, Lynx, speech) from the NOISEX-92 noise database (Varga and Steeneken 1993) were added to the test files at SNR levels varying from -6 dB to 18 dB with 6 dB steps. All utterances (training and test files) and noise signals were normalized to have equal energy in each utterance. Then, for the test files, noise was added with a suitable multiplier to have the desired SNR level. As mentioned in the previous section, noise was modeled with a single Gaussian. Although F16 and factory noises were not as stationary as Lynx and speech noises, this approach still estimates the noise sufficiently, as reported in the test results. However, more Gaussian components may be needed for more complex noise types, or more accurate results, at the expense of computational time.

26-dimensional MFCCs were extracted (13 static features including the zeroth coefficient, and their deltas). A UBM with 512 mixtures was trained with all available training data. 400 dimensional i-vectors were extracted. LDA was applied to reduce the i-vector dimensions to 200. PLDA was utilized for scoring stage. The system trained on the clean data with the given parameters was served as baseline. For the proposed noisy i-vectors, the model based methods were applied to obtain noisy first order statistics and noisy UBM.

As the success of the model compensation methods for GMM/HMM based systems are known from the related literature, results with the traditional GMM-UBM method are also given in order to prove that model compensation with the i-vectors can achieve better recognition performances. In the GMM-UBM method, speaker models were adapted from the clean UBM, then model compensation applied to both UBM and adapted speaker models. Since the training data is clean, this approach produced better results than adapting the speaker models from the noisy UBM. Table 1 shows the results for the GMM-UBM method, and Table 2 shows the results for the i-vectors.

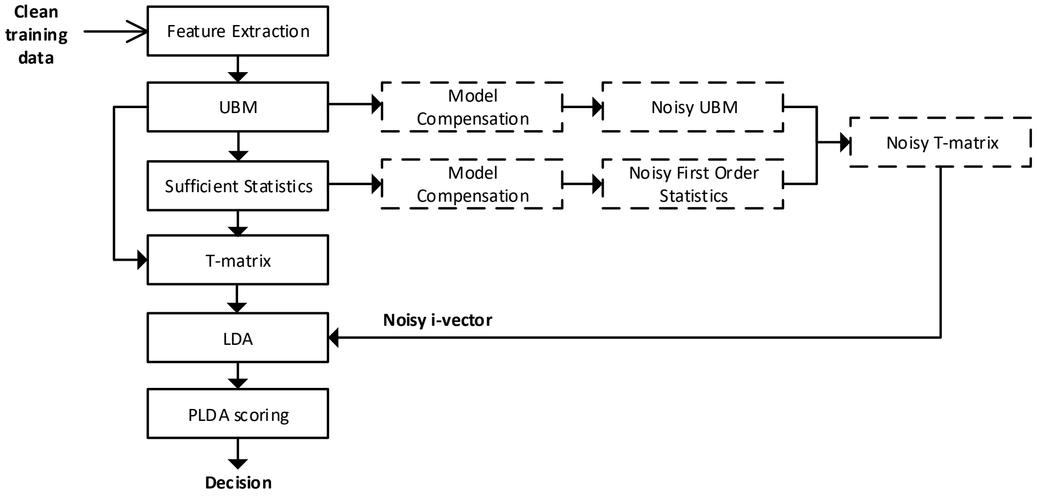

**Figure 1.** The conventional i-vector system (leftmost blocks) and the proposed modifications (dashed blocks). At the test stage, the noisy test data follows the same path for the conventional system. For the proposed system, noisy UBM and noisy T-matrix are used instead of their clean versions.

