Quantile Based Histogram Equalization for Noise Robust Speech Recognition

Florian Hilger and Hermann Ney

Lehrstuhl für Informatik VI
RWTH Aachen – University of Technology
Ahornstr. 55
D – 52056 Aachen, Germany
{hilger, ney}@informatik.rwth-aachen.de

Abstract

This paper describes an approach to increase the noise robustness of automatic speech recognition systems by, transforming the signal after Mel scaled filtering, to make the cumulative density functions of the signal's values in recognition match the ones that where estimated on the training data. The cumulative density functions are approximated using a small number of quantiles. Recognition tests on several databases showed significant reductions of the word error rates. On a real life database recorded in driving cars with a large mismatch between the training and testing conditions the relative reductions of the word error rates where over 60%.

1. Introduction

The distortion of a speech signal caused by noise usually deteriorates the recognition performance of automatic speech recognition systems considerably. The quantile based histogram equalization ("quantile equalization") that is introduced here can be applied during MFCC feature extraction to increase the noise robustness. Rather than attempting to remove the contribution of the noise from the spectrum by spectral subtraction or filtering [1], the cumulative density functions of the signal's values after critical band filtering are considered as a whole. If there is a mismatch between the training data and the test utterance, it is reduced by applying transformation functions.

The cumulative density functions on the test utterances are approximated in a non–parametric way by dividing the signal into quantiles (figure 1). Together with the corresponding values determined in training this gives a set of points that can be used to calculate the transformation functions.

2. Quantile equalization

The output of the k th Mel scaled filter before taking the logarithm shall be denoted $Y_k[t]$. A transformation function T_k is applied to each of these outputs to make the cumulative density functions of the equalized values $Y_k^{eq}[t]$ match the ones observed in training:

$$Y_k^{eq}[t] = T_k(Y_k[t]) \tag{1}$$

The cumulative density functions can be estimated non–parametrically using histograms. In [2] using a large number of histogram bins spaced equally between the minimum and maximum is proposed. To allow a robust estimation of the cumulative density functions on single (eventually short) utterances, without additional adaptation data, the use of a small number ($N_Q < 10$) of quantiles Q shall be proposed here (figure 1).

Training: In a pass over the training data the $Q_{k,i}^{train}$ for the k th filter are determined for every sentence and the average over all sentences is calculated. Instead of using these individual training quantiles for each filter they can also be averaged over all filters to get better recognition results (table 1):

$$Q_i^{train} = \frac{1}{K}\sum_{k=1}^{K} Q_{k,i}^{train} \tag{2}$$

Note that no transformation of the data is applied during training. If a reference set trained as usual on this data is already available it can be used as it is. The quantiles are just calculated in one pass over the data without modifying anything.

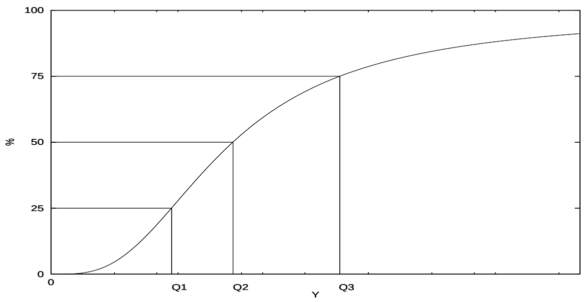

Figure 1: *A cumulative density function shown with three quantiles Q_i of 25% each ($N_Q = 4$).*

Recognition: During recognition the quantiles $Q_{k,i}$ are calculated only on the utterance that is to be recognized. In practical applications it would be possible to include preceding sentences to get better estimates, while the amount of data grows. The intention here was to show that even one short utterance is enough to get sufficiently good estimates for the transformation, if the number of quantiles and the type of the transformation function is chosen adequately.

