Domain Mismatch Compensation for Text-Independant Speaker Recognition

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

June 14, 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, like:

- 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 not relevant to the task.

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.

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

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

2 Compensation techniques

Several compensation techniques were developed to be applied at different steps of a speaker recognition system. Some will be applied on cepstral features (e.g. mel-frequency cepstral coefficients (MFCC)), others 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.

2.1 Cepstral features standardization and warping

Acoustic feature standardization is a common and longstanding method of normalization that was originally developed to provide noise robust speech recognition and was then applied to speaker recognition. The technique is applied by using parameters estimated over a sliding segmental window to normalize each element of a signal-processing front-end feature vector to have a zero mean and unit standard deviation. This method was intended to reduce mismatches between the training and testing examples.

Feature warping is an acoustic feature vector post-processing method designed to introduce robustness to linear channel mismatch and additive noise conditions. Robustness is achieved by warping individual cepstral feature streams over a short time interval to conform to a standard Gaussian target distribution. 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.

2.2 Hyperparameter model conditioning

Linear discriminant analysis (LDA) is a common method used to identify a subspace onto which i-vectors are projected so as to simultaneously maximize speaker discrimination while reducing inter-session variability. 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 TODO 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, 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.

2.3 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.

2.4 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.

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.

3 Library of Whiteners

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. In practice, however, the data available to train the whitening parameters may be 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, this paper proposes using a *withening library* whose individual components are estimated from labeled source collections. Each i-vector w is compared against a source-specific Gaussian model $\mathcal{N}(w|\mu_j, S_j)$, where parameters (μ_j, 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 $y = S_j^{-1/2}(w - \mu_j)$ where

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

4 Dataset-Invariant Covariance Normalization

5 Evaluation on DAC13

5.1 About the DAC13 dataset

It is very well known.

5.2 Metrics

6 Results

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

Glossary

i-vector	projection of the Baum-Welch statistics on the total variability space.
total variability matrix	projection matrix used to project Baum-Welch statistics on the total variability space.

Acronyms

IDVC	inter-dataset variability compensation.
LDA	linear discriminant analysis.
MFCC	mel-frequency cepstral coefficients.
NAP	nuisance attribute projection.
PLDA	probabilistic linear discriminant analysis.
SWB	Switchboard.
UBM	universal background model.