Whitening-Based Feature Space Transformations in a Speech Impediment Therapy System

András Kocsor[1,2], Róber Busa-Fekete[1], and András Bánhalmi[1]

[1] MTA-SZTE Research Group on Artificial Intelligence
H-6720 Szeged, Aradi vértanúk tere 1., Hungary
{kocsor, busarobi, banhalmi}@inf.u-szeged.hu
[2] Applied Intelligence Laboratory Ltd., Petőfi S. Sgt. 43., H-6725 Szeged, Hungary

Abstract. It is quite common to use feature extraction methods prior to classification. Here we deal with three algorithms defining uncorrelated features. The first one is the so-called whitening method, which transforms the data so that the covariance matrix becomes an identity matrix. The second method, the well-known Fast Independent Component Analysis (FastICA) searches for orthogonal directions along which the value of the non-Gaussianity measure is large in the whitened data space. The third one, the Whitening-based Springy Discriminant Analysis (WSDA) is a novel method combination, which provides orthogonal directions for better class separation. We compare the effects of the above methods on a real-time vowel classification task. Based on the results we conclude that the WSDA transformation is especially suitable for this task.

1 Introduction

The primary goal of this paper is twofold. First we would like to deal with a unique group of feature extraction methods, namely with the uncorrelated ones. The uncorrelation can be carried out by using the well-known whitening method. After whitening among the linear transformations precisely the orthogonal ones preserve the property that the data covariance matrix remains the identity matrix. Thus following the whitening process we can apply any feature extraction method, which resulted in orthogonal feature directions. This kind of method composition in every case leads to uncorrelated features. Among the possibilities we selected two methods from the orthogonal family. The first one is the Fast Independent Component Analysis proposed by Hyvärinen and Oja [8], while the second one, recently introduced, is the Springy discriminant Analysis [9]. In this paper we investigate a version of this method combined with the whitening process. Our second aim here is to compare the effects of the above methods on a speech recognition task. We try to apply them on a real-time vowel classification task, which is one of the basic building blocks of our speech impediment therapy system [10].

Now without loss of generality we shall assume that, as a realization of multivariate random variables, there are n-dimensional real attribute vectors in a compact set $\mathcal{X}$ over $\mathbb{R}^n$ describing objects in a certain domain, and that we have a finite $n \times k$ sample matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ containing k random observations. Actually, $\mathcal{X}$ constitutes the initial

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 222–229, 2007.

feature space and X is the input data for the linear feature extraction algorithms which defines a linear mapping

$$\begin{aligned} h : \mathcal{X} &\to \mathbb{R}^m \\ \mathbf{z} &\to V\mathbf{z} \end{aligned} \tag{1}$$

for the extraction of a new feature vector. The $m \times n$ ($m \leq n$) matrix of the linear mapping – which may inherently include a dimension reduction – is denoted by V, and for any $\mathbf{z} \in \mathcal{X}$ we will refer to the result $h(\mathbf{z}) = V\mathbf{z}$ of the mapping as $\mathbf{z}^*$. With the linear feature extraction methods we search for an optimal matrix V, where the precise definition of optimality can vary from method to method. Now we will decompose V in a factorized form, i.e. we assume that $V = WQ$, where W, Q are orthogonal matrices and Q transforms the covariance matrix into the identity matrix. We will obtain Q by the whitening process, which can easily be solved by an eigendecomposition of the data covariance matrix (cf. Section 2). For the W matrix, which further transforms the data, we can apply various objective functions. Here we will find each particular direction of the optimal W transformations *one-by-one*, employing a $\tau : \mathbb{R}^n \to \mathbb{R}$ objective function for each direction (i.e. row vectors of W) separately. We will describe the Fast Independent Component Analyses (FastICA), and the Whitening-based Springy Discriminant Analysis (WSDA) via defining different τ functions.

The structure of the paper is as follows. In Section 2 we introduce the well-known whitening process, which is followed in Section 3 and 4 by the description of Independent Component Analysis and Springy Discriminant Analysis, respectively. Section 5 deals with the experiments, than in Section 6 we round off the paper with some concluding remarks.

