INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Mitigating Effects of Recording Condition Mismatch in Speaker Recognition Using Partial Least Squares

Jeremiah J. Remus, Jenniffer M. Estrada, Stephanie A.C. Schuckers

Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY, USA
jremus@clarkson.edu, estradjm@clarkson.edu, sschucke@clarkson.edu

Abstract

Speaker recognition systems have been shown to work well when recordings are collected in conditions with relatively limited mismatch. Thus, a significant focus of the current research is techniques for robust system performance when greater variability is present. This study considers a diverse data set with recordings collected in multiple different rooms with different types of microphones. A technique recently introduced to the speaker recognition community, called partial least squares (PLS), is considered for decomposing the features and mitigating the degradation in performance due to room and/or microphone mismatch. Results of this study suggest that PLS decomposition can provide substantial improvements in performance in the presence of mismatched recording conditions. The outcomes of this study provide further validation for the partial least squares decomposition and encourage further consideration of PLS for reducing session and environment variability in speaker recognition tasks.

Index Terms: speaker recognition, partial least squares, subspace decomposition

1. Introduction

Speaker recognition techniques have long been capable of verifying a speaker's identity when two speech recordings come from similar or identical environments. A significant focus in recent years has been the development of methods for extending the performance of speaker recognition systems to scenarios where greater variation between the enrollment and test data may exist. State of the art automatic speaker recognition systems perform relatively well on channel mismatch, but other environmental factors including room variability may still pose a significant challenge. The MultiRoom8 data set that was used in this study includes recordings from 51 speakers in four different room environments using multiple microphones. The MultiRoom data set provides a significant level of diversity in room effects and noise type, confounded by the differences in microphone type as well.

The most common approach for managing these sources of non-speaker variation is adoption of a model that assumes the captured recording is a superposition of two elements: the speaker-specific features useful for speaker recognition and an additive non-speaker component. Several investigators have proposed linear subspace modeling techniques that can be used to estimate and factor out the non-speaker component in recorded audio. Several techniques have been considered recently in speaker recognition, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Joint Factor Analysis (JFA) and i-vectors. More recently, partial least squares (PLS) has been successfully applied in the speaker recognition community [1, 2]. This approach has a particular appeal for learning a supervised projection from a high-dimensional to low-dimensional subspace that represents only the differences between individual speakers.

The significance of this study is an assessment of the effectiveness of partial least squares as a subspace projection technique to improve performance of a speaker recognition system in the presence of environment-based variability. The remainder of this paper is organized as follows. Section 2 describes the data set and experiment approach including partial least squares and the classification methods. Section 3 contains results for several speaker recognition experiments, with conclusions presented in Section 4.

2. Experiment method and approach

This study utilized the MultiRoom8 data set, made available for this project by the sponsor. The MultiRoom8 data set consists of multi-session audio recordings with collection conditions designed to include a number of distinct environmental scenarios (e.g. noise and room acoustics). A total of 424 audio recordings were used in this study, each approximately three minutes in duration. These three-minute recordings were divided into two recordings of equal length to allow training and testing in the same condition, since the provided data included only a single session in each recording condition. The available data contained ten conditions (i.e. room-microphone combination) for analysis of the variability introduced by different room and microphone types. The audio files used in this study were collected from a group of 51 speakers, with 35 speakers common to all ten of the conditions. The conditions in MultiRoom8 include four distinct rooms of various sizes: a conference room (unknown size), small (206 ft^2, 19 m^2), medium (430 ft^2, 40 m^2), and large (2013 ft^2, 187 m^2). There were five microphone/recording setups available, although not all were available in each environment. Directional microphones were placed at two distances: 3 feet facing towards the speaker, and 5 feet facing away from the speaker. Omnidirectional microphones were placed at three distances: close (1 foot), mid-distance (1/3 across the room) and far (2/3 of distance across room).

