Sub-band level Histogram Equalization for Robust Speech Recognition

Vikas Joshi, Raghavendra Bilgi, S. Umesh

Department of Electrical Engineering
Indian Institute of Technology, Madras, India
ee10s001, ee10s009, umeshs@ee.iitm.ac.in

L. Garcia, C. Benitez

Dept of Signal Theory, Telematics and Communications
University of Granada, Spain
luzgm, carmen@urg.es

Abstract

This paper describes a novel modification of Histogram Equalization (HEQ) approach to robust speech recognition. We propose separate equalization of the high frequency (HF) and low frequency (LF) bands. We study different combinations of the sub-band equalization and obtain best results when we perform a two-stage equalization. First, conventional HEQ is performed on the cepstral features, which does not completely equalize HF and LF bands, even though the overall histogram equalization is good. In the second stage, an equalization is done separately on the HF and the LF components of the above equalized cepstra. We refer to this approach as Sub-band Histogram Equalization (S-HEQ). The new set of features has better equalization of the sub-bands as well as the overall cepstral histogram. Recognition results show a relative improvement of 12% and 15% over conventional HEQ in WER on Aurora-2 and Aurora-4 databases respectively.

Index Terms: Histogram Equalization, S-HEQ, Noise robust speech recognition

1. Introduction

Speech Recognition accuracy degrades significantly in noisy conditions. The effects of noise can be broadly classified into two categories. First, random effects of the noise, which result in the loss of information. The second is the change in the overall histogram from that of the clean signal. This causes distortion in the feature space, usually by means of a nonlinear transformation. The distortion leads to mismatch between train and test conditions resulting in poor recognition accuracy. There have been several approaches proposed in literature to compensate for this latter effect of the noise.

Cepstral Mean subtraction (CMS) and Cepstral Mean and Variance Normalization (CMVN) are two commonly followed approaches for reducing the effects of noise. CMS, apart from compensating for channel effects, transforms the train and test features into zero mean features, thus eliminating the effect of the noise on mean of the distribution. Going one step further, CMVN normalizes both mean and variance of the features, compensating the effect of noise on first and second order moments. Histogram Equalization [1, 2] (HEQ) is a more general technique which attempts to compensate even the higher order moments, by matching the distribution of the noisy and clean features. HEQ is a transformation of noisy features, mapping the histogram of the noisy data to match that of the clean speech histogram. Although, it compensates the statistical effects of the noise, the loss of the information due to the random effects of the noise is not recovered. The advantage of HEQ is that it is neither model based nor does it make any assumptions about the underlying distribution of noise or speech cepstra. More importantly, the computation complexity of HEQ is very small when compared to other approaches such as VTS [3, 4]. The effectiveness of the HEQ depends on correctly estimating the histogram, which essentially requires sufficiently large amount of data samples.

HEQ has been performed on Log Filter bank coefficients [5, 6], and on cepstral features [1, 2]. HEQ has also been applied on delta and delta-delta features [4]. Hilger *et.al* [5, 6] have shown the capability to compensate the distortions of certain specific frequencies by applying HEQ in the log filter bank domain. Hung and Fan [7] filter the cepstra along "time", essentially using modulation frequency bands and apply HEQ on them. In our method, we are referring to original frequency (i.e. Hz domain) and not modulation frequency.

Speech energy being often concentrated in low frequency regions, results in low SNR's at high frequency regions. Hence, the idea of equalization on individual filter bank energies seems more appropriate. However, many researchers have argued against equalization on log filter bank, since there is a strong correlation between the filter bank energies and, therefore, an independent equalization is not most appropriate. HEQ performed in cepstral domain, equalizes better as the cepstral features are fairly uncorrelated. However, there has been no attempt to apply *separate HEQ at different frequency bands*. As we show in this paper, even though histogram of the overall cepstra match for clean and noisy conditions, sub-band level histograms do not match well. In this paper, we use the term overall cepstra to clearly distinguish between cepstra and its filtered components which will be described in more detail later. The filtered components are obtained using a simple difference and mean operations on adjacent cepstral coefficients *within* a frame. For an ideal noise compensation algorithm, along with the overall histogram, the histogram at different frequency-bands should also match. In Sub-band HEQ (S-HEQ) approach proposed in this paper, we attempt to match histogram of low pass filtered (LPF) and high pass filtered (HPF) content of cepstral features, along with the overall cepstral histogram. Using S-HEQ features, we obtain significant improvement in recognition results.

