Evaluation of Quantile Based Histogram Equalization in Combination with Different Root Functions

Florian Hilger[1] and Hermann Ney[2]

[1] *Telenet GmbH Kommunikationssysteme, Marsstr. 33, 80335 Munich, Germany,* f.hilger@telenet.de

[2] *Lehrstuhl für Informatik VI, RWTH Aachen, Ahronstr. 55, 52056 Aachen, Germany,* ney@informatik.rwth-aachen.de

Abstract

This paper presents an evaluation of the RWTH large vocabulary speech recognition system on the Aurora 4 noisy Wall Street Journal database. First, the influence of different root functions replacing the logarithm in the feature extraction is studied. Then quantile based histogram equalization is applied, a parametric method to increase the noise robustness by reducing the mismatch between the training and test data distributions. Putting everything together, the word error rate could be reduced from 45.7% to 25.5% (clean training data) and from 19.5% to 17.0% (multicondition training data).

Logarithm and Root Functions

In a conventinal Mel-frequency cepstral coefficient (MFCC) feature extraction a logarithm is applied after the Mel-scaled filterbank to reduce the dynamic range of the signal. This logarithm can be replaced by a root function. The general relation between root/power functions and the logarithm can be expressed as follows:

$$f_r(x) = \frac{x^r - 1}{r} \tag{1}$$

A comparison between the Taylor series expansion for this expression and the one for the logarithm reveals that the limit of $f_r(x)$ for $r \to 0$ is the logarithm:

$$\lim_{r \to 0} f_r(x) = \log(x) \tag{2}$$

A special property of the logarithm is that a constant gain applied to the input will result in a simple shift of the output. Such a shift, typically introduced by the transmission channel, can be eliminated by mean normalization i.e. a subtraction of the longterm mean. This nice property is lost when using $r > 0$, but there is experimental evidence that this is no drawback. On the contrary, the noise robustness can be increased when replacing the logarithm by an appropriate root function. The experiments presented in [1] show that a value of r around 0.1 gave best recognition results in noisy conditions.

The approach of replacing the logarithm by a root can be generalized even more: the constant shift of -1 and scaling by $1/r$ in equation 1 will both be applied during training and recognition, so they will not affect the final recognition result. So, an expression of the type x^r was used for the experiments described below.

Quantile Equalization

Histogram based methods that remove an eventual mismatch between the distribution of the current test data and the distribution of the system's training data have been successfully used to increase the robustness of speech recognition systems. If some minutes of test data are available to estimate high resolution histograms, a non parametric transformation to reduce the mismatch can be calculated and applied.

An alternative which can be used for online systems that only allow short delays is quantile based histogram equalization [2]. This method approximates the cumulative distribution functions of the signals using a few quantiles and then optimizes the parameters of a transformation function based on these values.

Within this work a two step transformation scheme was applied [2]. First, a power function transformed the individual output channels of the Mel–scaled filter bank:

$$T_k(Y_k) = Q_{k,N_Q} \left(\alpha_k \left(\frac{Y_k}{Q_{k,N_Q}} \right)^{\gamma_k} + (1 - \alpha_k) \frac{Y_k}{Q_{k,N_Q}} \right) \tag{3}$$

here Y_k denotes the output of the kth filter (obviously Y_k also depends on the time frame, but the index for the is dropped). $Q_{k,i}$ is the ith quantile for the kth channel. With N_Q being number of quantiles, Q_{k,N_Q} is the maximal quantile. The transformation parameters α_k and γ_k are optimized to minimize the squared distance between the current recognition quantiles and the training quantiles. In the second step, neighboring filter channels are combined linearly ($\tilde{Y}_k = T_k(Y_k)$):

$$\tilde{T}_k(\tilde{Y}_k) = (1 - \lambda_k - \rho_k)\tilde{Y}_k + \lambda_k \tilde{Y}_{k-1} + \rho_k \tilde{Y}_{k+1} \tag{4}$$

The transformation parameters λ_k and ρ_k are again chosen in order to minimize the squared distance to the training quantiles. In the following experiments estimation of the quantiles is always carried out utternace wise.

Experimental Setup

Database: The Aurora 4 database [3] provided by ELRA was used for the recognition experiments. Aurora 4 is based on the Wall Street Journal 5k (WSJ) database that was used in the ARPA evaluations. Different noise samples at various SNRs were added, to turn the original data recorded in quiet studio conditions into a noisy database. There is a clean and a noisy training data set and a total of 14 test sets with different types of added noises.

Recognizer: a MFCC feature extraction front end with 20 Mel scaled filter bank channels, 16 cepstral coefficients and a linear discriminant analysis applied to 7 successive cespstral coefficient vectors was used. The resulting feature vector was 32 dimensional.

The setup of the RWTH speech recognizer was optimized on the clean test set: across–word triphone models, a phonetic classification and regression tree with 4001 tied states, 210k–230k Gaussian densities, 5k recognizer vocabulary and a trigram language model. The real time factor was about 2 on the clean test set and around 10 for the noisy data sets (1800MHz AMD Athlon).

Experimental Results

Table 1 shows the correlation *cor* between the clean and noisy test data sets given by

$$cor = \operatorname{cov}\left(Y_k^c, Y_k^n\right) / \sqrt{\operatorname{var}\left(Y_k^c\right) \operatorname{var}\left(Y_k^n\right)} \qquad (5)$$

were Y_k is the output of the k-th filterbank channel after logarithm or root, for the clean c and noisy n signal respectively. The covariance cov and variance var are calculated over all filter channels and time frames.

Obviously the result is $cor = 1$ for the clean data and the correlation decreases with growing mismatch. While the overall average is 0.67 for the logarithm, it is 0.72 when using 10th root and 0.76 when the 5th root is applied (table 1). Of course one can not expect that this increase of the correlation between the clean data set and the noisy data sets is directly related to the reduction of the average word error rate of the recognition.

The results for the square, 2nd root illustrate that: the initial error rate on the clean test set 1 is very high, so in this case the large correlation between that data set and the noisy ones is no indication for good overall recognition results. Nonetheless, the correlation can still be considered to be a measure for the mismatch between the clean and noisy data, so if the initial error rate on the clean data set is low and the average correlation is high, the average error rate is likely to be low too.

The complete recognition results are shown in table 2. For both training sets the average word error rate, before and after applying quantile equalization, averaged over all 14 test sets is given. In the multicondition case, the error rate reductions obtained by applying the root functions are not as large as in the clean training case, but consistent. Quantile equalization leads to a further improvement in all conditions. The best results were obtained with the 5th respectively 10th root, confirming the results of previous evaluations on smaller vocabulary databases [2].

Conclusions

Simply replacing the logarithm in the MFCC feature extraction by a root function can already significantly improve the noise robustness of automatic speech recognitions systems. On the Aurora 4 database, the combination of this approach with quantile based histogram equalization lead to an overall improvrement of the word error rate from 45.7% to 25.5% (clean training data) and from 19.5% to 17.0% (multicondition training data), when using the 10th root.

Table 1: Correlation (equation 5) between the clean (1) and noisy (2–14) test data sets of the Aurora 4 database compared to the average word error rates and the corresponding error rate on the clean subset (recognizer trained on clean data).

test set	Correlation				
	LOG	20th	10th	5th	2nd
1	1.00	1.00	1.00	1.00	1.00
2	0.73	0.76	0.78	0.83	0.92
3	0.71	0.74	0.76	0.81	0.91
...	...	...	...	...	...
12	0.55	0.57	0.59	0.64	0.74
13	0.60	0.62	0.65	0.69	0.78
14	0.56	0.58	0.61	0.65	0.75
average	0.67	0.70	0.72	0.76	0.84
average WER [%]	45.7	33.2	29.7	24.5	30.7
clean WER [%]	4.5	4.5	4.4	5.1	8.9

Table 2: Comparison of the logarithm in the feature extraction with different root functions on the Aurora 4 database. 2nd – 20th: root instead of logarithm, QEF: quantile equalization with filter combination.

	Word Error Rates [%]	
	clean training	multi. training
LOG	45.7	19.5
20th	33.2	18.0
10th	29.7	**17.8**
5th	**24.5**	18.0
2nd	30.7	26.3
20th QEF	28.8	17.0
10th QEF	25.5	**17.0**
5th QEF	**23.7**	17.4

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE 572/4-1&3.

References

[1] J. Tian and O. Viikki, "Generalized cepstral analysis for speech recognition in noise," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999, pp. 87–90.

[2] F. Hilger, H. Ney, O. Siohan, and F. K. Soong, "Combining neighboring filter channels to improve quantile based histogram equalization," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, Hong Kong, China, Apr. 2003, pp. 640–643.

[3] H.-G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front–ends on a large vocabulary task, Version 2.0, AU/417/02," ETSI STQ-Aurora DSR Working Group, Tech. Rep., Oct. 2002.