INTERSPEECH 2009 BRIGHTON

Dimension Reduction Approaches for SVM based Speaker Age Estimation

Gil Dobry [1], *Ron M. Hecht* [2], *Mireille Avigal* [1], *Yaniv Zigel* [3]

[1] Computer Science Dept., The Open University of Israel,
[2] PuddingMedia, Kfar-Saba, Israel
[3] Bio-medical Eng. Dept., Ben-Gurion University

Gil.Dobry@gmail.com, Hadasron@gmail.com, Miray@openu.ac.il, Yaniv@bgu.ac.il

10.21437/Interspeech.2009-584

Abstract

This paper presents two novel dimension reduction approaches applied on the gaussian mixture model (GMM) supervectors to improve age estimation speed and accuracy. The GMM supervector embodies many speech characteristics irrelevant to age estimation and like noise, they are harmful to the system's generalization ability. In addition, the support vectors machine (SVM) testing computation grows with the vector's dimension, especially when using complex kernels. The first approach presented is the weighted-pairwise principal components analysis (WPPCA) that reduces the vector dimension by minimizing the redundant variability. The second approach is based on anchor-models, using a novel anchors selection method. Experiments showed that dimension reduction makes the testing process 5 times faster and using the WPPCA approach, it is also 5% more accurate.
Index Terms: age estimation, anchor models, dimension reduction, GMM supervector, NAP, SVM

1. Introduction

Speaker age is part of non-verbal information of a speech session that gained increasing importance recently in speech-based applications. For interactive voice response (IVR) systems, this information is helpful to adapt to the user and give a more natural human-machine interaction. Speech recognition systems can select an appropriated language model to the speaker age and improve the detection accuracy. Estimate speaker age at call centers is also used to do user-profiling which is a basis to applications like market research, targeted advertising and service customization.

Several speech-based age and gender estimation systems have been proposed [1, 2, 3], these use and combine different kinds of acoustic features and classification algorithms. More recently, a support vector machine (SVM) framework over GMM supervectors was proposed by Bocklet et al. [4] for age and gender classification. This framework was used before in various speech analysis tasks and was found to be effective. However, the very high supervectors dimension makes the training and testing processes very heavy in term of computational resources. Also the irrelevant information like channel characteristics, spoken language, accent and emotion is a part of the supervector and is harmful to the system's performance. Dimension reduction techniques can project the supervectors into a lower dimension space and suppress noise to reach a faster and easier separability.

In this paper, we compare four dimension reduction approaches using the GMM supervectors framework: (1) principal components analysis (PCA). (2) Supervised PCA (SPCA). And two novel ones: (3) Weighted-pairwise PCA (WPPCA), a method based on nuisance attributes projection (NAP) [5], that uses labels to find and preserve the between-class variability. (4) A modification of the anchor-models technique [6], where anchor supervectors are selected to be distant from each other. This selection method avoids information loss and redundancy in the scores space. As shown in section 5, the testing computation decreases when applying dimension reduction, and the accuracy is improved. Using the Gaussian RBF-kernel, the computation time is significantly decreased, and comparing to other kernels it also performs better. The rest of the article is organized as follows: In section 2, the age estimation system by age-group classification is introduced and section 3 describes the dimension reduction approaches, including the two novel ones. In section 4, the SVM testing complexity is analyzed and the experimental results are presented in section 5.

2. Age estimation system

The age estimation system is gender dependant and trained to classify speakers to one of 3 age groups: Young people (Y), adults (A) and seniors (S). The GMM means supervector framework is used with the mel-frequency cepstrum coefficients (MFCC) acoustic features. Figure 1 shows the block diagram of the entire system with its sub-components. The training part consists of two phases. In phase *A*, the universal background GMM model (UBM) called "world model" is trained over a large speech database where speakers are uniformly distributed over ages and genders. In phase *B*, a GMM model is created for each training session using MAP adaptation on the UBM model. For each session *x*, a GMM model is trained and a GMM supervector $\mathbf{v}_x$ is formed by concatenating all the *g* Gaussian means:

$$\mathbf{v}_x = \left(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots \boldsymbol{\mu}_g\right)^T \quad (1)$$

Where $\boldsymbol{\mu}_i$ is the mean vector of the i^{th} Gaussian. The dimension reduction projection matrix is calculated on the training GMM-supervectors for each dimension reduction method. The projection matrix always comes in a matrix form, applied efficiently on the GMM-supervectors by a simple matrix multiplication. The reduced vectors are used to train binary SVM classifiers, to distinguish each age-group to the rest (one-versus-all). Thus, three classifiers are trained for each gender. Cross validation SVM scores are evaluated using the n-fold validation technique, and their distribution parameters are revaluated for each age group. These distribution parameters are used to normalize the scores to probability-of-membership values for each age-group.

