ROBUST SPEAKER CLUSTERING IN EIGENSPACE

R. Faltlhauser, G. Ruske

Inst. for Human-Machine-Communication,
Technische Universität München, Munich, Germany
{faltlhauser,ruske}@ei.tum.de

ABSTRACT

In this paper we propose a speaker clustering scheme working in 'Eigenspace'. Speaker models are transformed to a low-dimensional subspace using 'Eigenvoices'. For the speaker clustering procedure simple distance measures, e.g. Euklidean distance can be applied. Moreover, clustering can be accomplished with base models (for Eigenvoice projection) like Gaussian Mixture Models as well as conventional HMMs. In case of HMM models re-projection to original space readily yields acoustic models. Clustering in subspace produces well-balanced cluster and is easily to control. In the field of speaker adaptation several principal techniques can be distinguished. The most prominent among them are Bayesian adaptation (e.g. MAP), transformation based approaches (MLLR) as well as so-called Eigenspace techniques. Especially the latter have become increasingly popular, as they make use of a-priori information about the distribution of speaker models. The basic approach is commonly called the Eigenvoice (EV) approach. Besides these techniques, speaker clustering is a further attractive adaptation scheme, especially since it can be - and has been - easily combined with the above methods.

1. INTRODUCTION

In the topic of speaker adaptation several fundamental techniques can be distinguished: One of the most prominent is probably Bayesian adaptation (e.g. MAP), proposed by Gauvain in [1, 2]. Another important adaptation strategy is based on affine model transformation(s): Woodland and Leggetter (MLLR,[3]), Digalakis and Neumeyer [4, 5]. Recently another adaptation scheme has received remarkable interest: Eigenvoices, introduced by Kuhn et. al [6]. The approach is based on a transformation of the model parameters to a low-dimensional subspace. By restricting adapted models to this subspace, efficient use of the information about the distribution of speaker model parameters can be made. The latter approach is particularly effective for rapid speaker adaptation with sparse data.

Besides these adaptation techniques another principal adaptation scheme can be subsumed under the term 'speaker clustering'. The basic idea behind speaker clustering is to find groups of speakers in a database, who have similar acoustic properties - or at least similar properties in feature space. For each of these groups individual HMMs can be generated. As a first approach this can be used for speaker adaptation by selecting a model for a new speaker from the set of given speaker or cluster models - based on a similarity measure. In this basic form speaker clustering has a big advantage over other adaptation schemes: In case of unsupervised adaptation there is no need for a phonetic segmentation in order to allow model selection. If an adaptation step is based on a phonetic segmentation, it is commonly extracted from a 1st-pass recognition result using generic or unadapted models - always hoping the recognition was good enough. Such an assumption is usually prone to errors. As far as speaker clustering is concerned, in order to select a model for an unknown speaker from a pool of cluster (recognizer) models, the only need is a robust similarity measure. The pre-trained models can be either representing a whole cluster or can be taken from a representative for a cluster (reference speaker). Although the achievable improvements are limited, such a selection strategy serves very well as a first stage in a series of adaptation steps.

Various forms and extensions to the speaker clustering idea have been investigated, e.g. by Furui [7] or Gao [8]. Recently speaker clustering has received more attention since it can easily be combined with the above adaptation strategies, as shown e.g. by Johnson [9] or Gales [10].

A key problem in speaker clustering is the grouping procedure itself. Especially for high-dimensional problems the grouping process can be difficult to control. Sophisticated and often hand-crafted distance measures are mostly applied to avoid divergence of the clustering process and to achieve well-balanced cluster.

In this paper we propose a clustering procedure working in the low-dimensional Eigenvoice subspace of speaker model parameters. The transformation allows a high dimensionality reduction, together with a decorrelation of model parameters themselves. As a result of this transformation, simple speaker distance computation is possible.

