INVESTIGATIONS ON INTER-SPEAKER VARIABILITY IN THE FEATURE SPACE

R. Haeb-Umbach

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
haeb@pfa.research.philips.com

ABSTRACT

We apply Fisher variate analysis to measure the effectiveness of speaker normalization techniques. A trace criterion, which measures the ratio of the variations due to different phonemes compared to variations due to different speakers, serves as a first assessment of a feature set without the need for recognition experiments. By using this measure and by recognition experiments we demonstrate that cepstral mean normalization also has a speaker normalization effect, in addition to the well-known channel normalization effect. Similarly vocal tract normalization (VTN) is shown to remove inter-speaker variability. For VTN we show that normalization on a per sentence basis performs better than normalization on a per speaker basis. Recognition results are given on Wallstreet Journal and Hub-4 databases.

1. INTRODUCTION

A speech recognizer is called robust if it maintains its good recognition performance even if there is a mismatch between training and test conditions or if the acoustical environmental conditions are highly variable. One can try to achieve such robustness either by adaptation or normalization. In the first case a mismatch is reduced by adapting feature vectors or model parameters to the target environment. With normalization, on the other hand, the goal is to compute features or model parameters that are insensitive to undesired variations of the speech signal e.g. due to different channels or speakers. Normalization techniques are typically carried out in the front end of the recognizer. Examples include speech enhancement, such as spectral subtraction, robust feature extraction techniques, e.g. PLP [5], vocal tract normalization [7, 8] and various feature transformations, e.g. RASTA [6].

In this paper we present a statistical measure for the effectiveness of such speaker normalization techniques in the feature space. Applying Fisher Discriminant Analysis [3], we compute the ratio of feature variability due to different phonemes and due to different speakers. This tool allows an early assessment of a feature set without running time consuming recognition experiments.

By using this measure we can show that cepstral mean normalization also reduces inter-speaker variability, in addition to the channel normalization effect. Also, vocal tract normalization is shown to reduce inter-speaker variability. Since the measure is quite general, it can also be modified to measure other sources of variations.

2. A MEASURE OF INTER-SPEAKER VARIABILITY

Let $\xi_{r,c}$ denote a $D$-dimensional mean feature vector:

$$\xi_{r,c} = \frac{1}{N_{r,c}} \sum_{x(t) \rightarrow r,ph,c} x(t)$$

The sum is over all $N_{r,c}$ cepstral feature vectors $x(t)$, which are from speaker $r, r \in \{1, \ldots, R\}$ and which have been assigned by a time alignment to phoneme $c, c \in \{1, \ldots, C\}$. There is a total of $N = \sum_{c=1}^{C} \sum_{\{r|\rightarrow ph,c\}} 1 = \sum_{c=1}^{C} N_{c}$ such mean vectors. $\{r|\rightarrow ph,c\}$ denotes the set of speakers that have spoken phoneme $c$. There are $N_{c}$ such speakers. (For reasonable amounts of data per speaker, there is of course $N_{c} = R$ for all $c$, i.e. every speaker has spoken every phoneme.)

Now Fisher variate analysis [3] is applied, where we assume that a phoneme constitutes a class. The class-specific mean vectors are then

$$m_{c} = \frac{1}{N_{c}} \sum_{\{r|\rightarrow ph,c\}} \xi_{r,c}; c = 1, \ldots, C,$$

where the summation is over all speakers $r$ who have spoken phoneme $c$. The total mean is

$$m = \frac{1}{N} \sum_{c=1}^{C} \sum_{\{r|\rightarrow ph,c\}} \xi_{r,c} = \frac{1}{N} \sum_{c=1}^{C} N_{c} m_{c}$$
We now compute between-class and average within-
class sample covariance matrices:

\[
S_B = \sum_{c=1}^C \frac{N_c}{N} (m_c - m)(m_c - m)^T
\]

\[
S_W = \frac{1}{N} \sum_{c=1}^C \sum_{\{r \rightarrow ph,c\}} (\xi_{r,c} - m_c)(\xi_{r,c} - m_c)^T,
\]

where \( A^T \) denotes the transpose of a matrix \( A \).

One would like to have feature vectors such that all
vectors belonging to the same phoneme are close to-
gether in feature space, irrespective of the speaker, and
that they are well separated from the feature vectors
of the other phonemes.

An appropriate measure of this property is the deter-
minant or the trace of \( S_W^{-1} S_B \). The trace, for example,
is the sum of the eigenvalues \( \lambda_i \) of \( S_W^{-1} S_B \) and hence
the sum of the variances in the principal directions. It
can be interpreted as the radius of the scattering vol-
ume. In the experiments reported in this paper we com-
puted \( \sum_{i=1}^d \lambda_i \) for \( d = 1, \ldots, D \), i.e. the trace in a
\( d \)-dimensional subspace of the feature space, spanned
by the \( d \) principal discriminants.

The above class separability measure can now be
used to compare different feature sets. It may serve as
a first assessment of the quality of feature sets with-
out the need for recognition experiments. Indeed we
will see in the next sections that the higher the class
separability measure the lower is the recognition error
rate.

Note that we have derived a measure of inter-speaker
variability within phonemes. We could as well have
used hidden Markov Model states as classes, rather
than phonemes. Further one can study also other
sources of undesired variability, e.g. the variability
of phonemes between different databases or channels
by database- (channel-) specific rather than speaker-
specific mean vectors.

3. CEPSTRAL MEAN AND VARIANCE
NORMALIZATION

Cepstral mean and variance normalization are well-
known techniques to improve the insensitivity of the
feature vector to channel distortions.

The mean and variance normalized feature \( y_d(t) \) is
computed as follows:

\[
y_d(t) = \frac{x_d(t) - \bar{x}_d(t)}{\hat{\sigma}_d(t)}; d = 1, \ldots, D
\]

where \( d \) is the cepstral index, \( D \) being the number of
(static) features. \( \bar{x}_d(t) \) is an estimate of the mean and

Figure 1: Trace criterion versus vector dimension on
WSJ0 database (42 male and 42 female speakers) show-
ing the effect of cepstral mean and variance normaliza-
tion.

\( \hat{\sigma}_d(t) \) is an estimate of the standard deviation of
the input cepstral feature \( x_d(t) \). Both mean and variance
are computed over a block of frames, in our case over
the duration of one utterance.

Cepstral mean normalization (CMN) introduces a
spectral null at d.c. in the so-called modulation spec-
trum of speech \([6]\). It is well known that such an oper-
ation is able to remove an unknown constant channel
transfer function. This holds also for causal, recursive
forms of subband filtering, such as RASTA \([6]\).

However, it has also been observed that CMN can
improve recognition performance even if there are no
unknown channel transfer functions. Chen \([2]\) added a
bias term to the cepstral mean vectors and scaled the
cepstral variance to compensate for speaking style vari-
atations. In \([4]\) we reported 20% relative improvement of
the string error rate on the TI-Digits recognition task
due to CMN.

Figures 1 and 2 verify that CMN indeed suppresses
speaker characteristics: they show the value of the trace
criterion, on the WSJ0 (Wall Street Journal) database, \( \text{Fig. 1} \) for the whole 42 male and 42 female database,
and \( \text{Fig. 2} \) for the 42 male only subdatabase. One can
see that the trace criterion is much larger for feature
vectors after CMN compared to feature vectors with-
out CMN. These results are an indirect evidence for the
conjecture that low modulation frequencies, which are
suppressed by CMN, indeed contain speaker-specific
characteristics. For tasks where the speaker identity
should not be lost, e.g. speaker identification or speaker
clustering, these frequencies should therefore not be
suppressed.
Figure 2: Trace criterion versus vector dimension on WSJ0 database (42 male speakers) showing the effect of cepstral mean and variance normalization.

Comparing the two figures one observes that interspeaker variability is slightly larger on the 42m+42f database compared to the 42m male only subdatabase, which is in line with our expectations.

We also ran recognition experiments on the 5k development and evaluation test sets of Nov. 92 and 93 and noticed that CMN delivered a 25% relative error rate reduction, although these databases hardly contain any variation of the channel transfer function. All following investigations were therefore done with CMN switched on.

4. VOCAL TRACT NORMALIZATION

Vocal tract normalization (VTN) tries to reduce interspeaker variability by a speaker-specific frequency warping [7, 8]. Differences in vocal tract length are compensated for by a linear warping factor applied to the mel-frequency scale. We implemented the frequency warping by shifting the center frequencies of the mel-spaced filter bank. Let $k\Delta f_{mel}$, $k = 1, \ldots, K$, denote the $K$ center frequencies in mel. Then the center frequencies in Hz for a warping factor of $\alpha$ are:

$$f_{Hz}^{\alpha}(k\Delta f_{mel}) = \frac{1}{\alpha}700 \left(10^{\frac{k\Delta f_{mel}}{2000}} - 1\right)$$

The cepstral feature vectors are then computed from the accordingly arranged filter bank.

We employed VTN both in training and recognition. In training, the warping factor is estimated by computing the likelihood of the training data for feature sets obtained with different warping factors using HMM model sets with a low acoustic resolution. In recognition, we used a preliminary transcription of the test sentence obtained from a recognition based on unwarped features to estimate the warping factor, similar to [7]. For more details on the VTN setup, see [9].

Figures 3 and 4 show that the warped features exhibit lower inter-speaker variability, as measured by the trace criterion on the WSJ0 training database. Again, the gender-specific database (42 male speakers, Fig. 4) has less inter-speaker variability than the unspecific database (42 male and 42 female speakers, Fig. 3). The index "per speaker" denotes that a single warping factor is determined on all utterances of a training speaker.

Better results, however, were obtained when the warping factor was determined on a per sentence basis: for each sentence a separate warping factor is estimated, rather than a single warping factor for all sentences of a speaker. Information on which sentences stem from the same speaker is no longer required. Still the trace criterion is slightly larger. As shown in Table 1 the error rates achieved are also better.

Table 1: Word error rate (WER) and % deletions and insertions on WSJ 5k 92/93 dev/eval test sets for different VTN setups. 20 Male speakers only, bigram lm.

<table>
<thead>
<tr>
<th>VTN setup</th>
<th>del-ins</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>no VTN</td>
<td>2.1-0.9</td>
<td>9.9</td>
</tr>
<tr>
<td>$\alpha$ per speaker</td>
<td>2.1-0.9</td>
<td>9.5</td>
</tr>
<tr>
<td>$\alpha$ per sentence</td>
<td>2.0-0.8</td>
<td>9.3</td>
</tr>
</tbody>
</table>

Tables 2 and 3 summarize recognition results with VTN on the Hub-4 1996 development and evaluation
Figure 4: Trace criterion vs vector dimension on WSJ0 database (42 male speakers); comparison of unwarped and warped features (VTN).

test sets. In all cases we estimated the warping factor on a per segment basis (in Hub-4, the equivalent of a sentence). The achieved error rate reductions due to VTN were however considerably smaller, compared to WSJ.

Table 2: Word error rates in % on Hub-4'96 dev. set (male speakers only) for different vtn scenarios. Bigram lm, within-word models, no adaptation in recognition, partitioned evaluation.

<table>
<thead>
<tr>
<th>VTN setup</th>
<th>Focus condition</th>
<th>overall</th>
<th>F0</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>no VTN</td>
<td></td>
<td>36.5</td>
<td>17.1</td>
<td>36.5</td>
<td>45.1</td>
</tr>
<tr>
<td>VTN in trn and rec</td>
<td></td>
<td>35.3</td>
<td>16.4</td>
<td>35.3</td>
<td>42.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Focus condition</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>FX</th>
</tr>
</thead>
<tbody>
<tr>
<td>no VTN</td>
<td>33.7</td>
<td>29.3</td>
<td>36.6</td>
<td>61.2</td>
</tr>
<tr>
<td>VTN in trn and rec</td>
<td>30.5</td>
<td>29.7</td>
<td>34.1</td>
<td>62.4</td>
</tr>
</tbody>
</table>

While the results on the development test set were obtained in a so-called partitioned evaluation, the results on the evaluation test set were computed in an unpartitioned evaluation, i.e. the segment boundaries had been derived automatically (see [1] for a definition of the terms).

5. CONCLUSIONS

We have applied Fisher Discriminant Analysis to derive a measure of inter-speaker variability which "cor-

relates" well with the error rate. It allows a comparison of different feature sets at an early stage of the system development, i.e. without running extensive recognition experiments. Using this tool we have shown that cepstral mean normalization reduces inter-speaker variability, in addition to eliminating unknown channel transfer functions. This statistical result supports the conjecture from psychoacoustics that low modulation frequencies contain speaker-specific cues. Further, we have demonstrated that vocal tract normalization reduces inter-speaker variability and that a normalization on a per sentence basis performs better than a normalization on a per speaker basis. The tool we have used is very general and can also be modified to measure other sources of variability, e.g. variability due to different channels.

6. REFERENCES