In the test phase, a testing session is processed the same way it was done for the training sessions, to create a corresponding GMM supervector. The dimension reduction projection matrix is applied on it to create a reduced vector evaluated by the SVM models.

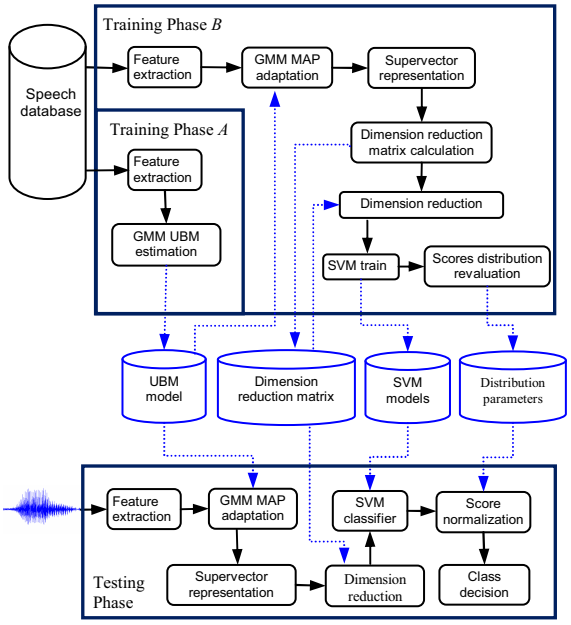

Figure 1: *GMM supervectors framework for age group classification.*

3. Dimension reduction approaches

In this section, we describe the four dimension reduction methods mentioned in section 1. They were implemented, evaluated and compared to a baseline system where the GMM supervectors are used directly as feature vectors.

3.1. Principle components analysis

PCA [7] is an orthogonal linear transformation that projects the feature vectors to a basis which components are linearly uncorrelated and arranged in a decreasing order of variance. This method assumes that most of the relevant information is found in the first coordinates in the projected space, where most of the variance is found. The PCA projection matrix columns are the eigenvectors of the feature vectors' covariance matrix. A dimension reduction is then made by using only the first m coordinates of the projected vectors, where m is much smaller than the original feature vectors dimension.

3.2. Supervised PCA

SPCA is a PCA variant where the feature vectors are preprocessed before applying PCA on them. The preprocessing consists to screen out coordinates having the lowest correlation with labels. First, the correlation vector $\mathbf{c}$ between the training supervectors and the labels is calculated:

$$\mathbf{c} = \hat{\mathbf{A}}\hat{\mathbf{y}} \tag{2}$$

where $\hat{\mathbf{A}}$ is the normalized input training supervectors matrix and $\hat{\mathbf{y}}$ is the normalized label vector, both of them are normalized by their mean and standard-deviation. The vector $\mathbf{c}$ contains the correlation coefficient between each coordinate of the feature vectors and the labels. This method is generally used in regression problems [8] where the label is continuous; we can then apply it in the age estimation system too since age is a continuous variable.

3.3. Weighted pairwise PCA (WPPCA)

In PCA we achieve a dimension reduction that preserves most of the vectors variance without taking in account the class labels. There is no guarantee that the directions of maximum variance will contain good features for discrimination. The proposed technique permits to shape the features variability at the projected space based on the label information. The NAP projection framework was proposed in [5] and was found useful to eliminate the inter-session variability for speaker verification. Here we extend this framework to preserve the variability needed to discriminate speakers by age groups while reducing the supervectors dimension.