2 The Whitening Process

Whitening is a traditional statistical method for turning the data covariance matrix into an identity matrix. It has two steps. First, we shift the original sample set $\mathbf{x}_1, \ldots, \mathbf{x}_k$ with its mean $E\{\mathbf{x}\}$, to obtain data

$$\mathbf{x}'_1 = \mathbf{x}_1 - E\{\mathbf{x}\}, \ldots, \mathbf{x}'_k = \mathbf{x}_k - E\{\mathbf{x}\}, \tag{2}$$

with a mean of $\mathbf{0}$. The goal of the next step is to transform the centered samples $\mathbf{x}'_1, \ldots, \mathbf{x}'_k$ via an orthogonal transformation Q into vectors $\mathbf{z}_1 = Q\mathbf{x}'_1, \ldots, \mathbf{z}_k = Q\mathbf{x}'_k$, where the covariance matrix $E\{\mathbf{z}\mathbf{z}^\top\}$ is the unit matrix. If we assume that the eigenpairs of $E\{\mathbf{x}'\mathbf{x}'^\top\}$ are $(\mathbf{c}_1, \lambda_1), \ldots, (\mathbf{c}_n, \lambda_n)$ and $\lambda_1 \geq \ldots \geq \lambda_n$, the transformation matrix Q will take the form $[\mathbf{c}_1\lambda_1^{-1/2}, \ldots, \mathbf{c}_t\lambda_t^{-1/2}]^\top$. If t is less than n a dimensionality reduction is employed.

Whitening transformation of arbitrary vectors. For an arbitrary vector $\mathbf{z} \in \mathcal{X}$ the whitening transformation can be performed using $\mathbf{z}^* = Q(\mathbf{z} - E\{x\})$.

Basic properties of the whitening process. i) for every normalized $\mathbf{v}$ the mean of $\mathbf{v}^\top\mathbf{z}_1, \ldots, \mathbf{v}^\top\mathbf{z}_k$ is set to zero, and its variance is set to one; ii) for any matrix W the covariance matrix of the transformed, whitened data $W\mathbf{z}_1, \ldots, W\mathbf{z}_k$ will remain a unit matrix if and only if W is orthogonal.

3 Independent Component Analysis

Independent Component Analysis [8] is a general purpose statistical method that originally arose from the study of blind source separation (BSS). An application of ICA is unsupervised feature extraction, where the aim is to linearly transform the input data into uncorrelated components, along which the distribution of the sample set is the least Gaussian. The reason for this is that along these directions the data is supposedly easier to classify.

For optimal selection of the independent directions, several objective functions were defined using approximately equivalent approaches. Here we follow the way proposed by A. Hyvärinen et al. [8]. Generally speaking, we expect these functions to be non-negative and have a zero value for the Gaussian distribution. Negentropy is a useful measure having just this property, which is used for assessing non-Gaussianity (i.e. the least Gaussianity). The negentropy of a variable η with zero mean and unit variance is estimated by using the formula

$$J_G(\eta) \approx \left(E\{G(\eta)\} - E\{G(\nu)\}\right)^2, \tag{3}$$

where $G : \mathbb{R} \to \mathbb{R}$ is an appropriate non-quadratic function, E again denotes the expectation value and ν is a standardized Gaussian variable. The following three choices of $G(\eta)$ are conventionally used: η^4, $\log(\cosh(\eta))$ and $-\exp(-\eta^2/2)$. It should be mentioned that in Eq. (3) the expectation value of $G(\nu)$ is a constant, its value only depending on the selected G function.

In Hyvärinen's FastICA algorithm for the selection of a new direction $\mathbf{w}$ the following τ objective function is used:

$$\tau_G(\mathbf{w}) = \left(E\{G(\mathbf{w}^\top \mathbf{z})\} - E\{G(\nu)\}\right)^2, \tag{4}$$

which can be obtained by replacing η in the negentropy approximant Eq. (3) with $\mathbf{w}^\top \mathbf{z}$, the dot product of the direction $\mathbf{w}$ and sample $\mathbf{z}$. FastICA is an approximate Newton iteration procedure for the local optimization of the function $\tau_G(\mathbf{w})$. Before running the optimization procedure, however, the raw input data X must first be preprocessed – by whitening it.

Actually property $i)$ of the whitening process (cf. Section 2) is essential since Eq. (3) requires that η should have a zero mean and variance of one hence, with the substitution $\eta = \mathbf{w}^\top \mathbf{z}$, the projected data $\mathbf{w}^\top \mathbf{z}$ must also have this property. Moreover, after whitening based on property $ii)$ it is sufficient to look for a new orthogonal base W for the preprocessed data, where the values of the non-Gaussianity measure τ_G for the base vectors are large. Note that since the data remains whitened after an orthogonal transformation, ICA can be considered an extension of PCA. The optimization procedure of the FastICA algorithm can be found in Hyvärinen's work [8].

Transformation of test vectors. For an arbitrary test vector $\mathbf{z} \in \mathcal{X}$ the ICA transformation can be performed using $\mathbf{z}^* = WQ(\mathbf{z} - E\{\mathbf{x}\})$. Here W denotes the orthogonal transformation matrix we obtained as the output from FastICA, while Q is the matrix obtained from whitening.

