Intersession Variability in Speaker Recognition: A behind the Scene Analysis

Daniel Garcia-Romero, Carol Y. Espy-Wilson

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD
dgromero@umd.edu, espy@umd.edu

Abstract

The representation of a speaker’s identity by means of Gaussian supervectors (GSV) is at the heart of most of the state-of-the-art recognition systems. In this paper we present a novel procedure for the visualization of GSV by which qualitative insight about the information being captured can be obtained. Based on this visualization approach, the Switchboard-I database (SWB-I) is used to study the relationship between a data-driven partition of the acoustic space and a knowledge based partition (i.e., broad phonetic classes). Moreover, the structure of an intersession variability subspace (IVS), computed from the SWB-I database, is analyzed by displaying the projection of a speaker’s GSV into the set of eigenvectors with highest eigenvalues. This analysis reveals a strong presence of linguistic information in the IVS components with highest energy. Finally, after projecting away the information contained in the IVS from the speaker’s GSV, a visualization of the resulting GSV provides information about the characteristic patterns of spectral allocation of energy of a speaker.

Index Terms: speaker recognition, intersession variability, model compensation, intersession variability subspace.

1. Introduction

The state-of-the-art in speaker recognition is mostly dominated by the use of Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) as classifiers working with short-time cepstral features [1], [2]. A fundamental problem with this feature extraction approach is that not only is the speaker information captured, but also other sources of variability (e.g., channel characteristics, environment, speaker emotional state, linguistic content) which severely affect the performance of the recognition systems. These undesired sources of information are commonly referred to as a whole under the name of intersession variability. A wide variety of techniques have been proposed in the past ten years to address this issue (see [3] for an overview). Recently, feature- and model-level intersession compensation techniques have been proposed in the form of factor analysis for GMMs [4]. These methods work by modeling the UBM-MAP adapted means of a speaker’s GMM using latent factors. The key idea behind these approaches is the construction of a Gaussian supervector by stacking the means of the mixture components. Since the MAP adaptation process only updates the means of the speaker’s GMM with respect to the UBM, all the discriminant information of the speaker captured by the GMM is contained in this high-dimensional GSV. In this way, a speaker’s utterance is represented by a point in a high-dimensional vector space (i.e., dimension ~ 50k). Moreover, this space is assumed to be spanned by the direct sum of two subspaces, namely the speaker information subspace (SIS) and the intersession variability subspace (IVS). Hence, any GSV belonging to the vector space has a unique decomposition in terms of speaker information components and intersession variability.

An important property of the GSV paradigm is that it provides a fixed-length feature vector to represent variable-length speech utterances. Due to this property, a new class of recognizers has been proposed in which the UBM-MAP adapted GMM paradigm is used as a front-end and a SVM classifier is used in the back-end [2]. In this way, the representation of a speaker’s identity by means of GSV plays an important role in GMM systems, as a way to compensate for session variability, and also in SVM systems as input feature vectors.

The excellent performance of ISV-compensated speaker recognition systems based on GSV motivates our interest in the structure of this representation space. It is for this reason that we propose a procedure to display GSV and use it to obtain a qualitative insight about the speaker information subspace and the intersession variability subspace.

2. Experimental setup

2.1. Switchboard-I database (SWB-I)

The Switchboard-I database is comprised of conversational speech between two speakers recorded over landline telephone channels with a sampling rate of 8 KHz [5]. The average duration of each conversation is 5 minutes (approx. 2.5 min per speaker) and each conversation side is recorded in a different file. The total number of speakers in the database is 520 with a balance in gender and recorded into 4856 speech files. The telephone handsets were either electret or carbon bottom with an approximate proportion of 70% and 30% respectively. The availability of manual phonetic transcriptions [6] along with a fairly limited amount of channel/handset variability makes this database a good candidate for the experiments in this paper.

2.2. UBM-GMM system

A UBM-GMM system with 2048 mixtures and diagonal covariance matrices [1] has been used throughout this work with two purposes. First to compute GSVs for the analysis presented in sections 3, 4 and 5. Second, to construct a GMM speaker recognition system to quantify the effects of intersession variability compensation.

3. Gaussian supervectors (GSVs)

3.1. GSV computation

A Gaussian supervector is constructed by stacking the means of the mixture components of a MAP mean-adapted GMM from a UBM. In the following we provide a detailed
A GSV can be understood as a summary of the average patterns of spectral allocation of energy of a particular speaker. However, the linguistic content of the speech signal imposes some constraints in these patterns (e.g., relative position of formants). In this way, it seems natural to think that the elements (i.e., mean vectors) of the GSV will exhibit some kind of clustering. To check this, a simple K-means procedure was used to partition the elements of the two UBM GSVs into a set of classes. The Euclidean distance between the mean vectors (i.e., 19 MFCC vectors) was used and the number of classes was set to 16. We followed the visualization methodology of Figure 1 to display the GSV of the UBM for \( P_1 \). The second panel of Figure 3 shows the result of processing the GSV of the UBM of the partition \( P_1 \) of SWB-I.
Clusters of Gaussian mixture’s mean vectors

Also, it is important to note, that the clustering was done prior to the pseudo-inversion stage and therefore no imprecision was introduce in the process.

