ENHANCED HISTOGRAM NORMALIZATION IN THE ACOUSTIC FEATURE SPACE

Sirko Molau, Florian Hilger, Daniel Keysers, and Hermann Ney

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

ABSTRACT

We describe two methods that aim at normalizing acoustic vectors at the filterbank level such that the test data distribution matches the training data distribution. They enhance the histogram normalization technique proposed earlier by taking care of the variable silence fraction for each speaker, and by rotating the feature space. We report a number of recognition tests under minor (different microphones in training and test, telephone data) and major (office vs. car recordings) mismatch conditions. Both methods give superior performance to the basic histogram normalization approach. The overall improvements in word error rate (WER) range between 6% and 85% relative.

1. INTRODUCTION

Large variability in the acoustic signal is a major error source for automatic speech recognition systems. Undesired variations result from variable recording environments (ambient noise, recording equipment, and transmission channels) and changing speakers (speaking styles, accents), for example. A number of techniques were suggested to cope with these variations, which fall into two broad categories: *Normalization* schemes try to remove the variability by transforming the acoustic vectors, and *adaption* techniques amount to a transformation of the acoustic model to adapt it to the specific test conditions. Normalization techniques can be further subdivided into two broad classes [4] depending on whether they are based on some specific model for speech production, transmission, or perception (*model-based*) or on the overall distribution of the training and test data (*data-distribution based*).

Dharanipragada et al. [1] suggested a histogram normalization method that falls into the latter category. In [4] we reported on a detailed analysis of this technique which was applied on a large vocabulary German conversational speech task. Here we discuss two extensions which overcome some of the limits of the basic histogram normalization approach.

The remainder of this paper is organized as follows: In the next section we recapitulate histogram normalization and its main assumptions. The third section describes how a variable silence fraction may be treated, and in section four we discuss a method to account for feature space rotations. Next we describe three test corpora with different degrees of mismatch between training and test data. Recognition test results are presented and discussed in section six, and the paper is summarized in the last section.

This work was partially funded by the European Commission under the Human Language Technologies project CORETEX (IST-1999-11876), and by the DFG (Deutsche Forschungsgemeinschaft) under contract NE 572/4-1.

2. HISTOGRAM NORMALIZATION

Here we briefly recapitulate the histogram normalization technique. A more detailed introduction can be found in [4].

The main idea is to remove undesired variations by transforming the acoustic test data such that their distribution matches the training data distribution. Histogram normalization relies on two basic assumptions:

1. The global statistics of the speech signal are the same independently of what was actually spoken.
2. The feature space dimensions are oriented such that the variations are independent in each dimension.

Given these assumptions, histogram normalization can be performed on each feature space dimension independently in a very efficient way:

- Compute the histograms $p_X(x)$ and $p_Y(y)$ of the training and test data
- Compute normalized cumulative histograms
$$P_X(x) = \frac{\int_{-\infty}^{x} dx' p_X(x')}{\int_{-\infty}^{\infty} dx' p_X(x')}$$
$$P_Y(y) = \frac{\int_{-\infty}^{y} dy' p_Y(y')}{\int_{-\infty}^{\infty} dy' p_Y(y')}$$
- Replace each test set value y by $\tilde{x} = \tilde{x}(y)$ that corresponds to the same point in the cumulative training data distribution:
$$P_Y(y) \stackrel{!}{=} P_X(\tilde{x})$$

In [4] we have reported on a detailed analysis of histogram normalization with the following results:

- Best performance is achieved when histogram normalization is applied to the log-filterbank coefficients.
- It is useful to apply the normalization both in training and test. That is, the overall distribution of all training data is used as reference (target histogram), and the data of each test *and* training speaker is transformed to match the target histogram.
- The target histogram can be replaced by a mixture of two Gaussian densities to smooth data scatter.

Histogram normalization can be implemented as a simple table look-up and is therefore computationally attractive. It is an off-line technique as it requires enough data from each speaker (typically a few sentences) to reliably estimate the distributions and ensure that these are not affected by the selection of words that were spoken.