2. EIGENVOICES

It was Kuhn et al. who first introduced the concept of Eigenvoices [6]. The idea was probably inspired by the concept of Eigenfaces, which has received much attention in image processing, more precisely in face recognition. The basic idea behind this approach is to apply a Principal Components Analysis (PCA) to speaker model parameters instead of - what is commonly done - applying it to the feature parameters. PCA achieves decorrelation by determining the principal axes (Eigenvoices) of the speaker covariance matrix.

PCA is basically a linear transformation aimed at finding new basis vectors, i.e. a new coordinate system in which the transformed model parameters are linearly independent. In this sense, PCA is highly effective for dimensionality reduction, since it allows to determine the most important axes, i.e. the axes, which produce the least error if all other basis vectors are omitted.

The algorithm can be outlined as follows:

- train a speaker-dependent (SD) model for each speaker in the database, using e.g. MAP
- for all speakers: align all model parameters of each speaker in one large supervector, i.e. N_S speakers $\rightarrow N_S$ supervectors
- compute mean and covariance matrix of all N_S supervectors
- apply PCA to determine principal axes
- select first K Eigenvoices

By the application of Eigenvoices, the dimensionality of free parameters can be reduced from several thousands to less than 100. A fact, which makes this idea very attractive for rapid speaker adaptation: if only sparse adaptation data are available, the model for the new, unknown speaker can be constrained to the subspace spanned by the first K Eigenvoices allowing robust estimation. For this task, Kuhn introduced an adaptation scheme, called MLED ('Maximum Likelihood Eigenspace Decomposition'). This algorithm estimates the constrained model parameters using a Maximum Likelihood approach. The maximization is done in original-space with respect to the weights $\mathbf{w}$ in Eigenspace. This can be achieved by re-projecting the speaker specific weights $\mathbf{w}$ back to the original space:

$$\mu_{jm} = \bar{\mu}_{jm} + \sum_{k=1}^{K} w(k)\mathbf{e}_{jm}(k)$$

$\mathbf{e}_{jm}(k)$ is that part of Eigenvoice k belonging to gaussian density j of state m. Usually only the gaussian mean parameters are adapted, nevertheless the algorithm should work with variances as well. Solving the ML equations, a reestimation formula can be derived for the parameters $\mathbf{w}$. The estimate is very robust, however it requires a phonetic segmentation.

As a main effect the transformation to a low-dimensional subspace reduces the number of parameters necessary for a similarity measure between speakers drastically. Furthermore speakers sharing similar Eigenspace coordinates can be assumed to be similar even if Eigenspace dimensionality is reduced. In Eigenspace the Euklidean distance between 2 speakers i and j is given by:

$$d_{ij}^2 = |\mathbf{w}_i - \mathbf{w}_j|^2 = (\mathbf{w}_i - \mathbf{w}_j)^T(\mathbf{w}_i - \mathbf{w}_j) =$$

$$\sum_{k=1}^{K}(w_{ik} - w_{jk})^2$$

The Eigenvoice method can be applied to Hidden Markov Models, as has been shown by Kuhn et al. In [11] Thyes used the Eigenvoice approach also for speaker identification based on Gaussian Mixture Models (GMM,[12]). He proposed 2 identification setups, one after backtransformation in original space and one in the reduced space. Thyes found that speaker identification in reduced space is not as effective as in original space. Nevertheless, for the task of speaker grouping the exact speaker ID is of no importance - as long as similarity conditions between speakers are preserved.

3. CLUSTERING ALGORITHM

3.1. Distance measures

The key task of the clustering algorithm is to automatically find groups (cluster) of speakers with similar properties in feature space. In order to be able to make a statement about 'similarity', a distance measure between 2 speakers has to be defined. The distance measure should work or at least depend on features which are somehow contained or accessible to the preprocessing stage and the model structure of the speech recognizer later on. Simple example: it would make no sense if the clustering is based on the color of eyes - since the recognizer does not make any use of this information (remark: just an example - not that color of eyes would have something to do with speech).

