FEATURE SPACE NORMALIZATION IN ADVERSE ACOUSTIC CONDITIONS

Sirko Molau, Florian Hilger, and Hermann Ney

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

ABSTRACT

We study the effect of different feature space normalization techniques in adverse acoustic conditions. Recognition tests are reported for cepstral mean and variance normalization, histogram normalization, feature space rotation, and vocal tract length normalization on a German isolated word recognition task with large acoustic mismatch. The training data was recorded in clean office environment and the test data in cars. Speech recognition failed completely without normalization on the highway dataset, whereas the word error rate could be reduced to 17% using an online setup and to 10% with an offline setup.

1. INTRODUCTION

Mismatch between training and test data is a major error source for automatic speech recognition systems. Variable environments (ambient noise, recording equipment, and transmission channels) result in a severe degradation of recognition performance [11]. Inter-speaker variations like different vocal tract lengths induce further variability to the speech signal that make the recognition task even more difficult.

In the literature a number of techniques were presented to cope with mismatch conditions. They fall into two broad categories: *normalization* schemes try to reduce the mismatch by transforming the acoustic vectors, *adaption* techniques amount to a transformation of the acoustic model to adapt it to the specific test conditions.

From a statistical point of view, reducing the mismatch between training and test conditions means to match the distributions of the signals' values. In the following we will study different normalization schemes with growing complexity, and sequential applications of these. Starting from simple mean and variance normalization, more elaborate histogram-based techniques will be described. A version suited for online applications will be compared to more complex techniques including feature space rotations. Finally vocal tract length normalization (VTN) will be applied to reduce remaining speaker dependent variations.

2. TRAINING AND TEST CONDITIONS

The normalization techniques will be studied on the CarNavigation corpus, a German isolated word database with a 2k-word closed vocabulary and strong mismatch conditions. Training data were recorded with a sampling rate of 16 kHz in a quiet office environment. The office test set was recorded under the same conditions (SNR 21 dB). Two further test sets were recorded in cars (city and highway traffic, average SNRs 9 dB and 6 dB, respectively). There was no overlap in vocabulary between the different test sets, and between the training and test sets. Statistics of the training and test corpora are summarized in Table 1. Turn duration gives the average amount of acoustic data used to estimate the histograms and/or rotation matrix in offline mode.

Table 1. Statistics of the CarNavigation corpus

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

Recognition tests will be carried out with the RWTH large vocabulary speech recognition system which was described in detail in [8] and [10]. The recognizer contains a standard MFCC front-end with subsequent linear discriminant analysis. Words were modeled with triphones using 700 decision tree based tied states plus one silence state. The acoustic models consist of approximately 20k Gaussian densities with globally pooled diagonal covariance matrices.

3. CEPSTRAL MEAN AND VARIANCE NORMALIZATION

The speech signal produced by a speaker is transmitted over some channel before it reaches the recording device. The channel disturbs the original speech signal. Convolutional distortions are multiplicative in the spectrum domain. Due to the logarithmic compression of the filterbank channels before the cosine transformation, multiplicative distortions become additive in the cepstrum domain [12].

Thus, a simple and effective way of channel normalization is to subtract the mean of each cepstrum coefficient (cepstral mean normalization, CMN) which will remove time-invariant distortions introduced by the transmission channel and the recording device. Furthermore it is known that normalizing the variance of cepstral coefficients (CVN) helps to improve recognition in adverse conditions.

Recognition test results for these techniques are summarized in Table 2. The baseline word error rate (#0) for clean test data

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.

is 2.8%. Under mismatch conditions it increases dramatically. Whereas in the city traffic test set at least a third of the words are still recognized correctly, essentially nothing is correctly recognized in the highway test set.

Cepstral mean normalization (Table 2, #1) has no impact on the office test set, since there is no channel mismatch. It more than halves the word error rate (WER) on the city data, but it is not sufficient for the highway data. Only when the variance is normalized as well (#2), the word error rate drops in all conditions below 50%.

Interestingly, cepstral variance normalization significantly lowers the recognition accuracy in the clean office condition. Furthermore, histogram normalization as described in the following sections gives consistently better results without subsequent CVN [7]. The variance is normalized implicitly when the feature space dimensions are mapped onto the same target histogram, which is why a further transformation to unity variance may be counterproductive. For these reasons, cepstral variance normalization has not been further pursued in subsequent tests.

4. QUANTILE EQUALIZATION

The aim of histogram normalization [1] [9] is to match the overall distribution of each feature space dimension in training and test, not just the mean and variance. It is based on the assumption that in the absence of mismatch at a certain stage of the feature extraction the global statistics of the speech signal are the same independently of what was actually spoken. As with cepstral mean and variance normalization, the feature space dimension are normalized independently of each other. Hence, only variations that are decorrelated at the normalization stage can be treated properly.

When sufficient amounts (minutes) of adaptation data are available, a non-parametric histogram based approach can be used to estimate the distribution of the training and test data and define an appropriate transformation. This approach will be evaluated in section 5. For real time applications that require short delays in signal analysis, a parametric transformation function should be used which allows for robust parameter estimation on a short data window.

Quantile equalization as introduced in [3] and [4] is a parametric type of histogram normalization. It relies on estimating the signals' cumulative density functions (CDFs) based on quantiles (typically four) instead of the full histograms. A transformation function is calculated that minimizes the mismatch between the quantiles of the current test utterance and those estimated on the training data (Figure 1). Depending on where in the feature extraction the transformation shall take place, different transformation functions may be appropriate. On the CarNavigation corpus, a power function applied to the Mel-scaled filterbank channels of reduced dynamic range gave good results [4]. To ensure that the coefficients are positive, the 10th root was used for dynamic range reduction instead of the logarithm. It turned out, that replacing the logarithm by the 10th root alone reduces the error rate similar to cepstral variance normalization (Table 2, #3).

As it was shown in [4], quantile equalization requires as little as one second of data to estimate the transformation function reliably. Furthermore it is possible to combine quantile equalization with mean normalization in a way that does not induce additional delay. Table 2 gives recognition results for joint quantile and mean normalization (#4). The total delay is 500ms with a window length of 1s to estimate the quantiles and the mean.

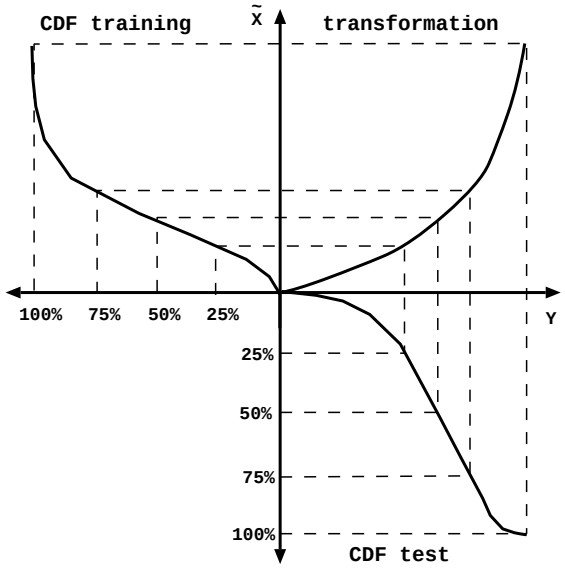

Fig. 1. Applying a parametrized transformation to make the four test quantiles match the training quantiles (CDF: cumulative density function).

So far, quantile equalization had been applied to the test data only. Better results can be obtained when the training data are normalized, too. First the target quantiles are estimated on all training data. Then quantile equalization is applied to match the distribution of each training utterance to the target distribution, before acoustic models are trained on the normalized data. Even though the training data of the CarNavigation corpus were recorded in a clean office environment, a significant reduction in word error rate under mismatch conditions of up to 17% relative was obtained by training data normalization (Table 2, #5).

5. HISTOGRAM NORMALIZATION

In offline applications all utterances of each speaker (on average about 7 min of data on the CarNavigation corpus) can be used to estimate detailed histograms. In previous work [6] we found that histogram normalization performs best at the log-filterbank level. As in the case of quantile equalization both training and test data are normalized with the overall distribution of all training data used as reference (target) histogram. Since the transformation is not restricted to positive values and since it is non-parametric, it can mimic any monotone function for reducing the dynamic range of the filterbank channels. Hence, root compression is not mandatory anymore. The choice of the compression function is only of interest at startup when the target histogram is estimated, since this histogram determines the distribution of training and test data after normalization.

Recognition test summarized in Table 2 show that estimating the full histogram (#6) alone does not yield better results in mismatch conditions than quantile equalization (#5). However, the offline approach does allow for some further refinements leading to significant improvements in recognition performance.

The assumption on the global statistics of the speech signal is sometimes violated. Even if enough speech data are available to ensure approximately equal phoneme frequency for each speaker, and even if the phoneme-dependent distributions are identical for

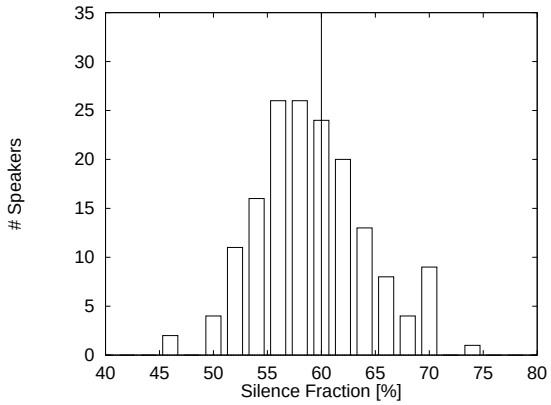

Fig. 2. Histogram over the silence fractions of individual speakers in the CarNavigation training corpus. The vertical line marks the average silence fraction of 60%.