**Table 1.** Speaker verification results in terms of percent EER for the GMM-UBM method.

| Noise Type | SNR Level (dB) | Baseline GMM-UBM | PMC GMM-UBM | VTS GMM-UBM | Average relative EER reduction with PMC | Average relative EER reduction with VTS |
|---|---|---|---|---|---|---|
| Lynx | -6 | 42.6991 | 18.8879 | 11.4853 | ~53% | ~57% |
| | 0 | 32.2446 | 7.881 | 7.6711 | | |
| | 6 | 20.742 | 5.3957 | 5.7279 | | |
| | 12 | 9.2896 | 4.4805 | 4.8593 | | |
| | 18 | 4.8387 | 4.2667 | 4.0667 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **F16** | -6 | 46.3478 | 26.4758 | 16.446 | ~65% | ~69% |
| | 0 | 41.5461 | 11.1111 | 9.6579 | | |
| | 6 | 33.6294 | 6.4645 | 6.7848 | | |
| | 12 | 20.9424 | 4.8346 | 5.2057 | | |
| | 18 | 9.0211 | 4.2636 | 4.5739 | | |
| **Factory** | -6 | 45.2763 | 22.8278 | 18.8776 | ~55 | ~60 |
| | 0 | 36.4122 | 10.163 | 9.3799 | | |
| | 6 | 23.913 | 6.4626 | 6.4181 | | |
| | 12 | 12.3656 | 4.7804 | 5.1453 | | |
| | 18 | 5.5556 | 4.3702 | 4.3928 | | |
| **Babble** | -6 | 38.8889 | 11.3965 | 11.0018 | ~44 | ~44 |
| | 0 | 25.8913 | 7.1288 | 7.4157 | | |
| | 6 | 12.7907 | 5.6355 | 5.5353 | | |
| | 12 | 5.8548 | 4.6914 | 4.563 | | |
| | 18 | 4.1995 | 4.0897 | 4.1775 | | |

The average relative EER reduction rates are also given in the last columns for each noise type. The VTS performed slightly better than the PMC. For the GMM-UBM method, reduction rates vary between 44% and 69%. The proposed method produced similar results with the i-vectors, 42% and 65%, the lowest and the highest reduction rates, respectively. This proves that compensating the first orders statistics along with the UBM, model compensation methods have fitted in the i-vector scheme seamlessly. Another important point is that the baseline i-vectors produced better EER values than the baseline GMM-UBM, which was expected. Therefore, despite the similar range, EER reductions within the i-vector framework are much more valuable.

**Table 2.** Speaker verification results in terms of percent EER for the i-vector method.

| Noise Type | SNR Level (dB) | Baseline i-vector | PMC i-vector | VTS i-vector | Average relative EER reduction with PMC | Average relative EER reduction with VTS |
|---|---|---|---|---|---|---|
| **Lynx** | -6 | 39.0395 | 14.5712 | 12.8728 | ~60 | ~65 |
| | 0 | 30.8655 | 6.8681 | 6.3043 | | |
| | 6 | 19.021 | 4.4348 | 4 | | |
| | 12 | 7.3783 | 3.5477 | 2.8278 | | |
| | 18 | 4.4346 | 3.0214 | 2.6992 | | |
| **F16** | -6 | 41.0819 | 23.1481 | 19.913 | ~58 | ~62 |
| | 0 | 33.6728 | 10.9322 | 9.0617 | | |
| | 6 | 23.8502 | 5.7018 | 5.0998 | | |
| | 12 | 11.7347 | 3.8069 | 3.6465 | | |
| | 18 | 5.5987 | 3.4404 | 3.3814 | | |
| **Factory** | -6 | 39.1631 | 19.087 | 15.1623 | ~53 | ~57 |
| | 0 | 29.6333 | 8.8207 | 8.5456 | | |
| | 6 | 19.5322 | 5.4581 | 4.5918 | | |
| | 12 | 8.2418 | 3.7037 | 3.3419 | | |
| | 18 | 4.0466 | 3.2651 | 3.2776 | | |
| **Babble** | -6 | 35.6015 | 14.961 | 15.5222 | ~42 | ~45 |
| | 0 | 22.7521 | 6.3662 | 6.402 | | |
| | 6 | 10.9442 | 4.2813 | 3.7275 | | |
| | 12 | 4.6784 | 3.6355 | 3.3259 | | |
| | 18 | 3.0435 | 3.0702 | 2.9933 | | |

### *4.2. Discussion*

The experimental results indicated that the i-vectors can benefit from the model compensation techniques, without extreme changes in the conventional procedure. Both the VTS and the PMC methods achieved very high relative reduction rates in terms of EER. The VTS performed slightly better than the PMC. This situation was expected since the previous literature showed that the VTS approximates the noise better than the PMC (Acero *et al.*, 2000).