The rest of paper is organized as follows. In the next section, we describe the conventional HEQ approach. This is followed by a description of our proposed Sub-band HEQ approach in Section 3. Section 4 contains experimental results and a discussion of the results. Finally, conclusions are presented in Sections 5.

2. Histogram Equalization (HEQ)

HEQ as applied in speech recognition, transforms both train and test features to match a common cumulative distribution func-

tion (CDF), known as reference CDF (ref. CDF). HEQ [1, 2] is based on the fact that a random variable x, with a CDF $C_x(x)$ can be transformed to another random variable y with CDF $C_y(y)$, such that, $C_y(y)$ is equal to reference CDF $C_{ref}(y)$.

$$y = T(x)$$

$$C_y(y_0) = C_x(x_0) = C_{ref}(y_0)$$

$$y_0 = C_{ref}^{-1}(C_x(x_0))$$

Some authors have tried using Normal-distribution as the reference distribution. Clean Speech CDF is another choice, provided sufficient samples are available to estimate the distribution. In our experiments, clean speech CDF from the training speakers is used as reference distribution. The transformation $C_{ref}^{-1}C_x$ is a monotonically non-decreasing function, and hence it can only compensate for noise which monotonically transforms the features. Further, only the mismatch in the distribution is compensated but not the random effects of noise.

Conventional HEQ matches the overall cepstral histogram, but does not attempt to compensate differently, the different frequency-bands. Fig. 1(g), 1(c) and 1(e), show the histograms of overall cepstra, low pass filter (LPF) cepstra and high pass filter (HPF) cepstra. Histogram of overall cepstral features match well, whereas there is a considerable mismatch for LPF and HPF components of the cepstra. We explore into ways to match overall cepstral histogram and their LP and HP histograms and provide details of how the filtered components are obtained.

3. Sub-band Histogram Equalization (S-HEQ)

An ideal noise compensation algorithm should not only match the overall cepstral histogram, but also histogram of features obtained by any transformation. For HEQ, low pass and high pass filtered cepstral histogram do not match as shown in Fig. 1(c) and Fig. 1(e).

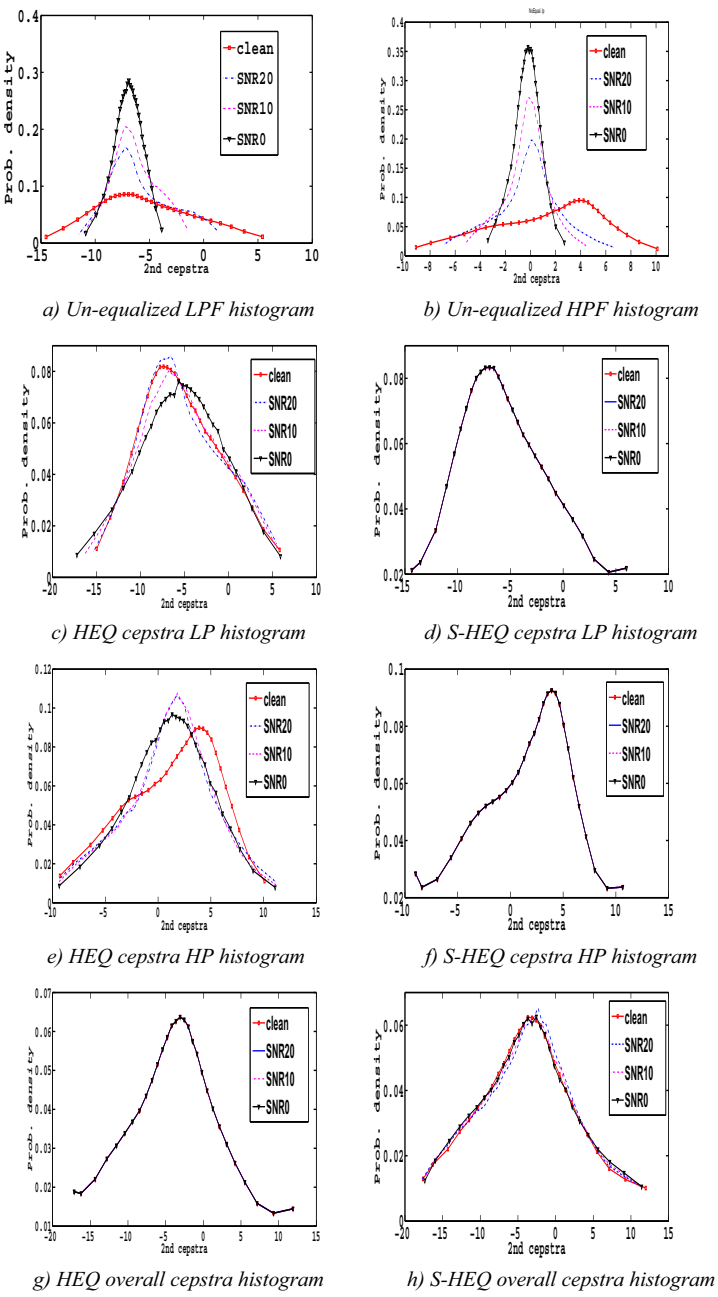
The proposed method aims to equalize individually the LPF and HPF cepstral histogram along with overall cepstral histogram. LP and HP regions have different SNR levels and independent equalization could be more effective. Here, we are assuming that the LPF and HPF cepstra are independent.

