Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Linearly Constrained Minimum Variance for Robust I-vector Based Speaker Recognition

A. Khosravani, M.M. Homayounpour

Laboratory for Intelligent Multimedia Processing
Department of Computer Engineering & Information Technology
Amirkabir University of Technology
a.khosravani@aut.ac.ir, homayoun@aut.ac.ir

Abstract

This paper aims at presenting our algorithm used to make submission for the NIST 2013-2014 speaker recognition i-vector challenge. The fixed dimensional i-vector representation of speech utterances has attracted attentions from other communities. This challenge focuses on the task of speaker detection using i-vectors derived from conversational telephony speech data. However, the unlabeled i-vectors provided for development purpose make the problem more challenging. The proposed method uses the idea of one of the popular robust beamforming techniques named Linearly Constrained Minimum Variance (LCMV), which has been presented in the context of beamforming for signal enhancement. We will show that LCMV can improve performance by building a model from different i-vectors of a given speaker so as to cancel inter-session variability and increase inter-speaker variability. Imposter covariance matrix modification and score normalization using a selection of imposter speakers have been proposed to improve performance. As measured by minimum decision cost function defined in the challenge, our result is 27% better relative to the baseline system.

1. Introduction

Over recent years, i-vector representation of speech segments has been widely used by the state-of-the-art speaker recognition systems [1]. This representation maps arbitrary duration speech segments into a fixed and low dimensional vector which encourages researchers to explore new ideas from other communities to be used in speaker recognition.

The main challenge of i-vector representation is the variability associated with different i-vectors of the same speaker. This variability is mainly due to the use of different handsets, environmental noise, speaker health and emotion or segment duration. Therefore, inter-session variability compensation techniques have been developed to directly remove unwanted variability from i-vectors [1-6]. The state-of-the-art Probabilistic Linear Discriminant analysis (PLDA) [4] with its variants (Gaussian and heavy-tailed) [2], Linear Discriminant Analysis (LDA), Within-Class Covariance Normalization (WCCN) and Nuisance Attribute Projection (NAP) [7] are the most important techniques which have shown their ability to successfully compensate for inter-session variability. However, their ability to compensate for this variability has a direct relationship with the quantity of developmental labeled i-vectors in which a typical speaker has as many i-vectors (acquired from variety of conditions) as it is sufficiently broad to cover most of the session variability. In contrary to the costly process of obtaining labeled speech data, unlabeled speech data are readily available. Therefore, unsupervised techniques need to be developed to make use of them to improve speaker recognition systems. The NIST 2013-2014 speaker recognition i-vector challenge tries to provide such opportunity [8].

The objective of this paper is to describe the method used to make submission for the NIST i-vector challenge. The proposed method is inspired by the theory of adaptive filtering. Adaptive beamforming is widely used in several application areas such as radar, sonar, wireless communication and even speech enhancement [9]. The objective of adaptive beamforming techniques is to maximize signal-to-interference-plus-noise ratio (SINR). Provided that the desired signal is not distorted, the problem leads to the minimum variance distortionless response (MVDR) or standard Capon beamformer (SCB) [10] which provides an excellent performance if we have data without desired signal component. The goal of MVDR is to find a filter that passes the desired signal while rejects noise and interferences. In practice, however, the desired signal is not known exactly and this will cause the performance of MVDR to degrade significantly. Multiple techniques have been proposed to improve the robustness of MVDR beamformer which include Linearly Constrained Minimum Variance (LCMV) [9], norm constraints [9], multidimensional covariance fitting [11] and more. LCMV improves performance by incorporating different signals which are in the vicinity of the desired signal.

The similarity of LCMV to the problem at hand encourages us to use this technique. In this way, we take all i-vectors of a target speaker in the evaluation set as signals in the vicinity of the unknown desired signal which is the expected i-vector of that target speaker, and all i-vectors in the development set as noise or interference signals since the speakers involved in these two sets are disjoint. Our experiments indicate that using those imposter i-vectors which are the most similar to the target model can significantly improve the performance. Moreover, to better compensate the inter-session variability, we have built a within-class covariance matrix using groups of similar i-vectors from the development set, each is supposed to be from the same speaker, and have used it along with the imposter covariance matrix to improve the performance. Finally, the score is normalized using a selection of imposter i-vectors.

The paper is organized as follows. In Section 2 we will explain how beamforming techniques can be used to compensate for the variability associated with speaker i-vectors and describe the techniques used to improve performance. In Section 3 we have carried out experiments to show the effect of different techniques on the speaker

recognition performance. Finally we have concluded the paper in Section 4.