The benefits of the proposed scheme were more observable for the lower SNR values. For instance, both baseline and the proposed systems performed similar when the SNR level was 18 dB. However, as the SNR drops to 6 dB, the baseline method's performance dropped dramatically while the proposed systems' EERs were not even doubled. For the -6 dB SNR level, the best performing baseline system produced 35.6% EER. The worst performing proposed system yielded 26.47% EER. The gap between the best and worst performing systems proves the effectiveness of the noise compensation methods.

No other data besides the clean training data were used in the experiments. This is a more practical approach then the multicondition training since the training data usually collected in clean, controlled environments. It should be noted that the noise was modeled with a single Gaussian. The F16 and factory noises are more volatile than the Lynx and babble noises, however, model compensation still increased the robustness against these noises. More Gaussians should lead to more accurate noise estimates, which will further improve the results. On the other hand, the system's complexity will increase in accordance with the number of Gaussians used to estimate the noise parameters.

In the experiments, we assumed that the noise type in the test data is known to focus solely on the performance of the model compensation within i-vector framework. For practical systems, various methods can be used to estimate the noisy sections on-line, if no prior information is available. A little adaptation data is required for the model compensation methods, and the proposed approach only includes compensation in the UBM and first order statistics, then the total variability matrix can be trained as usual. Hence, the system can be adapted to a new environment with a little noise data, and once the adaptation completed, the scoring is just as fast as the conventional i-vectors. For the noisy environments where the noise is highly non-stationary the proposed approach still can be effective providing enough mixture to model the noise off-line, but in a practical system adapting the system to a highly non-stationary environment will be much more time consuming. In fact, robustness against non-stationary noise is an active research area for speech related studies, and most solutions require complex systems.

## 5. Conclusion

In this paper, state-of-the-art model compensation methods, namely PMC and VTS, were combined with the i-vectors. The main purpose of the proposed method is to extract noisy i-vectors from the clean training speech, hence the mismatch between the clean training data and the noisy test data will be minimized. Contrary to the previous approaches, mutlicondition training was not required, since the compensation was directly applied to the UBM and the first order statistics. Hence, only a little noise data was used for modeling the parameters with a single Gaussian. The proposed approach does not change the training of the total variability matrix, hence the standard EM training was used. The LDA dimensionality reduction and PLDA scoring were also included without any modification, as in a standard i-vector system.

Speaker verification experiments were conducted to show the effectiveness of the proposed method. Four different noise types were considered with SNR level changing from -6 dB to 18 dB with 6 dB steps. Results with the GMM-UBM method were also given to indicate the effectiveness of the model compensation systems since they are mainly combined with GMM/HMM systems in robust speech recognition systems. Results with the i-vector method proved that the proposed method could produce as high EER reductions as the GMM-UBM system. Considering all noise types and all SNR levels, more than 50% relative EER reduction was achieved.

## 5. References

Acero, A., Deng, L., Kristjansson, T., & Zhang, J. 2000. HMM Adaptation Using Vector Taylor Series for Noisy Speech Recognition. In *Sixth International Conference on Spoken Language Processing* (pp. 869–872). Beijing, China.

Baby, R., Kumar, C. S., George, K. K., & Panda, A. 2017. Noise compensation in i-vector space using linear regression for robust speaker verification. In *2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)* (pp. 161–165). Aligarh, India: IEEE. https://doi.org/10.1109/MSPCT.2017.8363996

Bellot, O., Matrouf, D., Merlin, T., & Bonastre, J.-F. 2000. Additive and Convolutional Noises Compensation for Speaker Recognition. In *Sixth International Conference on Spoken Language Processing* (pp. 799–802). Beijing, China.