We use the idea that differencing adjacent cepstra within a frame in cepstral domain crudely corresponds to high-pass filtering (or high frequency components in the original frequency (Hz) domain). Similarly, averaging (or finding the sample mean) of adjacent cepstra corresponds to choosing the low-frequency spectral components.

Consider conventional HEQ equalization of MFCC vectors by the following equation.

$$c^{heq}(n) = T_{heq}(c(n)) \quad (1)$$

where $c(n)$ is the observed noisy n^{th} cepstral coefficient of MFCC vector, $n = 0, 1, ...12$ representing the 13 cepstral coefficients. $c^{heq}(n)$ is the equalized cepstra. $T_{heq} = C_{ref}^{-1}$.C is the equalizing transformation, where C_{ref} is clean speech CDF which is used as reference CDF and C represents the CDF of the observed noisy MFCC vectors that need to be transformed. This conventional HEQ will normalize the overall cepstra. Using this normalized cepstra, we now normalize the transformed cepstral-coefficients corresponding to high and low-frequency band as follows.



a) Un-equalized LPF histogram

b) Un-equalized HPF histogram

c) HEQ cepstra LP histogram

d) S-HEQ cepstra LP histogram

e) HEQ cepstra HP histogram

f) S-HEQ cepstra HP histogram

g) HEQ overall cepstra histogram

h) S-HEQ overall cepstra histogram

Figure 1: Histograms of 2^{nd} cepstral coefficient for Aurora-2 database, for a) Un-equalized LPF cepstra, b) Un-equalized HPF cepstra, c) HEQ equalized and then LPF cepstra, d) S-HEQ equalized and then LPF cepstra, e) HEQ equalized and then HPF cepstra, f) S-HEQ equalized and then HPF cepstra, g) HEQ equalized overall cepstra, h) S-HEQ equalized overall cepstra

The cepstral transformation corresponding to high-frequency band are implemented using simple differencing of adjacent cepstral coefficients within a frame, i.e.

$$c_{hp}^{heq}(n) = \begin{cases} \frac{[c^{heq}(n) - c^{heq}(n-1)]}{2} & n = 1, 2...12 \\ c^{heq}(0) & n = 0 \end{cases} \quad (2)$$

And, the cepstral transformation corresponding to low-frequency band is then obtained by

$$c_{lp}^{heq}(n) = c^{heq}(n) - c_{hp}^{heq}(n) \quad (3)$$

$c_{lp}^{heq}(n)$ and $c_{hp}^{heq}(n)$ are the transformed cepstral coefficients corresponding to low-frequency and high-frequency spectra *for a particular frame*.

We now apply HEQ separately on each of the high-pass (HP) and low-pass (LP) transformed cepstra

$$\hat{c}_{hp}^{heq}(n) = T_{hp}(c_{hp}^{heq}(n)) \qquad \hat{c}_{lp}^{heq}(n) = T_{lp}(c_{lp}^{heq}(n)) \quad (4)$$

In the above equation, $T_{hp} = C_{refhp}^{-1}.C$ and $T_{lp} = C_{reflp}^{-1}.C$ are equalizing transforms for HP and LP cepstra respectively. C_{refhp}^{-1} and C_{reflp}^{-1} are CDF of HPF and LPF unequalized cepstra. Reference CDF's are obtained from LPF and HPF components of clean unequalized cepstra.