2.1. Baseline GMM-UBM and supervector construction

The baseline classification technique utilized GMM supervectors constructed by concatenating mean vectors from each GMM component of a GMM-UBM model. The audio processing and feature extraction was performed using S-Pro 5.0 and the ALIZE/SpkDet open-source software platform for speech and speaker recognition [3]. The 16-element MFCCs and first-order deltas, resulting in a 32-element feature vector, were generated for each 20 millisecond frame (frames overlapped by 50%). After normalization and silence removal, the MFCC coefficients were used to estimate a universal background model (UBM) and Gaussian mixture models (GMM). The UBM was generated using 100 separate speaker files containing more than five hours of speech. A 750-component diagonal covariance GMM was adapted from the UBM for each speaker recording. The GMM supervector is generated by concatenating the means of each Gaussian component from the GMM-UBM for a corresponding .wav file. In this study, there were 32 cepstral coefficients with 750 GMM components, resulting in a GMM supervector with length 24,000. The GMM supervectors were the baseline feature set for the experiments conducted in this study.

2.2. Partial least squares decomposition

The GMM supervectors have been shown in previous studies to provide sufficient information for successful speaker discrimination; however, they are extremely high-dimensional, which introduces potential concerns about computational complexity and overfitting during the learning stages in the pattern recognition techniques. More significantly, these supervectors contain non-speaker artifacts introduced by the channel, environment, and session-to-session variability. These factors motivate the use of subspace decomposition techniques to find a lower-dimensional representation of the GMM supervector that represents only the speaker-specific attributes and will be robust to variability introduced by changes in channel, environment, and session. In the high-dimensional supervector space, several speakers may be indistinguishable due to non-speaker sources of variability. The ideal subspace decomposition would project the supervectors into a lower-dimensional space where all recordings from a single speaker cluster together, and different speakers are separable.

Partial least squares performs a linear projection to a lower-dimensional subspace, which allows use on high-dimensional data sets without running into the “large p, small n” problem (i.e. many feature dimensions and few data points). One advantage of partial least squares over other linear subspaces projection methods (such as principal component analysis) is that partial least squares performs a supervised decomposition, making use of the labels for the development data. Therefore, the resulting lower-dimensional subspace is more likely to maintain separability between classes, by using a criterion that seeks linear projections *w* and *q* that maximize the covariance between the independent and dependent variables *X* and *Y*, respectively, in the lower-dimensional projection space.

$$\max_{\|w\|=1,\|q\|=1} \operatorname{cov}\left(Xw,Yq\right) \qquad (1)$$

This contrasts with PCA, which maximizes the variance of the data under the constraint of a unit-norm weight vector, ignoring any available class labels for the training data.

The subspace projection from the high-dimensional supervector features to a lower-dimensional space requires a set of parameters which must be estimated from some data. To provide a robust experiment result, the GMM supervector decompositions were learned using a separate development data set. Given the available data, there are several options for how the development data set could be constructed: the speakers may be either the same or different from those in the training and test data sets, and the room-microphone combinations may either be the same and/or different from those in the training and test data sets. Two development data sets were considered in this study, and are described in Table 1 for the example where Condition A is used for training and Condition B for testing (there are ten conditions, A through J, in total). Note: the actual development data set is adjusted as necessary within the iterations of the cross-condition training and testing as all ten available conditions are eventually used for both training and testing. Development data set #1 included all available speaker data in conditions excluding the train and test conditions. Thus, the development data will contain recordings of the speakers in the test set. Development data set #2 reflects the less optimistic, but more general, scenario where the development data set contains none of the speakers in the test set.

2.3. Classification techniques

Two classification techniques were considered in this study. The first technique, the nearest neighbor classifier, assigns labels to feature vectors in the test set based on distances calculated between the unlabeled test sample and all of the available labeled training data. The label of the nearest training sample (i.e. nearest neighbor) is assigned to the test sample. This classification rule is supported by theoretical results that relate it to nonparametric modeling of probability distribution functions and the likelihood ratio test [4]. For the present study, the negated value of the correlation coefficient was used as the measure of distance between two GMM supervectors.