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

Ben Kheder, W., Matrouf, D., Bousquet, P.-M., Bonastre, J.-F., & Ajili, M. 2014. Robust Speaker Recognition Using MAP Estimation of Additive Noise in i-vectors Space. In *International Conference on Statistical Language and Speech Processing* (pp. 97–107). Grenoble, France.

Ben Kheder, W., Matrouf, D., Bousquet, P.-M., Bonastre, J.-F., & Ajili, M. 2017. Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition. *Computer Speech & Language*, ***45***, 104–122.

Chung, Y. 2016. Vector Taylor series based model adaptation using noisy speech trained hidden Markov models. *Pattern Recognition Letters*, ***75***, 36–40.

Chuwatthananurux, S., & Wanvarie, D. 2016. Improving noise estimation with RAPT pitch voice activity detection under low SNR condition. In *2016 8th International Conference on Knowledge and Smart Technology (KST)* (pp. 77–82). Chiangmai, Thailand.

Das, B., & Panda, A. 2016. Vector taylor series expansion with auditory masking for noise robust speech recognition. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 1–5). Tianjin, China.

Davis, S., & Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ***28*****(4)**, 357–366.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. 2011. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, ***19*****(4)**, 788–798.

Dişken, G., Tüfekci, Z., & Çevik, U. 2017. A robust polynomial regression-based voice activity detector for speaker verification. *EURASIP Journal on Audio, Speech, and Music Processing*, ***2017*****(1)**, 1-23.

Dişken, G., Tüfekçi, Z., Saribulut, L., & Çevik, U. 2017. A Review on Feature Extraction for Speaker Recognition under Degraded Conditions. *IETE Technical Review*, ***34*****(3)**, 321–332.

El Ayadi, M., S.O. Hassan, A.-K., Abdel-Naby, A., & A. Elgendy, O. 2017. Text-independent speaker identification using robust statistics estimation. *Speech Communication*, ***92***, 52–63. https://doi.org/10.1016/j.specom.2017.05.005

Gales, M.J.F. 1997. “NICE” Model-Based Compensation Schemes for Robust Speech Recognition. In *Robust Speech Recognition for Unknown Communication Channels* (pp. 55–64). Pont-a-Mousson, France.

Gales, M.J.F., & Young, S. J. 1993. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, ***12*****(3)**, 231–239.

Gales, M. J. F., & Young, S. J. 1995. A fast and flexible implementation of parallel model combination. In *1995 International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 133–136). Detroit, USA.

Gales, M. J. F., & Young, S. J. 1996. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, ***4*****(5)**, 352–359.

Gao, Z., Bao, C., Bao, F., & Jia, M. 2014. HMM-based speech enhancement using vector Taylor series and parallel modeling in Mel-frequency domain. In *2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (pp. 733–737). Guilin, China.

Garcia-Romero, D., Zhou, X., Espy-Wilson, C. Y. 2012. Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal*

*Processing (ICASSP)* (pp. 4257–4260). Kyoto, Japan.

Geng-Xin N., Shu-Hung L., Kam-Keung C., Gang W. 2006. A parallel model combination scheme with improved delta parameter compensation. In *2006 IEEE International Symposium on Circuits and Systems* (pp. 5535–5538). Island of Kos, Greece: IEEE. https://doi.org/10.1109/ISCAS.2006.1693888

Ghosh, P. K., Tsiartas, A., Narayanan, S. 2011. Robust Voice Activity Detection Using Long-Term Signal Variability. *IEEE Transactions on Audio, Speech, and Language Processing*, ***19*(3)**, 600–613.

Gong, Y. 2002. A COMPARATIVE STUDY OF APPROXIMATIONS FOR PARALLEL MODEL COMBINATION OF STATIC AND DYNAMIC PARAMETERS. In *7th International Conference on Spoken Language Processing* (pp. 1–4). Denver, Colorado, USA.

