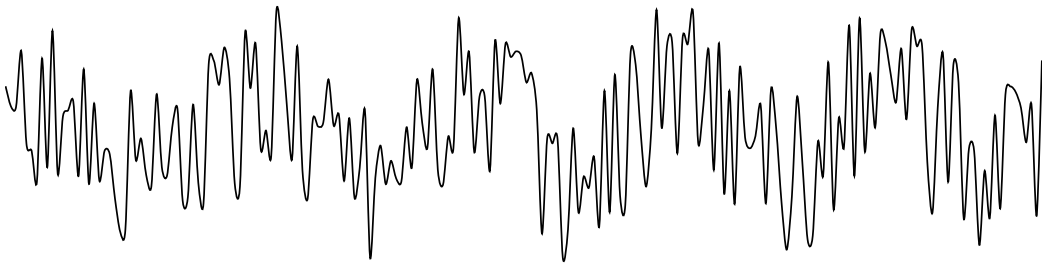
Domain Mismatch Compensation for Text-Independant Speaker Recognition



Valentin Iovene
Laboratoire de Recherche et Développement de l'EPITA
toogy@lrde.epita.fr

June 21, 2016

Abstract

Although the development of the **i-vector**-based probabilistic linear discriminant analysis (PLDA) systems led to promising results in speaker recognition, the impact of **domain mismatch** when the system training data and the evaluation data are collected from different sources remains a challenge. Johns Hopkins University (JHU) 2013 speaker recognition workshop, for which a domain adaptation challenge (DAC13) corpus was created, focused on finding solutions to address this problem.

This research report lays out the state-of-the-art techniques used for domain mismatch compensation ; such as a combination of various **whitening** transforms, and the use of a **dataset-invariant** covariance normalization to obtain domain-invariant representations of PLDA training data. Those techniques are evaluated on the DAC13 corpus and compared.

Résumé

Bien que le développement des systèmes d'analyse discriminante linéaire probabiliste (PLDA) basés sur les i-vecteurs a donné lieu à des résultats prometteurs en reconnaissance du locuteur, l'impact du *domain mismatch* lorsque les données d'entraînement du système et les données d'évaluation proviennent de sources différentes reste un défi. Le workshop de reconnaissance du locuteur de 2013 de l'Université Johns Hopkins (JHU), pour lequel un corpus d'adaptation du domaine (DAC13) a été créé, a travaillé à trouver des solutions pour résoudre ce problème.

Ce rapport de recherche présente les techniques de pointe utilisées pour la compensation du *domain mismatch* ; comme une combinaison de plusieurs transformées de blanchiment, et la normalisation de la covariance indépendante du jeu de données pour obtenir des représentations des données d'entraînement de la PLDA invariantes par rapport au domaine. Ces techniques sont évaluées sur le corpus DAC13 et comparées.

Contents

1 Understanding domain mismatch

This section gives a brief definition of what *domain mismatch* is. It attempts to make the context clear, ensuring the reader fully understands the objectives of this study and the content of the next sections.

* * *

An audio recording of a speech contains various information, e.g.:

- characteristics of the speaker's voice,
- environment specificities,
- the content of the speech (language and words used by the speaker),
- or behavioral information such as the accent or speech rate.

Depending on the task, some information can either be very valuable or completely useless (noise). It even becomes a **handicap** because the noise varying from one recording to another leads to an increased error rate when comparing two speakers. *Note this is not only true for the particular task of speaker recognition but also in general.* Errors are observed because the decision is not based *purely* on the speaker's vocal characteristics, which is the only valuable information in the case of *text-independent* speaker recognition, but also on information that is irrelevant.

Domain mismatch happens when the training examples (audio recordings) used for tuning the hyperparameters [1] of the speaker recognition system and the examples used for enrolment+testing do not come from the same dataset. They were recorded in different conditions: the microphone used for recording or the environment surrounding the speaker may not have been the same. Thus, to apply speaker recognition techniques to real-life conversation, it is necessary to find ways to make the system more robust to new conditions. The extra-information that comes from a new domain need to be compensated. A definition of *domain mismatch compensation* can then be given.

Compensating domain mismatch means *filtering out* the information in the data that is specific to the domain (dataset) while *emphasizing* (or at least leave unimpaired) the speaker-dependent information.