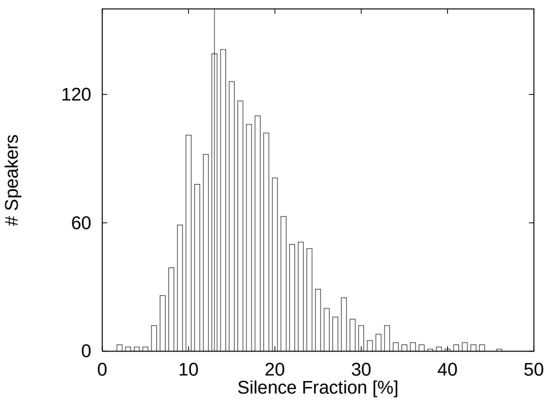

Fig. 1. Histogram over the silence fraction of individual speakers in the VerbMobil II corpus. The vertical line marks the average silence fraction of 13%.

3. SILENCE FRACTION TREATMENT

Assumption 1 on the global statistics of the speech signal is sometimes violated. Even if enough speech data is available to ensure that the phoneme frequency is about the same for each speaker, and even if the phoneme-dependent distributions are identical for all speakers, the histograms may still vary due to different silence fractions. This has a severe impact on speakers with a much lower or higher than average silence fraction. In the first case, histogram normalization will transform a number of acoustic speech vectors to silence and cause more deletions of words. In the latter case, some silence vectors will be transformed to speech and cause word insertions.

Figure 1 shows a histogram of the speaker-wise silence fraction in the VerbMobil II corpus. The average silence fraction on this corpus is 13%, but the number varies between 3% and 76% for individual speakers.

The solution is to estimate two independent target histograms for silence and speech (Figure 2). A forced alignment with the reference transcriptions is carried out on the training data, and all acoustic vectors mapped to the silence mixture are accumulated in the silence histogram p_{sil}, all other vectors in the speech histogram p_{speech}. It turned out that the bimodal structure (left histogram in Figure 2) we observed in most histograms [4] was in fact a manifestation of speech and silence, as the first peak can be attributed to silence frames, whereas the second peak was caused by more energetic speech frames.

In the normalization step, the silence fraction α of the actual training or test speaker has to be determined first. For the training speakers, it is obtained by a forced alignment as before. Since the correct transcription is unknown, we need to calculate the silence fraction for test speakers either in a preliminary recognition pass (two-pass recognition) or with a dedicated speech/silence detector (e.g. as described in [3]).

Next an adapted target histogram is computed for each speaker by linear interpolation between the speech and silence histograms. Note that the result is independent of the order of cumulation and interpolation if the histograms are normalized beforehand.

$$P_\alpha = \alpha \cdot P_{sil} + (1 - \alpha) \cdot P_{speech}$$

The adapted target histogram is then used for normalization as in the basic histogram normalization approach.

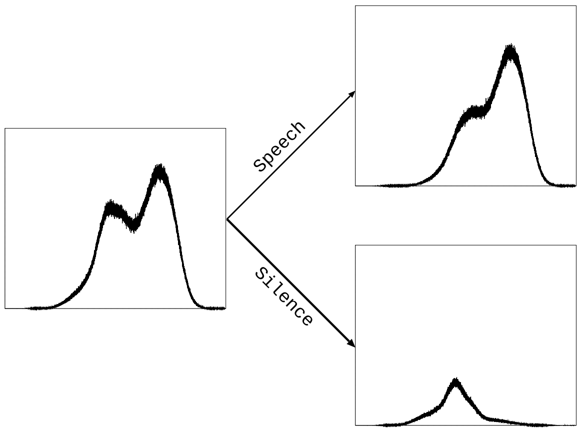

Fig. 2. Histogram over the third log-filterbank coefficient on the VerbMobil II corpus. The left side shows the original histogram, on the right side the histogram is split into speech and silence.

4. FEATURE SPACE ROTATIONS

A feature space in which the variations accounted for are approximately decorrelated is the second basic assumption of histogram normalization. Previous tests have suggested that this condition is best met at the filterbank level, i.e. that the variations are decorrelated best in the frequency domain, as histogram normalization of cepstrum and LDA-transformed vectors gave a much smaller gain in recognition performance [4]. Still the feature space as such might be rotated by a small amount, which could not be treated properly by histogram normalization.

To overcome this limitation, we have investigated explicit feature space rotations in addition to histogram normalization. At first, a covariance matrix was computed over all log-filterbank training vectors. A set of target eigenvectors was calculated and sorted in descending order of their corresponding eigenvalues. It turned out that the first eigenvalue was significantly larger than all others (Figure 3). Hence, the feature space has one preferred direction with large scatter, and along the other principal axes data scatter is much smaller. Note that the logarithm increases the scatter of the filterbank coefficients, as these are typically small.

Next the covariance matrix and eigenvector basis for each speaker were derived. The first eigenvector, i.e. the direction of the principal axis with largest data scatter, differed slightly from the direction of the first target eigenvector in the 20-dimensional log-filterbank feature space. Figure 4 shows a histogram over the speaker-wise deviation angles β on the VerbMobil II corpus.

To account for this deviation, we derived a matrix U for each speaker that rotates the feature space in the plane spawned by the first speaker-dependent and the first target eigenvector. The matrix was designed such that the speaker's feature space as such remains undistorted, but the principal axis with largest data scatter becomes identical for all speakers:

$$\begin{aligned} \tilde{x} &= Ux \\ U &= PR_\beta P^T + (I - PP^T) \end{aligned}$$

With the matrix P the filterbank vector x is projected onto the plane spawed by the two eigenvectors. It is rotated within this plane with R_β by the angle β between both eigenvectors, and projected back into the original space by P^T. The last term $I - PP^T$ with the identity matrix I restores the dimensions orthogonal to the plane of rotation lost in the first projection.

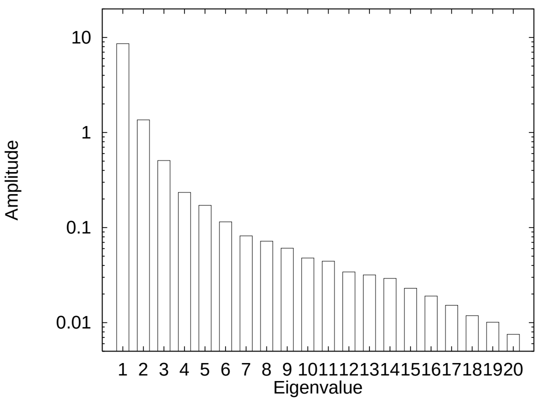

Fig. 3. Sorted eigenvalues of the covariance matrix computed on the log-filterbank coefficient of the VerbMobil II training corpus. Note the logarithmic scale of the ordinate. The first eigenvalue is about one order of magnitude larger than all others.

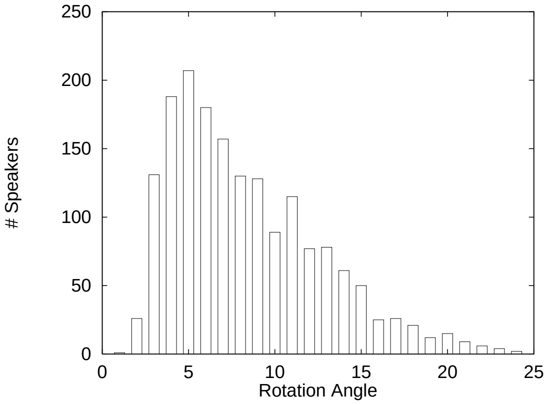

Fig. 4. Histogram over the deviation angles between the first eigenvectors of the speaker dependent covariance matrices and of the target covariance matrix obtained from log-filterbank coefficients on the VerbMobil II training corpus.

The procedure can be applied sequentially to match further pairs of eigenvectors. It turned out, however, that the directions of subsequent principal axes are not well-defined anymore and that the order of eigenvectors may change from speaker to speaker as their eigenvalues differ by only a small amount. The deviation angles between the second eigenvectors, for example, were already much larger and for some speakers close to 90 deg.

As usual the rotation matrix was computed and applied to both test and training data. We report on recognition tests with feature space rotation alone, and rotation before or after silence fraction adapted histogram normalization.

5. TRAINING AND TEST CORPORA

Recognition tests were carried out on three corpora with different degree of mismatch between the training and test data. The statistics of all three corpora are summarized in Table 1 and 2. Turn duration gives the average amount of acoustic data used to estimate the histograms and/or rotation matrix.

VerbMobil II is German conversational speech task with a 10k-word vocabulary. One part of the training data (49h) was collected with a head-set, the other (12h) with a room microphone.

Table 1. Statistics of the continuous speech corpora.

Corpus	VerbMobil II		EuTrans II	
	Training CD1-41	Test DEV99B	Training D1.3c/d	Test EVAL00
Bandwidth	microphone		telephone	
Duration	61.5h	0.5h	7.9h	0.8h
Sil. Fraction	13%	11%	32%	33%
Turn Duration	140s	112s	104s	119s
# Speakers	857	6	276	25
# Sentences	36,015	336	3,187	300
# Run. Words	701,512	4,346	52,511	5,555
Trigram PP.	-	74.6	-	28.6

EuTrans II is an Italian conversational speech task with a 2k-word closed vocabulary. Both training and test data were recorded over wireline telephone, but the channel quality varied significantly between different recording sessions.

CarNavigation is a German isolated word database with a 2k-word closed vocabulary. Training data were recorded in a quiet office environment. The office test set was recorded under the same conditions (SNR 21dB). Two further test sets were recorded in driving cars (city and highway traffic, average SNRs 9dB and 6dB, respectively). The test words did not occur in the training data set. A more detailed description can be found in [2].

Recognition tests were carried out with the RWTH LVCSR, which is described in detail in [5] and [6]. The baseline system was optimized for each task.

6. RECOGNITION TEST RESULTS

Recognition test results for the different corpora are summarized in Tables 3 and 4. Baseline histogram normalization gave a reduction in WER of 9% relative on the VerbMobil II corpus as already reported in [4]. The error rate could be further reduced by taking care of the silence fraction. It is important that acoustic vectors which are aligned with noise mixtures (7% of the training corpus) are treated like speech frames. Feature space rotation alone gave a somewhat smaller gain in recognition accuracy, and a sequential application of both techniques was not helpful either.

On the EuTrans II telephone corpus, basic histogram normalization gave no improvement in recognition accuracy. The transmission channel was more noisy and showed larger variations from one speaker 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, however, yielded a relative WER reduction of 5%. Feature space rotation alone improved the recognition accuracy by a similar amount,

Table 2. Statistics of the isolated word corpora. Each word was uttered once and the vocabulary was different for all four corpora.

CarNavigation	Training	Test		
	Office	Office	City	Highway
Bandwidth	microphone			
Duration	18.8h	1.7h	1.7h	1.8h
Sil. Fraction	60%	69%	73%	75%
Turn Duartion	785s	425s	450s	468s
# Speakers	86	14	14	14
# Run. Words	61,742	2,069	2,100	2,100
Zerogram PP.	-	2,100	2,100	2,100

Table 3. Recognition test results for histogram normalization and feature space rotation on the continuous speech corpora.

Normalization steps in the order of their application	WER [%]	
	VerbMobil	EuTrans
Baseline without normalization	24.6	16.5
Histogram without silence treatment	22.5	16.5
Histogram with silence treatment	**21.8**	15.6
Feature space rotation	23.0	15.8
Rotation & Hist. with sil. treatment	22.8	**15.5**
Hist. with sil. treatment & Rotation	22.4	15.6

but unfortunately the gain of histogram normalization and rotation was not additive. A sequential application of both techniques did not further reduce the word error rate significantly.