Guo, J., Xu, N., Qian, K., Shi, Y., Xu, K., Wu, Y., Alwan, A. 2018. Deep neural network based i-vector mapping for speaker verification using short utterances. *Speech Communication*, ***105***, 92–102.

Jinyu, L., Li D., Dong, Y., Yifan, G., Acero, A. 2007. High-performance hmm adaptation with joint compensation of additive and convolutive distortions via Vector Taylor Series. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* (pp. 65–70). Kyoto, Japan.

Kalinli, O., Seltzer, M.L., Droppo, J., Acero, A. 2010. Noise Adaptive Training for Robust Automatic Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, ***18*(8)**, 1889–1901.

Kalinli, O., Seltzer, M. L., Acero, A. 2009. Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3825–3828). Taipei, Taiwan.

Kenny, P. 2012. A Small Footprint i-Vector Extractor. In *Odyssey 2012-The Speaker and Language Recognition Workshop* (pp. 1–6). Singapore.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P. (2007). Speaker and Session Variability in GMM-Based Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing*, **15(4)**, 1448–1460.

Kheder, W. Ben, Matrouf, D., Ajili, M., Bonastre, J.-F. 2018. A Unified Joint Model to Deal With Nuisance Variabilities in the i-Vector Space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ***26*(3)**, 633–645.

Kim, W., Hansen, J.H.L. 2009. Feature compensation in the cepstral domain employing model combination. *Speech Communication*, ***51*(2)**, 83–96.

Kinnunen, T., Li, H. 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, ***52*(1)**, 12–40.

Krobba, A., Debyeche, M., Selouani, S.-A. 2019. Multitaper chirp group delay Hilbert envelope coefficients for robust speaker verification. *Multimedia Tools and Applications*, ***78*(14)**, 19525–19542.

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

Lei, Y., Burget, L., Scheffer, N. 2013. A noise robust i-vector extractor using vector taylor series for speaker recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6788–6791). Vancouver, BC, Canada.

Lei, Y., McLaren, M., Ferrer, L., Scheffer, N. 2014. Simplified VTS-based I-vector extraction in noise-robust speaker recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4037–4041). Florence, Italy.

Lei, Y., Scheffer, N., Ferrer, L., McLaren, M. 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1695–1699). Florence, Italy.

Li, M., Narayanan, S. 2014. Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification. *Computer Speech and Language*, ***28*(4)**, 940–958.

Li, N., Mak, M.W. 2015) SNR-Invariant PLDA Modeling in Nonparametric Subspace for Robust Speaker

Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ***23*****(10)**, 1648–1659. 7

Li, N., Mak, M.W., Chien, J.-T. 2016. Deep neural network driven mixture of PLDA for robust i-vector speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 186–191). San Diego, CA, USA.

Li, N., Mak, M.-W., Chien, J.T. 2017. DNN-Driven Mixture of PLDA for Robust Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ***25*****(6)**, 1371–1383.

Lin, Z., Goubran, R. A., Dansereau, R. M. 2007. Noise estimation using speech/non-speech frame decision and subband spectral tracking. *Speech Communication*, ***49*****(7)**, 542–557.

Lit Ping Wong, Russell, M. 2001. Text-dependent speaker verification under noisy conditions using parallel model combination. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Vol. 1, pp. 457–460). Salt Lake City, UT, USA.

Liu, G., Hansen, J.H.L. 2014. An Investigation into Back-end Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ***22*****(12)**, 1978–1992.

Mahto, S., Yamamoto, H., Koshinaka, T. 2017. i-Vector Transformation Using a Novel Discriminative Denoising Autoencoder for Noise-Robust Speaker Recognition. In *Interspeech 2017* (pp. 3722–3726). Stockholm, Sweden.

Mak, M.W. 2014. SNR-Dependent Mixture of PLDA for Noise Robust Speaker Verification. In *INTERSPEECH 2014* (pp. 1855–1859). Singapore.

Mak, M.W., Pang, X., Chien, J.T. 2016. Mixture of PLDA for Noise Robust I-Vector Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ***24*****(1)**, 130–142.

Martin, R. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, ***9*****(5)**, 504–512.

Martinez, D., Burget, L., Stafylakis, T., Lei, Y., Kenny, P., Lleida, E. 2014. Unscented transform for ivector-based noisy speaker recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4042–4046). Florence, Italy.

McLaren, M., Lei, Y., Scheffer, N., Ferrer, L. 2014. Application of convolutional neural networks to speaker recognition in noisy conditions. In *INTERSPEECH 2014* (pp. 686–690). Singapore.

Ming, J. 2007. Robust Speaker Recognition in Noisy Conditions. *IEEE Transactions on Audio, Speech and Language Processing*, ***15*****(5)**, 637–1723.

Moreno, P. J., Raj, B., Stern, R. M. 1996. A vector Taylor series approach for environment-independent speech recognition. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (Vol. 2, pp. 733–736). Atlanta, GA, USA.

Novotný, O., Plchot, O., Glembek, O., Černocký, J. Honza, Burget, L. 2019. Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition. *Computer Speech & Language*, ***58*****,** 403–421.

Rajan, P., Kinnunen, T., Hautamäki, V. 2013. Effect of Multicondition Training on i-Vector PLDA Configurations for Speaker Recognition. In *INTERSPEECH 2013* (pp. 3694–3697). Lyon, France.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, ***10*****(3)**, 19–41.

Ribas, D., Vincent, E. 2019. An Improved Uncertainty Propagation Method for Robust I-Vector Based Speaker Recognition. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6331–6335). Brighton, UK.

Sarkar, S., Sreenivasa R.K. 2014. A Novel Boosting Algorithm for Improved i-Vector based Speaker Verification in Noisy Environments. In *INTERSPEECH 2014* (pp. 671–675). Singapore.

Sim, K.C. 2013. Approximated Parallel Model Combination for efficient noise-robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7383–7387). Vancouver, BC, Canada.

Sim, K.C., Luong, M.T. 2011. A Trajectory-based Parallel Model Combination with a unified static and dynamic parameter compensation for noisy speech recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 107–112). Waikoloa, HI, USA.

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., Khudanpur, S. 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 165–170). San Diego, CA, USA.

Tao, Y., Li, X., Wu, B. 2008. An Effective PCM Based Environment Compensation Approach in Speech Processing for Mobile e-Learning Platform. In *2008 Third International Conference on Pervasive Computing and Applications* (pp. 772–775). Alexandria, Egypt.

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, ***90***, 250–271.

Tufekci, Z., Gowdy, J.N., Gurbuz, S., Patterson, E. 2006. Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition. *Speech Communication*, ***48*****(10)**, 1294–1307.

Varga, A., Steeneken, H.J.M. 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, ***12*****(3)**, 247–251.

Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4052–4056). Florence, Italy.

Wang, S., Huang, Z., Qian, Y., Yu, K. 2018. Deep Discriminant Analysis for i-vector Based Robust Speaker Recognition. In *11th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 195–199). Taipei, Taiwan.

Zhang, X., Zou, X., Sun, M., Wu, P., Wang, Y., He, J. 2020. On the complementary role of DNN multi-level enhancement for noisy robust speaker recognition in an i-vector framework. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, ***E103A*****(1)**, 356–360.

Zhang, X., Zou, X., Sun, M., Zheng, T. F., Jia, C., Wang, Y. 2019. Noise Robust Speaker Recognition Based on Adaptive Frame Weighting in GMM for i-Vector Extraction. *IEEE Access*, ***7*****(2019)**, 27874–27882.

Zhou, L., Li, H., Chen, Y., Wu, Z., Lu, Y. 2016. VTS feature compensation based on two-layer GMM structure for robust speech recognition. In *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)* (pp. 1–5). Yangzhou, China.