1. training the universal background model (UBM) and total variability matrix

2 I-vector based speaker verification

This section gives a brief overview of the i-vector based speaker verification system used in this study. The different steps of the process used to extract an i-vector from an audio signal are described.

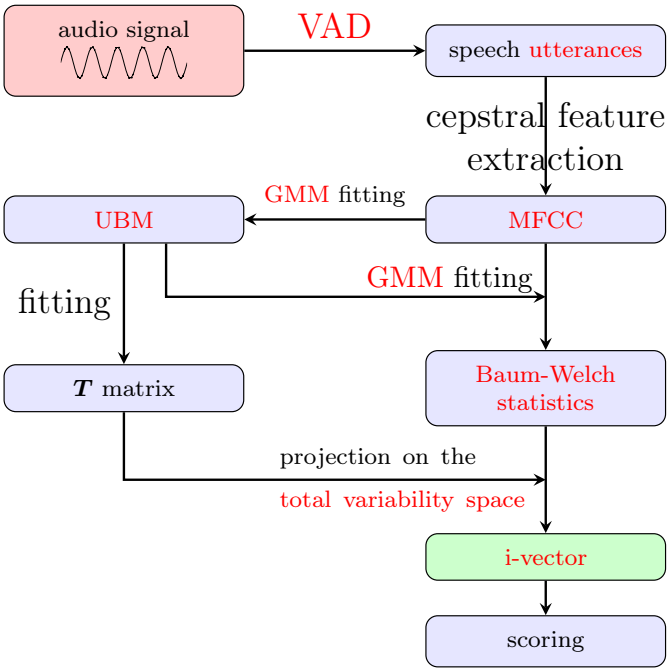


Figure 1 – Flowchart of an i-vector based speaker verification system

Once an i-vector is obtained from an utterance, the extraction mechanism is ignored and the i-vector is regarded as an observation from a probabilistic generative model.

3 Compensation techniques

Several compensation techniques were developed to be applied at different steps of a speaker recognition system. Some can be applied directly on cepstral features (mel-frequency cepstral coefficients (MFCC)) while others can be applied on i-vectors. *Note that multiple compensations strategies can be combined.*

This section lays out examples of strategies that were developed or used for compensating domain mismatch.

3.1 Cepstral mean substraction

Cepstral mean substraction (CMS) is a common and longstanding method of normalization. It was first developed to provide noise robust speech recognition [9] and was soon after applied to speaker recognition [3]. It is used to make the system robust to adverse effects that can distort the short-term distribution of the speech features, such as additive noise and non-linear effects attributed to handset transducers [4].

The recording of a conversation between multiple speakers is often the sum of two separately recorded channels with possibly different channel characteristics (e.g. two different phones, in the case of a phone conversation). Parameters should not be calculated over the whole conversation. Instead, parameters (μ, σ), where μ is the mean of the cepstral features and σ their standard deviation, are estimated over a limited sliding time window. The time window has to be long enough to have a reliable estimation while containing the speech of only one speaker.

This method is equivalent to data whitening assuming that the feature vector elements are uncorrelated.

3.2 Feature warping

Like CMS, feature warping is an acoustic feature post-processing method designed to introduce robustness to linear channel mismatch and additive noise conditions. [6] It showed improvements over a number of methods such as CMS.

The idea is to conform the speech statistics to a target distribution. The distribution of a cepstral feature stream is warped to a standardised distribution over a specified time interval.

Unlike feature standardization, feature warping guarantees that the warped feature distributions are Gaussian, although this may lead to a suboptimal mapping when the real distribution is non-Gaussian.

It has become the standard for cepstral feature normalization.

3.3 Hyperparameter model conditioning

3.3.1 Linear discriminant analysis

LDA is a channel compensation method used to define a subspace onto which i-vectors are projected that minimizes the intra-class variance caused by channel effects and maximize the variance between speakers through the eigenvalue decomposition of,

$$\boldsymbol{S}_b \boldsymbol{v} = \tau \boldsymbol{S}_w \boldsymbol{v} \tag{1}$$

where τ are the eigenvalues, $\boldsymbol{v}$ the eigenvectors and $\boldsymbol{S}_b, \boldsymbol{S}_w$ the between-classes and within-class matrices, computed as follows,

