ANISOTROPIC MAP DEFINED BY EIGENVOICES FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Henrik Botterweck

Philips GmbH Forschungslaboratorien
Weißhausstraße 2, 52066 Aachen, Germany
Henrik.Botterweck@philips.com

ABSTRACT

A general method is examined, which unifies the eigenvoice approach [1, 2, 4] and MAP adaptation. The *a priori* distribution for MAP is chosen to be anisotropic with the eigenvoices as preferred directions while still allowing adaptation into all other directions. This allows the exploitation of *a priori* knowledge about typical speaker variability within the MAP framework. This approach has two advantages: long term adaptation leads to the same good results as the MAP method, whereas for ultra-short adaptation in the range of 1–2 seconds an overfitting as for *maximum likelihood* techniques is avoided.

The method is applied to large vocabulary continuous speech recognition. Results are to be compared with our recent transfer of the *maximum likelihood* eigenvoice method to LVCSR [4].

Even after only one recognized word significant improvements of the WER of up to 6% relative are observed for gender independent recognition. 14% improvement are obtained after 5 seconds.

1. INTRODUCTION

The exploitation of *a priori* knowledge of typical speaker variations can considerably improve the fast adaptation of speaker independent models [1, 2].

In this paper, the eigenvoice method is revisited in the framework of a generalized *maximum a posteriori* adaptation technique, which includes as special cases the well known MAP-procedure as well as the maximum-likelihood algorithm used so far for eigenvoices.

The transfer of eigenvoice adaptation to continuous speech with a large vocabulary has been described in a previous article [4].

The general aim is to characterize speaker adapted models with a million and more parameters by some ten or hundred coefficients, which is only a fraction of the number of degrees of freedom of other fast adaptation methods like MLLR. The goal is a robust and generally applicable speaker adaptation within the first seconds or even the first spoken phonemes of continuous, unsupervised speech.

The main idea behind the eigenvoice method is the representation of speakers and their combined accoustic models as elements of a linear, affine space. A simple and successful approach is the concatenation of all parameters describing the speaker to a high dimensional vector. This is done here for all density means of all mixture distributions of a continuous HMM recognizer, hence leading to huge vectors of e.g. one million dimensions.

All speaker models of some training material being *points* in this space, it is now feasible to look for the principle axes of their distribution. These directions will describe the correlated variations of all model parameters for different types (e.g. the gender) of speakers. After some density observations for a yet unknown speaker all parameters can be adapted along these *a priori* known vectors.

However, trying to use a handful of recognized phonemes with their respective HMM densities to adapt millions of model parameters with a *maximum likelihood* criterion sometimes lead to considerable overfitting. On the other hand, by adapting only within the eigenvoice subspace, the WER nearly lowers to its optimum after 10–20s — further improvements have to be obtained by combining other long-term adaptation techniques such as MAP with the eigenvoices.

This is the reason to develop a MAP method, which incorporates *a priori* knowledge as represented by eigenvoices: it will allow a very fast and correlated movement of density means along some special (eigenvoice) directions, while still allowing arbitrary transformations into all other directions.

For this purpose the MAP adaptation formula has to be generalized for anisotropic gaussian *a priori* probability distributions in the high dimensional space of all density mean parameters.

2. ANISOTROPIC MAP ADAPTATION

Only the adaptation of the density means is examined here. The variances are assumed to be globally fixed, diagonal and — only for simplicity — equal.

Notation:

n_E is the number of used eigenvectors (*eigenvoices*).

n_M is the number of parameters (feature components) of one density mean.

$i(\tau)$ is the index of the recognized density at time τ.

$\underline{\mu}_\tau$: observation at time τ. This and the following vectors are interpreted as vectors in the high-dimensional space of all adapted parameters.

$\underline{\bar{\mu}}_i$: mean of all observations of the density i.

$\underline{\mu}_i^r$: mean of the density in the final vector, i.e. the adaptation result.

$\underline{\mu}_i^0$: mean of the density in the *a priori* vector, i.e. the speaker independent vector.

$\underline{E}_e$: 'eigenvoice' e.

$\beta = \frac{1}{\sigma^2}$: σ is the density variance (here globally constant).

$\alpha = \frac{1}{\sigma_0^2}$: adaptation parameter for all transformations which are orthogonal on the 'eigenvoices'. σ_0 is the assumed variance of speakers in this direction and should therefore be smaller than all observed eigenvoice variances.

$\epsilon_e = \frac{1}{\sigma_e^2}$: adaptation parameter for transformations parallel to the eigenvoice $\underline{E}_e$.