2. Inter-session Compensation

Numerous session compensation techniques have recently been proposed to remove the nuisance variability from i-vectors. The most effective techniques include Within-Class Covariance Normalization (WCCN) [12] which uses the inverse of within-class covariance matrix to weaken the effect of nuisance directions, Nuisance Attribute Projection (NAP) [13] which removes the nuisance direction, Linear Discriminant Analysis (LDA) which defines new axes to minimize inter-session variability of each speaker's i-vectors and maximize between-speaker variance, and Probabilistic Linear Discriminant Analysis (PLDA) [4] which incorporates speaker and channel subspaces. However, these techniques require a large quantity of labeled i-vectors which is not provided in this challenge. Therefore, in order to be able to use these supervised techniques, we need to find i-vectors belonging to each speaker in the development set through unsupervised techniques. In this way, their performance will be affected by the accuracy of the unsupervised technique.

In this work we have proposed a method based on the idea used in beamforming to build a model from different i-vectors of a target speaker as well as a selection of imposter i-vectors to cancel inter-session variability and increase inter-speaker variability. Further improvement has been achieved by incorporating a within-class covariance matrix built in an unsupervised manner.

2.1. Adaptive Filtering

The baseline system presented for this challenge can be described by the theory of adaptive filters if we just ignore the length normalization of i-vectors. Adaptive filtering is a means of adaptive extraction of a weak desired signal in the presence of noise signal. Let's take each i-vector as a signal of dimension N. Given a collection of target speaker i-vectors $i_1,...i_M$ and a collection of imposter i-vectors i_{imp}, we define a filter $w \in \mathbb{R}^{\mathrm{N}}$ that only passes target speaker i-vectors. The filter output is defined as

$$\tilde{d} = w^T i.$$

We need to find w so that $\tilde{d}$ equals 1 for all target speaker i-vectors,

$$w^T i_s = 1, \quad s = 1,...,\mathrm{M}$$

and 0 for all imposter speakers i_{imp},

$$w^T i_{imp} = 0.$$

Using minimum mean squared error (MSE), the optimum filter will be

$$\begin{aligned} &\mathrm{MSE} = \mathrm{E}\{|\, d - w^T i \,|^2\} \\ &\min_w \mathrm{MSE} \rightarrow \partial\,\mathrm{MSE}/\,\partial w = 0 \\ &\rightarrow \mathrm{E}\{i\, i^T\} w - \mathrm{E}\{i\, d\} = 0 \\ &\rightarrow w = \mathbf{R}^{-1}.i_{tar} \end{aligned}$$

where d is the desired output,

$$\mathbf{R} = \mathrm{E}\{i\, i^T\}$$

is the i-vector covariance matrix which can be estimated using the development set ($\tilde{\mathbf{R}}$) and

$$i_{tar} = \mathrm{E}\{i\, d\} \approx \frac{i_1 + ... + i_M}{M}$$

is the expected value of the target speaker's i-vector which can be estimated by the available samples ($\tilde{i}_{tar}$). This filter is the optimum model for the target speaker in mean squared sense and can also be achieved iteratively using algorithms such as Least Mean Square (LMS) or Recursive Least Squares (RLS) [14]. The output produced by the baseline system (the cosine similarity of the whitened i-vectors) that is given by

$$\tilde{d}_{baseline} = \left(\frac{\tilde{\mathbf{R}}^{-\frac{1}{2}}.\tilde{i}_{tar}}{\sqrt{\tilde{i}_{tar}^T \tilde{\mathbf{R}}^{-1}.\tilde{i}_{tar}}}\right)^T.\left(\frac{\tilde{\mathbf{R}}^{-\frac{1}{2}}.i_{test}}{\sqrt{\tilde{i}_{test}^T \tilde{\mathbf{R}}^{-1}.\tilde{i}_{test}}}\right)$$

$$= \frac{w^T.\tilde{i}_{test}}{\sqrt{\tilde{i}_{tar}^T \tilde{\mathbf{R}}^{-1}.\tilde{i}_{tar}}\sqrt{\tilde{i}_{test}^T \tilde{\mathbf{R}}^{-1}.\tilde{i}_{test}}},$$

is the same as the one produced by the filter if we ignore the projection of i-vectors into unit sphere. We should note that we have centralized i-vectors using the development data as a preprocessing step.

2.2. Minimum Variance Distortionless Response (MVDR)

The goal of MVDR is to pass signals impinging on an array from a desired angle while rejecting noise and interferences from all the other angles. An optimum beamformer provides maximum noise rejection while matching the signal from a desired angle. MVDR uses maximum of signal-to-Interference-plus-Noise-Ratio (SINR) as the criterion.