In scenarios with large mismatch between training and test data like on the CarNavigation task, there is much room for improvements. A standard normalization technique is cepstral variance normalization (CVN). On this task, it lowered the recognition accuracy in the clean office condition, but significantly improved the baseline result for the city and highway test sets (Table 4).

Basic histogram normalization was helpful both with and without CVN, but better results were obtained without this extra step. The variance is already implicitly normalized when the feature space dimensions are mapped onto the same target histogram, which is why a further transformation to unity variance may be counterproductive. The same holds for histogram normalization with silence fraction treatment, but here the gain in recognition accuracy was even higher. Without CVN, the word error rate was reduced by 10% relative on the office data, 74% relative on the city and 81% relative on the highway data.

In connection with variance normalization, feature space rotation outperformed histogram normalization with silence fraction treatment slightly. Without CVN, however, it gave only small improvements over the baseline system. Hence, feature space rotation is reducing the mismatch in a different way and does not normalize the variance similar to histogram normalization. The rotation angles for the test speakers increased with the mismatch. Whereas on the office data the average rotation angle was 6 deg, it increased to 23 deg on the city and 32 deg on the highway data.

If applied in the right order, feature space rotation and histogram normalization together performed always better than rotation and histogram normalization alone. The best result on the office test data was achieved when the rotation was applied before histogram normalization, and on the mismatch city and highway data when applied afterwards. The rotation angles were reduced by more than a factor of two on average when histogram normalization was applied before rotation.

Table 4. Recognition test results for histogram normalization and feature space rotation on the isolated word corpora. Results are reported with and without subsequent variance normalization.

CVN	Normalization steps in the order of their application	WER [%]		
		Office	City	High.
yes	Baseline without norm.	4.2	20.8	39.7
	Histogram without sil. tr.	4.2	14.7	25.7
	Histogram with sil. tr.	3.6	12.4	21.4
	Feature space rotation	3.8	11.7	21.0
	Rotation & Hist. with sil.	3.7	10.4	19.0
	Hist. with sil. & Rotation	**3.5**	**8.9**	**14.6**
no	Baseline without norm.	2.9	31.6	74.2
	Histogram without sil. tr.	2.8	10.2	16.6
	Histogram with sil. tr	2.6	8.2	14.3
	Feature space rotation	2.5	24.0	64.6
	Rotation & Hist. with sil.	**2.4**	9.5	18.0
	Hist. with sil. & Rotation	2.9	**7.1**	**11.1**

We find that in general one should apply the normalization method first that gives most gain in recognition performance alone.

7. SUMMARY

We applied histogram normalization to log-filterbank coefficients during signal analysis. We demonstrated that the basic approach improves the recognition performance on tasks with different degree of mismatch between training and test data. Better performance was achieved on all tasks when the silence fraction was treated explicitly. Each speaker was normalized with a different target histogram that was adapted to his specific silence fraction.

To overcome the indepence assumption of histogram normalization we have investigated feature space rotations that match the first eigenvector of the speakers' log-filterbank covariance matrices. On many tasks this technique gave a performance gain that was similar to histogram normalization with silence fraction treatment. A sequential application of both methods only sometimes reduced the word error rate further. Best results were achieved when the normalization method was applied first that gave better recognition results when used alone.

Further analyses of feature space rotations and their combination with histogram normalization are currently under way.

8. 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, Oct. 2000.

[2] F. Hilger, H. Ney: "Noise Level Normalization and Reference Adaptation for Robust Speech Recognition", *Proc. ASR2000 - Int. Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, Paris, France, pp. 64–68, Sept. 2000.

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

[4] S. Molau, M. Pitz, and H. Ney: "Histogram Based Normalization in the Acoustic Feature Space", *Proc. ASRU2001 - Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, Dec. 2001.

[5] 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.

[6] 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.