ABSTRACT

The principle of the eigenvoice method — using a priori knowledge on the speaker variability as collected during the training for a very fast adaptation — is applied to continuous speech recognition with large vocabulary. The handling of mixture density HMM models is discussed. For the case of gender independent models, a decrease of the word error rate of up to 15% is observed for unsupervised adaptation and even the first recognized phonemes lead to considerable improvements. The first two eigenvectors of adaptation can be characterized as classifying the gender and the recording environment. Comparisons of the method with MLLR are done as far as the latter is applicable at all.

1. INTRODUCTION

Recently it has been shown [1, 3] that the exploitation of a priori knowledge of typical speaker variations can considerably improve the fast adaptation of speaker independent models for single word recognition.

The transfer of this technique to continuous speech with a large vocabulary is described in this article. The question has to be answered, if it is possible to characterize speaker adapted models with even 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 acoustic 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, some difficulties have to be overcome before the method can be applied to continuous speech recognition with large vocabulary. These are mainly the assignments of densities belonging to the same mixture of different speakers to construct the linear space.

One of the central ideas presented in this paper is to adapt the training speakers from common speaker independent models to be able to identify corresponding densities of the same mixture for different speakers.

2. THE MAXIMUM LIKELIHOOD APPROACH TO EIGENVOICES

An unknown speaker is described by its estimated N eigenvoice coefficients. Given observations $\mathbf{q}(t)$ corresponding to densities with original means $\mu_j$. N coefficients $c_j$ are determined, which maximize the likelihood:

$$\frac{\partial}{\partial c_j} \sum_i \left( \mu_i(t) + \sum_j c_j \Delta_{\alpha j}(t) - \mathbf{q}(t) \right)^2 = 0 \quad \forall i \quad (1)$$

$\Delta_{\alpha j}$ denotes the part of the $j$th normalized eigenvoice vector describing the shift of the density $\alpha$.

Knowing the observation frequencies $N_\alpha$ of the densities results in a system of linear equations for the weights $c_j$:

$$\sum_j c_j \sum_\alpha N_\alpha \Delta^{ir}_{\alpha j} = \sum_\alpha \left( \mu_\alpha - \mu_i \right) N_\alpha \Delta^{ir}_{\alpha \alpha} \quad \forall i \quad (2)$$

This system of N unknowns can be solved without much computational effort during recognition.

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 $10^9$ density mean parameters available for the description and adaptation of speakers.

Some results on a very limited amount of adaptation data between 20 and 50 frames are given: one frame of length 25ms is calculated every 10ms.

3.1. The determination of eigenvoices

For training, 50 speakers per gender, each with one hour of speech, are combined with 200 speakers from WSJ ($\approx 15$ 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 [1, 2]. This takes approx. 1.5 hours on a 500 MHz Alpha CPU.

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-65s, of the first two spoken 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. Results are documented starting with 20 frames (300ms) of non-silence data.

Error rates are determined on the following 1,294 words per speaker. This results in a statistical uncertainty of the given error rates of at most 0.35% if the errors were independent.

The density variances are globally pooled and not taken into account during the adaptation. Generally, one could include them, but forming a positive rank-2 tensor (diagonal or not), they would have to be mapped into the used linear space in a useful manner.

4. RESULTS AND DISCUSSION

In Fig. 1 the eigenvalues of the principal axis decomposition of the training speakers are shown on a logarithmic scale. The overall order of magnitude of the eigenvalues as well as the eigen-speaker coefficients below depends on the chosen units in parameter space, only their relations are of interest here.

The ordered eigenvalues descend in good approximation like $1/x$ with two exceptions: the second eigenvalue is quite large and will be shown to discriminate between the two sources of training material. This fast decay allows to use only the (e.g. ten) most important eigenvectors. The last 100 eigenvalues drop off even faster due to the finite number of sample speakers.

![Figure 1: The PCA eigenvalues for the 300 adapted training speakers in descending order (double-logarithmic scale).](image1)

Fig. 2 shows the first two eigenvoice coefficients of the 300 training speakers. The first coefficient appears to discriminate between the genders, the second one between the WSJ sources and the additional speakers. Because of the applied time normalization on all squared FFT-features in the signal analysis this effect does not result from different energy densities in the frequency channels.
The adaptation of the first eigenvoice coefficients significantly lowers the error rate (Table 1) after three seconds. The use of hundred eigen vectors leads to further improvements after 12s and 40s (Table 2), but in the case of very fast adaptation ten vectors are sufficient.

| Table 1: Mean WER for various numbers of short-time adapted eigen voices (EV) after three seconds. The test speakers are chosen to differ significantly in speaking style and resulting error rates, but in no case there is a degradation by the eigenvoice adaptation. In all cases the sign of the first eigenspeaker coefficient after three seconds of adaptation is in accordance with the observations on the training material (Fig 2), hence the gender is correctly classified. The mean results for 0 eigen vectors and a longer adaptation time are given in Table 2. The obtained improvements of error rate can be compared to MLLR results as given in Table 3. For 10s of MLLR only one regression class has been determined successfully. After 40s the result is comparable to the 3s experiments with 10 eigen voices on the same amount of data.

**Table 1:** Mean WER for short-time adaptation with 10 eigen voices.

<table>
<thead>
<tr>
<th>EV</th>
<th>3s</th>
<th>12s</th>
<th>40s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10%</td>
<td>16%</td>
<td>20%</td>
</tr>
<tr>
<td>10</td>
<td>10%</td>
<td>12%</td>
<td>16%</td>
</tr>
<tr>
<td>50</td>
<td>10%</td>
<td>12%</td>
<td>16%</td>
</tr>
</tbody>
</table>

**Table 2:** Mean WER for long-time adaptation with 10 eigen voices.

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>12s</th>
<th>40s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>10</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>50</td>
<td>10%</td>
<td>12%</td>
</tr>
</tbody>
</table>

**Table 3:** Mean WER for long-time adaptation with 10 eigen voices.

<table>
<thead>
<tr>
<th>EV</th>
<th>12s</th>
<th>40s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>10</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>50</td>
<td>10%</td>
<td>12%</td>
</tr>
</tbody>
</table>

**Figure 1:** The short-time information of the training material.

**Figure 2:** The short-time information of the training material.

**Figure 3:** The short-time information of the training material.

**Figure 4:** The long-time information of the training material.

**Table 4:** Mean WER for long-time adaptation with 10 eigen voices.

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>12s</th>
<th>40s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>10</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>50</td>
<td>10%</td>
<td>12%</td>
</tr>
</tbody>
</table>
lated, but not reproducible by a linear or at least smooth transformation in parameter space (e.g. s–z versus ch–j). It has to be noted, that the other LDA features lack a simple, phonetical interpretation as it is the case for these two ones. For them the eigenvoice adaptations of the phonemes are less smoothly correlated.

The mean relative WER for the test speakers after up to 10s of adaptation is shown in Fig. 5. For the ten eigenvoice degrees of freedom used, most of the improvement is obtained during the first four seconds. A slight degradation at the very beginning can be explained by the use of the maximum likelihood projection criterion.

Figure 4: The means of all densities belonging to one center phoneme (●) and the adaptation of their two most significant LDA parameters for a female after 4 seconds (●).

Figure 5: The mean WER for the test speakers during the first ten seconds of adaptation with ten eigenvoices.

5. SUMMARY

It has been shown that a priori knowledge on speaker dependent correlated movements of densities can successfully be applied to continuous, large vocabulary speech recognition although there are far more model parameters than in the case of single word recognition.

The approach of assigning densities of different training speakers together, which have their origin in the same non-adapted density, has proven to work. The word error rate can be reduced by up to 15% relative, depending on the number of eigenvoice vectors and the adaptation time.

An appealing feature of the proposed adaptation method is its general principle: there are no a priori restrictions on the possible transformations as it is the case for MLLR. With this work it has been shown, that al-

6. REFERENCES