The idea in the solution of this problem inspired us to apply it to our problem. In this way we aim at finding a filter w which maximizes the SINR formulated as

$$\max_w \mathrm{SINR} = \frac{E\{|\, w^T i_{tar} \,|^2\}}{E\{|\, w^T i_{imp} \,|^2\}},$$

where i_{tar} is the desired target i-vector and i_{imp} is the imposter i-vector. The maximization of SINR can be simplified to

$$\min_w w^T \mathbf{R} w, \quad s.t. \quad w^T i_{tar} = 1,$$

where we have defined imposter covariance matrix as

$$\mathbf{R} = \mathrm{E}\{i_{imp} i_{imp}^T\}.$$

The solution to this minimization problem can easily be derived and is given by [14]

$$w_{MVDR} = \frac{\tilde{\mathbf{R}}^{-1}\tilde{i}_{tar}}{\tilde{i}_{tar}^{T}\tilde{\mathbf{R}}^{-1}\tilde{i}_{tar}},$$

where we have used the development data to estimate $\mathbf{R}$ and $\tilde{i}_{tar}$ to represent the desired target i-vector as defined in previous section. The relation between the output of MVDR and that of the baseline system can be formulated as

$$\tilde{d}_{MVDR} = \tilde{d}_{baseline}.\frac{\sqrt{\tilde{i}_{test}^{T}\tilde{\mathbf{R}}^{-1}.\tilde{i}_{test}}}{\sqrt{\tilde{i}_{tar}^{T}\tilde{\mathbf{R}}^{-1}.\tilde{i}_{tar}}},$$

where we can see that MVDR, in contrary to the baseline system, does not take into account the norm of test i-vectors.

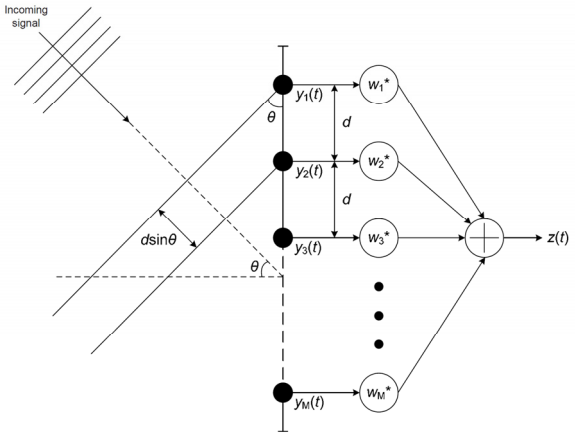

Figure 1: Schematic diagram of a plane wave impinging on a uniform linear array and the beamforming operation.

2.3. Linearly Constrained Minimum Variance (LCMV)

Multiple techniques have been proposed to improve the robustness of MVDR beamformer. Linearly constrained minimum variance beamforming is a popular one. LCMV improves performance by specifying the array response for different angles in the vicinity of the desired angle. This can be formulated as a linear constraints on the filter coefficients. We can use this idea in order to incorporate the uncertainty associated with the target speaker i-vector into the model. Thus, the minimization problem will be formulated as

$$\min_{w} w^{T}\mathbf{R}w, \quad s.t. \quad w^{T}\mathbf{C} = f.$$

where $\mathbf{C} = [i_1, i_2, ..., i_M] \in \mathbb{R}^{N\times M}$ is the constraints' matrix, $f = [1,1,...,1]^{T} \in \mathbb{R}^{M\times 1}$ is the constraints' vector, and $\mathbf{R}$ is the imposter covariance matrix. The solution to this problem is given by [14]

$$w_{LCMV} = \tilde{\mathbf{R}}^{-1}\mathbf{C}\,(\mathbf{C}^{\mathrm{T}}\tilde{\mathbf{R}}^{-1}\mathbf{C})^{-1}f,$$

where we have used the sample estimate for $\mathbf{R}$. This model ensures that the target i-vectors get a score of one by removing their nuisance directions (inter-session variability) while providing a maximum rejection of imposter i-vectors (increase inter-speaker variability). In fact LCMV uses the correlation matrix of the target i-vectors for normalization.