Since the structure of models is usually very complex (i.e. many high dimensional densities; in case of HMM: states), simple one-to-one distance measures are often not directly applicable between models. Instead, likelihood or modified nearest/furthest-Neighbor measures have to be used. E.g. in case of comparing HMM models, on the one hand alignments between HMM states have to be found, and on the other hand the distance measures between the multimodal distributions within the states have to be defined.

However, by the application of the Eigenvoice transformation a whole complex speaker model simply can be reduced to a low-dimensional set of coordinates $\mathbf{w}$, leading to:

$$d_{ij} = |\mathbf{w}_i - \mathbf{w}_j|$$

The above equation gives the distance (or similarity) between 2 speakers. In order to measure the distance between 2 speaker cluster (cluster-to-cluster distance) we used a furthest-Neighbor criterion. In our case the cluster-to-cluster distance is given by the furthest speaker-to-speaker distance in Eigenspace from a speaker of cluster m to a speaker from cluster n:

$$D = D_{mn} = \max_{i \in m, j \in n} d_{ij}$$

At this point other definitions could be applied, e.g the introduction of reference speakers, which would lead to:

$$D_{mn} = \max_{j \in n} d_{ij}$$

with $i = ref(m)$, whereby here: $D_{mn} \neq D_{nm}$.

3.2. Clustering Algorithm in Eigenspace

Bottom-Up Clustering:

Concerning clustering, 2 basic approaches are very common: top-down and bottom-up. In the following we will focus on a bottom-up scheme. Nevertheless, top-down approaches such as LBG [13] are applicable as well.

The bottom-up clustering procedure in Eigenspace can basically be outlined as follows:

```
start: assign 1 speaker to each
       initial cluster
repeat
  - calculate distances in Eigenspace
    between all M cluster
  - find 2 cluster with minimal
    distance D
  - join the 2 cluster
    M -> M-1
until M=C or
    termination criterion fulfilled
```

Clearly the algorithm can be further optimized, since - for example - not all cluster distances have to be recomputed in every iteration. The basic idea behind this algorithm is to iteratively join the most similar cluster. At the beginning each speaker is forming an initial unique cluster. In each iteration all distances from each cluster to all others have to be evaluated searching for the minimal distance. The 2 cluster having minimum distance are joined giving a new cluster containing the speakers from both cluster. The whole process is iterated until the desired number of cluster C is reached or some termination criterion is satisfied.

Binary Clustering:

The application of the Eigenvoice reduction implies another quite simple clustering scheme. After transforming the speaker coordinates are zero-mean and are distributed along the principal axes. This allows a binary clustering approach in Eigenspace. Along each axis a separating function can be introduced:

$$c_i = \begin{cases} 1 & \text{if } w_i \geq 0 \\ 0 & \text{else} \end{cases}$$

K Eigenvoices give K separating functions. Accordingly, K Eigenvoices allow the definition of 2^K cluster. All cluster boundaries are given by a change in sign of the Eigenvoice coefficients. The cluster id $I = 1 \dots K$ can be computed by:

$$I = 1 + \sum_{k=1}^{K} c_i 2^{k-1}$$

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

As base models for speaker clustering we used Gaussian Mixture Models [12, 14] with one GMM per speaker and 64 densities with diagonal variances per model. Model training was performed using MAP enrolment starting from a generic base model. Preprocessing included 12 MFCCs as well as delta-coefficients. Speech data were taken from the German Verbmobil database. In total, there were 613 male and female speakers. Eigenvoices were computed on the Gaussian mean parameters only, giving a original supervector dimension of $64 * 24 = 1536$.

4.2. Clustering Process in Eigenspace

In order to show the cabability of the algorithm, we evaluated a $K = 2$ Eigenvoices task. Figure 1 shows a scatterplot of these first 2 Eigenvoice coefficients. As clustering algorithm we used the bottom-up approach described beforehand. The clustering was strictly forward, i.e. no reassignment of speakers - except by joining of cluster - was allowed.