The MAP formulae to estimate means are known. Here they are derived in such a way, that it does not become necessary to make use of a complete base of the giant space of all parameters where any coordinate transformation is impracticable. The relations are derived exactly for the case of gaussian densities.

Applying MAP means the search for an element $\underline{\mu}^r$ of the parameter space P which maximizes the *a posteriori* probability of the observations X. The *a priori* distribution is defined by the eigenvoices $\underline{E}_e$ and variances σ_e:

$$p\left(\underline{\mu}^r\right) \quad \propto \quad \prod_i e^{\left\{-\frac{1}{2}\left(\underline{\mu}_i^r - \underline{\mu}_i^0\right) \underline{\underline{A}} \left(\underline{\mu}_i^r - \underline{\mu}_i^0\right)^{\mathrm{tr}}\right\}} \tag{1}$$

$$\underline{\underline{A}} \quad = \quad \alpha \underline{\underline{1}} + \sum_e \underline{E}_e^{\mathrm{tr}} \left(\epsilon_e - \alpha\right) \underline{E}_e \tag{2}$$

Therefore:

$$\underline{\mu}^r{}_{\mathrm{MAP}} = \arg\max_{\underline{\mu}^r} \; p\left(X|\underline{\mu}^r\right) \; p\left(\underline{\mu}^r\right) \tag{3}$$

$$= \arg\min_{\underline{\mu}^r} \sum_{\text{Obs. } \tau} \left(\underline{\mu}_{i(\tau)}^r - \underline{\mu}_\tau\right) \beta \underline{\underline{1}} \left(\underline{\mu}_{i(\tau)}^r - \underline{\mu}_\tau\right)^{\mathrm{tr}} + \sum_i \left(\underline{\mu}_i^r - \underline{\mu}_i^0\right) \underline{\underline{A}} \left(\underline{\mu}_i^r - \underline{\mu}_i^0\right)^{\mathrm{tr}} \tag{4}$$

A necessary condition for the searched minimum is, that the derivation with respect to a variational parameter ε vanishs:

$$\underline{\mu}^r \to \underline{\mu}^r + \varepsilon \Delta \underline{\mu}^r, \tag{5}$$

at $\varepsilon = 0$ for all densities i and arbitrary $\Delta\underline{\mu}^r$.

Introducing counts and mean values of single densities

$$N_j = \sum_{\text{Obs.} i(\tau)=j} 1 \qquad N_j \underline{\bar{\mu}}_j = \sum_{\text{Obs.} i(\tau)=j} \underline{\mu}_\tau \tag{6}$$

into these conditions leads to vector equations:

$$\sum_i \Bigg[N_i \beta \left(\underline{\mu}_i^r - \underline{\mu}_\tau\right) \; \underline{\underline{1}} + \alpha \left(\underline{\mu}_i^r - \underline{\mu}_{i(\tau)}^0\right) \; \underline{\underline{1}} \tag{7}$$

$$+ \sum_e \left(\epsilon_e - \alpha\right) \left(\underline{\mu}_i^r - \underline{\mu}_i^0\right) \; \underline{E}_e^{\mathrm{tr}} \underline{E}_e \Bigg] \Delta \underline{\mu}^{r\,\mathrm{tr}} = 0 \tag{8}$$

At first an arbitrary vector which is orthogonal on all eigenvoices is chosen for $\Delta\underline{\mu}^r$:

$$\underline{v}^\perp = \underline{v} \underbrace{\left(\underline{\underline{1}} - \sum_e \underline{E}_e^{\mathrm{tr}} \underline{E}_e\right)}_{:=\underline{\underline{P}}} \tag{9}$$

The operator $\underline{\underline{P}}$ projects onto the space orthonormal to all eigenvoice vectors. Hence it eliminates all contributions in (7) which contain the projector $\sum_e \underline{E}_e^{\mathrm{tr}} \underline{E}_e$. The residue can be separated into components as usual.

So it follows for each density, as in the case of standard MAP with exception of the additional projector $\underline{\underline{P}}$:

$$\left[N_i \beta \left(\underline{\mu}_i^r - \underline{\bar{\mu}}_i\right) + \alpha \left(\underline{\mu}_i^r - \underline{\mu}_i^0\right)\right] \underline{\underline{P}} = 0 \quad \forall i. \tag{10}$$