3.3.1. Projection matrix

We create an $n \times m$ linear projection matrix $\mathbf{P}$ that projects the n dimensional supervectors to an m-dimensional subspace where $m << n$. $\mathbf{P}$ must be chosen to maximize the pairwise variability criterion:

$$\delta = \sum_{i,j=1}^{n} \mathbf{W}_{i,j} \left\| \mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j \right\|^2 \tag{3}$$

Where $\mathbf{W}$ is the symmetric weight matrix containing weight values for all the vector pairs: $(\mathbf{x}_i, \mathbf{x}_j) | i, j = 1,2,...,n$. According to [5], the variability criterion δ is maximized by choosing $\mathbf{P}$ to be the m eigenvectors with the largest eigenvalues of matrix $\mathbf{S}$:

$$\mathbf{S} = \mathbf{AZ(W)A}^t \tag{4}$$

$\mathbf{A}$ is the matrix whose columns are the training supervectors, $\mathbf{Z(W)}=diag(\mathbf{W1})-\mathbf{W}$, where $diag(\mathbf{x})$ is the matrix whose diagonal is $\mathbf{x}$, and $\mathbf{1}$ is the vector of all ones. Since the supervectors dimension is generally higher than the number of training sessions, the resulting matrix $\mathbf{S}$ is singular. However, the matrix S can be decomposed to $\mathbf{S} = \mathbf{VV}^t$ where $\mathbf{V} = \mathbf{AZ(W)}$, so the eigenvectors of $\mathbf{S}$ can be determined by singular value decomposition (SVD) of $\mathbf{V}$.

3.3.2. Weights matrix

We first introduce the preprocessing algebraic logistic function:

$$\psi_{\theta,\beta}(x) = \frac{(x-\theta)}{\sqrt{\beta + (x-\theta)^2}} \tag{5}$$

Where θ is the center of the logistic function and β is its width factor. Applied on age values, the logistic function is monotonically increasing and emphasizes the age-group membership: Its value applied on ages below and above θ is much more different than if applied on ages from the same side. For a binary model classifying between speakers above and below a certain age limit l, we chose the center θ to be l. The weight matrix is built using the following formula:

$$\mathbf{W}_{i,j} = \left| \psi_{l,\beta}(a_i) - \psi_{l,\beta}(a_j) \right| \tag{6}$$

Where a_i is the i^{th} speaker age. The idea is to have a maximal weight for speakers from different classes and a minimal weight for speakers from to the same class. This weights matrix is applied in (4) to obtain a classifier-specific

projection matrix **P**. The desired between-class variability is preserved in the projected space while the within-class variability is minimized.

3.4. Anchor models

Anchor models representation is a technique generally used for speaker verification [9] to project a given session into a scores space. This technique uses anchor models trained on a predefined set of speech sessions. In our framework these models are obtained by MAP adaptation of the UBM. The anchor models representation of a candidate session x is by the scores it obtained from the anchor models.

3.4.1. Projection matrix

It was shown in [6] that the log-likelihood values obtained by the anchor models can be approximated by:

$$\mathbf{s}(X) \approx \hat{\mathbf{\Lambda}}^t \hat{\mathbf{v}}_x \tag{7}$$

Where $\hat{\mathbf{\Lambda}}$ is the normalized anchors supermatrix whose columns are all the normalized supervectors of the anchor models and $\hat{\mathbf{v}}_x$ is the normalized GMM supervector of session x. The supervector normalization is applied on the concatenated Gaussian means in (1), by the formula:

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i \sqrt{w_i / \boldsymbol{\Sigma}_i} \tag{8}$$

Where $\boldsymbol{\mu}_i, w_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector, weight and covariance matrix of the i^{th} Gaussian respectively.

3.4.2. Anchor-supervectors selection

The anchor-supervectors need to be diversified and represent speakers from all class labels to ensure minimal information loss in the projected space. Moreover, projecting using adjacent anchor-supervectors in (7) gives highly correlated values, which leads to redundancy in the projected space. To avoid that, we use a selection method such that the anchor-supervectors are chosen to be distant from each other, considering them as points in a high-dimensional space. Finding a subset of the most distant points in a given set is an NP-complete problem, but there are approximation algorithms giving good results. Zigel and Cohen [10] proposed the close impostor clustering (CIC), an iterative algorithm to find distant cohort models used for score normalization. We use a similar selection method by clustering the candidate supervectors using K-means. Note that because of the normalization in (8), the log-likelihood approximation in (7) becomes also an upper-bound to the KL distance [11] between GMM models:

$$K(\Phi_x, \Phi_y) \le \sum_{i=1}^{g} (\boldsymbol{\mu}_i^x \sqrt{w_i / \boldsymbol{\Sigma}_i})^t \cdot (\boldsymbol{\mu}_i^y \sqrt{w_i / \boldsymbol{\Sigma}_i}) = \hat{\mathbf{v}}_x^t \cdot \hat{\mathbf{v}}_y \tag{9}$$