Calculating individual quantiles $Q_{k,i}$ for each filter is mandatory in recognition to be able to cope with the different spectral characteristics of occuring noises. Given the set of points $\{Q_{k,i}, Q_i^{train}\}$ the transformation functions can be calculated. In the following piecewise linear and power function transformations shall be introduced. In both cases the functions are restricted to the area below the diagonal in figure 2 by defining lower bounds for the $Q_{k,i}$:

$$\textbf{if} \quad Q_{k,i} < Q_i^{train} \quad \textbf{then} \quad Q_{k,i} = Q_i^{train} \tag{3}$$

This has to be done to avoid unintended raising of a noise that might be lower than in training.

2.1. Piecewise linear equalization

The piecewise linear transformation function simply connects the given points $\{Q_{k,i}, Q_i^{train}\}$, it has the general form:

$$Y_k^{eq}[t] = a_{k,i} \cdot Y_k[t] + b_{k,i} \tag{4}$$

The values of $a_{k,i}$ and $b_{k,i}$ depend on the segment that has to be used to transform the current $Y_k[t]$.

$$\textbf{if } Y_k[t] \in [Q_i, Q_{i+1}[\textbf{ then } a_{k,i} = \frac{Q_{i+1}^{train} - Q_i^{train}}{Q_{k,i+1} - Q_{k,i}} \tag{5}$$

$$b_{k,i} = Q_i^{train} - a_{k,i} \cdot Q_{k,i} \tag{6}$$

If short utterances are used in recognition there might be no or little speech activity in some of the filters. The signal would then be distorted in an undesired way if $a_{k,i}$ and $b_{k,i}$ where calculated with the estimated quantiles. To increase the robustness of the approach in such cases, the 0th and the N_Qth quantiles can be set to predefined values:

$$Q_0 \; = Q_0^{train} = 0 \tag{7}$$

$$Q_{N_Q} = Q_{N_Q}^{train} = \infty \tag{8}$$

In the practical implementation ∞ would be some value that is much larger than any maximum that could occur in the data.

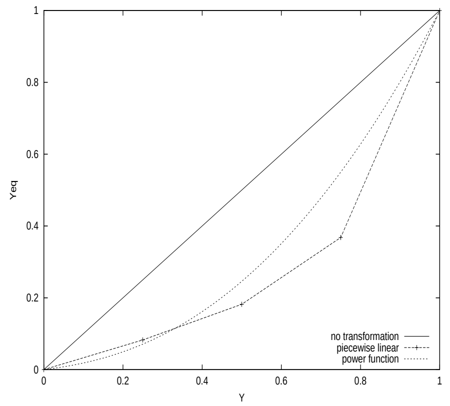

Figure 2: *Transformation from Y to Y^{eq} with a piecewise linear and a power function. Scaling: Y/Q_{N_Q}*

2.2. Power function equalization

The discontinuous derivative of the piecewise linear transformation function can cause problems on data with a large dynamic range. In this case the use of a power function gives better recognition results. The filter outputs are scaled to the interval [0, 1], then the power function is applied and the resulting values are scaled back to the original range:

$$Y_k^{eq}[t] = T(Y_k[t]) =$$
$$Q_{k,N_Q} \left(\alpha \left(\frac{Y_k[t]}{Q_{k,N_Q}} \right)^{\gamma} + (1-\alpha) \frac{Y_k[t]}{Q_{k,N_Q}} \right) \tag{9}$$

The parameters α and γ are chosen to minimize the squared distance to the points $\{Q_{k,i}, Q_i^{train}\}$:

$$\{\gamma, a\} = \underset{\{\gamma,a\}}{\operatorname{argmin}} \left(\sum_{i=1}^{N_Q-1} \left(T(Q_{k,i}) - Q_i^{train} \right)^2 \right) \tag{10}$$

Introducing a second degree of freedom with the parameter α is absolutely necessary to reach good recognition results. A function of the form $Y^{eq}[t] = (Y[t])^{\gamma}$ is not sufficient.