The solution of (10) is unique up to the constituents parallel to the $\underline{E}_e$. *One* choice is the MAP-solution:

$$\underline{\tilde{\mu}_i^r} \quad = \quad \frac{1}{N_i \beta + \alpha} \left(N_i \beta \underline{\bar{\mu}}_i + \alpha \underline{\mu}_i^0\right) \tag{11}$$

$$= \quad \underline{\mu}_i^0 + \left(1 - \frac{1}{1 + \frac{\beta}{\alpha} N_i}\right) \left(\underline{\bar{\mu}}_i - \underline{\mu}_i^0\right) \tag{12}$$

The projection onto the eigenvoice space is separated from this expression, consequently the complete solution of the original system of equations (7) can be written as:

$$\underline{\mu}^r = \underline{\mu}^0 + \left(\underline{\tilde{\mu}^r} - \underline{\mu}^0\right) \underline{\underline{P}} + \sum_e x_e \underline{E}_e \tag{13}$$

To determine the remaining n_E unknowns x_e, the variational equations (7) are now projected onto an eigenvoice $\underline{E}_f$. Once again the projection expressions are simplified, but now the system does not separate into independent components for single densities:

$$\sum_i \left[N_i \beta \left(\underline{\mu}_i^r - \underline{\bar{\mu}}_i\right) + \epsilon_f \left(\underline{\mu}_i^r - \underline{\mu}_i^0\right)\right] \; \underline{E}_f^{\mathrm{tr}} = 0 \quad \forall f \tag{14}$$

Here the *ansatz* (13) is included, resulting in a system of equations for n_E unknowns. Let $\underline{\underline{N}}$ be the diagonal matrix which contains in its components the number of observations N_i of the corresponding densities i:

$$\beta \left(\sum_e x_e \underline{E}_e + \underline{\mu}^0 + \left(\underline{\tilde{\mu}^r} - \underline{\mu}^0\right) \underline{\underline{P}} - \underline{\bar{\mu}}\right) \underline{\underline{N}} \underline{E}_f^{\mathrm{tr}} + \epsilon_f \sum_e x_e \underline{E}_e \; \underline{E}_f^{\mathrm{tr}}$$

$$= \left[\sum_e x_e \underbrace{\left(\beta \underline{E}_e \, \underline{\underline{N}} \, \underline{E}_f^{\mathrm{tr}} + \delta_{ef} \epsilon_f\right)}_{:=B_{ef}}\right] - \underbrace{\beta \left(\underline{\bar{\mu}} - \underline{\mu}^0 - \left(\underline{\tilde{\mu}^r} - \underline{\mu}^0\right) \underline{\underline{P}}\right) \underline{\underline{N}} \underline{E}_f^{\mathrm{tr}}}_{:=a_f}$$

$$= 0 \quad \forall f \tag{15}$$

Combining the unknowns to a (only n_E-dimensional) vector $\underline{x}$ leads to an expression for the solution:

$$\underline{x} = \underline{a} \, \underline{\underline{B}}^{-1} \tag{16}$$

The equations (11),(13) and (16) allow the explicit determination of the references which maximize the observation probability taking into account the known (gaussian) speaker variability.

3. EXPERIMENTS

The experiments are conducted using a combination of Wall Street Journal data and a collection of training and test speakers dictating general texts from an internal database. This material contains more data per speaker than WSJ and is well suited for the training of adapted models for each speaker.

Signal analysis was done applying LDA (Linear Discriminant Analysis) in order to obtain 33 features per frame. The frame shift of 10ms results in 100 frames per second available for adaptation. All tests are done with context dependent phonemes, generalized with CART. Using HMM-mixture models with laplacian densities this results in 10^6 density mean parameters available for the description and adaptation of speakers.

3.1. The determination of eigenvoices

For training, 50 speakers per gender, each with one hour of speech, are combined with 200 speakers from WSJ1 ($\approx$ 15 minutes per speaker). The means of 26.838 laplacian distributions, describing 1.975 mixture models, are obtained after a gender and speaker independent ('SI'-) training .

In the second step, optimized references for each training speaker are determined: three (WSJ speakers) resp. seven (additional speakers) iterations of supervised, combined MAP and MLLR adaptation are done, starting with the SI-models obtained before. For the final 300 ('SD')-results, each laplacian density can be related to its counterparts for the other speakers which have their origin in the same SI-density.

The principal axes of the distribution of these 300 vectors are determined with standard methods.

The obtained eigenvectors are related to the mean of the adapted models as their origin, not to the SI-models. Note that even if all speakers spoke the same texts these means will differ from the SI means, which are found as maximum likelihood solutions for the *combination* of all speakers. Consequently, adaptation will result in a density shift away from this origin. It will be shown below that these mean models are already superior to the SI ones. Because in this paper the behaviour of the eigenvoice method itself is examined, WER improvements are to be given relative to this lower rate.

3.2. Tests with unknown speakers

For the tests, samples of eight speakers not included in the training set are used, five females and three males. The adaptation is done with the first parts of varying length, 500ms–10s, of the first sentences of the test material. A lexicon with 34.000 entries is used for recognition.

The dependency of the WER on the amount of adaptation material is examined by extracting the models adapted so far after the first second of input speech and each following second until the tenth. Because only completely recognized words are used for the adaptation, the exact amount of data available for the first iterations differs from speaker to speaker.

Error rates are determined on the following 1.294 words per speaker. This results in a statistical incertainty of the given error rates of at most 0.35% if the errors were independent.

4. RESULTS AND DISCUSSION

The WER without adaptation are shown in tab.1 both for gender dependent and independent models. The initial eigenvoice models — the means of all models adapted to the training speakers — are already superiour to the *maximum likelihoo* trained ones. Consequently the former ones have been chosen to start the eigenvoice adaptation with.

Initial WER (%)			
	Max.-lik. train.	Origin of ev.	rel. (%)
gender indep.	20.23	18.90	-6.58
gender dep.	17.19	16.49	-4.04

Table 1. WER without adaptation for standard *maximum likelihood* training compared to the initial models for eigenvoice adaptation. All other results are compared to the better second ones.

To study its short-time behaviour the adaptation was performed in a quasi continuous manner: every 100 frames (1s) the uniquely recognized words up to that time have been used to update the eigenvoice coefficients.

	GI					GD	
EV:	10			100		10	100
w:	∞	300	100	300	100	300	300
0s	18.90					16.49	
1s	19.31	17.99	17.72	18.13	17.74	16.70	16.61
2s	17.62	17.57	17.34	17.39	17.18	16.23	16.02
3s	17.56	17.41	17.33	17.04	16.82	16.44	15.95
4s	17.43	17.40	17.39	16.53	16.73	16.43	16.00
5s	17.31	17.24	17.33	16.27	16.66	16.38	15.79
6s	17.23	17.15	17.14	16.12	16.51	16.41	15.75
7s	17.15	17.16	17.17	16.07	16.49	16.30	15.70
8s	17.16	17.20	17.04	15.92	16.33	16.30	15.55
9s	17.36	17.22	17.13	15.81	16.17	16.38	15.46
10s	17.25	17.19	17.23	15.90	16.13	16.37	15.44

Table 2. Absolute error rates for gender (in-)dependent (GI/GD) adaptation with 10 or 100 eigenvoices (EV) and different MAP-adaptation weights (w, ∞ means *maximum likelihood*) after 1–10s of unsupervised adaptation.

The mean relative WER for the test speakers after up to 10s of adaptation with ten eigenvoice degrees of freedom is shown in fig.1. The absolute error rates for this and the other tests are collected in tab.2. The results for different settings of the MAP-adaptation parameters are compared: a *maximum likelihood* adaptation criterion, obtained by setting all eigenvoice dependent MAP-parameters ϵ_e to values *near* infinity, and two varieties of weighting the PCA eigenvalues to get MAP-parameters in the order of 10^3. For this purpose the eigenvalues are simply scaled by appropriate factors.

In all cases most of the improvement is obtained during the first four seconds. But the MAP *damping* significantly improves the WER even after one and two seconds, corresponding to the very first recognized words. The *slow* MAP adaptation into directions orthogonal to the eigenvoices does not play a role on this time scale.

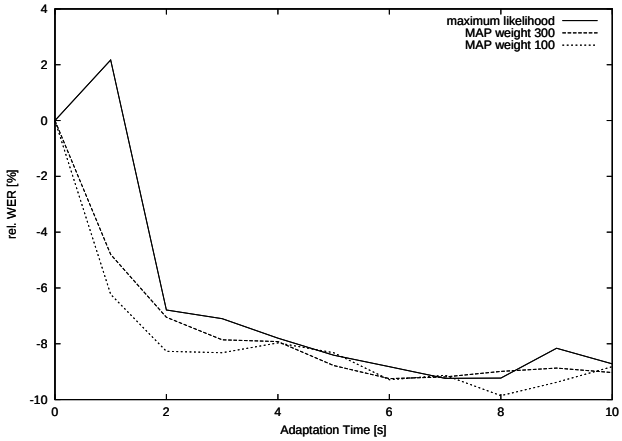

Fig. 1. *The mean WER for the test speakers during the first ten seconds of adaptation with ten eigenvoices with different weighting of the MAP parameters.*

The results for adaptation with 100 eigenvoices are shown in fig.2. The exploitation of these additional degrees of freedom improves the WER after 3–10s considerably, whereas during the first seconds the results are comparable to those obtained with ten eigenvoices. This reflects the fact, that the additional PCA-eigenvalues and hence the direction dependent MAP-adaptation coefficients are smaller and more time is needed to optimize them.