Where Φ_x and Φ_y are the MAP-adapted GMM models having the same weight values and covariance matrices, but different means vectors: $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$. $\hat{\mathbf{v}}_x$ and $\hat{\mathbf{v}}_y$ are the normalized GMM supervectors of Φ_x and Φ_y respectively. The KL distance approximation is then calculated with the normalized supervectors, and used as a distance measure for the K-means clustering. From each resulting cluster, only the supervector which is closest to the cluster's mean is chosen. By the nature of clustering, close vectors will be grouped within the same cluster, so the chosen anchor-supervectors will be distant and span efficiently the supervectors' space.

4. SVM testing complexity

The SVM score evaluation of a d dimensional feature vector $\mathbf{x}$ consists to calculate the weighted sum of the kernel function $k(\mathbf{x},\mathbf{z})$ calculated over all the n support vectors:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{z}_i) + b \tag{10}$$

Where f is the test score function, y_i are the target values, a_i the model parameters and $\mathbf{z}_i$ the support vectors. This formulation shows that the complexity depends on the number of support vectors and the kernel function calculation. The chosen type of kernel affects on its complexity, but it is generally linear to the dimension d like the Gaussian RBF-kernel involves a distance calculation between two vectors. The complexity of the test score function calculation f is in this case $O(nd)$, which is sensitive to the vectors dimension d. Also, the storage needed for the support vectors is $O(nd)$ which is also reduced with a smaller d.

5. Experimental Setup and Results

5.1. Database

Speech data used to train the UBM model was taken from LDC's Switchboard corpus annotated with age and gender labels. Six minutes long sessions from 2430 speaker were taken summing up to around 50 hours of speech. For the training and testing sessions, LDC's Fisher corpus was used. This database consists of 11699 spontaneous phone conversations lasting 10 minutes. 12000 different speakers are recorded, ~5000 males and ~7000 females with age ranging from 15 till 85 years old. Age-groups were defined as follows:

- Young people: 18-25 years (Y)
- Adults: 26-54 years (A)
- Seniors: 55-80 years (S)

Training and testing sets were selected as shown in table 1.

Table 1. *Session sets (number of sessions).*

Gender	Age group	Training-set	Testing-set
Female	Y	1250	1250
	A	1400	1400
	S	750	750
	Total	3400	3400
Male	Y	1250	1250
	A	1400	1400
	S	500	500
	Total	3150	3150

5.2. Setup

The acoustic features used are MFCC with 12 coefficients + C0 and their first derivatives forming a 26 dimension acoustic feature vector. The UBM is trained to 512 Gaussians, so the GMM supervectors dimension is 13312, (26×512). The dimension reduction approaches were applied at different reduction levels on the supervectors and trained using the

Gaussian RBF kernel. The system was programmed in Python, the acoustic feature extraction and the GMM model training were executed with the HTK toolset. The SVM model training and testing were executed with LIBSVM [12].

5.3. Results and Discussion

5.3.1. Performance evaluation

The average equal error rate (EER) of the SVM classifiers is calculated for each approach per target dimension. This value is the average of the EER values obtained by each one of the three classifiers, classifying each age-group from all the rest. Figure 2 shows the performance on female and male speakers. It can be seen that WPPCA, using β=100 for the sigmoid function performed the best on both genders at target dimension 600. This is ~5% better than the baseline system, which uses vectors of dimension 13312. The second best performer is the anchor-model approach with a similar improvement seen only on female speakers.