The second classification technique was the support vector machine, or SVM. This technique was applied to the raw GMM supervector to establish baseline performance as the GSL-SVM method, as well as to the PLS-decomposed supervectors. In this research effort, the LIB-SVM [5] implementation was used with two kernel configurations: a linear kernel when operating on the GMM supervectors as features and a radial-basis function (RBF) kernel with unit variance when operating in a low-dimensional subspace generated by PLS projection of the GMM supervectors. The SVM is natively configured for binary classification (where only two classes of data are present). To extend the SVM to the current application where many speakers are present, a set of (N(N+1))/2 SVMs was constructed, with each SVM

Table 1: *Development data sets for estimating PLS projections. Note that different subjects are present in each development data set.*

Training Data Set	Condition A, Speakers 1 to 10	Condition A, Speakers 1 to 10
Testing Data	Condition B, Speakers 1 to 10	Condition B, Speakers 1 to 10
Development Data Set	Conditions C to J, Subjects 1 to 51	Conditions C to J, Subjects 11 to 51
Development Data Set includes:	*All speakers, different conditions (i.e. excludes train/test)*	*Different speakers, different conditions (exclude train/test)*

discriminating between a pair of the N total speakers in the data set. The (N(N+1))/2 classifiers then vote on the final classification of a test sample.

2.4. Experiment setup

Results for the MultiRoom8 data were generated using five approaches. The baseline technique is the support vector machine applied to supervectors constructed from the GMM-UBM. Alternatively, rather than utilizing the raw, high-dimensional GMM supervectors as inputs to pattern classification algorithms, the GMM supervectors were decomposed using partial least squares. The PLS projections were learned using each of the two development data sets described in Table 1. After reducing the dimensionality of the feature vectors using PLS, the SVM and nearest neighbor classifier using a correlation-based distance metric were each used for speaker identification.

To partition the data into separate training, testing, and development data sets, only the first ten speakers (organized by speaker ID) were used for training and testing. This preserved the higher-numbered speakers for development data set #2, which contains a different set of speakers than are present in training and testing. Another common element of the experiment setup was the use of cross-condition testing. Each of the ten room/microphone conditions (A through J) was used both for training and testing against all of the other ten conditions. Thus, results in the form of 100 detection-error trade-off (DET) curves can be generated and equal-error rates (EER) can be calculated. These 10x10 matrices of equal-error rates were the common basis in this study for comparison of the five speaker recognition techniques.

3. Results

The common approach for evaluating each of the five techniques included a set of 100 experiment configurations, systematically using each of the available ten recording conditions for training and/or testing. The equal-error rates (EER) for all 100 experiment configurations can be calculated, and a distribution can be generated. Table 2 lists statistics (20th percentile, median, and 80th percentile) of the distribution of EERs for each of the five methods: baseline SVM applied to the raw GMM supervector (GSL-SVM), the SVM applied to PLS-decomposed supervectors using two different development data sets, and the distance-based neighborhood classifier applied to PLS-decomposed supervectors using the same two development data sets. The statistics shown in Table 2 indicate a consistent improvement in overall performance through the use of PLS supervector decomposition. All four approaches that use the PLS-decomposition outperform the baseline GSL-SVM. Table 2 also suggests improved performance across the distribution of EERs when using development data set #1, which reflects the more optimistic scenario where recordings (from different recording conditions) for speakers in the test set are available for inclusion in the development set.

An additional observation from Table 2 is that there appears to be a greater range of scores when using the SVM versus nearest neighbor on the PLS-decomposed supervectors. A follow-up to the results presented in Table 2 examined the consistency of results across the four methods that included PLS decomposition in an effort to identify advantages in certain train/test conditions when using a particular classifier or development data set. Correlation coefficients were calculated between the length-100 vectors of EERs for each of the four methods. Higher correlations are indicative of more consistent ordering of conditions in the distribution of EERs across the 100 experiment configurations. The correlation coefficients indicate that the order of conditions from “easiest” to “most difficult”, in terms of EER, was relatively consistent when the same development data set was used. The EERs for the SVM and neighborhood classifier has correlation coefficients of 0.87 and 0.77 for development data set #1 and #2, respectively. Thus, the differences in the classifiers may have resulted in different levels of performance, but did not significantly change the ordering of conditions when sorted by difficulty.

Table 2: *Statistics from the distribution of equal-error rates for each of the five classification approaches when evaluated on the entire suite of 100 cross-condition experiment configurations.*

	20th percentile	Median	80th percentile
GSL-SVM	11.7%	23.3%	28.6%
PLS SVM, *Dev. Set #1*	6.8%	10.2%	18.0%
PLS Distance, *Dev. Set #1*	5.1%	8.2%	13.4%
PLS SVM, *Dev. Set #2*	4.7%	14.1%	25.0%
PLS Distance, *Dev. Set #2*	5.0%	12.5%	22.7%

Comparing the effect of the different development data sets on performance, the SVM and neighborhood classifier had correlation coefficients of 0.74 and 0.56, respectively. Thus, the most variation in ranking conditions by performance occurs when comparing the neighborhood classifier performance across the two different development data sets. For the neighborhood classifier, the difference in performance for individual conditions when comparing development data sets #1 and #2 appears largely arbitrary. The most consistent attribute is that experiment configurations involving the Large, Far Omni recording condition perform poorest with either development data set. Another exception observed in the results is that, while development data set #1 typically provided improved performance, the experiment that paired “Large, Far Omni” and “Large, Dir@3ft” for train/test was substantially better (9.4%) using development data set #2.

An additional avenue of investigation examined the effect of PLS subspace dimensionality on performance, and potential improvement in EER, for the speaker recognition system. The results reported in Table 2 were calculated with PLS projections into a 25-dimensional subspace. In Figure 1, results are shown for the PLS neighborhood classifier as a function of the number of PLS subspace dimensions. The boxplots in each subplot represent the distribution of EERs within a single cross-condition EER matrix. Center lines in each box are the median of the distribution, upper and lower edges of the box identify the 75th and 25th percentile (N = 100), and the hash symbols indicate outliers that are more than 1.5 standard deviations beyond the edge of the box. Cross-condition EER matrices were generated as the number of PLS subspace dimensions was varied from D = 5 to D = 30, with the upper limit imposed by the amount of

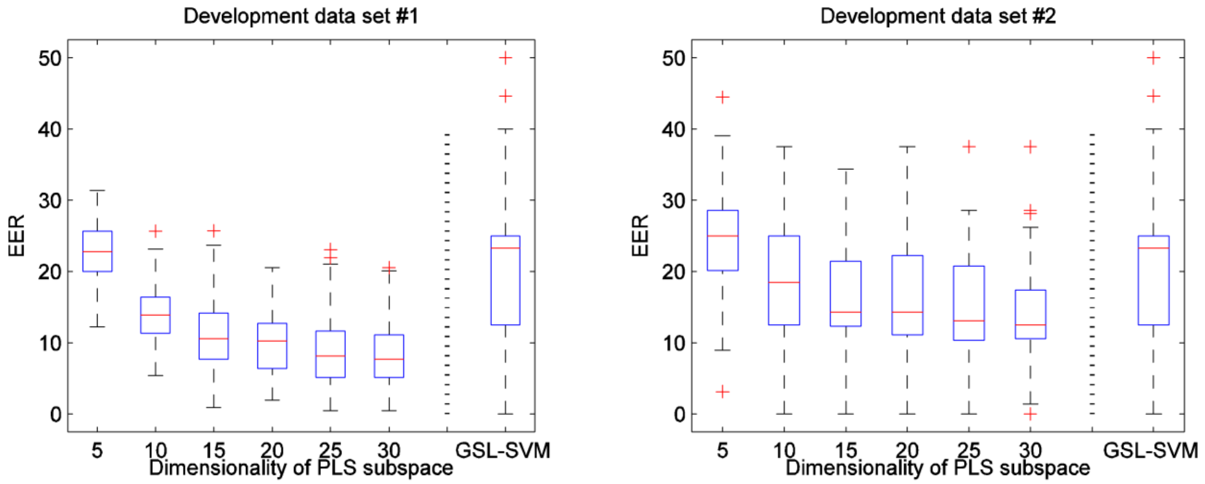

Figure 1: *Effect of PLS subspace dimensionality on the distribution of EERs generated using the neighborhood classifier for all 100 MultiRoom8 cross-condition experiment configurations. Also shown are the distributions of EERs using the baseline GSL-SVM.*

available data. The right side of each subplot also shows distributions of EERs for the baseline GLS-SVM.

Figure 1 reveals a clear and consistent trend between the median EER and the dimensionality of the PLS subspace. For all classifiers, the median value of the EER distribution using a PLS subspace with at least 10 dimensions is lower than the value attained using the GLS-SVM. The study of this trend was limited by the size of the available development data set. Further investigation with a larger development data set might provide a clearer understanding of the relationship between characteristics of the data set and improvements observed with increasing dimensionality of the PLS subspace.

4. Conclusions

This paper describes a study of subspace decomposition techniques to improve performance of a speaker recognition system in data conditions that contain both room and microphone variability. Consistent with recent trends in the research community, the primary focus was on dimensionality reduction techniques applied to the GMM supervector, which attempt to find a lower-dimensional subspace that only represents individual speakers and removes the variability introduced by the room and environment. The results of this study indicated that the partial least squares (PLS) subspace consistently provided an improved feature set for discrimination between speakers. A combination of partial least squares decomposition of the GMM supervector and nearest neighbor classification using a correlation-based distance metric provided the overall best performance for 100 different experiment configurations created by using each of the ten conditions in the MultiRoom8 data. The combination of these techniques provided significant improvements in equal-error rate when compared to the SVM applied to the GMM supervector, and consistently outperformed the SVM applied to the PLS-decomposed supervector.

The results of this study provide further evidence to support the validity of partial least squares decomposition for mitigating certain sources of variability in speaker recognition tasks. Previous studies have also shown that partial least squares decomposition provides a lower-dimensional subspace that is appropriate for discriminating between speakers. The outcomes of this research effort encourage further consideration of supervised subspace decomposition techniques (e.g. partial least squares) to address scenarios where speaker recognition must be performed in the presence of significant room and microphone variability.

5. Acknowledgements

The authors would like to thank Anthony Larcher of the Institute for Infocomm Research (I2R) in Singapore for his help with the ALIZE toolbox. This material is based on research sponsored by Air Force Research Laboratory under agreement number FA8750-10-1-0231. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

6. References

[1] B. V. Srinivasan, D. Garcia-Romero, D. N. Zotkin and R. Duraiswami, "Kernel partial least squares for speaker recognition," in *INTERSPEECH,* 2011, pp. in press.

[2] B. V. Srinivasan, D. N. Zotkin and R. Duraiswami, "A partial least squares framework for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on,* 2011, pp. 5276-5279.

[3] J. F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve and J. Mason, "ALIZE/SpkDet: A state-of-the-art open source software for speaker recognition," in *Proceedings of Odyssey,* 2008, .

[4] J. J. Remus, K. D. Morton, P. A. Torrione, S. L. Tantum and L. M. Collins, "Comparison of a distance-based likelihood ratio test and k-nearest neighbor classification methods," in *IEEE Workshop on Machine Learning for Signal Processing,* 2008, pp. 362-367.

[5] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, pp. 27:1-27:7, 2011.