3. Recognition results

The first optimization tests where conducted with the RWTH speech recognition system on the MoTiV – CSDC2 database. To show the genericity of the approach the settings that where found to be optimal in those tests where also applied on a larger vocabulary car navigation database and on SpeechDat–Car databases, using the HTK speech recognition toolkit in accordance with the ETSI Aurora evaluation settings [3] [4].

3.1. MoTiV – CSDC2 database

Database definitions: MoTiV – CSDC2 [5] is a database with training and testing data recorded in car environment. Sampling rate is 16kHz. The German digit string utterances where used for the following tests. 582 utterances (28min) for training and 195 utterances (9min) for testing.

Baseline recognizer setup:

- RWTH speech recognition system
- MFCC feature extraction (RWTH implementation [6])
  20 filters;
  16 cepstral coeff. + 16 first deriv. + 1 second deriv.
  = 33 – dimensional feature vector
- Cepstral mean normalization
- Linear discriminant analysis:
  Reduction from 3 × 33 components to 33
- Word models for 11 German digits, including "zwo"
- Gender independent models
- Gaussian mixtures
  - Pooled diagonal covariance matrix
  - Number of densities: 7.3k

Table 1: *Recognition results on the MoTiV – CSDC2 car database. Optimization of the quantile number N_Q for piecewise linear quantile equalization. INDIVIDUAL: individual training quantiles for each filter, AVERAGED: averaged training quantiles.*

MoTiV – CSDC2		
	Word Error Rate [%]	
N_Q	INDIVIDUAL	AVERAGED
2	5.7	5.5
3	5.5	5.6
4	5.3	4.9
6	5.3	5.1
12	5.5	5.6

Table 2: *Recognition results on the MoTiV – CSDC2 car database. NLN + RA: noise level normalization and reference adaptation [6], LINEAR: piecewise linear quantile equalization, POWER: power function equalization.*

MoTiV – CSDC2				
SNR	Word Error Rate [%]			
[dB]	baseline	NLN + RA	LINEAR	POWER
8	6.6	5.1	4.9	4.8

Recognition results: The number of quantiles used has to be optimized. When using too few quantiles the transformation is not optimal. On the other hand it is not reasonable to calculate too many quantiles on short sentences. On this database with an average sentence length of 2.7 seconds the optimal number was found to be four (table 1).

Individual training quantiles for each filter $Q_{k,i}^{train}$ where used in the first column of table 1. All results are better than the baseline of 6.6% word error rate (table 2), but the 5.1% obtained using a combination of noise level normalization and reference adaptation [6] are not reached. The recognition performance can be increased by averaging the training quantiles over all filters Q_i^{train} (Equation 2). As the second column of table 1 shows the best result then is 4.9% word error rate. This corresponds to a relative reduction of 26% compared to baseline.

When using four quantiles with the power function transformation one error less occurs (51 vs. 52), the word error rate is 4.8%. Apparently on this database it does not make a significant difference if piecewise linear or power function transformation is used.

3.2. Isolated word car navigation database

Database definitions: The database contains isolated German words. The training data (13h 37min) was recorded in a quiet office environment, with the microphone at 30cm distance directly in front of the speaker. Sampling rate was 16kHz. The testing sets of approximately 75min each where recorded in the office (avg. SNR 21dB) and driving cars (city– and highway traffic, avg. SNRs 9dB and 6dB), with the speaker sitting in the passenger seat and the microphone mounted above the speaker on the visor. See [6] for more details. Note that none of the 2100 equally probable words used for the recognition tests occurred in training data.

Baseline recognizer setup:

- RWTH speech recognition system
- MFCC feature extraction (RWTH implementation [6])
  20 filters;
  16 cepstral coeff. + 16 first deriv. + 1 second deriv.
  = 33 – dimensional feature vector
- Cepstral mean normalization
- Linear discriminant analysis:
  Reduction from 3 × 33 components to 33
- Triphone classification and regression tree:
  700 tied states
- Gender independent models
- Gaussian mixtures
  - Pooled diagonal covariance matrix
  - Number of densities: 21k
- Single word recognition, vocabulary 2100 words

Recognition results: Here the utterances of the testing data only consist of single words, but still one utterance is sufficient to estimate four quantiles and obtain good recognition results (table 3). Piecewise linear and power function transformation perform similarly again.