$$\boldsymbol{S}_b = \sum_{s=1}^{S} n_s (\bar{\boldsymbol{w}}_s - \bar{\boldsymbol{w}})(\bar{\boldsymbol{w}}_s - \bar{\boldsymbol{w}})^t \tag{2}$$

$$\boldsymbol{S}_w = \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\bar{\boldsymbol{w}}_i^s - \bar{\boldsymbol{w}}_s)(\bar{\boldsymbol{w}}_i^s - \bar{\boldsymbol{w}}_s)^t \tag{3}$$

where S is the total number of out-domain speakers, n_s is the number of sessions of speaker s, $\bar{\boldsymbol{w}}_s$ is the mean i-vector for each speaker and $\bar{\boldsymbol{w}}$ is the mean of all speakers which are defined by,

$$\bar{\boldsymbol{w}}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \boldsymbol{w}_i^s \tag{4}$$

$$\bar{\boldsymbol{w}} = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \boldsymbol{w}_i^s \tag{5}$$

where N is the total number of sessions.

3.3.2 LDA variants

Several domain compensation techniques have been proposed that modify the conventional computation of the within-speaker and between-speaker scatter matrices used in applying LDA. In source normalization, the between-speaker scatter matrix is computed as the average of the individual in-domain between-speaker scatter matrices, thereby removing cross-domain bias from the final matrix and avoiding an overestimation of between-speaker variability. The within-speaker scatter is then computed as the difference between the total variability scatter and the between-speaker scatter, thus attempting to introduce cross-domain variability into the within-speaker scatter matrix. The authors in [5] showed that this method provided improved performance in NIST SRE12 task, which contained cross-domain utterances from telephony, microphone and interview recordings.

A related method, referred to as within-class covariance correction (WCCN), relies on a conventional computation of the between-speaker scatter matrix but adds a correction term to the within-scatter matrix using the cross-domain scatter obtained from a source-labeled dataset. Improvements are reported for both the RATS task and the 2013 domain adaptation challenge data. Note that both source normalization and within-class covariance correction rely on the availability of a large speaker-labeled dataset (although not necessary in-domain) to compute the LDA transformation.

3.4 Inter-dataset variability compensation

Inter-dataset variability compensation (IDVC) was introduced during the JHU 2013 summer workshop as a means of conditioning i-vectors to compensate for dataset mismatches deliberately introduced into the domain adaptation challenge task (DAC13). As will be described in further detail in Section III, the DAC13 corpus consisted of a fully labeled Switchboard (SWB) dataset, an enroll and test dataset derived from MIXER data, and an unlabeled SRE dataset available for domain adaptation. Results presented in TODO show that domain mismatch compensation can be achieved without using any in-domain adaptation data by applying nuisance attribute projection (NAP) to the i-vectors to remove known between-dataset variability. An extension of the IDVC concept, described in TODO, showed additional improvements in performance on the DAC13 challeng task. This method was not evaluated

in the present study.

3.5 Whitening and length normalization

Another i-vector conditioning technique is length normalization, which was introduced in TODO as a means of Gaussianizing i-vectors so that they conform to the Gaussian modeling assumptions made in probabilistic linear discriminant analysis (PLDA). The length-normalization approach requires that i-vectors be whitened before being projected onto the unit sphere. This relatively simple process allows conventional i-vector recognizers to match the performance of the more computationally demanding heavy-tailed PLDA model.

Whitening is applied to the data via an affine transformation $y = S^{-1/2}(w - \bar{w})$ where w is the raw i-vector, $\bar{w}$ is the i-vector mean, S is the i-vector covariance matrix, and y is the resulting whitened i-vector. Ideally $\bar{w}$ and S are estimated from i-vectors derived from the same source as w.

Length normalization is a common component of published state-of-the-art systems, although descriptions do not always make clear whether the i-vector were properly whitened prior to normalization.

4 Library of Whiteners

The idea of using different whiteners was introduced in [7]. In practice, the whitening parameters $(\bar{\boldsymbol{w}}, \boldsymbol{S})$ are derived from a combination of sources, making the underlying distribution multi-modal and thus leading to an improper whitening transformation and suboptimum Gaussianization via length normalization.

