HISTOGRAM BASED NORMALIZATION IN THE ACOUSTIC FEATURE SPACE

Sirko Molau, Michael Pitz, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department,
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{molau,pitz,ney}@informatik.rwth-aachen.de

ABSTRACT

We describe a technique called histogram normalization that aims at normalizing feature space distributions at different stages in the signal analysis front-end, namely the log-compressed filterbank vectors, cepstrum coefficients, and LDA-transformed acoustic vectors. Best results are obtained at the filterbank, and in most cases there is a minor additional gain when normalization is applied sequentially at different stages.

We show that histogram normalization performs best if applied both in training and recognition, and that smoothing the target histogram obtained on the training data is also helpful.

On the VerbMobil II corpus, a German large-vocabulary conversational speech recognition task, we achieve an overall reduction in word error rate of about 10% relative.

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 variations in the speech signal which are irrelevant for the recognition process. Sources of irrelevant variability are for example varying transducer and transmission channels, different speakers, speaking styles, or accents, or a varying ambient or channel noise. In a more general view this can be regarded as a mismatch between training and test conditions, which will deteriorate the performance of the speech recognizer.

There are two ways to cope with the mismatch: We may either transform the acoustic vectors (**normalization**) or the acoustic model (**adaptation**). The different concepts are presented in detail in Figure 1. The feature space is on the left, the model space on the right side. Given one particular attribute of the speech signal (e.g. the vocal tract length of the speaker) we can distinguish between three levels.

The training data (first level) typically contain a collection of different values of that attribute (e.g. different vocal tract lengths). A particular test utterance (third level), on the contrary, has one specific value of the attribute, which may or may not be present in the training data set. Finally there is the intermediate canonical level, where the variations caused by the particular attribute are ideally removed.

In this framework, speech recognition can be regarded as a combination of a specific feature and model space condition. There is a mismatch if the acoustic features and model do not belong to the same level (cf. Figure 1). Without normalization and adaptation there is a strong mismatch between the test features Y and the acoustic model Λ_X. The aim of normalization and adaptation is to transform either the acoustic vectors or the acoustic model, respectively, to reduce or overcome the mismatch.

With adaptation (e.g. MLLR, [6]) the acoustic model can be transformed directly to a specific test condition ($\Lambda_X \rightarrow \Lambda_Y$),

This work was partially funded by the European Commission under the Human Language Technologies project CORETEX (IST-1999-11876).

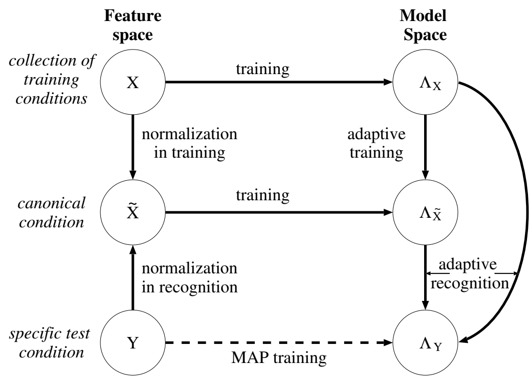

Figure 1: Overview of normalization and adaptation concepts. Histogram normalization affect only the feature space.

which is why adaptation is usually quite successful even when carried out in recognition only. Normalization (e.g. VTN, [2][5]) of acoustic vectors, however, results in a transformation into the canonical form. For this reason there is often only a moderate gain if normalization is applied in recognition only (a small mismatch between $\tilde{X}$ and Λ_X remains), whereas the best performance is typically achieved if both training and test data are normalized (no mismatch between $\tilde{X}$ and $\Lambda_{\tilde{X}}$).

Here we will describe a technique that aims at normalizing feature space distributions at different stages of the signal analysis front-end. The remainder of the paper is organized as follows. In the second section we will classify normalization techniques in more detail and show how histogram normalization fits into that framework. The third section discusses how histogram normalization works in generell, and which algorithmic decisions have to be made. Then we will describe test conditions and results from recognition tests on the German VerbMobil task, a large-vocabulary conversational speech corpus. The results are summarized and an outlook is given in section five.

2. NORMALIZATION TECHNIQUES

There are different ways to classify normalization schemes. Based on whether they are derived from some physical model we can distinguish between **model based** and **data distribution based** normalization.

In the model based case one tries to normalize a specific variability with a known functional form and an expected effect on the speech signal. The Normalization is based on some model for speech production, transmission, or perception. A small number of model parameters are estimated on the test data, and applied according to the underlying model to normalize the acoustic vectors. Well-known speaker normalization techniques like vocal tract [2][5] and speaking rate normalization [8] are model based approaches. Other algorithms that fall into this category are channel and environment normalization schemes like cepstral mean normalization (CMN, [7]) and noise supression techniques that rely on a SNR estimate.

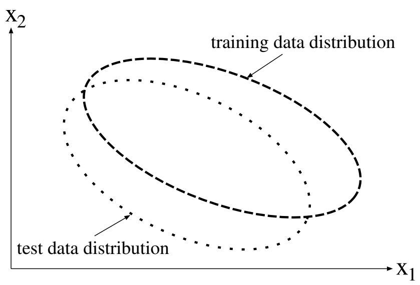

Figure 2: Schematic distribution of training and test data in a two-dimensional example feature space.

Data distribution based normalization, on the contrary, aims at transforming the acoustic vectors into a domain that is more suitable for automatic speech recognition independent of any model for speech production, transmission, or perception. The transformation parameters are obtained from the distribution of the training and test data. A typical example is the Gaussianization technique proposed by Gopinath et al. [3], which aims at transforming acoustic vectors in a way that they better follow a Gaussian distribution. Also the histogram normalization proposed by Dharanipragada et al. [1] falls into this category. The idea is to transform test vectors such that their distribution matches the distribution of the training data. The quantile based equalization of Hilger et al. [4] is a special case of histogram normalization with coarse binning that was successfully applied to small test data samples.

3. HISTOGRAM NORMALIZATION

Figure 2 shows the distribution of training and test data in a two-dimensional example feature space. The mismatch between the data sets may have different reason (see section one), but will be especially large if there are major differences in the recording environments, i.e. different microphones or bandwidths.

Histogram normalization tries to transform the test data so that their distribution matches that of the training data. Under the assumption that the process which is responsible for the mismatch has an independent effect on the different acoustic vector components, each feature space dimension can be normalized independently of the others, which simplifies the problem a lot.

For each dimension, the distribution $p(x)$ (histogram) of the training and test data is computed. A cumulative histogram $P(x) = \int_{-\infty}^{x} dx' p(x')$ is derived, and the test data distribution is transformed to the training data distribution as shown in Figure 3. Each test set value x_t is replaced by the value $\tilde{x}_t$ that corresponds to the same point in the cumulative training data distribution ($P(x_t) = P(\tilde{x}_t)$). It is obvious that histogram normalization is computationally attractive, as it can be implemented by a simple table lookup.

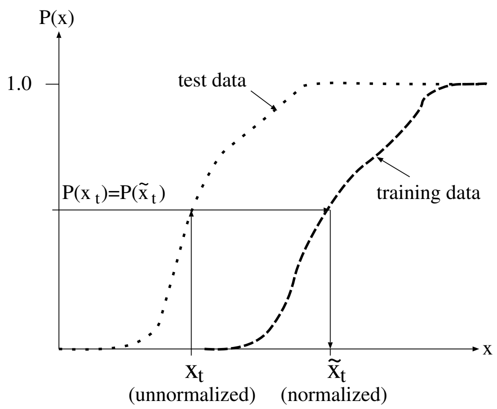

Figure 3: Principle of histogram normalization. The test data are transformed such that their cumulative histogram matches the cumulative histogram of the training data distribution.

Due to the independence assumption histogram normalization can account for scaling, shifting, or any type of non-linear distortion of each feature space dimension, but not for rotations of the feature space. In the example case (Figure 2) histogram normalization will reduce the mismatch significantly but not remove it completely.

3.1. What should be normalized?

Dharanipragada et al. [1] propose a normalization of cepstral features. However, there are different stages in the signal analysis front-end (Figure 4) where histogram normalization may be applied.

- The FFT transforms the speech signal into a sequence of discrete-frequency line spectra. Each individual spectral line could be regarded as an independent distribution that needs to be normalized. For computational reasons it will be better to apply histogram normalization after the filterbank, which leaves in the order of 20 distributions.
  In theory it makes no difference whether the normalization is applied before or after the logarithm, but in practice spectral log-compression will help to keep histogram discretization errors small. Histogram normalization of the log-filterbank coefficients may help to reduce spectral distortions that could be limited to certain frequency bands. It also normalizes the energy distribution for each frequency band.
- The mean of cepstral coefficients is typically subtracted (CMN) in order to remove constant channel transfer functions. In many tasks it also helps to transform cepstral coefficients to unity variance. Histogram normalization after the discrete cosine transform (DCT), however, has a larger degree of freedom. It may not only shift and scale the distribution of each cesptral coefficient, but also distort it non-linearly.
- LDA transformation of cepstral coefficients and their time derivatives is a standard feature of the RWTH large vocabulary speech recognition system (LVCSR) [9], since it consistently reduces the WER on all tasks.

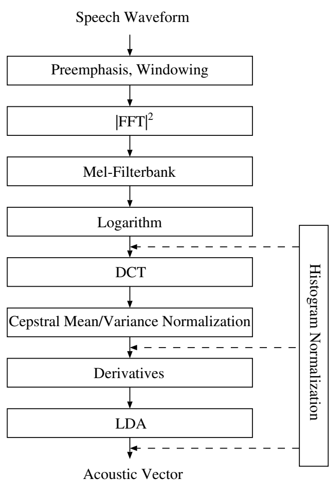

Figure 4: Typical signal analysis front-end. There are a number of stages where histogram normalization can be applied.

> The LDA-transformed vector is the one that is finally presented to the speech recognizer. Hence, applying histogram normalization to LDA-transformed acoustic vectors will normalize the distributions of test vector to those observed during training of the corresponding acoustic model.

It is even possible to apply histogram normalization sequentially in a multi-pass scheme: In a first signal analysis pass the histogram of the first normalization stage is derived. In the next pass the acoustic vectors are normalized at the first stage, and the histogram for the second stage is accumulated. In the end the distributions of the acoustic vector components are normalized at all stages.

As it is beforehand unknown at which stage of signal analysis histogram normalization performs best or if there is a gain by sequential normalization at different stages, we have tested all three methods and combinations of them.

3.2. Histogram normalization in training

As described so far, we take the distribution of training data as the reference (target) histogram and normalize the test data accordingly. However, normalization of the test data alone results in moderate gain of recognition performance only (see section 1). It is usually necessary to normalize the training data in the same way to avoid a mismatch and achieve the full performance.

In the context of histogram normalization we have to decide how to pool the training data. In VTN, for example, all data from the same speaker are pooled, a warping factor is computed, and the data are normalized. For histogram normalization on the VerbMobil corpus, which consists of a large number of dialogs between two or more speakers, we propose to pool the data from each speaker in each dialog. This will give enough data (typically several minutes) to estimate the distributions reliably. It also allows to normalize both speaker-dependent effects and channel distortions which may differ from dialog to dialog.

In practice, we first compute the overall distribution of each acoustic vector component on all training data. The cumulative histogram of that distribution becomes the target for the following normalization of training and test data. Then, the distribution for each training data pool is calculated and transformed to match the target distribution. On the normalized data a standard acoustic model training is performed. The test data are transformed in the same way. A histogram for each acoustic vector component is computed on all data from the same test speaker and dialog, and transformed to match the target distribution.

It is clear that histogram normalization as described here is not suitable for on-line recognition or small data samples. For these cases a histogram with only a few bins and interpolation inbetween is required which can be reliably derived on very short speech samples [4].

3.3. Histogram smoothing

As an example, Figure 5 shows the histogram of the third log-filterbank coefficient over the full VerbMobil training corpus. It turns out that the distributions of many components of the filterbank vectors, the cepstral coefficients, and the LDA-transformed vectors have a bimodal shape. They can be well approximated by a mixture of two Gaussians. Data scattering is efficiently smoothed and outliers at the tails of the distribution are better modelled. So in a final test we computed a histogram over the full training corpus as before and estimated a two-density Gaussian mixture with minimum mean squared error to the training data histogram. The cumulative probability density function of the mixture was then used as the target histogram.

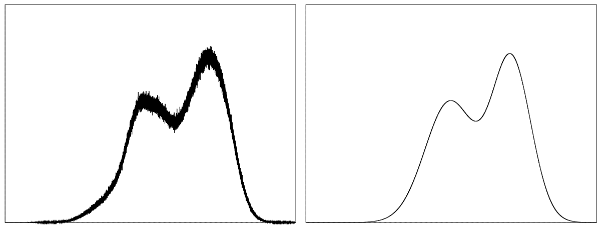

Figure 5: Histogram over the third log-filterbank coefficient obtained on the whole VerbMobil training corpus. The bimodal distribution (left) can be well approximated by a Gaussian mixture with two densities (right).

4. RECOGNITION TESTS

Histogram normalization was evaluated on the VerbMobil II corpus with the RWTH LVCSR [9][10] that can be characterized as follows:

- 16 cepstral features with first derivatives and the second derivative of the energy, 10 ms frame shift
- cestral mean and variance normalization in a sliding window of two second length
- LDA transformation of three consecutive cepstrum and derivative vectors, reduction to 33 dimensions
- 2500 decision tree based within-word triphone states including noise plus one state for silence
- gender independent acoustic models with globally pooled diagonal covariance
- 3-state-HMM topology with skip
- class-trigram language model

VerbMobil II is German spontaneous speech task with a 10k-word vocabulary. One part of the training data (49h) was collected with a head-set microphone, the other (12h) with a room microphone. The training and test data were collected at different sites, so there are minor differences in the recording environment as well. The statistics of the training and test corpus are summarized in Table 1.

In a first set of experiments we compared the recognition performance when applying histogram normalization at different signal analysis stages in recognition only. The results are summarized in Table 2. It turns out that histogram normalization gives minor improvements of up to 3% relative at all investigated signal analysis stages.

Table 1: Statistics of the training and test corpus

VerbMobil II	Training CD1-41	Test DEV99B
Duration	61.5h	0.5h
Sil. Portion	13%	11%
# Speakers	857	6
# Sent.	36,015	336
# Words	701,512	4,346
Trigram PP.	-	74.6

Table 2: Recognition results for histogram normalization at different signal analysis stages in recognition only.

Histogram Normalization			Overall [%]	
Filterbank	Cepstrum	LDA	Del - Ins	WER
no	no	no	4.9 - 4.4	24.6
yes	no	no	5.0 - 4.0	**23.8**
no	yes	no	4.5 - 4.3	24.0
no	no	yes	4.6 - 4.3	24.2

Table 3: Recognition results for histogram normalization in training and recognition.

Histogram Normalization			Overall [%]	
Filterbank	Cepstrum	LDA	Del - Ins	WER
no	no	no	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 4: Recognition results for histogram normalization with a smoothed target histogram in training and recognition.

Histogram Normalization			Overall [%]	
Filterbank	Cepstrum	LDA	Del - Ins	WER
no	no	no	4.9 - 4.4	24.6
yes	no	no	4.6 - 3.8	**22.5**
no	yes	no	5.2 - 4.3	23.9
no	no	yes	4.9 - 4.2	23.7
yes	yes	no	4.9 - 4.1	23.2
yes	no	yes	4.7 - 3.9	22.8
no	yes	yes	4.8 - 4.4	23.8
yes	yes	yes	4.8 - 3.8	**22.5**

In a next set of experiments we applied histogram normalization both in training and recognition as explained in section 3.2. The results for normalization at individual signal analysis stages and combinations of them are presented in Table 3. For each test, the number of densities in the acoustic model was optimized with respect to WER.

It turns out that filterbank normalization gains significantly from normalized training data, whereas there is only little or no gain for cepstrum and LDA-transformed acoustic vectors. The performance improvement of histogram normalization at different stages is to some extend additive. The best result is observed when applying filterbank and LDA feature transformation, which yields a relative WER reduction of almost 10%.

Finally we have repeated the experiments with a smoothed target histogram as described in section 3.3. The results are summarized in Table 4. In most cases the recognition performance was higher compared to the original histogram (Table 3). Even though the lowest word error rate (WER) of 22.5% could not be reduced any further, this result was now also obtained by normalizing the filterbank components alone. The normalization is much faster and easier if cepstrum and LDA feature vector normalization can be omitted.

5. CONCLUSIONS AND OUTLOOK

In this paper we applied histogram normalization at different stages of the signal analysis front-end. We demonstrated that normalizing the cepstral coefficients or the LDA-transformed acoustic vector components helps a little, but most gain was achieved when transforming the log-filterbank coefficients. The gain obtained by histogram normalization at different signal analysis stages was to some extend additive.

An explanation for the superior performance of filterbank normalization could be, that most of the variations compensated for by histogram normalization have an independent effect on the individual filterbank components, but not on the cepstrum and LDA coefficients.

As expected, normalization of training and test data yieled better results than normalization of the test data alone. Smoothing the target histogram gave a further gain in recognition accuracy. The overall largest reduction in WER of about 10% relative was achieved by applying filterbank normalization with a smoothed target histogram both in training and test.

We intend to conduct a number of further experiments in the context of histogram normalization. So far we have only smoothed the target histogram. The individual histograms for each training and test data pool, however, were mapped to the target without any preprocessing. Smoothing may help here as well, since there is much less data available for the estimation of the distribution, which will result in significantly larger histogram scatter and stronger influence of outliers.

Histogram normalization as presented here handles the vector components independently of each other. Rotating the feature space first may help to overcome this limitation.

Finally, there is only a minor mismatch between training and test conditions in the VerbMobil corpus. It will be interesting so see how the technique performs under stronger mismatch conditions.

6. REFERENCES

[1] S. Dharanipragada, M. Padmanabhan: "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition", *Proc. Int. Conf. on Spoken Language Processing*, Bejing, China, pp. 556–559, October 2000.

[2] E. Eide, H. Gish: "A Parametric Approach to Vocal Tract Length Normalization", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 346–349, May 1996.

[3] R. A. Gopinath: "Gaussianization", *IMA Workshop: Mathematical Foundations of Speech Processing and Recognition*, Minneapolis, MN, September 2000.

[4] F. Hilger, H. Ney: "Quantile Based Histogram Equalization for Noise Robust Speech Recognition", to appear in *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, September 2001.

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

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

[7] F. H. Liu, R. M. Stern, X. Huang, A. Acero: "Efficient Cepstral Normalization for Robust Speech Recognition", *Proc. ARPA Speech and Nat. Language Workshop*, Princeton, NJ, pp. 69–74, March 1993.

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

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

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