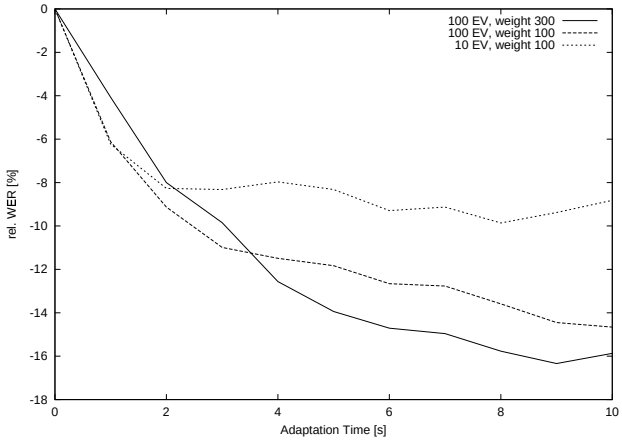

Fig. 2. *The mean WER for the test speakers during the first ten seconds of adaptation with 100 eigenvoices (compared with the best 10-eigenvoice test).*

For gender dependent adaptation 100 eigenvoices are necessary to get a significant improvement after 10s (fig.3). Just one second of speech (including some initial silence) does not lead to a lowered WER as it has been in the gender independent case. Obviously the ultra-fast adaptation within the first spoken phonemes is mostly due to the first eigenvoices, which have been shown to classify according to gender [4].

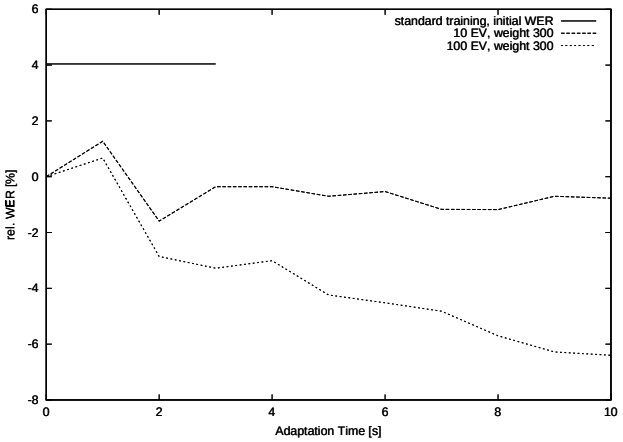

Fig. 3. *The mean WER for the test speakers with gender dependent initial models, 10 and 100 eigenvoices.*

5. SUMMARY

MAP adaptation with anisotropic gaussian priors has proven to significantly improve the recognition even after 1–2 seconds of unsupervised adaptation by up to 8% relative. This very fast adaptation seems to be effected by a nearly instantaneous optimization of the few gender depending eigenvoice coefficients.

For *long term* (10s) adaptation, it becomes worthwhile to use 100 instead of 10 or less eigenvoices. Then the word error rate can be reduced by 14% relative after $5s$ for gender independent recognition and 16% after $10s$.

The main advantage of the method as compared with e.g. fast MLLR techniques shows up during the first few seconds of adaptation. But obviously the huge main memory consumption is a problem for applications.

With this combination of MAP and eigenvoices within *one* adaptation method *a priori* knowledge on speaker dependent correlated movements of densities can succsessfully be applied to continuous, large vocabulary speech recognition although there are far more model parameters than in the case of single word recognition.

6. REFERENCES

[1] R.Kuhn, P.Nguyen, J.-C.Junqua, L.Goldwasser, N.Niedzielski, S.Fincke, K.Field, M.Contolini, *Eigenvoices for Speaker Adaptation*, Proc. ICSLP 1998, pp.1771–1774.

[2] R.Kuhn, P.Nguyen, J.-C.Junqua, R.C. Boman, N.A. Niedzielski, S.C. Fincke, K.L. Field, M. Contolini, *Fast Speaker Adaptation Using a priori Knowledge*, Proc. ICASSP 1999, pp.749–752.

[3] P.Nguyen, C.Wellekens, J.-C.Junqua, *Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environments*, Proc. Eurospeech 1999, pp.2519–2522.

[4] H.Botterweck, *Very fast Adaptation for large vocabulary continuous Speech Recognition using Eigenvoices*, Proc. ICSLP 2000 vol. IV, pp.354–357.