4 Whitening-Based Springy Discriminant Analysis

Springy discriminant analysis (SDA) is a method similar to Linear Discriminant Analysis (LDA), which is a traditional supervised feature extraction method [4,9]. Because SDA belongs to the supervised feature extraction family, let us assume that we have r classes and an indicator function $\mathcal{L} : \{1, \ldots, k\} \to \{1, \ldots, r\}$, where $\mathcal{L}(i)$ gives the class label of the sample $\mathbf{x}_i$. Let us further assume that we have preprocessed the data using the whitening method, the new data being denoted by $\mathbf{z}_1, \ldots, \mathbf{z}_k$.

The name Springy Discriminant Analysis stems from the utilization of a spring & antispring model, which involves searching for directions with optimal potential energy using attractive and repulsive forces. In our case sample pairs in each class are connected by springs, while those of different classes are connected by antisprings. New features can be easily extracted by taking the projection of a new point in those directions where a small spread in each class is obtained, while different classes are spaced out as much as possible. Now let $\delta(\mathbf{w})$, the potential of the spring model along the direction $\mathbf{w}$, be defined by

$$\delta(\mathbf{w}) = \sum_{i,j=1}^{k} \left((\mathbf{z}_i - \mathbf{z}_j)^\top \mathbf{w} \right)^2 [M]_{ij} \,, \tag{5}$$

where

$$[M]_{ij} = \begin{cases} -1, \text{ if } \mathcal{L}(i) = \mathcal{L}(j) \\ 1, \text{ otherwise} \end{cases} \qquad i, j = 1, \ldots, k. \tag{6}$$

Naturally, the elements of matrix M can be initialized with values different from ± 1 as well. The elements can be considered as a kind of force constant and can be set to a different value for any pair of data points.

It is easy to see that the value of δ is largest when those components of the elements of the same class that fall in the given direction $\mathbf{w}$ ($\mathbf{w} \in \mathbb{R}^n$) are close, and the components of the elements of different classes are far at the same time.

Now with the introduction of the matrix

$$D = \sum_{i,j=1}^{k} (\mathbf{z}_i - \mathbf{z}_j) (\mathbf{z}_i - \mathbf{z}_j)^\top [M]_{ij} \tag{7}$$

we immediately obtain the result that $\delta(\mathbf{w}) = \mathbf{w}^\top D \mathbf{w}$. Based on this, the objective function τ for selecting relevant features can be defined as the Rayleigh quotient $\tau(\mathbf{w}) = \delta(\mathbf{w}) / \mathbf{w}^\top \mathbf{w}$. It is straightforward to see that the optimization of τ leads to the eigenvalue decomposition of D. Because D is symmetric, its eigenvalues are real and its eigenvectors are orthogonal. The matrix W of the SDA transformation is defined using those eigenvectors corresponding to the m dominant eigenvalues of D.

Transformation of test vectors. For an arbitrary vector $\mathbf{z} \in \mathcal{X}$ the Whitening-Based SDA transformation can be performed using $\mathbf{z}^* = WQ(\mathbf{z} - E\{\mathbf{x}\})$.

5 Experiments and Results

In the previous sections three linear feature space transformation algorithms were presented. *Whitening* concentrates on those uncorrelated directions with the largest variances. *FastICA* besides keeping the directions uncorrelated, chooses directions along which the non-Gaussianity is large. *WSDA* creates attractive forces between the samples belonging to the same class and repulsive forces between samples of different classes. Then it chooses those uncorrelated directions along which the potential energy of the system is maximal. In this section we discuss these methods on the real-time vowel recognition tests. The motivation for doing this is to improve the recognition accuracy of our speech impediment therapy system, the 'SpeechMaster'. Besides reviewing 'SpeechMaster' here we will talk about the extraction of the acoustic features, the way the transformations were applied, the learners we employed and, finally, about the setup and evaluation of the real-time vowel recognition experiments.

The 'SpeechMaster'. An important clue to the process of learning to read for alphabetical languages is the ability to separate and identify consecutive sounds that make words and to associate these sounds with its corresponding written form. To learn to read in a fruitful way young learners must, of course, also be aware of the vowels and be able to manipulate them. Many children with learning disabilities have problems in their ability to process phonological information. Furthermore, phonological awareness teaching has also great importance for the speech and hearing handicapped, along with improving the corresponding articulatory strategies of tongue movement.