To reduce the effects of domain mismatch on i-vector length normalization, [7] proposes using a *withening library* whose individual components are estimated from labeled source collections. Each i-vector $\boldsymbol{w}$ is compared against a source-specific Gaussian model $\mathcal{N}(\boldsymbol{w}|\mu_j, \boldsymbol{S}_j)$, where parameters $(\mu_j, \boldsymbol{S}_j)$ are estimated from source-specific data. The parameters of the model that produces the maximum likelihood are used to whiten the i-vector w and applied via $y = \boldsymbol{S}_j^{-1/2}(\boldsymbol{w} - \mu_j)$ where

$$j = \underset{i \in (1,\ldots,K)}{\operatorname{argmax}} \mathcal{N}(\boldsymbol{w}|\mu_i, \boldsymbol{S}_i)$$

In other words, the parameters of a Gaussian are estimated independently for each dataset and a new i-vector is whitened using the parameters of the Gaussian that most likely produced the i-vector.

5 Dataset-Invariant Covariance Normalization

5.1 Length-normalized GPLDA system

6 Evaluation on DAC13

6.1 About the DAC13 dataset

It is very well known.

6.2 Metrics

7 Results

Performance is shown in terms of equal error rate (EER), minimum decision cost function (minDCF) and detection error tradeoff (DET) curves.

Glossary

Baum-Welch algorithm	algorithm used to fit a hidden markov model (HMM).
Baum-Welch statistics	emission matrix of a HMM fitted by the Baum-Welch algorithm.
i-vector	projection of the Baum-Welch statistics on the total variability space.
scatter matrix	A scatter matrix S is a statistic used to make estimates of a covariance matrix. In the case of a multivariate normal distribution, S can be written as: $$S = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t \in \boldsymbol{R}^{d \times d}$$ where n is the number of samples, d the number of features, $\boldsymbol{x}_i$ are the samples and $\bar{\boldsymbol{x}}$ their mean. S is positive definite if there exists a subset of the data consisting of d affinely independent observations (which we will assume).

total variability matrix

a channel-dependant gaussian mixture model (GMM) supervector $\boldsymbol{M}$ can be modeled as follows

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{6}$$

where $\boldsymbol{m}$ is a speaker- and channel-independent supervector (the UBM supervector is a good estimate of $\boldsymbol{m}$, $\boldsymbol{T}$ is a low rank matrix, which represents a basis of the deduced total variability space and $\boldsymbol{w}$ is a standard normally distributed vector. $\boldsymbol{T}$ is the name of the total variability matrix; the components of $\boldsymbol{w}$ are the total factors and they represent the coordinates of the speaker in the reduced total variability space. These feature vectors are referred to as *identity vectors* or i-vectors for short.

total variability space

lower dimensional space on which Baum-Welch statistics are projected.

utterance

speech sequence consisting of one or more words and preceded and followed by silence or a change in speaker.

Acronyms

CMS	cepstral mean substraction.
DET	detection error tradeoff.
EER	equal error rate.
GMM	gaussian mixture model.
HMM	hidden markov model.
IDVC	inter-dataset variability compensation.
LDA	linear discriminant analysis.
MFCC	mel-frequency cepstral coefficients.
minDCF	minimum decision cost function.
NAP	nuisance attribute projection.
PLDA	probabilistic linear discriminant analysis.
SWB	Switchboard.
UBM	universal background model.
VAD	voice activity detection.
WCCN	within-class covariance correction.

References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011.

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

[3] J. Koolwaaij and L. Boves. Local normalization and delayed decision making in speaker detection and tracking. *Digital Signal Processing*, 10(1-3):113–132, January/April/July 2000.

[4] N. Kurosawa, H. Kobayashi, and K. Kobayashi. Channel linearity mismatch effects in time-interleaved adc systems. In *Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on*, volume 1, pages 420–423 vol. 1, May 2001.

[5] M. McLaren and D. van Leeuwen. Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):755–766, March 2012.

[6] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, pages 213–218, Crete, Greece, 2001. International Speech Communication Association (ISCA).

[7] E. Singer and D. A. Reynolds. Domain mismatch compensation for speaker recognition using a library of whiteners. *IEEE Signal Processing Letters*, 22(11):2000–2003, Nov 2015.

[8] J. H. University. 2013 speaker recognition workshop. http://www.clsp.jhu.edu/workshops/archive/ws13-summer-work-shop/groups/spk-13/. [Online].

[9] O. Viikki and K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Comumunication*, 25(1-3):133–147, August 1998.