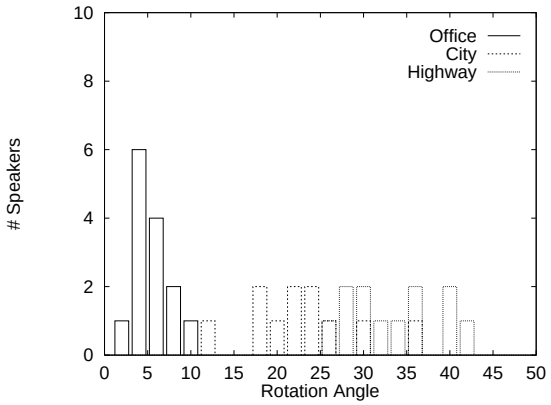

Fig. 3. Histogram over the deviation angles between the first eigenvectors of the speaker dependent covariance matrices and of the target covariance matrix obtained from log-filterbank coefficients on the different CarNavigation test sets.

all speakers in the absence of mismatch, 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.

Figure 2 shows a histogram of the speaker-wise silence fractions on the CarNavigation training corpus. The average silence fraction on this corpus is 60%, but the number varies between 45% and 75% for individual speakers.

The solution is to estimate two independent target histograms for silence and speech [7]. A forced alignment with the reference transcriptions is carried out on the training data, and all acoustic vectors mapped to the silence mixture are assigned to the silence histogram, all other vectors to the speech histogram.

In the normalization step, the silence fraction 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, the silence fraction of test speakers is determined either in a preliminary recognition pass (two-pass recognition) or using a speech/silence detector.

Next an adapted target histogram is computed for each speaker by linear interpolation between the cumulative speech and silence histograms. The adapted target histogram is used for normalization as before. Recognition tests (Table 2) show that explicit silence fraction treatment (#7) reduces the word error rate by another 7% to 20% relative to baseline histogram normalization (#6).

6. FEATURE SPACE ROTATIONS

The second basic assumption of histogram normalization is a feature space in which the considered variations are approximately decorrelated. Previous tests have suggested that this condition is best met at the filterbank level, i.e. that the variations are approximately decorrelated in the frequency domain [6]. 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, the feature space may be rotated in addition to histogram normalization [7].

At first, a covariance matrix is computed over all training vectors. A set of target eigenvectors is calculated and sorted in descending order of their corresponding eigenvalues. It turns out that the first eigenvalue is significantly larger than all others. Hence, the feature space has one preferred direction with large scatter, and along the other principal axes data scatter is much smaller.

Next the covariance matrix and eigenvector basis for each speaker is derived. The first eigenvector, i.e. the direction of the principal axis with largest data scatter, usually differs from the direction of the first target eigenvector in the 20-dimensional log-filterbank feature space. On the CarNavigation corpus, the rotation angles increase with the mismatch. Whereas on the office test data the average rotation angle is 6 deg, it increases to 23 deg on the city and 32 deg on the highway data (Figure 3).

To account for this deviation, a speaker-dependent transformation matrix is calculated that rotates the feature space in the plane spawned by the first speaker-dependent and the first target eigenvector. The matrix is designed such that the speaker's feature space remains undistorted, but the principal axis with largest data scatter becomes identical for all speakers. Details on the transformation matrix can be found in [7].

In the experiments the rotation matrix was computed and applied to both test and training data. Results are reported in Table 2 in combination with silence fraction adapted histogram normalization and cepstral mean normalization (#8). If applied in the right order, feature space rotation and histogram normalization together perform better than rotation and histogram normalization alone. We find that in general one should apply the normalization method first that gives most gain in recognition performance alone, i.e. histogram normalization first. We observed significantly reduced rotation angles after histogram normalization which might therefore be estimated more reliably.

7. VOCAL TRACT LENGTH NORMALIZATION

The shape and size of the human vocal tract differs from speaker to speaker, with female speakers having on average shorter vocal tracts than male. These differences result in a shift of formant frequencies which is approximately inverse proportional to the length of the vocal tract [2]. The idea of VTN is to warp the frequency axis during signal analysis and shift the formants to their "canonical" position.

Vocal tract length normalization has been studied by numerous groups (e. g. [2],[5]), the RWTH setup was described in [10]. Here we applied piecewise-linear frequency warping in training and test. Warping factors for training speakers were estimated in a maximum likelihood framework using a low-resolution (single

Table 2. Recognition test results for the CarNavigation corpus with different normalization techniques: Cepstral mean normalization (CMN), cepstral variance normalization (CVN), filter bank mean normalization (FMN), quantile equalization (QE) in training and test (QETT), histogram normalization (HN) with silence fraction treatment (HNSIL), feature space rotation (ROT), and vocal tract length normalization (VTN). Logarithm (log) and 10th root (root) were used to reduce the dynamic range of the filterbank channels.

#	Normalization	WER [%]		
		Office	City	High.
0	log, no normalization	2.8	68.0	99.0
1	log, CMN	2.9	31.6	74.2
2	log, CMN, CVN	4.2	20.8	39.7
3	root, FMN	2.8	19.9	40.1
4	root, FMN, QE	3.2	11.7	20.1
5	root, FMN, QETT	3.3	10.1	16.7
6	log, CMN, HN	2.8	10.2	16.6
7	log, CMN, HNSIL	2.6	8.2	14.3
8	log, CMN, HNSIL, ROT	2.4	7.1	11.1
9	log, CMN, HNSIL, ROT, VTN	2.2	6.6	10.4

densitiy) acoustic model. For test, speaker-wise warping factors were obtained in a two-pass scheme. An acoustic model trained with all normalizations excluding VTN (Table 2, #8) was used in a first recognition pass to obtain a preliminary transcription of the utterances. The transcription was then used in connection with a fully normalized acoustic model to find the warping factor with maximum likelihood.

VTN typically yields a reduction of word error rate in the order of 10% relative. Larger reductions were reported for "simple" tasks or acoustic models, whereas the WER reduction fell typically well below 10% relative for large vocabulary systems with advanced acoustic modeling trained on a large amount of data. On the CarNavigation corpus we achieved between 6% and 8% relative reduction of WER (Table 2, #9) compared to the best offline system including cepstral mean subtraction, histogram normalization with silence fraction treatment, and feature space rotation.

8. SUMMARY

In this paper various normalization techniques applied during feature extraction where compared. Although the approaches have different levels of complexity the common idea behind all of them is the reduction of an eventual mismatch between the data distribution of the current test utterances and the data the system was trained on. A database recorded in cars was used to investigate how the techniques perform under adverse acoustic conditions. Recognition test results are summarized in Table 2.

The word error rates of the baseline system with cepstral mean normalization were 2.8% (office), 31.6% (city), and 74.2% (highway). The best online approach using quantiles to estimate the cumulative density functions of the signal yielded 3.3%, 10.1%, and 16.7%. An offline setup with non-parametric histogram normalization including silence fraction treatment, feature space rotation, and vocal tract length normalization lead to the best result of 2.2%, 6.6%, and 10.4%.

We have shown that normalization techniques are vital for conditions with major mismatch between training and test conditions. Whereas essentially no word was correctly recognized on the highway data without normalization, the word error rate could be reduced to a reasonable level of 17% for an online and 10% for an offline recognition setup.

9. REFERENCES

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

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

[3] F. Hilger, H. Ney: Quantile Based Histogram Equalization for Noise Robust Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, pp. 1135–1138, Aalborg, Denmark, Sept. 2001.

[4] F. Hilger, S. Molau, H. Ney: Quantile Based Histogram Equalization For Online Applications. Proc. *Int. Conf. on Spoken Language Processing*, vol. 1, pp. 237–240, Denver, CO, USA, Sept. 2002.

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

[6] S. Molau, M. Pitz, 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.

[7] S. Molau, F. Hilger, D. Keysers, H. Ney: Enhanced Histogram Normalization in the Acoustic Feature Space. To appear in: Proc. *Int. Conf. on Spoken Language Processing*, vol. 1, pp. 1421–1424, Denver, CO, USA, Sept. 2002.

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

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

[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*, pp. 1671–1674, Istanbul, Turkey, June 2000.

[11] J. de Veth, B. Cranen, L. Boves: Acoustic Features and Distance Measure to Reduce Vulnerability of ASR Performance due to the Presence of a Communication Channel and/or Background Noise. In: J.-C. Junqua, G. van Noord (Eds.), *Robustness in Language and Speech Technology*. Kluwer Academic Publishers, Doordrecht, the Netherlands, pp. 9–45, 2001.

[12] M. Westphal: The Use of Cepstral Means in Conversational Speech Recognition. Proc. *European Conf. on Speech Communication and Technology*, pp. 1143–1146, Rhodes, Greece, Sept. 1997.