2.4. Covariance Matrix Modification

The imposter covariance matrix $\mathbf{R}$ has a considerable effect on the performance of speaker recognition. Many modern beamforming techniques such as Norm Constraints Capon Beamforming (NCCB) [15], Robust Capon Beamformer (RCB) [16] and Doubly Constrained Robust Capon Beamformer (DCRCB) [15] work using diagonal loading of interference-plus-noise covariance matrix. In order to have a better rejection of imposter i-vectors, we have modified $\tilde{\mathbf{R}}$ by computing the covariance matrix of only those imposter i-vectors that are most similar to the target model. In this way, we first produce an LCMV model of each target speaker using $\tilde{\mathbf{R}}$ computed on all imposter i-vectors. We then select only those imposters getting higher score when compared to the target model and use them to compute a new imposter covariance matrix $\hat{\mathbf{R}}$. Finally, the target model will be updated using this new matrix. We have selected the first 6,000 imposters for our submissions to the NIST 2013-2014 speaker recognition i-vector challenge. Our experiments indicate the effectiveness of the aforementioned approach.

In order to better compensate for the effect of inter-session variability we incorporated a within-class covariance matrix $\mathbf{W}$ computed using groups of similar i-vectors in the development set which are assumed to belong to the same speaker. In order to find these groups, each i-vector is treated as the target model using MVDR technique to find the most similar i-vectors. The within-class covariance matrix is then computed as

$$\mathbf{W} = \frac{1}{N}\sum_{g=1}^{N}\frac{1}{n_g}\sum_{i=1}^{n_g}(w_i^g - \overline{w}^g)(w_i^g - \overline{w}^g)^{\mathrm{T}},$$

where N is the number of i-vectors in the development set, n_g is the number of i-vectors in group g and $\overline{w}^g$ is the mean of i-vectors in group g. The within-class covariance matrix is then added to the modified imposter covariance matrix $\hat{\mathbf{R}}$ to produce the final model

$$\hat{w}_{LCMV} = (\hat{\mathbf{R}} + \mathbf{W})^{-1}\mathbf{C}(\mathbf{C}^{\mathrm{T}}(\hat{\mathbf{R}} + \mathbf{W})^{-1}\mathbf{C})^{-1}f$$

2.5. Score Normalization

Normalization of the scores has shown to be effective in speaker recognition systems [17]. The well-known z-norm and t-norm score normalization work well with cosine similarity scoring. Our experiments indicate that the best improvement can be obtained using the following normalization

$$zscore = \hat{w}_{LCMV}^{T}(i_{test} - \overline{i}_{imp}),$$

where $\overline{i}_{imp}$ is the mean of those imposters having higher scores when compared to the target model. We used the first

300 imposters for our submissions to the NIST i-vector challenge.

3. Experiments

The challenge aims at developing new methods to improve speaker detection performance in the context of conversational telephony speech using i-vector representation of speech segments. The task is to determine whether or not a given speaker is in the test segment.

3.1. Data Description

The 600-dimensional i-vectors used in this challenge have been derived from conversational telephony speech of NIST Speaker Recognitions (SRE's) from 2004 to 2012 using variety of telephone handsets. A set of 36,572 unlabeled i-vectors have been provided for system development. These i-vectors can be used as imposter i-vectors since different speakers take part in the evaluation set. A collection of five i-vectors for each target speaker has been provided for the enrolment purpose and the total target speakers are 1,306 which comprise 6,530 i-vectors. Each target speaker will be tested against a test set containing 9,634 i-vectors. Therefore, the total number of trials will be 12,582,004. These trials are then divided into a progress subset comprising 40% of the total trials and an evaluation subset with the remaining 60% of trials. We will report the scores of different techniques on the progress subset. We will also report the final official score on the evaluation subset for the best proposed model.

3.2. Performance Measure

Performance measure used in this challenge is based on a decision cost function (DCF) representing a weighted combination of false alarm and miss probability at a given threshold. A threshold t which minimizes the following cost function will be used for scoring each submission

$$\mathrm{DCF}(t) = P_{miss}(t) + 100 \times P_{fa}(t).$$

3.3. Results

In this section we present scores for different techniques as evaluated by NIST i-vector challenge website on progress subset as well as the final official score for the best technique on the evaluation subset. Table 1 summarizes the results obtained using different techniques. The best score, 0.280 on the progress subset and 0.270 on the evaluation subset, is achieved by the modified LCMV model with score normalization which indicates an improvement of 27% relative to the baseline system. The results indicate that MVDR model outperforms the baseline system. This is due to the fact that MVDR does not take into account the norm of the test i-vector. This is in contrary to the classical cosine scoring which is symmetric with respect to the target and test. The better performance of LCMV compared to MVDR resides on the ability of LCMV to incorporate the uncertainty associated with the target i-vector through removing the intra-speaker variability. The results given in the table show that the covariance matrix modification (labeled CMM in the table) of the LCMV model could lead to a relative improvement of about 12%. This is mainly due to the fact that the new model can better reject imposters that are most similar to the target model. As expected and indicated in the table, the normalization of scores can improve the performance of speaker recognition systems.

	Score	
	progress subset	**evaluation subset**
Baseline	0.386	-
MVDR	0.356	-
LCMV	0.343	-
LCMV+Znorm	0.321	-
LCMV+CMM	0.303	-
LCMV+CMM+Znorm	0.295	-
LCMV+CMM+W+Znorm	0.280	0.270

Table 1: Comparison of results for different techniques on the progress subset of the challenge.

The incorporation of the within-class covariance matrix (W) computed using groups of similar i-vectors in the development subset indicates a significant improvement. The reason for this effect is that the model not only rejects imposter i-vectors but also decrease the intra-speaker variability through weakening the nuisance directions.

4. Conclusions

The state-of-the-art i-vector representation of speech segments has attracted interests from other communities. With a fixed length and low dimensional vector representation of speakers, it is now possible to apply machine learning and pattern analysis techniques. In this paper we used an idea from the theory of adaptive filtering namely Linearly Constrained Minimum Variance (LCMV) beamforming and applied it to i-vector based speaker recognition. We have shown that this technique is effective in modeling a target speaker based on a set of i-vectors of that speaker and a collection of unlabeled imposter i-vectors. We also proposed a modification of imposter covariance matrix and score normalization in order to improve the speaker recognition performance. We believe that there are more effective techniques from this field that can be applied to i-vector based speaker recognition.

5. References

[1] Dehak N., Kenny P., Dehak R., Dumouchel P., and Ouellet P., "Front-end factor analysis for speaker verification", *IEEE Transaction on Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[2] Kenny P., "Bayesian speaker verification with heavy tailed priors", *Proc. Odyssey 2010: The speaker and Language Recognition Workshop*, Brno, Czech Rebublic, June 2010.

[3] Senoussaoui M., Kenny P., Brummer N., and Dumouchel P., "Mixture of PLDA models in i-vector

space for gender independent speaker recognition", *Proc. Interspeech 2011*, Florence, Italy, Aug. 2011.

[4] Burget L., Plchot O., Cumani S., Glembek O., Matejka P., and Brummer N., "Discriminatively trained probabilistic linear discriminant analysis for speaker verification", *Proceedings ICASSP*, pp. 4832–4835, 2011.

[5] Dehak N., Dehak R., Glass J., Reynolds D., and Kenny P., "Cosine similarity scoring without score normalization techniques", *Proc. Odyssey 2010: The speaker and Language Recognition Workshop*, Brno, Czech Rebublic, June 2010.

[6] Kenny, P., Stafylakis, T., Ouellet, P., Alam, J., and Dumouchel, P., "PLDA for speaker verification with utterances of arbitrary duration", *Proc. Int. Conf. Acoust., Speech, Signal Process,* 2013.

[7] Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., and Dumouchel, P., "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification", *Proc. Interspeech 2009*, Vol. 9, pp. 1559-1562, 2009.

[8] The 2013-2014 Speaker Recognition i-vector Machine Learning Challenge available at https://ivectorchallenge.nist.gov.

[9] Van Trees H. L., "Optimum Array Processing," Wiley, New York, NY, USA, 2002.

[10] Capon J., "High-resolution frequency-wavenumber spectrum analysis", *Proc. IEEE,* vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[11] Rubsamen, M., and Gershman, A. B., "Robust adaptive beamforming using multidimensional covariance fitting", *Signal Processing, IEEE Transactions on*, *60*(2), pp. 740-753, 2012.

[12] Hatch A., Kajarekar S., and Stolcke A., "Within-Class Covariance Normalization for SVM-Based Speaker Recognition", *International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 2006.

[13] Campbell W. M., Sturim D. E., Reynolds D. A., and Solomonoff A., "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, vol. 1, pp. 97–100, 2006.

[14] Manolakis D. G., Ingle V. K., and Kogon S. M., *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, Boston: McGraw-Hill, 2000.

[15] Li, J., Stoica, P., and Wang, Z., "Doubly constrained robust Capon beamformer", *IEEE Transactions on Signal Processing*, *52*(9), pp. 2407-2423, 2005.

[16] Li J., Stoica P., and Wang Z., "On robust Capon beamforming and diagonal loading", *IEEE Transaction on Signal Processing*, vol. 51, pp. 1702–1715, July 2003.

[17] Kenny P., Ouellet P., Dehak N., Gupta V., and Dumouchel P., "A study of interspeaker variability in speaker verification", *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(5), pp. 980-988, 2008.