A quick inspection of the UBM’s GSV reveals that the mean vectors that get grouped together share in common their most predominant regions of spectral allocation of energy (i.e., formants). This raises the following question: Is there any relationship between a data-drive partition of the acoustic space and a knowledge-based partition such as the broad phonetic classes? In order to answer this question the following experiment was conducted in each of the SWB-I partitions independently. First, for each file, the manual phonetic transcriptions of SWB-I [5] were used to align each feature vector with a broad phonetic class. The following set of phonetic classes was used: {liquids (LIQ), nasals (NAS), voiced/unvoiced fricatives (V/U-F), voiced/unvoiced stops (V/U-S), diphthongs (DIP) and back/center/front vowels (B/C/F-V)}. Then a probabilistic alignment of each feature vector with their corresponding UBM was performed. Only the top-1 scoring mixture was used for each feature vector. During this process, we kept track of the number of times each one of the 2048 UBM’s mixtures was used to score a frame with a particular phonetic class label. As a result, we obtained a probabilistic alignment of each UBM mixture with the aforementioned set of broad phonetic classes. As an example, if a given mixture was used 80% of the time to score frames in nasal regions, the process would assign a 0.8 probability mass to that mixture with respect to nasals.

Two important observations were made. First, every mixture had a non-zero probability mass for each broad phonetic class. Second and most important, the probability mass was not uniformly distributed and was highly concentrated on one or two phonetic classes for each mixture. Moreover, in order to establish a connection between the data-driven clusters and the broad phonetic classes we averaged the probabilistic assignments among all the mixtures in the same data-driven cluster. The top panel of Figure 3 shows the result of thresholding this averaged probabilistic alignment to keep approximately 90% of the probability mass. Each data-driven cluster gets aligned with at most 2 broad phonetic classes. After a close analysis of the resulting pairings between data-driven clusters and phonetic classes the authors believe that there is a good matching between the formant regions of the clusters and the canonical formant regions of the phonetic classes (see [7] for examples of these). The reader is encouraged to perform its own comparison.

Due to the lack of space, Figure 3 only depicts the results for the partition $P_1$ of SWB-I. However, the same
observations were made by analyzing the results of P2. This supports the generality of our results.

Based on the experiments presented in this section we can claim that not only GSVs capture average patterns of spectral allocation of energy, but also a phonetic meaning can be attached to partitions of the GSV.

5. GSV space structure

As a motivation for our interest in the structure of the IVS, we run the MIT-LL UBM-GMM recognition system [1] on SWB-I to quantify the effects of intersession compensation. The same protocol used in the Super-SID evaluation was followed [6]. We focused our attention on the 1conv-1conv recognition task. Each speaker model was trained with approximately 2.5 minutes of speech and tested against segments of similar duration. The total number of trials was 16280 from which around 10% were non-target trials. The solid line of Figure 4 shows the performance for the baseline system for which no intersession compensation was used.

The methodology presented in [8] was used to compute the IVS and also to constrain the adaptation of the speaker models to the test files into the intersession subspace. The data in partition P1 of SWB-I was used for the IVS of P2 and vice versa. Table 1 shows the influence of the dimension of the IVS in the recognition performance in terms of Equal Error Rate (EER). The best results were obtained for a 32-dimensional IVS. This number of dimensions is a little bit smaller than those reported for other databases with more sources of intersession variability [3]. The dashed line of Figure 4 shows an impressive improvement of more than 50% in the performance of the intersession variability compensated system with respect to the baseline.

The great improvement obtained indicates that projecting away the components of the IVS from the speaker’s GSVs result in a much more discriminative representation. However, the exact composition of the sources of variability removed from the GSVs is still an unexplored area.

In order to get a first look into this issue, we selected a female speaker from the partition P1 and obtained her GSV by MAP adaptation of the means of the corresponding UBM. Panel (c) of Figure 3 shows the result. It is interesting to observe how similar to the UBM’s GSV the speaker’s GSIs remains after adaptation. A more surprising result is the one obtained by projecting the speaker’s GSV into a 32-dimensional IVS (panel (d) in Figure 3). This projection resembles a smoothed version of the GSV in which the location of the formants becomes less defined. Therefore, the components of the projection into the IVS seem to mostly capture the variability in the spectral allocation of energy due to linguistic constraints. To this point, it is still not clear to the authors how the channel/handset variability of SWB-I is reflected in this representation. Further analysis will be carried out in the future to try to understand this issue.

Finally, by projecting away the IVS components from the GSV (see bottom panel in Figure 3), the most discriminative representation of the speaker’s identity is obtained. A lot of the smoothness of the GSV is removed. The components of the compensated GSV seem much more irregular than those of the original GSV. The improved discriminative power of this representation seems to indicate that this less smooth GSV captures the most idiosyncratic characteristics of the speaker’s spectral allocation of energy. Further analysis of this claim will be conducted in future work by contrasting the GSVs of different speakers.

6. Conclusions

A procedure to display GSVs was introduced in this paper. The potential benefits of this visualization tool were explored by using it as a visual aid to establish a probabilistic relation between a partition of the acoustic space and a set of broad phonetic classes. A meaningful relation was obtained that reveals useful information about the components of the GSV.

Moreover, a qualitative analysis of the structure of the speaker information subspace and the intersession variability subspace provided a first look into the actual composition of the sources of variability present in the IVS.

7. References