The results show how the relative reduction of the word error rate rises with a growing mismatch between training and testing conditions. On the clean testing data there is a slight loss in the recognition performance. Using power function transformation the relative reduction of the error rate compared to baseline is 62% on the city testing set, it rises to 72% on the highway data. These numbers are much higher than the one obtained on the MoTiV data where the mismatch was not as big.

Table 3: *Recognition results on the isolated word car navigation database. NLN + RA: noise level normalization and reference adaptation [6], LINEAR: piecewise linear quantile equalization, POWER: power function equalization.*

German Car Navigation Database				
	Word Error Rate [%]			
	baseline	NLN + RA	LINEAR	POWER
office	3.8	4.3	4.0	4.0
city	35.5	14.7	14.4	13.6
highway	78.0	24.1	21.4	21.8

Compared to the combination of noise level normalization and reference adaptation [6], quantile equalization on its own always performs better. It could be expected that a combination of quantile equalization in the front–end and reference adaptation in the back–end would decrease the error rate even further, but tests showed that this is not the case.

3.3. Aurora SpeechDat–Car databases

Database definitions: The databases are digit string subsets of the SpeechDat–Car data in Danish [7], German [8] and Spanish [9]. The sampling rate is 8kHz. Training and test sets where used as defined for the ETSI Aurora evaluations [3] [4].

Baseline recognizer setup:

- HTK speech recognition toolkit (Aurora evaluation settings [3])
- MFCC feature extraction (Aurora WI007 [3])
  23 filters;
  13 cepstral coeff. + 13 first deriv. + 13 second deriv.
  = 39 – dimensional feature vector
- Word models of fixed length (16 states) for the digits
- Gender independent models
- Gaussian mixtures
  - Pooled diagonal covariance matrix
  - Number of densities: 552

Recognition results: Please note that the results reported in table 4 are word accuracies, to be in accordance with the Aurora evaluation scheme.

The most significant difference between the SpeechDat databases and the previous databases is the poor performance of the piecewise linear transformation. On most of the test sets the resulting word error rates are much higher than baseline. A look at the values the Y_k take shows the reason for this problem: due to the large dynamic range of the signal, the $Q_{k,i}$ in recognition can be several hundred times larger than the corresponding Q_i^{train}. This results in a transformation function that changes the signal significantly and has large differences between the $a_{k,i}$ (equation 2) for the different linear segments. These major discontinuities in the derivative of the transformation function apparently deteriorate the recognition performance.

The power function transformation avoids this problem. Its derivative is continuous and the reductions of the word error rates are in the order of magnitude that would be expected regarding the previous experiments. The observation that the relative reduction of word error rate rises with growing mismatch is confirmed, for example on the Danish data these numbers are 19% for the well matched condition, 27% for medium mismatch and 48% for the high mismatch. Looking at the average relative improvement the numbers for Spanish (28.3%) and

Danish (29.4%) are similar, while the result for German is lower (17.6%). This can be explained by the higher baseline accuracy on the German data.

Table 4: *Recognition results on the Aurora SpeechDat databases. LINEAR: piecewise linear quantile equalization, POWER: power function equalization. The weighted average of the word accuracies and the relative reduction of word error rate are calculated according to the Aurora evaluation scheme (0.40·WellMatched+0.35·MediumMism.+0.25·HighMism.).*

Danish SpeechDat–Car			
	Word Accuracy [%]		
	baseline	LINEAR	POWER
WM	77.8	79.59	82.06
MM	47.4	49.48	61.72
HM	31.9	20.44	65.08
wght. avg.	55.69	54.26	70.70
rel. impr.		0.40	29.38