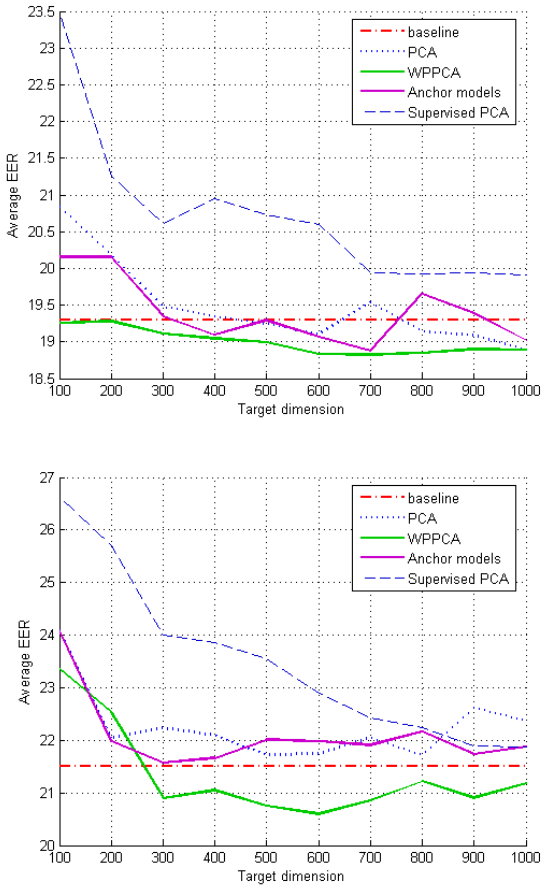

Figure 2: *Average EER obtained on female (top) and male (bottom) speakers vs. target dimension.*

5.3.2. Speed measurements

The supervectors dimension reduction has a great impact on the testing time. Table 2 shows the time measurement of a GMM supervector testing, including the dimension reduction process. At target dimension of 600 (which best performed), the testing time drops by ~79% comparing to the baseline, from 468 to 97.5 milliseconds.

Table 2. *Supervector testing time per target dimension (Running on an Intel™ Pentium IV, 3GHZ with 2GB of RAM).*

Target dimension	Dimension reduction (in ms)	SVM testing (in ms)	Total time (in ms)
Baseline: 13312	Not applied	468	468
1000	90	46.7	136.7
600	55	42.5	97.5
100	9	37.5	46.5

6. Conclusions

We proposed two novel dimension reduction approaches and applied them on the GMM supervectors for a speaker's age estimation system. We showed that using low dimensional vectors, we can decrease the SVM testing time and improve its accuracy. We compared our approaches to standard ones and found they perform better, WPPCA was found to be the most effective with consistent accuracy improvement of 5% on both genders. The drastic reduction (by 97%) of the GMM supervector reveals that the relevant information is just a tiny part of it, extracted by linear projection. In the future we plan to explore non-linear kernel-based dimension reduction techniques for different age applications like precise-age estimation systems.

7. References

[1] C. Muller, F. Burkhardt, "Combining Short-term Cepstral and Long-term Pitch features for Automatic Recognition of Speaker Age", in Interspeech, Pittsburgh, PA, 2007.

[2] F. Metze, J. Ajmera, "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications", in ICASSP, Honolulu, Hawaii, 2007.

[3] M. Nobuaki, S. Mariko, "Automatic Estimation of one's Age with his/her Speech Based Upon Acoustic Modeling Techniques of Speakers". University of Tokyo, 2002.

[4] T. Bocklet, E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines". In ICASSP, vol. 1, 2008, pp. 1605-1608.

[5] M. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation", in ICASSP, vol. 1, 2006, pp. 97-100.

[6] E. Noor, H. Aronowitz, "Efficient language Identification using Anchor Models and Support Vector Machines," in IEEE Odyssey, 2006, pp. 1-6.

[7] S. Lindsay, "A tutorial on Principal Components Analysis", 2002.

[8] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, "Prediction by supervised principal components". Journal of the American Statistical Association, 2006, pp. 119-137.

[9] Y. Yang, M. Yang, and Z. Wu, "A Rank based Metric of Anchor Models for Speaker Verification", in IEEE ICME, 2006, pp. 1097-1100.

[10] Y. Zigel, A. Cohen, "On cohort selection for speaker verification", in EUROSPEECH, 2003, pp. 2977-2980.

[11] R. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, "Linear and Non Linear Kernel GMM SuperVector Machines for Speaker Verification", in Interspeech, Antwerp, Belgium, 2007.

[12] C. Chih-Chung, and L. Chih-Jen, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm