Speaker Discrimination Using Several Classifiers and a Relativistic Speaker Characterization

Siham Ouamour, Zohra Hamadache, and Halim Sayoud(✉)

USTHB University, Algiers, Algeria
{siham.ouamour,halim.sayoud}@uni.de,
zohra.hamadache@yahoo.fr

Abstract. Automatic Speaker Discrimination consists in checking whether two speech signals belong to the same speaker or not. It is often difficult to decide what could be the best classifier to use in some specific circumstances. That is why, we implemented nine different classifiers, namely: Linear Discriminant Analysis, Adaboost, Support Vector Machines, Multi-Layer Perceptron, Linear Regression, Generalized Linear Model, Self Organizing Map, Second Order Statistical Measures and Gaussian Mixture Models. Moreover, a special feature reduction was proposed, which we called Relativistic Speaker Characteristic (*RSC*). On the other hand we further intensified the feature reduction by adding a second step of feature transformation using a Principal Component Analysis (*PCA*). Experiments of speaker discrimination are conducted on Hub4 Broadcast-News. Results show that the best classifier is the SVM and that the proposed feature reduction association (*RSC-PCA*) is extremely efficient in automatic speaker discrimination.

Keywords: Speaker discrimination · Speaker verification · Relativistic speaker characteristic · PCA reduction · Classification models

1 Introduction

Speaker discrimination consists in checking whether two different pronunciations (speech signals) are uttered by the same speaker or by two different speakers [1]. This research domain has several applications such as automatic speaker verification [2], speech segmentation [3] or speaker based clustering [4]. All these tasks can be performed either by generative classifiers or by discriminative classifiers.

However, existing approaches are not robust enough in noisy environment or in telephonic speech. Any new model must therefore improve the reliability of existing discriminative systems, without altering their architectures.

To address the above issue, we implemented 9 different classifiers and applied the PCA with these different classifiers. Furthermore, a new relativistic characteristic is proposed: we called it "Relativistic Speaker Characteristic" [5]. Basically, the introduction of the relative notion in speaker modelization allows getting a flexible relative speaker template, more suitable for the task of speaker discrimination in difficult environments. Moreover, to further intensify the feature reduction, a PCA reduction is

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 203–212, 2016.
DOI: 10.1007/978-3-319-33618-3_21

applied to reduce again the RSC feature. For that purpose, several speaker discrimination experiments are conducted on a subset of Broadcast-News dataset.

The overall structure of this paper is organized as follows: In Sect. 2, we describe some related works and explain the motivation of this investigation. Section 3 defines the nine used classifiers. Section 4 describes the RSC notion employed for the task of speaker discrimination and feature reduction. Experiments of speaker discrimination are presented in Sect. 5 and finally a general conclusion is presented at the end of this manuscript.

2 State of the Art in Feature Reduction Based Speaker Recognition

Speaker discrimination is the ability to check whether two utterances come from the same speaker or from different speakers, but in a broader sense, speaker recognition is the task of recognizing the true speaker of a given speech signal. Hence, in this section, we will shortly quote some recent works of speaker recognition using feature reduction (*such as PCA reduction*).

In 2008, Li *et al.* [6] proposed a novel hierarchical speaker verification method based on PCA and Kernel Fisher Discriminant (*KFD*) classifier. Later on, Zhao *et al.* [7] presented a new method which takes full advantage of both vector quantization and PCA. Also in 2009, Jayakurnar *et al.* [8] presented an effective and robust method for speaker identification based on discrete stationary wavelet transform (*DSWT*) and principal component analysis techniques. Ingeniously, Zhou *et al.* [9] proposed a method to reduce feature dimension based on Canonical Correlation Analysis (*CCA*) and PCA. In the same period, Mehra *et al.* (Mehra, 2010) presented a detailed comparative analysis for speaker identification by using lip features, PCA, and neural network classifiers: it was a multimodal feature combination. Then, Xiao-chun *et al.* [10] proposed a text-independent (*TI*) speaker identification method that suppresses the phonetic information by a subspace method: a Probabilistic Principle Component Analysis (*PPCA*) is utilized to construct these subspaces. Recently, Jing *et al.* (Jing, 2014) introduced a new method of extracting mixed characteristic parameters using PCA. This speaker recognition technique is based on the performance of the PCA on the Linear Prediction Cepstral Coefficients (*LPCC*) and Mel Frequency Cepstral Coefficients (*MFCC*). All of these works (*or at least most of them*) used the principal component analysis to reduce the feature space dimensionality without altering the recognition performances.

In this investigation, we not only propose a completely different feature reduction technique, but we also combine it with PCA reduction to further enhance both the memory size and recognition precision. Moreover, we evaluate the RSC-PCA efficiency in real environment (*Broadcast News*) and with 9 different classifiers.

3 Description of the Classifiers and Classification Process

The choice of the optimal classifier is crucial before any application of pattern recognition that is why we have decided to implement 9 classifiers and evaluate them in the same experimental conditions.

The different classification methods are described in the following sub-sections. However, since we are limited by the pages number of the article, we will only give the general definition of the different classifiers; the details could be found in the cited references.

3.1 LDA: Linear Discriminant Analysis

Linear discriminant analysis (*LDA*) is a method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events.

Consider a set of observations $\overrightarrow{x}$ (also called features, attributes, variables or measurements) for each sample of an object or event with known class y. This set of samples is called the training set. The classification problem is then to find a good predictor for the class y of any sample of the same distribution (not necessarily from the training set) given only an observation $\overrightarrow{x}$.

LDA approaches the problem by assuming that the conditional probability density functions $p(\vec{x}|y = 0)$ and $p(\vec{x}|y = 1)$ are both normally distributed with mean and covariance parameters $(\vec{\mu}_0, \Sigma_0)$and $(\vec{\mu}_1, \Sigma_1)$, respectively. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the log of the likelihood ratios is below a threshold T [11].

3.2 AdaBoost: Adaptive Boosting

AdaBoost, short for "Adaptive Boosting", is a machine learning meta-algorithm. It can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms (*'weak learners'*) is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

AdaBoost refers to a particular method of training a boosted classifier. A boost classifier is a classifier in the form

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{1}$$

where each f_t is a weak learner that takes an object x as input and returns a real valued result indicating the class of the object. The sign of the weak learner output identifies the predicted object class and the absolute value gives the confidence in that classification. Similarly, the T-layer classifier will be positive if the sample is believed to be in the positive class and negative otherwise [12].

3.3 SVM: Support Vector Machines

In machine learning, support vector machines (*SVMs*) are supervised learning models with associated learning algorithms that analyze data and recognize patterns. They are used for classification and regression analysis.

The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.

Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [13].

3.4 MLP: Multi Layer Perceptron

MLP is a feed-forward neural network classifier that uses the errors of the output to train the neural network: it is the "training step" [14].

MLP is organized in layers, one input layer of distribution points, one or more hidden layers of artificial neurons (*nodes*) and one output layer of artificial neurons (*nodes*).

Each node, in a layer, is connected to all other nodes in the next layer and each connection has a weight (*which can be zero*). MLPs are considered as universal approximators and are widely used in supervised machine learning classification. The MLP can use different back-propagation schemes to ensure the training of the classifier.

3.5 LR: Linear Regression

Linear regression is the oldest and most widely used predictive model. The method of minimizing the sum of the squared errors to fit a straight line to a set of data points was published by Legendre in 1805 and by Gauss in 1809. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norms (*as with least absolute deviations regression*), or by minimizing a penalized version of the least squares loss function as in ridge regression [15, 16].

In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models.

Usually, the predictor variable is denoted by the variable X and the criterion variable is denoted by the variable y. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median of the conditional distribution of y given X is expressed as a linear function of X.

3.6 GLM: Generalized Linear Model

In statistics, the generalized linear model (*GLM*) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value [17].

3.7 SOM: Self Organizing Map

A self-organizing map (*SOM*) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space.

This makes SOMs useful for visualizing low-dimensional views of high-dimensional data. The model was first described as an artificial neural network by Kohonen, and is sometimes called a Kohonen map. A Self-organizing Map is a data visualization technique developed by Kohonen in the early 1980's [18, 19].

Like most artificial neural networks, SOMs operate in two modes: training and mapping. The training builds the map using input examples, while the mapping automatically classifies the input vector.

3.8 SOSM: Second Order Statistical Measure

The proposed method uses mono-gaussian models based on the second order statistics, and provides some similarity measures able to make a comparison between two speakers (*speech segments*) according to a specific threshold. We recall bellow the most important properties of this approach [20].

Let $\{x_t\}_{1 \le t \le M}$ be a sequence of M vectors resulting from the P-dimensional acoustic analysis of a speech signal uttered by speaker $\boldsymbol{x}$. These vectors are summarized by the mean vector $\overline{x}$ and the covariance matrix X.

Similarly, for a speech signal uttered by speaker $\boldsymbol{y}$, a sequence of N vectors $\{y_t\}_{1 \le t \le N}$ can be extracted. These vectors are summarized by the mean vector $\overline{y}$ and the covariance matrix Y.

The Gaussian likelihood based measure μ_G is defined by:

$$\mu_G(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log\left(\frac{\det(Y)}{\det(X)}\right) + tr(YX^{-1}) + (\overline{y}-\overline{x})^T X^{-1}(\overline{y}-\overline{x})\right] - 1 \qquad (2)$$

we have:

$$\underset{\boldsymbol{x}}{Arg\max}\, \overline{G}\boldsymbol{x}(y_1^N) = \underset{\boldsymbol{x}}{Arg\min}\, \mu_G(\boldsymbol{x},\boldsymbol{y}) \qquad (3)$$

where "det" represents the determinant and "tr" represents the trace of the matrix.

The μ_g mesasure is widely used in speech analysis for the task of speaker recognition.

3.9 GMM: Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-*Maximization* algorithm or *Maximum A Posteriori* estimation from a well-trained prior model.

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(\mathrm{X}|\lambda) = \sum_{i=1}^{M} w_i \, g(\mathrm{X}|\mu_i, \Sigma_i), \tag{4}$$

where x is a D-dimensional continuous-valued data vector, wi, i = 1,…, M, are the mixture weights, and g(x|μi,_i), i = 1,…, M, are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g(X|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)' \, \Sigma_i^{-1}(X-\mu_i)\right\}, \tag{5}$$

with mean vector μi and covariance matrix $\sum_i$. The mixture weights satisfy the constraint that PM i = 1 wi = 1.

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. Herein, we wish to estimate the parameters of the GMM, which in some sense best matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM [21]. By far the most popular and well-established method is the *maximum likelihood* estimation.

In general GMMs are considered as the state-of-the-art classifier in speaker recognition.

3.10 PCA Features Reduction

PCA provides an interesting way to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it [22]. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the

data comes to lie on the first coordinate (i.e. the first principal component), the second greatest variance on the second coordinate, and so on.

In our investigation, PCA has been intensively used to further reduce the dimensionality of the features of the relative characteristic RSC. This fact has three advantages: reducing the processing time, avoiding the need of high training data and making the features more pertinent with regards to the discrimination task.

4 Relativity in Speaker Discrimination: Notion of RSC

We propose a new relative characteristic called RSC derived from the Mel Frequency Spectral Coefficients MFSC and which is used for the task of speaker discrimination.

This relativity approach is proposed in order to reduce the features dimension, optimize the learning classifiers training, without modifying the classifier architecture and without changing the input features either (Fig. 1).

The previous formula 1 gives us a similarity measure between a speech signal uttered by a speaker $\boldsymbol{y}$ and the reference model of the speaker $\boldsymbol{x}$.

We derive from this formula the following one:

$$\psi^*(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}[-\log(\det(\Re) + tr(\Re))] - 1 \tag{6}$$

We have called the $\Re$ ratio: **Relative Speaker Characteristic** (RSC)

$$\Re = RSC(\boldsymbol{x},\boldsymbol{y}) = \frac{\mathrm{Y}}{\mathrm{X}} = \mathrm{Y}^*\mathrm{X}^{-1} \tag{7}$$

Hence, $\psi^*(\boldsymbol{x},\boldsymbol{y})$ appears to be a function of the RSC.

5 Experiments of Speaker Discrimination on Hub4 Broadcast News

Several speech segments are extracted from "Hub4 Broadcast-News 96" dataset, containing some recordings from "CNN early edition". They are composed of clean speech, music, telephonic calls, noises, etc. The sampling frequency is 16 kHz and the speech signals are extracted and arranged into segments of about 4 s or 3 s. The corpus contains 14 different speakers (*most of them journalists, speaking about the news*) organized into 259 speaker combinations for the training and 195 speaker combinations for the testing.

The training dataset is composed of speech segments of 4 s each, whereas the testing set consists of speech segments of 3 s each. The choice of 3 s is due to the fact that previous works showed that the minimal speech duration for good speaker recognition is 3 s.

The general experimental protocol is described as follows:

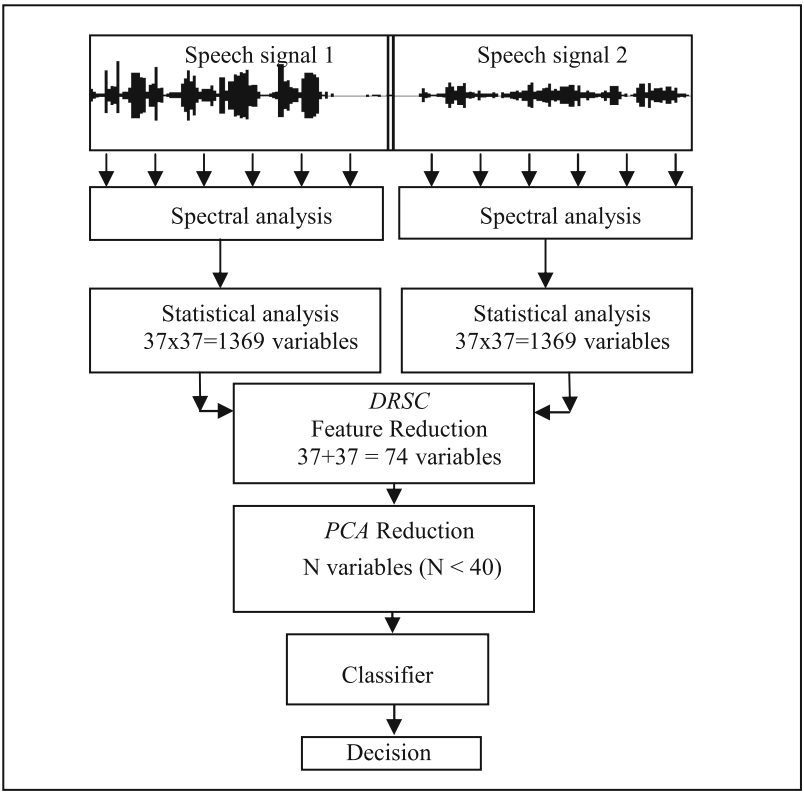

Fig. 1. The general experimental classification protocol

And the different results obtained during these experiments are summarized in the following table (*Table 1*).

Table 1. Scores of good speaker discrimination in %

Method	without PCA Reduction	after PCA Reduction
1. SOM map	76,92	80,51
2. SOSM-muGc	82,56	82,56 *(without PCA)*
3. Adaboost	68,21	84,62
4. MLP	77,95	85,64
5. GMM	*failure*	87,18
6. LDA	75,38	89,23
7. Lin Regress	75,38	89,74
8. GLM Regress	75,90	90,26
9. SVM	83,08	91,28

The first observation, we can make, is that the PCA reduction allowed the improvement of the accuracy for almost all the classifiers and the reduction of the features size (*and then the computation time too*).

The second observation is related to the comparative evaluation of the 9 classifiers: we can see that the 4 classifiers namely: LDA, Linear regression, GLM regression and SVM are the most accurate classifiers with a score about 90 % of good discrimination. The SVM is the best one by providing a score of 91.28 % of good classification. The 3 classifiers namely: Adaboost, MLP and GMM are relatively less accurate with a score of about 85 %. Although those 3 classifiers are known to be quite robust, however, the lack of training data made them not very efficient. Finally, for the 2 remaining classifiers, namely: SOM and SOSM, the performances noticed during the discrimination experiments, show that those classifiers are not suitable for the task of speaker discrimination, since the score of good classification is about only 80 %. However, one must note that the SOSM approach remains quite interesting since it does not require any training step (distance measure). Moreover, in the present experiments, no PCA reduction was applied for the SOSM (*technically not possible*). That is, if we observe the scores without PCA, we do notice that the SVM and SOSM provide the best performances (about 83 %).

6 Conclusion

In this paper, we dealt with the problem of speaker discrimination on Broadcast News speech segments. For that purpose, we implemented 9 different classifiers and proposed a new reduced pertinent characteristic called RSC, which has been successfully employed in association with the PCA (Principal Component Analysis).

The association RSC-PCA was applied to reduce the size of the features and speed up the training/testing process. This investigation has shown that this association did not only reduce the features dimensionality but it also further improved the classification accuracy.

The experiments of speaker discrimination were conducted on a subset of HUB-4 Broadcast News. Results have shown, on one hand, an important enhancement of the discrimination accuracy by using the new RSC characteristic; on the other hand, this investigation has allowed us to compare the performances of the different classifiers on the same experimental conditions.

The best speaker discrimination score is over 91 %, reached by the SVM, and which appears to be the best classifier used in our experiments.

As perspectives, we intend to implement some fusion architectures between the different classifiers in order to further enhance the discrimination performances.

References

1. Rose, P.: Forensic speaker discrimination with Australian English vowel acoustics. In: ICPhS, XVI (2007)
2. Matrouf, D., Bonastre, J.F.: Accurate log-likelihood ratio estimation by using test statistical model for speaker verification. In: The Speaker and Language Recognition Workshop, Odyssey (2006)

3. Meignier, S., et al.: Step- by- step and integrated approaches in broadcast news speaker diarization. Comput. Speech Lang. **20**, 303–330 (2006)
4. Meignier, S.: Indexation en locuteurs de documents sonores: segmentation d'un document et Appariement d'une collection. Ph.D. thesis, LIA Avignon, France (2002)
5. Ouamour, S., Guerti, M., Sayoud, H.: A new relativistic vision in speaker discrimination. Can. Acoust. J. **36**(4), 24–34 (2008). ISSN 0711-6659
6. Li, M., Xing, Y., Luo, R.: Hierarchical speaker verification based on PCA and kernel fisher discriminant. In : 4th International Conference on Natural Computation, pp. 152–156 (2008)
7. Zhao, Z.D., Zhang, J., Tian, J.F., Lou, Y.Y.: An effective identification method for speaker recognition based on PCA and double VQ. In: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, pp. 1686–1689 (2009)
8. Jayakurnar, A., Vimal Krishnan, V.R., BabuAnto, P.: Text dependent speaker recognition using discrete stationary wavelet transform and PCA. In: International Conference on the Current Trends in Information Technology CTIT, pp. 1–4 (2009)
9. Zhou, Y., Zhang, X., Wang, J., Gong, Y.: Research on speaker feature dimension reduction based on CCA and PCA. In: International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–4 (2010)
10. Xiao-chun, L., Jun-xun, Y., A.: Text-independent speaker recognition system based on probabilistic principle component analysis. In: 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization, pp. 255–260 (2012)
11. Contributors of Wikipedia: Linear discriminant analysis. https://en.wikipedia.org/wiki/Linear_discriminant_analysis. Accessed Nov 2015
12. Contributors of Wikipedia: Adaboost. https://en.wikipedia.org/wiki/AdaBoost. Accessed Nov 2015
13. Contributors of Wikipedia: Support vector machine. https://en.wikipedia.org/wiki/Support_vector_machine. Accessed Nov 2015
14. Sayoud, H.: Automatic speaker recognition – connexionnist approach. Ph.D. thesis, USTHB University, Algiers (2003)
15. Contributors of Wikipedia: linear discriminant analysis. https://en.wikipedia. Last Accessed Nov 2015, Wikipedia, "Linear regression". http://en.wikipedia.org/wiki/Linear_regression. From Wikipedia, Last Accessed 28 Mar 2013
16. Huang, X., Pan, W.: Linear regression and two-class classification with gene expression data. Bioinformatics **19**(16), 2072–2078 (2003)
17. Contributors of Wikipedia, 2015. Generalized linear model. Last Accessed Nov 2015. https://en.wikipedia.org/wiki/Generalized_linear_model
18. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990). doi:10.1109/5.58325. Invited Paper
19. Tambouratzis, G., Hairetakis, G., Markantonatou, S., Carayannis, G.: Applying the SOM Model to Text Classification According to Register and Stylistic Content. Int. J. Neural Syst. **13**(1), 1–11 (2003)
20. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L.: Second-Order Statistical Measures for text-independent Broadcaster Identification. Speech Commun. **17**(1–2), 177–192 (1995)
21. Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. **17**(1–2), 91–108 (1995)
22. Shlens, J.: A tutorial on principal component analysis - Derivation, Discussion and Singular Value Decomposition. Version 1, (2003). www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf