Matching Training And Test Data Distributions For Robust Speech Recognition

Sirko Molau (molau@informatik.rwth-aachen.de)*

Daniel Keysers (keysers@informatik.rwth-aachen.de)

Hermann Ney (ney@informatik.rwth-aachen.de)

Lehrstuhl für Informatik VI

Computer Science Department

RWTH Aachen – University of Technology

52056 Aachen, Germany

*Corresponding Author

This manuscript has 81 pages and contains 10 figures and 10 tables.

Keywords

normalization , feature transformation , feature extraction , noise robustness , histogram normalization , feature space rotation

Abstract

In this work normalization techniques in the acoustic feature space are studied that aim at reducing the mismatch between training and test by matching their distributions. Histogram normalization is the first technique explored in detail. The effect of normalization at different signal analysis stages as well as training and test data normalization are investigated. The basic normalization approach is improved by taking care of the variable silence fraction. Feature space rotation is the second technique that will be introduced. It accounts for undesired variations in the acoustic signal that are correlated in the feature space dimensions. The interaction of rotation and histogram normalization is analyzed and it is shown that the recognition accuracy is significantly improved by both techniques on corpora with different complexity, acoustic conditions, and speaking styles. The word error rate is reduced from 24.6% to 21.8% on *VerbMobil II*, a German large vocabulary conversational speech task, and from 16.5% to 15.5% on *EuTrans II*, an Italian speech corpus of conversational speech over telephone. On the *CarNavigation* task, a German isolated-word corpus recorded partly in noisy car environments, the word error rate is reduced from 74.2% to 11.1% for heavy mismatch conditions between training and test.

Zusammenfassung

In dieser Arbeit werden Normalisierungsverfahren im akustischen Merkmalsraum untersucht, die zur Erhöhung der Robustheit von automatischen Spracherkennungssystemen führen. Zunächst wird die Histogrammnormalisierung behandelt. Der Effekt der Normalisierung auf verschiedenen Ebenen der Signalanalyse sowie auf Trainings- und Testdaten wird untersucht. Die Berücksichtigung des Anteils an Sprechpausen relaxiert eine der Grundannahmen der Histogrammnormalisierung. Ein Verfahren zur Merkmalsraumrotation beseitigt unerwünschte Variationen im Sprachsignal, die in den einzelnen Dimensionen des Merkmalsraumes korreliert sind. Zuletzt wird die Interaktion von Histogrammnormalisierung und Rotation untersucht. Beide Verfahren erhöhen die Erkennungsleistung auf Korpora mit unterschiedlicher Komplexität, unterschiedlichen akustischen Bedingungen und Sprechstilen. Auf *VerbMobil II*, einem deutschen Spontansprachkorpus mit großem Vokabular, sinkt die Fehlerrate von 24,6% auf 21,8%, und auf *EuTrans II*, einem italienischen spontansprachlichen Telefonkorpus, von 16,5% auf 15,5%. Beim *CarNavigation* Korpus, einem verrauschten deutschen Einzelwortkorpus, der zum Teil in fahrenden Autos aufgenommen wurde, sinkt die Fehlerrate bei starker Diskrepanz zwischen Trainings- und Testdaten von 74,2% auf 11,1%.

1 Introduction

The acoustic signal contains a lot of variability. On the one hand this is necessary to discriminate between different speech units (e.g. phonemes), but on the other hand there are also many variations in the speech signal which are irrelevant or even harmful for the recognition process. In a more general view these variations can be regarded as a mismatch between training and test conditions.

The main idea of the two normalization techniques discussed in this work is to reduce the mismatch between training and test by matching the distributions of the training and test data. In *histogram normalization*, the acoustic vector is transformed such that the cumulative histogram of each vector component matches a given reference histogram. In *feature space rotation*, the acoustic vector is rotated such that the principal axes with largest data scatter become identical for all conditions. Novel contributions of this paper are:

- *Signal analysis stages:* Unlike earlier investigations (e.g. Dharanipragada & Padmanabhan, 2000) both normalization techniques are applied at different signal analysis stages. It turns out that normalization at the filter bank stage is most efficient.

- *Training data normalization:* In previous works histogram normalization has been applied to test data only. In this paper we show in

theory and practice that normalization of training data in the spirit of speaker-adaptive training further improves the recognition accuracy.

- *Silence fraction treatment:* A method to enhance histogram normalization by treating the variable silence fraction of different conditions is presented. It yields consistently better recognition results than the basic normalization approach.

- *Feature space rotations:* A rotation based normalization technique is proposed and studied in detail which aims at matching principal feature space axes. In particular, it overcomes the assumption of uncorrelated features of histogram normalization.

- *Normalization under different mismatch conditions:* Histogram normalization, feature space rotation, and a combination of both techniques are applied on corpora with different degrees of mismatch between training and test. It is shown that the gain in recognition performance increases with the mismatch. In the case of major mismatch between training and test, the reduction of word error rate by both techniques is to some extent additive.

The paper is organized as follows: In Section 2, normalization and adaptation techniques are introduced and compared from a theoretical point of view. A short review of different feature space matching techniques is given, and the goals of this work are formulated.

The algorithms to be introduced in this work are evaluated on three corpora with a different degree of mismatch between training an test. These corpora and the recognition setup are presented in Section 3.

Histogram normalization is discussed in Section 4. It is investigated at what signal analysis stage histogram normalization should be applied, and how training data normalization helps to improve the recognition performance. An extension is proposed that treats variable silence fractions in the speech signal and thereby enhances the basic approach.

A technique that overcomes the assumption of uncorrelated features of histogram normalization by rotating the feature space is introduced and evaluated in Section 5.

The interaction between histogram and feature space rotation is analyzed in Section 6, and in Section 7 it is shown how both techniques and combinations of them perform on different test corpora.

A summary is given in Section 8.

2 Motivation

2.1 Normalization and Adaptation

There are different sources that add variability to the speech signal which is irrelevant for the speech recognition process (Sankar & Lee, 1995):

- varying transducers and transmission channels

- different speakers, speaking styles, or accents
- a varying ambient or channel noise

The aforementioned mismatch between training and test can be handled by means of *normalization* or *feature transformation*, i.e. by reducing the variability of the speech signal during signal analysis, or by *adaptation* or *model transformation*, which amounts to a transformation of the acoustic model to the conditions found in the test data. A schematic view is given in Figure 1. The left side depicts the feature space, i.e. the sequence of acoustic vectors X and its generation, and the right side shows the model space, i.e. the acoustic model θ and its training.

[Figure 1 about here.]

Three abstract data levels can be distinguished:

- the training data (first level) contain typically a collection of different conditions (e.g. different speakers, speaking styles, transmission channels, etc.)
- a particular test utterance (third level) has usually one specific condition (e.g. a specific speaker with a particular accent and vocal tract length, a specific transmission channel, etc.) which may or may not be present in the training data set and usually differs from the mixture of training conditions
- at the intermediate reference stage (second level) the variations caused

by different conditions are ideally removed (e.g. vocal tract length normalized or speaker-adapted data and models)

In this framework, speech recognition can be regarded as a combination of acoustic vectors and an acoustic model from specific data levels, and there is a mismatch if they do not belong to the same level.

In non-adaptive acoustic modeling the mismatch between test data X_{Test} and the acoustic model θ_{Train} can be important. Normalization transforms the acoustic vectors to a different level, whereas adaptation amounts to transforming the acoustic model onto a different level to overcome the mismatch.

Adaptation schemes like Maximum Likelihood Linear Regression (Leggetter & Woodland, 1995) and Maximum A-Posteriori estimation (Lee & Gauvain, 1986) are capable to adapt an acoustic model trained on different conditions directly to one specific test condition ($\theta_{Train} \rightarrow \theta_{Test}$). For this reason, adaptation is usually successful even when carried out in test only.

Normalization of acoustic vectors (e.g. Vocal Tract Length Normalization, Lee & Rose, 1996) results in a transformation into the reference form. For this reason, there is often a moderate gain in recognition accuracy if normalization is applied in test only, since a minor mismatch between $\tilde{X}$ and θ_{Train} remains. The best performance is typically achieved if both training and test data are normalized (no mismatch between $\tilde{X}$ and $\tilde{\theta}$, Lee & Rose, 1996, Welling & Kanthak$^+$, 1999).

Adaptation or normalization in training alone is counterproductive. In this case, the acoustic model is adapted to a reference condition, but cannot cope well with test conditions that deviate from the average (mismatch between X_{Test} and $\tilde{\theta}$, Welling & Kanthak$^+$, 1999, Gales, 2001).

A detailed mathematical formulation of normalization and adaptation is presented in the following section. It will show how adaptive acoustic modeling fits into the framework of statistical speech recognition.

2.2 Mathematical Framework for Adaptive Acoustic Modeling

Bayes' decision rule states that in automatic speech recognition the word sequence W should be chosen that maximizes the posterior probability of the observed sequence of acoustic vectors X. Using Bayes' identity and omitting the a-priori probability of the vector sequence, which has no impact on the optimization, the decision rule can be written as:

$$W = \arg\max_{W'} \{p(W') \cdot p(X|W')\} \tag{1}$$

During recognition, the product of the acoustic probability $p(X|W)$ and the language model probability $p(W)$ has to be maximized over all possible word sequences.

Typically a *Hidden Markov Model* (HMM) framework with parameters

θ is used for acoustic modeling (Eqn. 2). The acoustic probability of a vector sequence X given the HMM state sequence S is represented by the sum of the probabilities of all possible alignments s_1^T between HMM states and acoustic vectors. The probability of each alignment is given by the product over all time frames of the transition probability $p(s_t|s_{t-1}, W)$ and the emission probability $p(x_t|s_t, W; \theta)$ (Eqn. 3). In practice, the sum is often replaced by the maximum (*Viterbi* or *maximum approximation*, Eqn. 4):

$$
\begin{aligned}
p(X|W) &\rightarrow p(X|W;\theta) && (2)\\
&= \sum_{s_1^T} \prod_{t=1}^{T} [p(s_t|s_{t-1}, W) \cdot p(x_t|s_t, W; \theta)] && (3)\\
&\cong \max_{s_1^T} \prod_{t=1}^{T} [p(s_t|s_{t-1}, W) \cdot p(x_t|s_t, W; \theta)] && (4)
\end{aligned}
$$

The equations derived so far are employed in conventional non-adaptive modeling where no distinction is made under which condition the acoustic signal was recorded. It was shown before, however, that often the acoustic data from training and test do not match. They often originate from different conditions (different speakers, speaking styles, transmission channels, etc.) which can be expressed mathematically by a new condition-dependent parameter α (Eqn. 5). For simplicity it is assumed that only the emission probabilities are affected by the variable recording condition (Eqn. 6):

$$
\begin{aligned}
p(X|W;\theta) &\rightarrow p(X|W;\theta,\alpha) && (5)\\
&\cong \max_{s_1^T} \prod_{t=1}^{T} [p(s_t|s_{t-1}, W) \cdot p(x_t|s_t, W; \theta, \alpha)] && (6)
\end{aligned}
$$

To handle the unknown parameter, it is treated as a continuous valued hidden variable that has to be integrated out (Eqn. 7). To avoid integration problems, the maximum approximation is applied at this stage as well (Eqn. 8):

$$\begin{aligned} p(X|W;\theta) &= \int d\alpha \; p(X,\alpha|W;\theta) \\ &= \int d\alpha \; p(\alpha|W;\theta) \cdot p(X|W;\theta,\alpha) \qquad (7) \\ &\cong \max_{\alpha} \{p(\alpha|W;\theta) \cdot p(X|W;\theta,\alpha)\} \qquad (8) \end{aligned}$$

Thus, for adaptive modeling Bayes' decision rule in the maximum approximation can be rewritten as:

$$W = \arg\max_{W'} \left\{ p(W') \cdot \max_{\alpha} \left\{ p(\alpha|W';\theta) \cdot p(X|W';\theta,\alpha) \right\} \right\} \qquad (9)$$

The training corpus contains a number of conditions $r = 1, \ldots, R$ (i.e. different speakers, speaking styles, transmission channels, etc.). For each training condition r, we are given acoustic data X_r along with the transcriptions W_r. The parameter $\hat{\alpha}_r$ of the condition are determined by some function $h(\cdot)$ which is text-independent in the case of the normalization schemes proposed in this paper:

$$\hat{\alpha}_r = h(X_r) \qquad (10)$$

When the parameters $\hat{\alpha_r}$ are determined, the training data are normalized and the acoustic model $\tilde{\theta}$ is trained as usual by maximum likelihood:

$$\tilde{\theta} = \arg\max_{\theta} \prod_{r=1}^{R} p(X_r|W_r;\theta,\hat{\alpha}_r) \qquad (11)$$

In practice, the dependency of the acoustic model on the condition r is typically implemented by transformations, whereby the functional form of the acoustic model is usually fixed and only the transformation parameters are estimated from data. There are two possible realizations to match the actual condition with the reference condition and derive adapted probabilities $\tilde{p}(\cdot)$:

In normalization, the transformation $f_\alpha(\cdot)$ is applied to the acoustic vectors:

$$X \to \tilde{X} = f_\alpha(X) \tag{12}$$

$$\begin{aligned} p(X|W;\tilde{\theta},\alpha) &= \tilde{p}(f_\alpha(X)|W;\tilde{\theta}) \cdot \det\left(\frac{df_\alpha(X)}{dX}\right) \\ &= \tilde{p}(\tilde{X}|W;\tilde{\theta}) \cdot \det\left(\frac{d\tilde{X}}{dX}\right) \end{aligned} \tag{13}$$

Here, the Jacobian determinant of the transformation is included. However, in a classification task the Jacobian determinant can be omitted because the transformation is assumed to be independent of the word sequence W.

In adaptation, the inverse transformation is applied to the acoustic model (Eqn. 14). For notational simplicity we use the symbol $f_\alpha^{-1}(\cdot)$ for the inverse transformation:

$$\tilde{\theta} \to \theta = f_\alpha^{-1}(\tilde{\theta}) \tag{14}$$

$$\begin{aligned} p(X|W;\tilde{\theta},\alpha) &= \tilde{p}(X|W;f_\alpha^{-1}(\tilde{\theta})) \\ &= \tilde{p}(X|W;\theta) \end{aligned} \tag{15}$$

Even though adaptation and normalization are equivalent in this frame-

work, both techniques are relevant in practice. The main challenge of adaptive modeling is to find suitable transformation functions $f_\alpha(\cdot)$ or $f_\alpha^{-1}(\cdot)$ that can compensate for the effects of a certain condition, and to estimate their parameters reliably on adaptation data. In some cases, the transformation function $f_\alpha(\cdot)$ has a simple functional form and allows for efficient parameter estimation (e.g. spectral warping as in vocal tract length normalization, which depends on a single parameter to be estimated from data), whereas the corresponding inverse transformation function $f_\alpha^{-1}(\cdot)$ for acoustic model adaptation cannot be derived easily or is much more complex.

The techniques studied in this paper are normalization techniques, i.e. during signal analysis the acoustic vectors are mapped to the reference condition.

2.3 Normalization Techniques based on Feature Space Matching

There are a number of environmental and speaker-dependent variations whose impact on the speech signal are to some extent predictable. *Model based* techniques like vocal tract length normalization try to account for such variability. They are based on some model for speech production, transmission, or perception. A small number of model parameters are estimated on the adaptation data and applied according to the underlying model to account for the undesired variability.

If the effect of the environment is not predictable or too complex, normalization techniques that are independent of any model may be applied. These *data distribution based* techniques aim at transforming the acoustic vectors into a domain that is more suitable for automatic speech recognition. The transformation parameters are obtained from the distribution of the training and test data.

A number of data distribution based normalization techniques rely on the principal idea of mapping acoustic vectors from the test data space into the training data space to minimize the mismatch. The techniques proposed in the literature differ mainly in the following ways:

- the domain in which the mismatch is determined (feature or model space)
- the functional form of the transformation (parametric or non-parametric)
- the method for estimating the transformation function parameters (supervised or unsupervised)
- the signal analysis stage at which the feature space is matched

There have been numerous publications on supervised mapping in the spectral and cepstral domain. Matsukoto & Hirowo (1992), for example, proposed a piecewise-linear mapping of cepstrum vectors from test speakers into a reference space. The transformation was computed phoneme-wise and later smoothed to maintain continuity in the mapped space.

Neumeyer & Weintraub (1994) proposed a piece-wise linear mapping in the cepstrum domain that was unsupervised but relied on a small amount of simultaneous recordings on different channels (e.g. clean and noisy). The transformation was based on a set of multi-dimensional linear least-squares filters. Almost the same improvements over the baseline system were found regardless whether clean training data were mapped to the noisy test domain or vice versa.

Sankar & Lee (1995) investigated transformations in the feature and model space to minimize the mismatch between test utterances and the acoustic model. Their stochastic matching technique was unsupervised and did not require simultaneous recordings. However, it relied on knowledge about the functional form of the mapping. Sankar and Lee showed how the parameters of the transformation function that maximizes the likelihood of the data given the model or vice versa can be estimated by the Expectation Maximization algorithm. In connection with cepstral mean normalization, feature and model space transformations yielded approximately the same reduction in word error rate in their recognition tests.

Giuliani (1999) proposed another unsupervised technique to match the acoustic space of the training and test data, which could be used in on-line recognition. The main idea was to describe the training and the test data space by two Gaussian mixture models (GMMs). If the training data model was used as an initial estimate for the test data model, the update

of the densities during training could be interpreted as the mismatch between training and test, and used for subsequent feature space matching. In matched and mismatch cases, the performance of this technique was slightly inferior to incremental adaptation.

An unsupervised histogram-based mapping technique that makes no assumption about the functional form of the transformation was proposed by Dharanipragada & Padmanabhan (2000). It was based on the idea of mapping the cumulative density function of the test data to the cumulative distribution of the training data. Under certain assumptions this resulted in a simple text-independent histogram matching procedure which was non-parametric, non-linear and computationally inexpensive. The gain in recognition performance was of the same order as improvements achieved by unsupervised MLLR, and it was to a large extent additive.

Two years earlier, the same basic technique of mapping cumulative test data to training data distributions was successfully applied in speaker identification by Balchandran & Mammone (1998). Unfortunately, they evaluated the procedure not on real data but on artificially distorted data only. An additional smoothing factor had to be introduced to avoid over-compensation when the training and test speaker were not identical.

In another publication by Padmanabhan & Dharanipragada (2001), the histogram matching technique was further extended. Linear interpolation between the points of the non-linear mapping function reduced quantization

errors. In addition, a text-dependent extension was proposed, in which the mapping function was estimated in a maximum likelihood framework. The aim was to get robust estimates of the transformation with only a few adaptation sentences.

Hilger & Ney (2001) applied a parametric histogram normalization technique at the filter bank stage. Only four bins (quantiles) of the cumulative histograms were estimated, and piece-wise linear and power transformation functions were fitted to these bins according to a minimum squared error criterion. Normalization was applied in test only, and the reference histogram was averaged over all filter channels on the training data. It was shown that even single word utterances were sufficient to estimate the transformation function reliably, which made this technique useful for real-time applications. In informal experiments, other noise suppression techniques such as spectral subtraction, noise level normalization and reference adaptation were found to be less effective than histogram normalization.

In summary, the following conclusions can be drawn from the previous work:

- Many of the proposed feature space matching procedures were either supervised or required simultaneous recordings from the different environments. Under these idealized conditions, typically a large gain in recognition performance could be obtained.

- The transformation parameters were estimated in different spaces:

Most supervised techniques as well as the histogram-based methods relied on distributions of the training and test data. The stochastic matching of Sankar & Lee (1995) transforms the test data to better match the acoustic model, and the Gaussian mixture model based approach of Giuliani (1999) derives the transformation function from simplified acoustic models for training and test data.

- The stochastic matching technique of Sankar and Lee is unsupervised but makes assumptions about the functional form of the transformation.

- The Gaussian mixture model based approach of Giuliani relies on the assumption that the mismatch can be expressed by deviations of prototype vectors describing the training and test data space.

- The histogram-based method of Dharanipragada & Padmanabhan (2000) is unsupervised and non-parametric. As it relies on global statistics of the speech data, a large amount of adaptation data is required to estimate the transformation reliably.

- If the number of histogram bins is reduced and a parametric transformation function is fitted to the discrete histogram points, histogram normalization can be applied successfully even if only little adaptation data (e.g. single words) are available (Hilger & Ney, 2001).

- Performance gain by feature space matching is to some extent addi-

tive to adaptation schemes like maximum likelihood linear regression (Dharanipragada & Padmanabhan, 2000).

2.4 Goals of this Work

It has been shown that histogram normalization is a powerful technique that effectively improves the recognition accuracy in mismatch conditions (Dharanipragada & Padmanabhan, 2000). However, there are still a number of open issues and limitations that are addressed in this work.

So far, histogram normalization has been applied at either the filter bank stage or in the final acoustic feature space (in this case at the cepstrum stage). However, the particular stage of signal analysis at which it is most appropriate to apply histogram normalization has not been investigated before. Hence, in this work, the technique is applied at all possible stages to find out where it is most effective. Sequential normalization at different stages is also investigated.

The results published thus far were achieved with only the test data being normalized. However, an additional gain in recognition accuracy can be expected when the training data are also normalized (cf. Section 2.1). Hence, the effect of histogram normalization in both training and test is investigated in this study.

Histogram normalization is based on the assumption that the global statistics of the speech signal are independent of what is actually spoken,

and that the feature space is oriented such that the variations are uncorrelated in each dimension. The first assumption is relaxed by considering the variable silence fraction in the utterances. Furthermore, a new rotation based normalization scheme is introduced that matches the principal feature space axes with the largest data scatter and overcomes the second assumption of histogram normalization.

In previous work, feature space matching has proved to be an effective way of coping with large mismatch between training and test data (e.g. different microphones, recordings in clean vs. noisy environments). It might also reduce speaker and channel dependent variations in the speech signal. For these reasons, histogram normalization and feature space rotation is evaluated on corpora with different degrees of mismatch.

3 Corpora and Recognition Setup

In the following sections, recognition tests are presented for three corpora with different complexities (isolated words vs. continuous speech), acoustic conditions (office vs. telephone and car recordings), and speaking styles (planned vs. spontaneous speech). The statistics of the training and test corpora are summarized in Tables 1 and 2.

VerbMobil was a German speech-to-speech translation research project for spontaneous speech in the domain of appointment scheduling and information desk (Wahlster, 2000). In the framework of the VerbMobil project, a

corpus of German spontaneous speech was collected and annotated. Speech data were gathered at different sites over three channel types:

- close-talking microphones
- room microphones
- various telephone channels (mobile, wireline, wireless)

The training corpus used in this study consists of 49 hours of speech data collected with a close-talking microphone, and 14 hours collected with a room microphone. Recognition tests were carried out with a 10k-word vocabulary on the 1999 development test corpus DEV99B (out-of-vocabulary rate = 2.0%) that was recorded with close-talking microphones. The experiment is characterized by a minor domain and acoustic mismatch (most training data were collected in a different domain with a different close-talking microphone type).

EuTrans was a research project on example-based machine translation techniques for text and speech input in a traveler task domain (Casacuberta & Llorens$^+$, 2001). One work package involved the collection of an Italian spontaneous speech corpus over wireline telephone. Training and test data were recorded in the same environment, but the channel quality varied significantly between different recording sessions. Recognition tests were carried out with a 2k-word closed vocabulary on the final evaluation test set.

[Table 1 about here.]

CarNavigation is a German isolated-word database that was collected by the *Lehrstuhl für Informatik VI* of *RWTH Aachen* (Hilger & Ney, 2001). The training data were recorded in a quiet office environment with a close-talking microphone, and they consist of isolated words and spelling sequences. The closed-vocabulary test sets consist of isolated-word utterances recorded in various environments. The office test set was collected in the same conditions as the training data. Two other test sets were recorded in cars (city and highway traffic). Each test set consists of 2,100 equally probable unique words which were uttered only once. There is no overlap in vocabulary among the test sets, and between the training and test corpora.

[Table 2 about here.]

Recognition tests were carried out with the RWTH large vocabulary conversational speech recognition system that has been described in detail by Ney & Welling$^+$ (1998) and Sixtus & Molau$^+$ (2000). The optimized baseline setup can be summarized as follows:

- 15 / 20 channel filter bank (4 kHz / 8 kHz bandwidth)
- 12 / 16 Mel-frequency cepstral coefficients (4 kHz / 8 kHz bandwidth), full first derivatives, second derivative of the energy
- cepstral mean and variance normalization
- Linear Discriminant Analysis (LDA) on three adjacent cepstrum vectors including the derivatives, reduction to 25 / 33 dimensions (4 kHz

/ 8 kHz bandwidth)

- 2500 (VerbMobil II) / 1500 (EuTrans II) / 700 (CarNavigation) decision-tree based generalized triphone states plus one silence state
- gender-independent within-word acoustic models with up to 456k (VerbMobil II) / 96k (EuTrans II) / 22k (CarNavigation) Gaussian mixture densities
- 3-state (VerbMobil II) / 6-state (EuTrans II, CarNavigation) HMM topology with skip
- class-trigram (VerbMobil II) / trigram (EuTrans II) / zerogram (CarNavigation) language model
- in the case of CarNavigation, pruning was deactivated (full search) and the recognizer was forced to recognize exactly one word for each test utterance

4 Histogram Normalization

4.1 Principle

Histogram normalization (or histogram equalization) is a widely used technique in image processing, object recognition and computer vision (e.g. Ballard & Brown, 1982, pp. 70–71), but there have been only few applications in speech recognition so far.

The principal idea is as follows: Suppose the training and test data are distributed as depicted in the two-dimensional example feature space in Figure 2. There is a mismatch between the data sets (for the reasons discussed in Section 2.1. The mismatch will be especially prominent if there are major differences in the recording environments. Histogram normalization transforms the test to the training data distribution by mapping the depicted marginal distributions (Dharanipragada & Padmanabhan, 2000). In the generalized approach presented here, both training and test data are mapped to some pre-defined reference distribution.

[Figure 2 about here.]

Histogram normalization relies on two basic assumptions:

1. The global statistics of the speech signal are independent of what was actually spoken, i.e. the phoneme frequencies in training and test are similar.

2. The feature space dimensions are oriented such that the variations that are tackled by histogram normalization are uncorrelated in each dimension.

Under these conditions, each feature space dimension can be mapped independently of the others - a significant simplification.

The basic normalization algorithm is as follows: First, the reference histogram to which all data is mapped has to be defined. Usually the overall

distribution of the training data is used for reference. This choice is somewhat arbitrary, as various other distributions could be used as well. It can be argued, however, that the inherent distribution of the training data is a good choice to start with. For each feature space dimension (note that for notational simplicity the dimension index is omitted in all equations):

1. Compute a normalized histogram $\tilde{p}(x)$ on the full training corpus.

2. Compute the cumulative training data histogram $\tilde{P}(x)$ which becomes the reference histogram:

$$\tilde{P}(x) = \int_{-\infty}^{x} dx' \tilde{p}(x') \tag{16}$$

In the normalization step, the parameter set α_r (Eqn. 10) of the transformation function $f_\alpha(x)$ (Eqn. 12) has to be determined for each condition. In the case of histogram normalization, the condition-dependent distributions $p_r(x)$ and $P_r(x)$ are calculated. For each condition $r = 1, \ldots, R$ and each feature space dimension:

(3) Compute a normalized histogram $p_r(x)$ from all data X_r.

(4) Compute the cumulative condition-dependent histogram $P_r(x)$:

$$P_r(x) = \int_{-\infty}^{x} dx' p_r(x') \tag{17}$$

Finally, the transformation is applied to all data X_r from condition r:

(5) Replace each value x by $\tilde{x}$ that corresponds to the same point in the cumulative reference histogram (Figure 3):

$$\begin{aligned} x \to \tilde{x} &= f_\alpha(x) \\ P_r(x) &\stackrel{!}{=} \tilde{P}(\tilde{x}) \\ \tilde{x} &= \tilde{P}^{-1}(P_r(x)) \end{aligned} \tag{18}$$

Since the normalization depends on the acoustic data only, it amounts to an additional signal analysis step that is independent of training and test. From the transformed training data a normalized acoustic model $\tilde{\theta}$ is derived (Eqn. 11), and the transformed test data are used for recognition.

[Figure 3 about here.]

Histogram normalization has a number of convenient properties:

- it is text-independent and relies only on global statistics of the speech data
- it is a non-parametric, discrete approximation of a complex non-linear transformation function and makes no assumption about the functional form of the transformation
- once the histograms are calculated, histogram normalization can be implemented by a simple table lookup, so it is also computationally attractive

Histogram normalization can compensate for any scaling, shifting, or

non-linear distortion of each feature space dimension but, due to the assumption of uncorrelated features, it cannot compensate for possible feature space rotations. In the case depicted in Figure 2, basic histogram normalization will reduce the mismatch significantly but not remove it completely, because the feature space is rotated by a small amount.

4.2 Definition of the Acoustic Conditions

An important aspect of histogram normalization is the definition of the acoustic conditions r, for which a particular histogram $P_r(x)$ is estimated. The definition is task-dependent and has to meet the following requirements:

- there has to be enough data for each condition (typically one or more minutes) to estimate the histogram reliably
- each condition should contain data for only a single speaker in order to allow for the normalization of possible speaker-dependent variations in the speech signal
- the channel conditions should be constant to allow for the normalization of possible channel-dependent distortions

Estimating one histogram on the full test corpus meets the first requirement but violates the other two, whereas sentence-wise normalization would meet only the latter two requirements but not the first. Hence, in the following analyses, histogram normalization is applied either turn-wise,

i.e. a condition contains all utterances from one speaker in one conversation (VerbMobil II), or speaker-wise, if all data from a speaker were collected under identical channel conditions (EuTrans II, CarNavigation). The average amount of data available for estimating the condition-dependent histograms is listed under "average condition duration" in the corpus descriptions (Section 3).

The first requirement prevents the use of histogram normalization in online recognition tasks or on small data samples. There are two solutions if only a few seconds worth of adaptation data are available: Either a coarse histogram with fewer bins and appropriate interpolation in-between is estimated, or a parametric transformation function is applied whose parameters are estimated from histogram statistics. These approaches have been investigated in detail by Padmanabhan & Dharanipragada (2001), Hilger & Ney (2001) and Hilger & Molau$^+$ (2002) and are not pursued further in this work.

4.3 Histogram Normalization in Test only

In previous work, histogram normalization has been applied in test only. This is a special case of the generalized approach presented here. The overall distribution of the training data is used for reference as well, but only the test data are mapped to the reference histogram. The data from the individual training conditions and therefore also the acoustic model remain

unnormalized.

In Section 2.1 a theoretic explanation was given why normalization of the test data alone results often in moderate gain of recognition performance only, whereas full performance is achieved when both test and training data are normalized. The corresponding recognition tests are summarized in Table 3. Results are presented for the normalization at different signal analysis stages (discussed in more detail in the following section). As expected, the best results on the VerbMobil II corpus are obtained when both training and test data are normalized.

[Table 3 about here.]

4.4 Normalization Stages

Dharanipragada & Padmanabhan (2000) proposed a normalization of cepstral features. There are, however, a number of stages in the signal analysis front-end where histogram normalization may be applied:

- In the course of signal analysis, the speech waveform is transformed into a sequence of spectra by means of a Fourier transform. Each individual spectral line could be regarded as an independent distribution that needs to be normalized. For computational reasons it is more practical to apply histogram normalization after the filter bank, which leaves typically 15 or 20 (4 / 8 kHz bandwidth) distributions for normalization. As the logarithm is a monotone function, it makes no

difference whether histogram normalization is applied before or after the logarithm. In practice, spectral log compression before normalization helps to keep quantization errors small.

Histogram normalization of the log filter bank coefficients may help to reduce spectral distortions that are limited to certain frequency bands. It also normalizes the energy distribution in each frequency band.

- The mean of cepstral coefficients is typically subtracted in order to remove time-invariant channel transfer functions. In some tasks it also helps to scale cepstral coefficients to unity variance. Histogram normalization at the cepstrum stage, however, has a larger degree of freedom. It may not only shift and scale the distribution of each cepstral coefficient, but also distort it non-linearly.

- Linear discriminant analysis of cepstral coefficients and their time derivatives is a standard feature of the RWTH large vocabulary speech recognition system, since it consistently improves the recognition accuracy on all tasks. The LDA-transformed vector is the one that is finally presented to the speech recognizer. Hence, applying histogram normalization after linear discriminant analysis will normalize the distribution of acoustic test vectors to that observed during training of the corresponding acoustic model.

In addition, it is possible to apply histogram normalization sequentially

at different stages in a multi-pass scheme: After the reference distributions are defined, the condition-dependent histograms of the first normalization stage can be derived in a first signal analysis pass. In the next pass, the acoustic vectors can be normalized at the first stage, and the condition-dependent histograms for the second stage can be accumulated, etc. In the end the distributions of the acoustic vector components can be normalized at all stages.

As it is a-priori unknown at which stage of signal analysis histogram normalization performs best, or if there is a gain by sequential normalization at different stages, all three stages and combinations were tested. The results are summarized for the VerbMobil II corpus in Table 4.

[Table 4 about here.]

It turns out that histogram normalization performs well at the filter bank stage, and that there are only marginal improvements when normalization is performed on cepstrum or LDA-transformed vectors. A possible explanation is that most of the variations compensated for by histogram normalization are uncorrelated in the spectral domain - note that not the individual filter bank channels are supposed to be uncorrelated, but that the spectral distortions seem to be restricted to certain frequency bands.

The performance improvement of histogram normalization at different stages is to some extent additive, but the computational effort increases significantly due to the multi-pass signal analysis.

It can be observed that the histograms of most filter bank vectors, cepstral coefficients, and LDA-transformed vectors have a bimodal shape. So the original reference distributions may be replaced by mixtures of two Gaussian densities which smoothes data scatter efficiently and results in better modeling of outliers (Molau & Pitz$^+$, 2001). It could be shown that in this case normalization of log filter bank coefficients alone yields the same error rate of 22.5% as the sequential normalization at different stages, which is why further histogram normalization tests have been carried out at the log filter bank stage only.

4.5 Treatment of Silence

The first assumption of histogram normalization about the global statistics of the speech signal (cf. Section 4.1) is often violated. Even if enough speech data is available to ensure that the phoneme frequency is about the same for each condition, and even if the acoustic realization of the phonemes is identical, the histograms may still vary due to differences in the proportions of silence present in different data sets. This has a severe impact on conditions with a much lower or higher than average fraction of silence. On the one hand, histogram normalization will transform a number of acoustic speech vectors to silence and cause more deletions of words. On the other hand, some silence vectors will be transformed to speech and cause word insertions.

[Figure 4 about here.]

Figure 4 shows a histogram of the condition-wise silence fractions in the VerbMobil II corpus. Non-speech events like hesitations or transcribed noise items are considered as "speech" in this context. The average silence fraction is 17%, but the number varies between 3% and 76% for individual conditions.

Two possible solutions to the problem rely on having separate reference histograms for speech and silence. In the first solution, two streams of acoustic vectors are fed into the speech recognizer. One of them is adapted to the speech, the other to the silence histogram. During recognition it is known at each point in time, if the current state hypothesis belongs to speech or not. The corresponding acoustic vector can be chosen for likelihood calculations. A disadvantage of this approach is the discontinuity of the acoustic vectors at each speech/silence boundary introduced by the different reference histograms. The same problem occurs if a speech/silence detector is used prior to recognition.

A conceptually simpler solution pursued here is to determine the silence fraction of each condition r beforehand and create condition-dependent reference histograms $\tilde{P}_r(x)$ from the speech and silence histogram that are adapted to the observed silence fraction. In this approach, the discontinuity is avoided and the speech recognizer needs no modifications.

To obtain the speech and silence histograms, a forced alignment with

the reference transcription is carried out on the training data. All acoustic vectors mapped to the silence mixture are accumulated in the silence histogram $\tilde{p}_{sil}(x)$. All other vectors are accumulated in the speech histogram $\tilde{p}_{sp}(x)$. It can be seen that the bimodal structure of most histograms observed earlier (Molau & Pitz$^+$, 2001) is in fact a manifestation of speech and silence (Figure 5). The first peak can be almost completely attributed to silence frames, whereas the second peak is mainly caused by more energetic speech frames.

[Figure 5 about here.]

In the normalization step, the silence fraction γ_r of the actual training or test condition r has to be determined first. For the training data, this is obtained as before by forced alignment with the reference transcription. Since in test the correct transcription is unknown, the silence fraction has to be calculated either in a preliminary recognition pass (two-pass recognition) or with a dedicated speech/silence detector (e.g. as described in Macherey & Ney, 2002).

For each condition $r = 1, \ldots, R$, an adapted reference histogram $\tilde{P}_r(x)$ is computed by linear interpolation between the speech and silence histograms. Note that the same result is obtained whether the normalized histograms $\tilde{p}_{sil}$ and $\tilde{p}_{sp}$ are interpolated before the cumulative histogram is computed, or whether the cumulative histograms $\tilde{P}_{sil}$ and $\tilde{P}_{sp}$ are interpolated (Eqn. 19).

The latter approach is computationally more efficient, though:

$$\tilde{P}_r(x) = \int_{-\infty}^{x} dx' \tilde{p}_r(x') = \gamma_r \cdot \tilde{P}_{sil}(x) + (1 - \gamma_r) \cdot \tilde{P}_{sp}(x) \quad (19)$$

$$\tilde{p}_r(x) = \gamma_r \cdot \tilde{p}_{sil}(x) + (1 - \gamma_r) \cdot \tilde{p}_{sp}(x) \quad (20)$$

$$\tilde{P}_{sil}(x) = \int_{-\infty}^{x} dx' \tilde{p}_{sil}(x') \qquad \tilde{P}_{sp}(x) = \int_{-\infty}^{x} dx' \tilde{p}_{sp}(x') \quad (21)$$

The adapted reference histograms $\tilde{P}_r(x)$ are used for normalization of training and test data as in the basic histogram normalization approach (cf. Section 4.1). As an example, Figure 6 shows the reference histogram of the third log filter bank coefficient for three different silence fractions. The left histogram adapted to a silence fraction of 10% is most similar to the original histogram for speech and silence (Figure 5, left), because this value is closest to the average silence fraction of the training corpus. The larger the silence fraction, the more prominent becomes the first "silence" peak, whereas the second "speech" peak loses significance.

[Figure 6 about here.]

Recognition test results for histogram normalization of log filter bank coefficients are summarized in Table 5. They show that the treatment of the silence fraction helps to further improve the recognition performance. In fact, on some corpora, histogram normalization only improved the recognition accuracy in connection with silence fraction treatment (see Section 7).

[Table 5 about here.]

5 Feature Space Rotation

5.1 Motivation

The second basic assumption of histogram normalization is that the feature space dimensions are uncorrelated with respect to the variations accounted for. Previous experiments (cf. Section 4.4) have suggested that this requirement is best met at the filter bank, since histogram normalization performs best at this signal analysis stage. Still the feature space might not only be distorted and translated, but also rotated by a small amount (e.g. Figure 2), which would not be treated properly by histogram normalization. In the following we will propose a transformation that is able to handle this type of mismatch between training and test data.

Just as in histogram normalization, training and test data of different conditions shall be transformed to some reference condition in order to reduce undesired variations in the speech signal. However, instead of mapping the axes of the feature space independently of each other, a linear transformation shall be applied to the complete acoustic vector. The aim is to reduce the differences between the condition-dependent covariance matrices in training and test.

To account for the type of mismatch depicted in Figure 2, the transformation will be restricted to be a rotation, which changes the orientation of the feature space axes but preserves Euclidean distances. We will further re-

strict the rotation to consist of elementary rotations that only map principal feature space axes.

Let us first consider a pathological case of an approximately "circular" feature space where the reference and the condition-dependent covariance matrices are nearly diagonal with identical values. In this case, the eigenvectors will be oriented arbitrarily and the eigenvalues are all similar, which would result in undesired arbitrary rotations for different conditions. If, on the other hand, the feature space is elongated, i.e. if the scatter is non-uniform in different directions, at least some eigenvectors are well-defined. For this reason, we will sort the eigenvectors in descending order of their eigenvalues and apply a number of elementary rotations. Only the first condition-dependent eigenvectors with dominantly larger eigenvalues will be mapped to their corresponding reference eigenvectors. Note that if all eigenvectors are considered at the same time, the transformation is identical to a principal component analysis, computed independently for training and test conditions.

5.2 Principle

Just as in histogram normalization, the reference condition has to be defined first. We will use the covariance matrix $\tilde{\Sigma}$ obtained from the full training corpus as reference. The corresponding D orthonormal reference eigenvec-

tors $\tilde{v}_1, \ldots, \tilde{v}_D$ and eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_D$ are defined by:

$$\tilde{\Sigma}\tilde{v}_d = \tilde{\lambda}_d\tilde{v}_d \qquad \tilde{v}_d \in \mathbb{R}^D, \quad ||\tilde{v}_d||^2 = 1, \quad d = 1, \ldots, D \tag{22}$$

$$\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \ldots \geq \tilde{\lambda}_D \geq 0 \tag{23}$$

The eigenvectors are sorted in descending order of their corresponding eigenvalues (Eqn. 23).

During normalization, the covariance matrix Σ_r of each training and test condition $r = 1, \ldots, R$ is computed from data X_r. Note that for improved readability the condition index r is omitted in all following equations.

The condition-dependent orthonormal eigenvectors and eigenvalues are calculated and sorted in the same way as the reference:

$$\Sigma v_d = \lambda_d v_d \qquad v_d \in \mathbb{R}^D, \quad ||v_d||^2 = 1, \quad d = 1, \ldots, D \tag{24}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D \geq 0 \tag{25}$$

Note that the eigenvectors are unique except for a scale factor of ± 1. If the sign of each component of an eigenvector is inverted, the same eigenvector basis is obtained with one axis pointing into the opposite direction. Here we choose the condition-dependent eigenvectors v_d such that the angles to the corresponding reference eigenvectors $\tilde{v}_d$ are less than or equal to 90 degrees, i.e. such that the dot product of all eigenvector pairs is positive:

$$\tilde{v}_d \cdot v_d \geq 0 \qquad d = 1, \ldots, D \tag{26}$$

A transformation matrix U_D that rotates all D condition-dependent eigenvectors to their corresponding reference eigenvectors is obtained by the

product of two eigenvector matrices $\tilde{V}$ and V (Eqn. 27). The first matrix is made of the reference eigenvectors $\tilde{v}_1, \ldots, \tilde{v}_D$, and the second is made of the condition-dependent eigenvectors $v_1, \ldots, v_D$:

$$\begin{aligned} U_D &= \tilde{V} \cdot V^T \qquad U_D, \tilde{V}, V \in \mathbb{R}^{D \times D} \\ \tilde{V} &= \left(\tilde{v}_1, \ldots, \tilde{v}_D \right) \\ V &= \left(v_1, \ldots, v_D \right) \end{aligned} \tag{27}$$

The matrix U_D is of little use, however, because we expect that only the direction of the first few eigenvectors is well-defined as described in the previous section. A transformation matrix that maps the first eigenvectors only will be constructed stepwise. First, the rotation matrix $\hat{U}_1$ to map the first condition-dependent eigenvector v_1 to the first reference eigenvector $\tilde{v}_1$ is derived. The rotation angle η_1 between the two eigenvectors is computed from their dot product, as they are of unit length:

$$\eta_1 = \arccos(\tilde{v}_1 \cdot v_1) \tag{28}$$

Since the two eigenvectors are not orthogonal, the Gram-Schmidt algorithm is applied to v_1 in order to obtain an orthonormal basis vector $\hat{v}_1$ lying in the same plane of rotation:

$$\hat{v}_1 = \frac{v_1 - (\tilde{v}_1 \cdot v_1) \cdot \tilde{v}_1}{||v_1 - (\tilde{v}_1 \cdot v_1) \cdot \tilde{v}_1||^2} \tag{29}$$

Next, the acoustic vector is projected onto the plane spanned by $\tilde{v}_1$ and

$\hat{v}_1$ with the projection matrix J_1^T:

$$J_1^T = \left(\hat{v}_1, \tilde{v}_1\right)^T \qquad J_1^T \in \mathbb{R}^{2\times D} \tag{30}$$

It is rotated within the plane with the rotation matrix R_1 (Eqn. 31) by the angle η_1, and projected back into the original $\mathbb{R}^{D\times D}$space with the transposed projection matrix $J_1 \in \mathbb{R}^{D\times 2}$:

$$R_1 = \begin{pmatrix} \cos\eta_1 & \sin\eta_1 \\ -\sin\eta_1 & \cos\eta_1 \end{pmatrix} \qquad R_1 \in \mathbb{R}^{2\times 2} \tag{31}$$

Finally, a correction term $I - J_1 J_1^T$ with the identity matrix I has to be applied that restores the dimensions orthogonal to the plane of rotation lost in the first projection. It ensures that all these dimensions remain unchanged. The full rotation matrix $\hat{U}_1$ is derived by:

$$\hat{U}_1 = J_1 R_1 J_1^T + I - J_1 J_1^T \tag{32}$$

Since eigenvectors are orthogonal, it is possible to repeat the procedure sequentially for further eigenvector pairs. Each new transformation will have no impact on previous feature space rotations.

To compute the rotation matrix for the second pair of eigenvectors, the condition-dependent eigenvector v_2 is rotated by $\hat{U}_1$, and the corresponding orthonormal basis vector $\hat{v}_2$ is computed (Eqn. 33). Next the rotation angle η_2 (Eqn. 34) and the second rotation matrix $\hat{U}_2$ are derived (Eqn. 35). It rotates the feature space in the plane spanned by the second condition-

dependent and the second reference eigenvector after the application of $\hat{U}_1$:

$$\hat{v}_2 = \frac{\hat{U}_1 v_2 - (\tilde{v}_2 \cdot \hat{U}_1 v_2) \cdot \tilde{v}_2}{||\hat{U}_1 v_2 - (\tilde{v}_2 \cdot \hat{U}_1 v_2) \cdot \tilde{v}_2||^2} \tag{33}$$

$$\eta_2 = \arccos(\tilde{v}_2 \cdot \hat{U}_1 v_2) \tag{34}$$

$$\hat{U}_2 = J_2 R_2 J_2^T + I - J_2 J_2^T \tag{35}$$

$$J_2 = \left(\hat{v}_2, \tilde{v}_2\right) \tag{36}$$

$$R_2 = \begin{pmatrix} \cos\eta_2 & \sin\eta_2 \\ -\sin\eta_2 & \cos\eta_2 \end{pmatrix} \tag{37}$$

The third and further eigenvectors can be mapped in the same way:

$$\hat{v}_d = \frac{U_{(d-1)} v_d - (\tilde{v}_d \cdot U_{(d-1)} v_d) \cdot \tilde{v}_d}{||U_{(d-1)} v_d - (\tilde{v}_d \cdot U_{(d-1)} v_d) \cdot \tilde{v}_d||^2} \tag{38}$$

$$\eta_d = \arccos(\tilde{v}_d \cdot U_{(d-1)} v_d) \tag{39}$$

$$\hat{U}_d = J_d R_d J_d^T + I - J_d J_d^T \tag{40}$$

$$J_d = \left(\hat{v}_d, \tilde{v}_d\right) \tag{41}$$

$$R_d = \begin{pmatrix} \cos\eta_d & \sin\eta_d \\ -\sin\eta_d & \cos\eta_d \end{pmatrix} \tag{42}$$

$$U_d = \hat{U}_d \hat{U}_{(d-1)} \hat{U}_{(d-2)} \ldots \hat{U}_1 \tag{43}$$

The product of all rotation matrices $\hat{U}_1, \ldots, \hat{U}_d$ (Eqn.43) gives the resulting condition-dependent transformation matrix U_d that maps the first d eigenvectors. U_d is equivalent to the parameter set α_r (Eqn. 10) of the transformation function $f_\alpha(x)$ (Eqn. 12). It is applied to normalize the acoustic

vectors by:

$$\begin{aligned} x \to \tilde{x} &= f_\alpha(x) \\ &= U_d \cdot x \end{aligned} \tag{44}$$

In the D-dimensional feature space, up to $D-1$ rotations may be carried out. The last dimension matches automatically due to the orthogonality constraint, which is a nice consistency check for the procedure. The deviation angle η_D between the rotated Dth condition-dependent eigenvector and the Dth reference eigenvector needs to be zero, and the resulting rotation matrix $U_{(D-1)}$ must be identical to the matrix U_D derived by equation Eqn. 27.

5.3 Experimental Results

Feature space rotation may be applied at the same signal analysis stages as histogram normalization (cf. Section 4.4). If applied at the filter bank it now makes a difference whether the feature space is normalized before or after log compression. Rotation before log compression may result in negative coefficients, which prevents the successive application of the logarithm. For this reason, rotation was only applied after log compression.

[Figure 7 about here.]

To calculate the reference condition, the covariance matrix $\tilde{\Sigma}$ and the eigenvector basis $\tilde{v}_1, \ldots, \tilde{v}_D$ are computed on the full training corpus (Eqn. 22). It turns out that at the log filter bank stage the first eigenvalue is signif-

icantly larger than all others as shown in Figure 7 for the VerbMobil II training corpus. Note that the logarithm *increases* the scatter of the filter bank coefficients, as these are typically small. The feature space has apparently one preferred direction with large scatter, and along the other principal axes data scatter is much smaller (Molau & Hilger$^+$, 2002). For this reason, we started with recognition tests where only the first condition-dependent eigenvector v_1 is mapped to the first reference eigenvector $\tilde{v}_1$.

During normalization, the covariance matrix Σ and the eigenvector basis $v_1, \ldots, v_D$ are calculated for each training and test condition $r = 1, \ldots, R$ (cf. Section 4.2). The first condition-dependent eigenvector v_1 deviates typically by a few degrees from the direction of the first reference eigenvector $\tilde{v}_1$ as shown in Figure 8. The figure depicts a histogram over the condition-wise deviation angles η_1 (Eqn. 28) calculated for log filter bank vectors on the VerbMobil II training corpus.

[Figure 8 about here.]

Finally, the condition-dependent rotation matrix U_1 for the first eigenvector is derived (Eqn. 32) and the training and test data are transformed (Eqn. 44). A normalized acoustic model is trained (Eqn. 11), and the normalized test vectors are used in recognition.

Recognition test results for feature space rotation at different signal analysis stages are summarized in Table 6. Given are the word error rate, the mean ratio of the first two reference eigenvalues $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$, and the mean

deviation angle $\overline{\eta_1}$ between the first condition-dependent eigenvectors v_1 and the first reference eigenvector $\tilde{v}_1$.

[Table 6 about here.]

Rotation of log filter bank vectors yielded a clear improvement in recognition accuracy. The gain was as large as for basic histogram normalization, but smaller than for histogram normalization with silence fraction treatment.

At the cepstrum stage, the ratio of the first two reference eigenvalues was much smaller than for the log filter bank. Consequently, the principal axes were not as well-defined, the deviation angles increased significantly, and the recognition performance dropped. A detailed analysis of recognition errors revealed that the performance improved for most conditions with minor rotation angles, but the word error rate almost doubled for a few test conditions with rotation angles η_1 close to 90 degrees, indicating a change in the order of eigenvectors.

When rotating the feature space after linear discriminant analysis, the mean ratio of the first and second reference eigenvalues and the average deviation angle of the first eigenvectors were similar to the values observed for the log filter bank stage. The reference covariance matrix $\tilde{\Sigma}$ was diagonal and the corresponding eigenvector matrix $\tilde{V}$ was the identity matrix, which results from the property of linear discriminant analysis to decorrelate the feature space dimensions. Normalization after LDA gave the

same minor performance improvement than histogram normalization at this stage (cf. Table 4), but it was inferior to rotation at the log filter bank stage. Hence, the outcome was comparable to histogram normalization which performed best at the log filter bank stage as well, and further tests were carried out at this signal analysis stage only.

In a next set of experiments, the number of reference eigenvectors mapped to their corresponding reference eigenvectors was increased. Matching more than the first eigenvector further reduces the mismatch between the condition-dependent covariance matrices Σ and the reference covariance matrix $\tilde{\Sigma}$. However, limits are set by the discrete order of eigenvectors. The smaller the differences between subsequent eigenvalues, the larger is the chance that the order of eigenvectors changes and for some conditions principal axes are mapped that represent different acoustic characteristics.

[Figure 9 about here.]

In practice it turned out that even the direction of the second principal axis is not well defined, and that the order of eigenvectors changes for different conditions. The rotation angles η_d for the second and further pairs of eigenvectors increase significantly as shown for the VerbMobil II corpus in Figure 9. On other corpora they soon became as large as 90 degree, and a rotation was not sensible.

The corresponding recognition test results are summarized in Table 7. They show that mapping more than the first eigenvector does not increase

the recognition accuracy any further. In fact, whereas on the VerbMobil II corpus the word error rate was of the same order when the first two or three eigenvectors were mapped, the performance significantly deteriorated in these cases on the other corpora.

[Table 6 about here.]

6 Combination of Histogram Normalization and Rotation

6.1 Motivation

Since feature space rotation overcomes one of the principal limits of histogram normalization, it is interesting to see if the gain in recognition performance obtained by applying both techniques at the log filter bank stage is additive.

The natural order of normalization would be to rotate the acoustic vectors first for optimal orientation of the feature space dimensions, and then normalize the distribution of each dimension. On the other hand, histogram normalization with silence fraction treatment gives a larger gain in recognition performance than feature space rotation. The deviation angles η_1 between the first condition-dependent eigenvectors v_1 and the first reference eigenvector $\tilde{v}_1$ are significantly reduced when histogram normalization is applied before rotation (Figure 10), which could make the estimation of the

rotation plane and angle more reliable and give superior results. From this perspective, feature space rotation would be concerned with mismatch in the condition-dependent distributions that remains after histogram normalization.

[Figure 10 about here.]

6.2 Experimental Results

[Table 8 about here.]

Recognition test results with either normalization technique in different order are summarized in Table 8. Better results were achieved when histogram normalization was applied before feature space rotation. The word error rate was lower than with feature space rotation alone (23.0%), but not as low as with histogram normalization with silence fraction treatment (21.8%). Both techniques seem to account for the same speech signal variations (which can be seen by the reduced rotation angles, Figure 10), but histogram normalization is more efficient.

7 Normalization under Different Mismatch Conditions

Histogram normalization with silence fraction treatment, feature space rotation to map one eigenvector, and a combination of both techniques at the log filter bank stage has been evaluated on different corpora to analyze the

performance of these techniques under various degrees of mismatch between training and test data.

Recognition test results for the VerbMobil II corpus with a minor acoustic mismatch were presented in Tables 5, 6 and 8. Histogram normalization with silence fraction treatment reduced the word error rate by 11% relative and feature space rotation by 7% relative. A combination of both techniques gave no further gain in recognition performance.

[Table 9 about here.]

Recognition test results for the EuTrans II corpus are summarized in Table 9. Both the training and test data were recorded over wireline telephone, so that there is no explicit acoustic mismatch. Basic histogram normalization without silence fraction treatment gave no improvement in recognition accuracy on this corpus. The transmission channel was more noisy and showed larger variations from one condition to the next, which is why deviations from the average silence fraction may have had a larger impact on the recognition accuracy. Silence fraction adapted histogram normalization yielded a relative error rate reduction of 5%, and feature space rotation improved the recognition performance by a similar amount. The reductions of both techniques were again not additive.

The CarNavigation database is a task with large mismatch conditions. The training data were recorded in a quiet office environment, and two of the test sets were recorded in cars (city and highway traffic). In scenarios

with such a mismatch there is much room for improvements. A standard normalization technique is cepstral variance normalization. On this task, it lowered the recognition accuracy in the clean office condition, but clearly improved the baseline result for the city and highway test sets (Table 10).

Histogram normalization reduced the word error rate significantly both with and without variance normalization. Better results were obtained without this extra normalization step. The variance of the filterbank channels is already implicitly normalized when the feature space dimensions are mapped onto the same reference histogram, which is why a further transformation to unity cepstral variance may be counterproductive. Without variance normalization, the word error rate was reduced between 10% relative (office) and 81% relative (highway).

[Table 10 about here.]

When feature space rotation was applied, the rotation angles for the test data increased with the mismatch. Whereas on the office data the mean rotation angle $\overline{\eta_1}$ was 6 degrees, it increased to 23 degrees on the city and 32 degrees on the highway data.

In connection with cepstral variance normalization, feature space rotation even outperformed histogram normalization slightly. The reduction in word error rate varied between 10% relative (office) and 47% relative (highway). Without variance normalization, however, rotation gave only small improvements over the baseline system. This supports the notion that even

though similar variations are accounted for by histogram normalization and feature space rotation, the mismatch is reduced in a different way. In particular the variance of the acoustic signal cannot be handled properly by feature space rotation alone. This comes as no surprise, as the feature space axes are only rotated but not scaled.

Note that cepstral variance normalization was always applied *after* filter bank normalization (e.g. histogram normalization and feature space rotation) in the tests reported here. Variance normalization of filter bank channels performs in general significantly worse.

If applied in the right order, feature space rotation and histogram normalization together performed always better than both normalization techniques alone. The best result on the office test data was achieved when rotation was applied before histogram normalization, and on the mismatch city and highway data when applied afterwards. The experiments show that in general the normalization method that gives most gain in recognition performance should be applied first.

8 Summary

Histogram normalization and feature space rotation are normalization techniques in the acoustic feature space. They aim at reducing the mismatch between training and test data by mapping different conditions (different speakers, speaking styles, transmission channels, etc.) to some reference

condition. They are model-free and text-independent, i.e. they only rely on global statistics of the speech signal.

Histogram normalization is widely used in image processing, but the application in automatic speech recognition has been largely unexplored. In this work it was shown that histogram normalization is conceptually simple but improves the recognition accuracy on a variety of corpora. It can be applied to different signal analysis stages and performs best when log filter bank vectors are transformed. Normalization of training and test data was superior to normalization of test data alone, for which a theoretic explanation was given. The larger the acoustic mismatch between the recording conditions in training and test, the larger was the gain in recognition performance. This suggests that histogram normalization can reduce channel and environmental variations very efficiently.

Histogram normalization relies on the assumptions that global statistics of the acoustic signal are the same independently of what is spoken. This requirement was relaxed by the new approach of explicit silence fraction treatment. It was shown that the recognition accuracy is significantly improved if the reference histogram is adapted to the silence fraction of each condition.

Feature space rotation is a normalization technique proposed to relax the assumption of histogram normalization regarding the orientation of the feature space axes. It aims at reducing the mismatch between condition-

dependent covariance matrices and a reference covariance matrix. For this purpose, transformation matrices were derived that map the principal axes with largest data scatter.

It was shown in this work that the feature space is not uniform at different signal analysis stages, but has one preferred direction with especially large scatter. Feature space rotation to match the first eigenvector performed best at the log filter bank stage similar to histogram normalization. Matching subsequent principal axes with large scatter did not improve the recognition accuracy any further. On the three corpora under investigation, the reduction in word error rate by feature space rotation was typically somewhat lower than the reduction by histogram normalization with silence fraction treatment.

In the case of a major mismatch between training and test data, further improvements of recognition accuracy were achieved by a combination of histogram normalization and feature space rotation. Best results were obtained when the normalization method that performs best on its own was applied first.

References

R. Balchandran, R. J. Mammone, 1998: Non-Parametric Estimation and Correction on Non-Linear Distortion in Speech Systems. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 749–752,

Seattle, WA, May 1998.

D. H. Ballard, C. M. Brown, 1982: *Computer Vision*. Prentice-Hall, Englewood Cliffs, NJ.

F. Casacuberta, D. Llorens, C. Martnez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, J. M. Vilar, 2001: Speech-To-Speech Translation based on Finite-State Transducers. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 613–616, Salt Lake City, UT, May 2001.

S. Dharanipragada, M. Padmanabhan, 2000: A Nonlinear Unsupervised Adaptation Technique for Speech Recognition. Proc. *Int. Conf. on Spoken Language Processing*, Vol. VI, pp. 556–559, Bejing, China, Oct. 2000.

M. J. F. Gales, 2001: Adaptive Training for Robust ASR. Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, 6 pages, CD ROM, IEEE Catalog No. 01EX544, Dec. 2001.

D. Giuliani, 1999: An On-Line Acoustic Compensation Technique for Robust Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, Vol. VI, pp. 2487–2490, Budapest, Hungary, Sept. 1999.

F. Hilger, H. Ney, 2001: Quantile Based Histogram Equalization for Noise Robust Speech Recognition. Proc. *European Conf. on Speech Communi-*

cation and Technology, Vol. II, pp. 1135–1138, Aalborg, Denmark, Sept. 2001.

F. Hilger, S. Molau, H. Ney, 2002: Quantile Based Histogram Equalization For Online Applications. Proc. *Int. Conf. on Spoken Language Processing*, Vol. I, pp. 237–240, Denver, CO, Sept. 2002.

C.-H. Lee, J.-L. Gauvain: Bayesian Adaptive Learning and MAP Estimation of HMM. In: C.-H. Lee, F. K. Soong, K. K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition*, Kluwer Academic Publishers, Boston, MA, pp. 83–107, 1996.

L. Lee, R. Rose, 1996: Speaker Normalization using Efficient Frequency Warping Procedures. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 353–356, Atlanta, GA, May 1996.

C. J. Leggetter, P. C. Woodland, 1995: Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer, Speech and Language*, Vol. 9, pp. 171–185, April 1995.

W. Macherey, H. Ney, 2002: Towards Automatic Corpus Preparation for a German Broadcast News Transcription System. Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. I, pp. 733–736, Orlando, Florida, May 2002.

H. Matsukoto, I. Hirowo, 1992: A Piecewise Linear Mapping for Supervised

Speaker Adaptation. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 449–452, San Francisco, CA, March 1992.

N. Mirghafori, E. Fosler, N. Morgan, 1995: Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes. Proc. *European Conf. on Speech Communication and Technology*, Vol. I, pp. 491–494, Madrid, Spain, Sept. 1995.

S. Molau, M. Pitz, H. Ney, 2001: Histogram Based Normalization in the Acoustic Feature Space. Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, 4 pages, CD ROM, IEEE Catalog No. 01EX544, Dec. 2001.

S. Molau, F. Hilger, D. Keysers, H. Ney, 2002: Enhanced Histogram Normalization in the Acoustic Feature Space. Proc. *Int. Conf. on Spoken Language Processing*, Vol. I, pp. 1421–1424, Denver, CO, Sept. 2002.

L. Neumeyer, M. Weintraub, 1994: Probabilistic Optimum Filtering for Robust Speech Recognition. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 417–420, Adelaide, Australia, April 1995.

H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel, 1998: The RWTH Large Vocabulary Continuous Speech Recognition System. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 853–856, Seattle, WA, May 1998.

M. Padmanabhan, S. Dharanipragada, 2001: Maximum Likelihood Non-linear Transformation for Environment Adaptation in Speech Recognition Systems. Proc. *European Conf. on Speech Communication and Technology*, Vol. IV, pp. 2359–2362, Aalborg, Denmark, Sept. 2001.

A. Sankar, C.-H. Lee, 1995: Robust Speech Recognition Based on Stochastic Matching. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, pp. 121–124, Adelaide, Australia, April 1995.

A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, H. Ney, 2000: Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. III, pp. 1671–1674, Istanbul, Turkey, June 2000.

W. Wahlster (Ed.), 2000: *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer Verlag: Berlin, Heidelberg, New York, 2000.

L. Welling, S. Kanthak, H. Ney, 1999: Improved Methods for Vocal Tract Normalization. Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. II, pp. 761–764, Phoenix, AZ, March 1999.

List of Figures

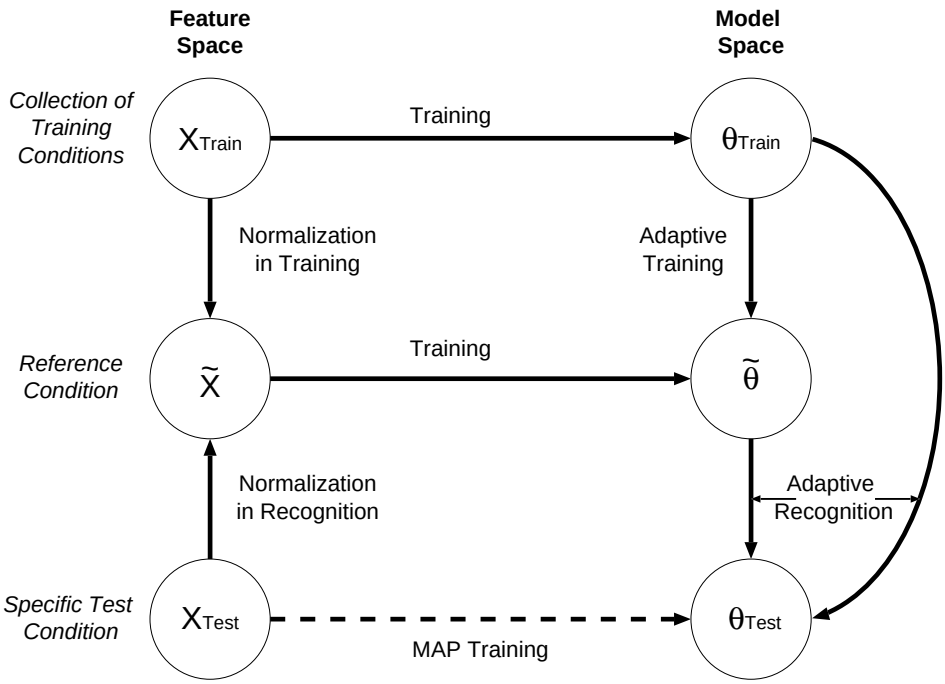

Figure 1: Overview of normalization and adaptation concepts.

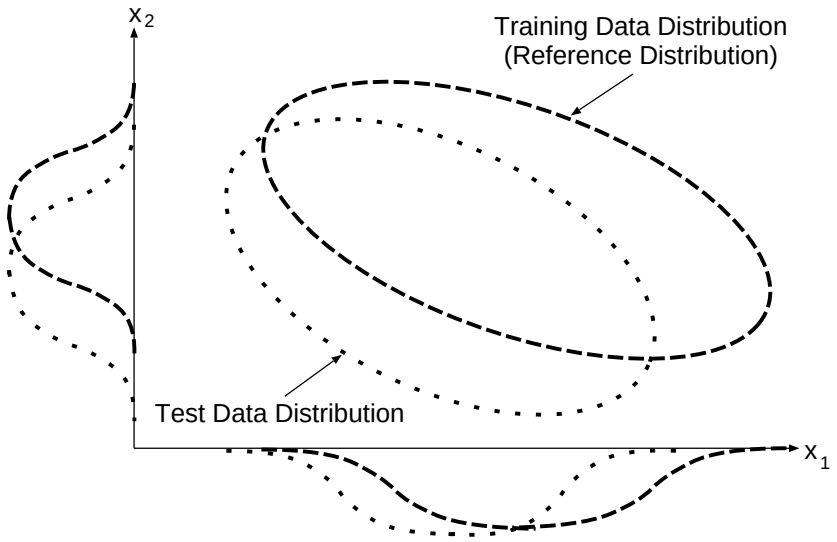

Figure 2: Schematic distribution of training and test data in a two-dimensional example feature space. The marginal distributions are plotted along both axes.

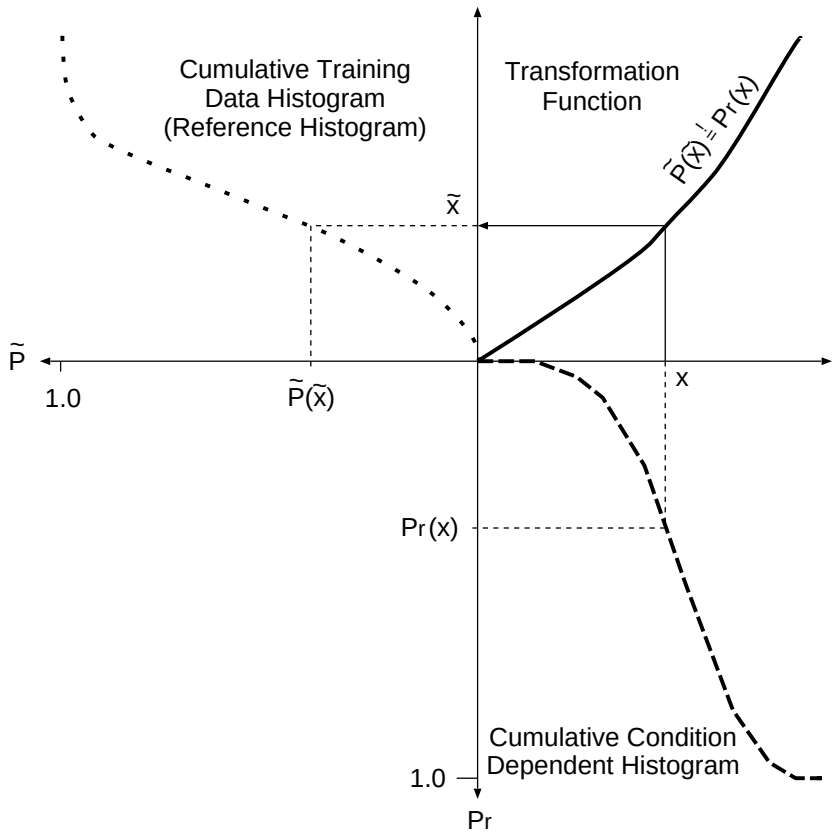

Figure 3: Principle of histogram normalization: Data X_r from condition r are transformed such that the cumulative condition-dependent histogram $P_r(x)$ matches the cumulative reference histogram $\tilde{P}(\tilde{x})$.

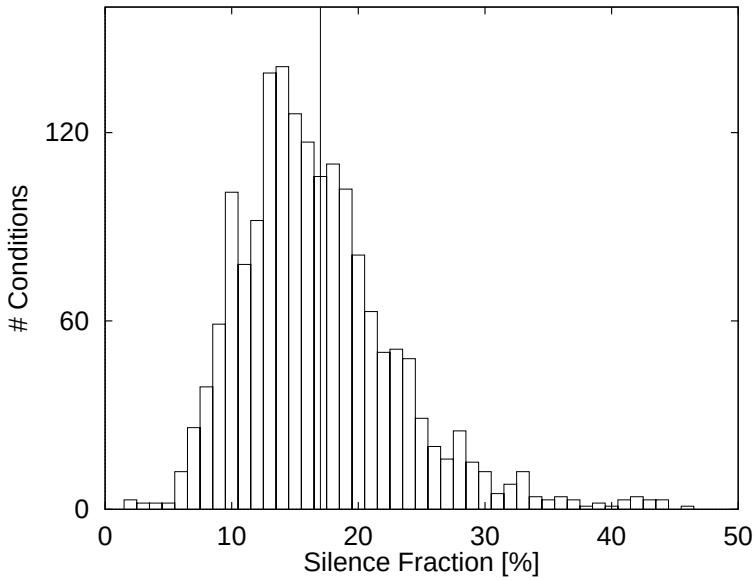

Figure 4: Histogram over the silence fractions of individual conditions in the VerbMobil II training corpus. The vertical line marks the average silence fraction of 17%.

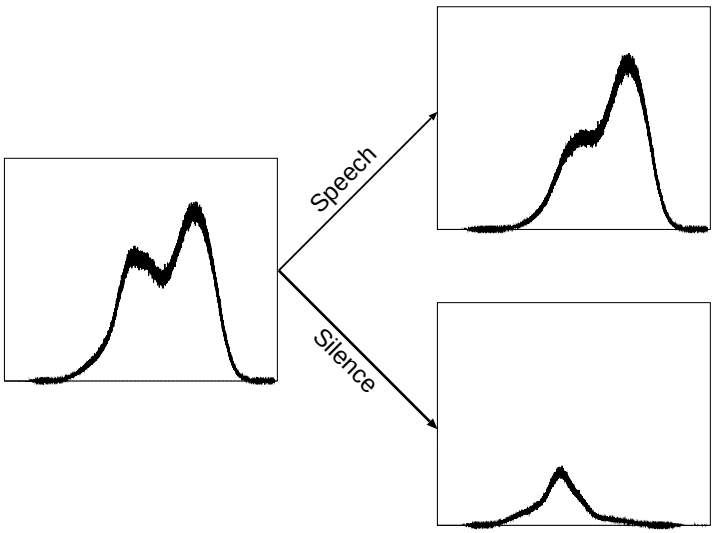

Figure 5: Histogram over the third log filter bank coefficient on the VerbMobil II training corpus. The left side shows the original reference histogram, on the right side the histogram is split into speech and silence. The speech and silence histograms are not yet normalized.

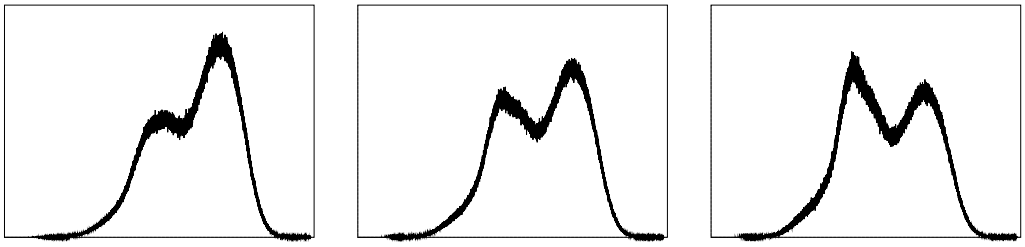

Figure 6: Reference histogram for the third log filter bank coefficient on the VerbMobil II training corpus adapted to three different silence fractions.

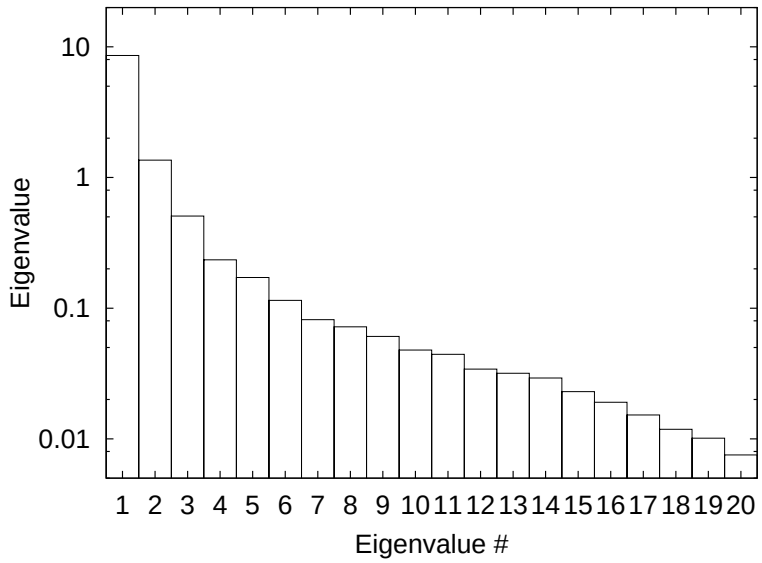

Figure 7: Sorted eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_D$ of the reference covariance matrix $\tilde{\Sigma}$ computed on log filter bank coefficients of the VerbMobil II training corpus. Note the logarithmic scale of the ordinate. The first eigenvalue is about one order of magnitude larger than all others.

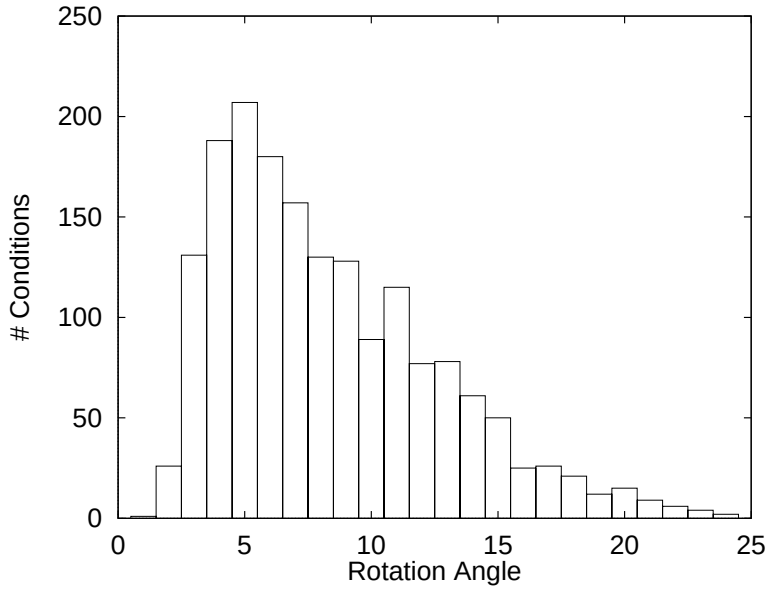

Figure 8: Histogram over the deviation angles η_1 between the first eigenvectors v_1 of the condition-dependent covariance matrices and the first reference eigenvector $\tilde{v}_1$ computed on log filter bank coefficients of the VerbMobil II training corpus.

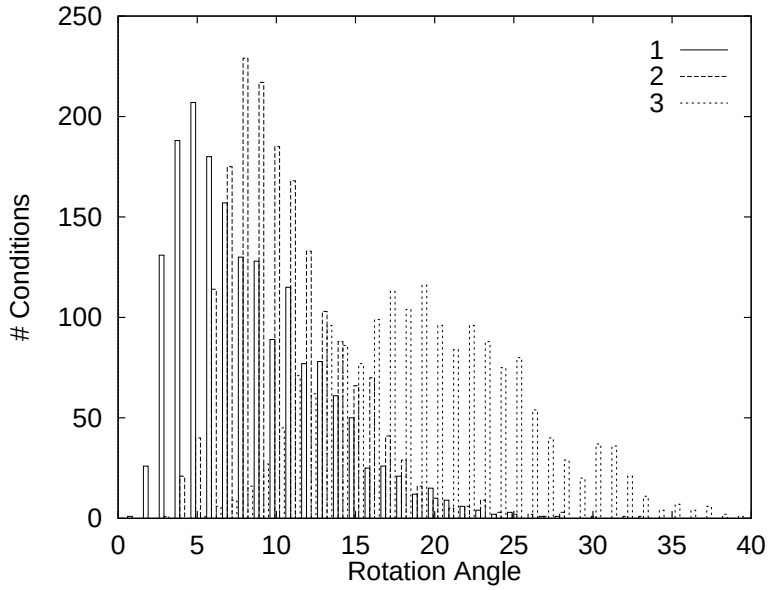

Figure 9: Comparison of the first three deviation angles η_1, η_2, η_3 between the first condition-dependent eigenvectors $v_1, \ldots, v_3$ and the first reference eigenvector $\tilde{v}_1, \tilde{v}_2, \tilde{v}_3$ computed on log filter bank coefficients of the VerbMobil II training corpus.

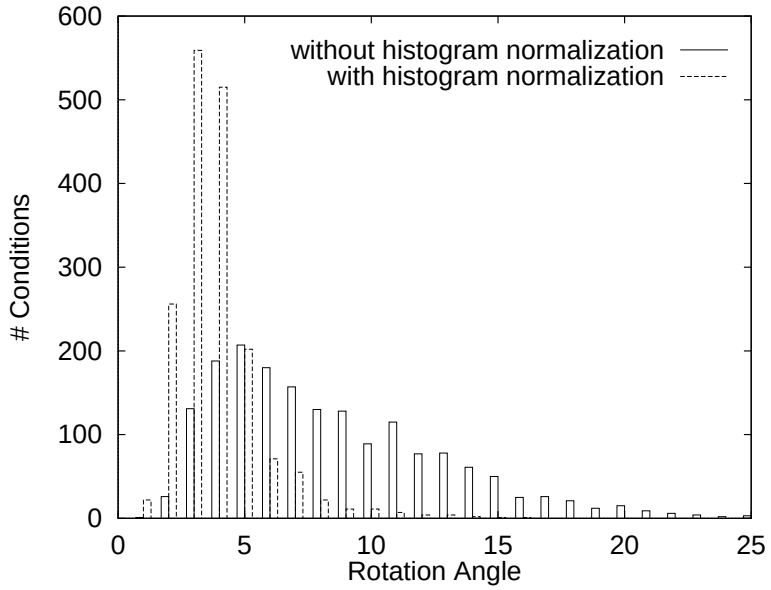

Figure 10: Histogram over the deviation angles η_1 between the first condition-dependent eigenvectors v_1 and the first reference eigenvector $\tilde{v}_1$ estimated on log filter bank vectors of the VerbMobil II training corpus. Results are given both with and without histogram normalization before rotation.

List of Tables

Table 1: Statistics of the continuous speech corpora VerbMobil II and EuTrans II.

Corpus	VerbMobil II		EuTrans II	
	Training	Test	Training	Test
	CD1-41	DEV99B	D1.3c/d	EVAL00
Language	German		Italian	
Speaking Style	spontaneous			
Bandwidth [kHz]	8		4	
Overall Duration [h]	61.5	0.5	7.9	0.8
Silence Fraction [%]	13	11	32	33
Average Condition Duration [s]	140	112	104	119
# Speakers	857	6	276	25
# Sentences	36 010	336	3 187	300
# Running Words	560 837	4 346	52 700	5 555
# Running Phonemes	2 308 741	18 040	250 749	26 853
Trigram LM Perplexity	-	74.6	-	28.6

Table 2: Statistics of the CarNavigation training and test corpora.

Corpus	Training	Test		
	Office	Office	City	Highway
Language	German			
Speaking Style	planned			
Bandwidth [kHz]	4			
Overall Duration [h]	18.8	1.7	1.7	1.8
Silence Fraction [%]	60	69	73	75
Average Condition Duration [s]	785	425	450	468
Average SNR [dB]	21	21	9	6
# Speakers	86	14	14	14
# Running Words	61 742	2 069	2 100	2 100
# Running Phonemes	189 996	16 842	17 184	17 117
Vocabulary	-	2 100	2 100	2 100

Table 3: Recognition test results on the VerbMobil II corpus for basic histogram normalization with and without training data normalization. The first line lists the baseline result without histogram normalization.

Histogram Normalization		Overall [%]	
Stage	Training Data Norm.	Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
log filter bank	no	5.0 - 4.0	23.8
	yes	4.9 - 4.4	**23.0**
cepstrum	no	4.5 - 4.3	24.0
	yes	5.0 - 4.7	24.3
after LDA	no	4.6 - 4.3	24.2
	yes	4.9 - 4.4	24.1

Table 4: Recognition test results on the VerbMobil II corpus for basic histogram normalization at different signal analysis stages. The first line lists the baseline result without histogram normalization.

Histogram Normalization			Overall [%]	
Log Filter Bank	Cepstrum	after LDA	Del - Ins	WER
baseline without normalization			4.9 - 4.4	24.6
yes	no	no	4.9 - 4.4	23.0
no	yes	no	5.0 - 4.7	24.3
no	no	yes	4.9 - 4.4	24.1
yes	yes	no	4.9 - 4.4	22.9
yes	no	yes	4.3 - 4.0	**22.5**
no	yes	yes	4.9 - 4.2	24.0
yes	yes	yes	4.9 - 4.3	22.7

Table 5: Recognition test results on the VerbMobil II corpus for histogram normalization of log filter bank vectors with and without silence fraction treatment.

Histogram Normalization Silence Fraction Treatment	Overall [%] Del - Ins	 WER
baseline without normalization	4.9 - 4.4	24.6
no	4.6 - 3.8	23.0
yes	4.2 - 3.9	**21.8**

Table 6: Recognition test results on the VerbMobil II corpus for feature space rotation at different signal analysis stages. Given are the mean ratio of the first two reference eigenvalues, the mean deviation angle between the condition-dependent first eigenvectors and the first reference eigenvector, and the word error rate. The first line lists the baseline result without normalization.

Normalization Stage	Mean Eigenvalue Ratio $\tilde{\lambda}_1/\tilde{\lambda}_2$	Mean Deviation Angle $\overline{\eta_1}$ [deg]	Overall [%] Del - Ins	WER
baseline without normalization			4.9 - 4.4	24.6
log filter bank	6.3	8.4	4.6 - 4.3	**23.0**
cepstrum	1.4	32.9	5.3 - 5.0	27.8
after LDA	5.4	8.6	5.1 - 4.2	24.1

Table 7: Recognition test results on the VerbMobil II corpus for feature space rotations to match up to four eigenvectors. The given mean deviation angles refer to the highest mapped eigenvector. The first line lists the baseline result without normalization.

Feature Space Rotation	Mean Deviation Angle $\overline{\eta_d}$ [deg]	Overall [%]	
		Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
first eigenvector	8.4	4.6 - 4.3	**23.0**
first two eigenvectors	10.7	4.6 - 4.3	23.2
first three eigenvectors	19.8	4.0 - 4.6	23.0
first four eigenvectors	21.6	4.3 - 4.8	23.6

Table 8: Recognition test results on the VerbMobil II corpus for the combination of feature space rotation with one eigenvector and histogram normalization with silence fraction treatment. The first line lists the baseline result without normalization.

Normalization		Overall [%]	
First Stage	Second Stage	Del - Ins	WER
baseline without normalization		4.9 - 4.4	24.6
rotation	histogram normalization	4.6 - 4.1	22.8
histogram normalization	rotation	4.6 - 3.8	**22.4**

Table 9: Summary of recognition test results on the EuTrans II corpus for histogram normalization with silence fraction treatment, feature space rotation to map one eigenvector, and a combination of both techniques. The first line lists the baseline result without normalization.

Normalization		Overall [%]	
First Stage	Second Stage	Del - Ins	WER
baseline without normalization		4.2 - 3.1	16.5
histogram normalization	-	3.8 - 3.0	15.6
rotation	-	3.6 - 3.1	15.8
rotation	histogram normalization	3.7 - 3.1	**15.5**
histogram normalization	rotation	3.5 - 3.1	15.6

Table 10: Summary of recognition test results on the CarNavigation test corpora for histogram normalization with silence fraction treatment, feature space rotation to map one eigenvector, and a combination of both techniques. Results are reported with and without subsequent cepstral variance normalization (CVN). The first line lists the baseline result without normalization.

Normalization			WER [%]		
CVN	First Stage	Second Stage	Office	City	Highway
yes	baseline without normalization		4.2	20.8	39.7
	histogram	-	3.6	12.4	21.4
	rotation	-	3.8	11.7	21.0
	rotation	histogram	3.7	10.4	19.0
	histogram	rotation	**3.5**	**8.9**	**14.6**
no	baseline without normalization		2.9	31.6	74.2
	histogram	-	2.6	8.2	14.3
	rotation	-	2.5	24.0	64.6
	rotation	histogram	**2.4**	9.5	18.0
	histogram	rotation	2.9	**7.1**	**11.1**