German SpeechDat–Car			
	Word Accuracy [%]		
	baseline	LINEAR	POWER
WM	90.58	89.16	92.21
MM	79.06	68.37	81.19
HM	74.28	67.48	81.59
wght. avg.	82.47	76.46	85.70
rel. impr.		-30.51	17.59

Spanish SpeechDat–Car			
	Word Accuracy [%]		
	baseline	LINEAR	POWER
WM	86.87	82.98	89.11
MM	73.74	46.07	80.48
HM	42.23	28.72	71.01
wght. avg.	71.11	56.50	81.56
rel. impr.		-54.57	28.26

4. Conclusions

This paper has described an approach to increase the noise robustness of automatic speech recognition systems by reducing the mismatch between training and testing conditions with a transformation of the Mel scaled filter bank outputs. The transformation functions are estimated on the basis of the cumulative density function's quantiles in training and recognition.

Piecewise linear and power function transformations were investigated. While the piecewise linear transformation turned out to be problematic for signals with a high dynamic range, the power function transformation performed well on all the databases that where tested. The relative reduction of the word error rates was highest for scenarios with a large mismatch between training and testing. Since the approach does not require any retraining of the references, its use can be recommended when porting a system, that was originally trained on clean data, to applications in a noisy environment.

It was shown that the approach works on low complexity digit recognition tasks (0.5 – 7k references), as well as on higher complexity tasks (21k references), such as the recognition of 2100 equally probable words modeled with triphones. The baseline results with the RWTH speech recognition system always included cepstral mean normalization and linear discriminant analysis, still the relative reductions of the word error rates by adding quantile equalization where high. Apparently quantile equalization interacts well with these approaches. Future work will show in detail to which extent quantile equalization can contribute to a combination of different approaches as it is presented in [10].

Quantile equalization is computationally inexpensive. Regarding the use in real–time online applications, it was shown that one short utterance is sufficient to estimate transformation functions when the proposed settings are used. This raises the hope that the approach does not even need an entire utterance for the estimation. Future investigations will have to show if the delay can be pushed down from one utterance to a few time frames, if incremental estimation is used.

5. References

[1] J.-C. Junqua, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, chapter 9, Kluwer Academic Publishers, Boston, 1996.

[2] S. Dharanipragada and M. Padmanabhan, "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition," in *Proceedings of the International Conference on Spoken Language Processing 2000*, Beijing, China, Oct. 2000, vol. IV, pp. 556 – 559.

[3] H.-G. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," in *ASR2000 – International Workshop on Automatic Speech Recognition*, Paris, France, Sept. 2000, pp. 181 – 188.

[4] D. Pearce, "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends," in *Applied Voice Input/Output Society Conference (AVIOS2000)*, San Jose, CA, May 2000.

[5] D. Langmann et al., "CSDC – The MoTiV Car Speech Data Collection," in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, May 1998.

[6] F. Hilger and H. Ney, "Noise Level Normalization And Reference Adaptation For Robust Speech Recognition," in *ASR2000 – International Workshop on Automatic Speech Recognition*, Paris, France, Sept. 2000, pp. 64 – 68.

[7] B. Lindberg, "Danish SpeechDat–Car Database for ETSI STQ Aurora Advanced DSR," in *Documentation on the Danish Aurora Project Database CD–ROMs*, Jan. 2001.

[8] L. Netsch, "Description and Baseline Results for the Subset of the Speechdat–Car German Database used for ETSI STQ Aurora WI008 Advanced Front–end Evaluation," in *Documentation on the German Aurora Project Database CD–ROMs*, Jan. 2001.

[9] D. Macho, "Spanish SDC–Aurora Database for ETSI STQ Aurora WI008 Advanced Front–End Evaluation: Description and Baseline Results," in *Documentation on the Spanish Aurora Project Database CD–ROMs*, Nov. 2000.

[10] B. Andrassy F. Hilger and C. Beaugeant, "Investigations on the Combination of Four Algorithms to Increase the Noise Robustness of Automatic Speech Recognition Front–Ends for Real World Car Data," submitted to the *7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept. 2001.