The 'SpeechMaster' software developed by our team seeks to apply speech recognition technology to speech therapy and the teaching of reading. Both applications require a real-time response from the system in the form of an easily comprehensible visual feedback. With the simplest display setting, feedback is given by means of flickering letters, their identity and brightness being adjusted to the speech recognizer's output [10]. In speech therapy it is intended to supplement the missing auditive feedback of the hearing impaired, while in teaching reading it is to reinforce the correct association between the phoneme-grapheme pairs. With the aid of a computer, children can practice without the need for the continuous presence of the teacher. This is very important because the therapy of the hearing impaired requires a long and tedious fixation phase. Experience shows that most children prefer computer exercises to conventional drills. In the 'SpeechMaster' system the real-time vowel recognition module has a great importance, this is why we chose this task for testing the uncorrelated feature extraction methods.

Evaluation Domain. For training and testing purposes we recorded samples from 160 normal children aged between 6 and 8. The ratio of girls and boys was 50% - 50%.The speech signals were recorded and stored at a sampling rate of 22050Hz in 16-bit quality. Each speaker uttered all the 12 isolated Hungarian vowels, one after the other, separated by a short pause. The recordings were divided into a train and a test set in a ratio of 50% - 50%.

Acoustic Features. There are numerous methods for obtaining representative feature vectors from speech data, but their common property is that they are all extracted from 20-30 ms chunks or "frames" of the signal in 5-10 ms time steps. The simplest possible

feature set consists of the so-called bark-scaled filterbank log-energies (FBLE). This means that the signal is decomposed with a special filterbank and the energies in these filters are used to parameterize speech on a frame-by-frame basis. In our tests the filters were approximated via Fourier analysis with a triangular weighting, as described in [6].

It is known from phonetics that the spectral peaks (called formants) code the identity of vowels [11]. To estimate the formants, we implemented a simple algorithm that calculates the gravity centers and the variance of the mass in certain frequency bands [1]. The frequency bands are chosen so that they cover the possible place of the first, second and third formants. This resulted in 6 new features altogether.

A more sophisticated option for the analysis of the spectral shape would be to apply some kind of auditory model. We experimented with the In-Synchrony-Bands-Spectrum of Ghitza [5], because it is computationally simple and attempts to model the dominance relations of the spectral components. The SBS model analysis the signal using a filterbank that is approximated by weighting the output of a FFT - quite similar to the FBLE analysis. In this case, however, the output is not the total energy of the filter, but the frequency of the component that has the maximal energy.

Feature Space Transformation. When applying the uncorrelated feature extraction methods (see Section 2, 3 and 4) we invariably kept only 8 of the new features. We performed this severe dimension reduction in order to show that, when combined with the transformations, the classifiers can yield the same scores in spite of the reduced feature set. Naturally, when we applied a certain transformation on the training set before learning, we applied the same transformation on the test data during testing.

Classifiers. Describing the mathematical background of the learning algorithms applied is beyond the scope of this article; in the following we specify only the parameters applied.

Gaussian Mixture Modeling (GMM). In the GMM experiments, three Gaussian components were used and the expectation-maximization (EM) algorithm was initialized by k-means clustering [4]. To find a good starting parameter set we ran it 15 times and used the one with the highest log-likelihood. In every case the covariance matrices were forced to be diagonal.

Artificial Neural Networks (ANN). In the ANN experiments we used the most common feed-forward multilayer perceptron network with the backpropagation learning rule [2]. The number of neurons in the hidden layer was set to 18 in each experiment (this value was chosen empirically, based on preliminary experiments). Training was stopped based on the cross-validation of 15% of the training data.

Projection Pursuit Learning (PPL). Projection pursuit learning is a relatively little-known modelling technique [7]. It can be viewed as a neural net where the rigid sigmoid function is replaced by an interpolating polynomial. In each experiment, a model with 8 projections and a 5th-order polynomial was applied.

Support Vector Machines (SVM). Support vector machines is a classifier algorithm that is based on the ubiquitous kernel idea [12]. In all the experiments with SVM the radial basis kernel function was applied.

Experiments. In the experiments 5 feature sets were constructed from the initial acoustic features described above. *Set1* contained the 24 FBLE features. In *Set2* we combined

Table 1. Recognition errors for each feature set as a function of the transformation and classification applied

feature set	**classifier**	none(all)	Whitening(8)	FastICA(8)	WSDA(8)
	GMM	16.38	14.21	16.45	14.32
	ANN	10.34	9.85	9.93	9.42
Set1 (24)	PPL	11.04	10.46	10.69	10.02
	SVM	9.93	10.12	8.95	**8.05**
	GMM	13.33	11.21	13.33	12.33
	ANN	7.43	7.35	7.36	**5.25**
Set2 (30)	PPL	9.37	8.41	6.54	6.23
	SVM	8.33	6.85	6.66	5.43
	GMM	25.90	22.34	25.90	23.67
	ANN	20.00	**18.41**	19.58	19.65
Set3 (24)	PPL	20.48	19.43	19.58	19.33
	SVM	19.65	20.08	18.88	19.48
	GMM	13.95	12.21	15.90	13.67
	ANN	10.27	9.79	8.05	8.48
Set4 (48)	PPL	10.48	8.80	9.37	9.31
	SVM	9.09	9.46	8.26	**7.41**
	GMM	15.48	12.46	13.33	12.72
	ANN	8.68	7.31	6.45	7.41
Set5 (54)	PPL	8.26	9.05	7.36	7.09
	SVM	9.37	9.11	5.76	**5.64**

Set1 with the gravity center features, so *Set2* contained 30 measurements. *Set3* was composed of the 24 SBS features, while in *Set4* we combined the FBLE and SBS sets. Lastly, in *Set5* we added all the FBLE, SBS and gravity center features, thus obtaining a set of 54 values.

In the classification experiments every transformation was combined with every classifier on every feature set. The results are shown in Table 1. In the header Whitening, FastICA, WSDA stand for the linear uncorrelated feature space transformation methods. The numbers shown are the recognition errors on the test data. The number in parentheses denotes the number of features preserved after a transformation. The best scores of each set are given in bold.

Results and Discussion. Upon inspecting the results the first thing one notices is that the SBS feature set (*Set3*) did about twice as badly as the other sets, no matter what transformation or classifier was tried. When combined with the FBLE features (*Set1*) both the graity center and the SBS features brought some improvement, but this improvement is quite small and varies from method to method.

When focusing on the performance of the classifiers, we see that ANN, PPL and SVM yielded very similar results. They, however, consistently outperformed GMM, which is still the method most commonly used in speech technology today. This can be attributed to the fact that the functions that a GMM (with diagonal covariances) is able to represent are more restricted in shape than those of ANN or PPL.

As regards the transformations, an important observation is that after the transformations the classification scores did not get worse compared to the classifications when no

transformation was applied. This is so in spite of the dimension reduction, which shows that some features must be highly redundant. Removing some of this redundancy by means of a transformation can make the classification more robust and, of course, faster. Comparing the methods, we may notice that WSDA brought significant improvement on the recognition accuracy. Maybe this is due to the supervised nature of the method.

6 Conclusions

In this paper three linear uncorrelated feature extraction algorithms (Whitening, FastICA and WSDA) were presented, and applied to real-time vowel classification. After inspecting the test results we can confidently say that it is worth experimenting with these methods in order to obtain better classification results. The Whitening-based Springy Discriminant Analysis brought a notable increase in the recognition accuracy despite applying a severe dimension reduction. This transformation could greatly improve our phonological awareness teaching system by offering a robust and reliable real-time vowel classification, which is a key part of the system.

Acknowledgments

A. Kocsor was supported by the János Bolyai fellowship of the Hungarian Academy of Sciences.

References

1. Albesano, D., Mori, R.D., Gemello, R., Mana, F.: A study on the effect of adding new dimensions to trajectories in the acoustic space. In: Proc. of Eurospeech'99, pp. 1503–1506 (1999)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford Univerisity Press Inc., New York (1996)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, New York (2001)
4. Fukunaga, K.: Statistical Pattern Recognition. Academic Press, New York (1989)
5. Ghitza, O.: Auditory Nerve Representation Criteria for Speech Analysis/Synthesis. IEEE Transaction on ASSP 35, 736–740 (1987)
6. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice Hall, Englewood Cliffs (2001)
7. Hwang, J.N., Lay, S.R., Maechler, M., Martin, R.D., Schimert, J.: Regression Modeling in Back-Propagation and Projection Pursuit Learning, IEEE Trans. on Neural Networks 5, 342–353 (1994)
8. Hyvärinen, J., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Comp. 9, 1483–1492 (1997)
9. Kocsor, A., Tóth, L.: Application of Kernel-Based Feature Space Transformations and Learning Methods to Phoneme Classification. Appl. Intelligence 21, 129–142 (2004)
10. Kocsor, A., Paczolay, D.: Speech Technologies in a Computer-Aided Speech Therapy System. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 615–622. Springer, Heidelberg (2006)
11. Moore, B.C.J.: An Introduction to the Psychology of Hearing, Acad. Pr. (1997)
12. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons Inc., NY (1998)