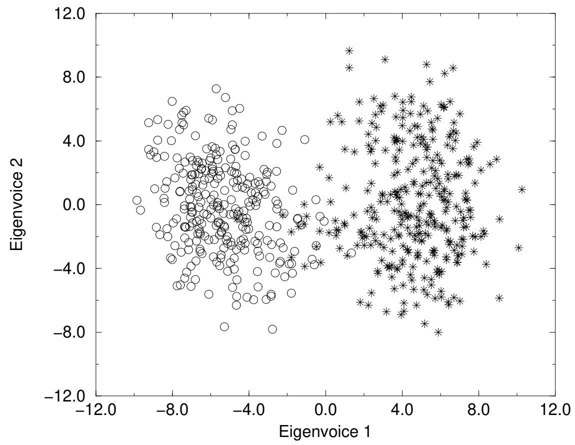

Fig. 1. Eigenspace with 2 dimensions shows strong separation in male ($*$) and female ($\circ$) speakers.

Figure 1 shows that 2 broad primary cluster are given by gender. Major information about gender is contained already in the first Eigenvoice coefficient. Similar results were reported in [15] and [16], although they used HMMs as base models for Eigenvoice computation.

The scatterplots in figure 2 depict some sample iterations of the clustering process. Depicted are iterations with $C = 4$ (upper figure) and $C = 8$ cluster. Cluster sizes seem well balanced.

4.3. Speech Recognition Performance

In order to examine clustering quality we performed several speech recognition experiments. Experiments were conducted on the German Verbmobil database using the Eval96 testset. For each speaker cluster an individual HMM set was trained for the recognizer using MAP. For each test utterance an individual HMM set was selected for the recognition pass. Our baseline setup allows basically several selection techniques. We were using an implicit selection based on the best HMM score. This setup requires parallel recognition passes with all HMM sets. Alternatively, a selection by a GMM classifier (1 GMM per speaker cluster) would also be possible.

Table 1 shows word error rates (WER) for different clustering approaches. "Bottom-up" depicts the proposed speaker clustering technique in Eigenspace, whereas "binary" indicates the binary Eigenvoice clustering scheme. The entries marked "SI" and "GMM" are used for comparison. SI is the baseline system with speaker-independent HMM models (only 1 HMM set). "GMM-orig" stands for a comparable standard clustering scheme working with speaker GMM models in original space. It uses a likelihood distance measure. Recognition results show equal performace for the bottom-up clustering approach in comparison to the standard scheme. Even slight improvements can be observed. Moreover,

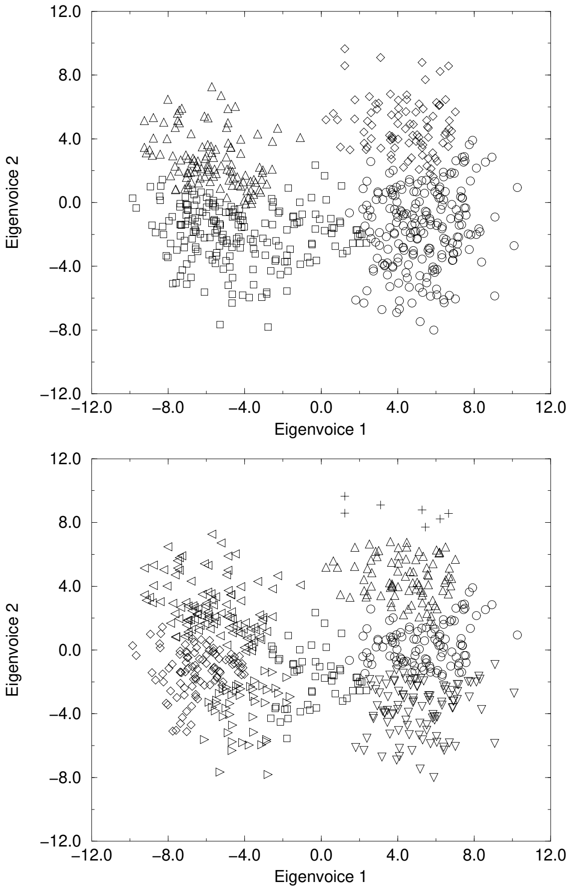

Fig. 2. Sample iterations in 2-dim. Eigenspace produce well balanced speaker cluster (4 and 8 cluster resp.).

the baseline system could be outperformed in any case. The key advantage is given by the speaker distance computation, which is far faster, much simpler and more robust.

clustering technique	#EV	#cluster	WER [%]
bottom-up	2	2	29.7
	2	4	29.2
	2	8	29.1
binary	2	4	30.7
none, SI (baseline)	-	(1)	31.2
GMM-orig	-	12	29.2

Table 1. Word error rates (WER) for different clustering techniques and different number of cluster.

5. CONCLUSION

In our paper we presented a speaker clustering approach working in speaker Eigenspace. By the application of the Eigenvoice transformation into a low-dimensional parameter space, a parameter reduction can be achieved allowing simple and robust distance computation between speakers. The clustering process of the algorithm has been shown on a sample task with 2 Eigenvoices. Overall performance has been compared with standard GMM-based clustering technique working in original model space. Speech recognition experiments have shown good performance, even improvements of the proposed clustering technique.

6. REFERENCES

[1] J.-L. Gauvain, C.-H. Lee, "Improved Acoustic Modeling with Bayesian Learning", Proc. ICASSP 92, pp. 481-484.

[2] J.-L. Gauvain, C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. Speech and Audio Processing, vol. 2, pp. 291-298, 1994.

[3] C. Leggetter, P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Proc. Comp. Speech Lang., vol. 9, pp. 171-185, 1995.

[4] V. Digalakis, L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods", Proc. ICASSP 95, pp. 680-683.

[5] V. Digalakis, D. Rtischev, L. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures", IEEE Trans. Speech and Audio Processing, vol. 3, no. 5, pp. 357-366, 1995.

[6] R. Kuhn, J. C. Junqua, P. Nguyen, N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space", IEEE Trans. Speech and Audio Processing, vol. 8, pp. 695-707, 2000.

[7] S. Furui, "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering", Proc. ICASSP 89, pp. 286-289.

[8] Y. Gao, M. Padmanabhan, M. Picheny, "Speaker Adaptation Based on Pre-Clustering Training Speakers", Proc. Eurospeech 97, pp. 2091-2094.

[9] S.E. Johnson, P.C. Woodland, "Speaker Clustering Using Direct Maximisation of the MLLR-Adapted Likelihood", Proc. ICSLP 98, Vol. 5, pp. 1775-1779.

[10] M.J.F. Gales, "Cluster Adaptive Training of Hidden Markov Models", Trans. on Speech and Audio Proc., vol. 8, no. 4, July 2000.

[11] O. Thyes, R. Kuhn, P. Nguyen, J.C. Junqua, "Speaker Identification and Verification using Eigenvoices", Proc. ICSLP 2000, paper no. 1155.

[12] D. A. Reynolds, R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech and Audio Processing, vol. 3, pp. 72-83, 1995.

[13] Y. Linde, A. Buzo, R. Gray, "An Algorithm for Vector Quantizer Design", IEEE Trans. on Communications, vol. 28, no. 1, pp. 84-95, 1980.

[14] R. Faltlhauser, G. Ruske, "Improving Speaker Recognition Using Phonetically Structured Gaussian Mixture Models", Proc. Eurospeech 2001, paper no. 1009.

[15] P. Nguyen, C. Wellekens, J.C. Junqua, "Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environments", Proc. Eurospeech 99, pp. 2519-2522.

[16] H. Botterweck, "Very Fast Adaptation for Large Vocabulary Speech Recognition using Eigenvoices", Proc. ICSLP 2000, paper no. 934.