These two separately equalized cepstra are then added to obtain the final equalized cepstra:

$$\hat{c}_{com}(n) = \hat{c}_{lp}^{heq}(n) + \hat{c}_{hp}^{heq}(n) \quad (5)$$

where $\hat{c}_{hp}^{heq}(n)$ and $\hat{c}_{lp}^{heq}(n)$ are equalized HP and LP of $c^{heq}(n)$. $\hat{c}_{com}(n)$ are the features used in the recognition.

The block diagrams that implement the equations discussed above are shown in the Fig. 2 and 3. Fig. 2 depicts the generation of reference CDF's. Fig. 3 shows S-HEQ strategy.

Therefore, our proposed S-HEQ method is a two-stage equalization process. First, HEQ is applied on the conventional MFCC to get the HEQ equalized cepstra. This does not optimally equalize the sub-bands of cepstra. Hence, LP and HP filtered components of HEQ equalized cepstra (of each frame) are obtained and a second level of equalization is done. LP and HP filtering is done on each frame and equalized. LP and HP filtering is done on equalized frames by a simple averaging and differencing operation shown in Eqn. 2 and 3. LP and HP equalized cepstra are then added to obtain a combined set of features. These combined set of features are LP and HP equalized and also overall histograms match closely as shown in Fig. 1(d), 1(f) and 1(h). Overall cepstra match closely because of the first level HEQ being applied. Applying second level of HEQ distorts the overall histogram by a small amount. Obuchi and Stern [4] have performed equalization by taking time derivative across the frames (delta features), *where as, in S-HEQ differencing is performed within each frame*.

Comparing Fig. 1(c), 1(e), 1(g) with Fig. 1(d), 1(f), 1(h), HEQ shows a very good match for overall cepstral histogram (between clean and noisy conditions), while a considerable mismatch is seen for LP and HP components. On the contrary, S-HEQ shows a very good match for LP and HP components of HEQ equalized cepstra, however, a small amount of mismatch is present in overall histogram. The amount mismatch seen in LP and HP histogram for HEQ is considerably large, compared to mismatch in overall histogram for S-HEQ. Thus a new set of features obtained from S-HEQ could potentially have a better match between clean and noisy conditions compared to plain HEQ features. The effectiveness of S-HEQ features is reflected in the recognition results.

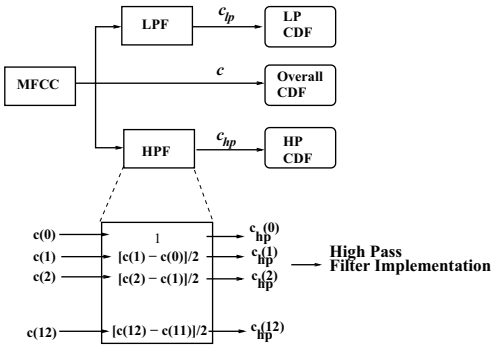


Figure 2: Generation of Reference CDF

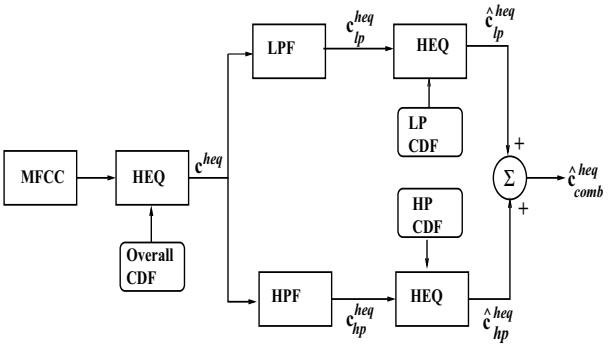


Figure 3: Block diagram describing the steps in performing S-HEQ

4. Experimental Results

4.1. Experimental Set-up

The proposed scheme has been tested on AURORA-2 and AURORA-4 databases. All procedures for training and recognition are identical to the reference experiments, except the noise normalization block (S-HEQ) in the front-end as discussed in the paper. In Aurora-4, recognition system is based on continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state. The language model is the standard bigram for the WSJ0 task. For AURORA2 connected digits task, each digit is modelled as a left to right continuous density HMM with 16 states and 6 Gaussians per state. 13 MFCC features is used as the basic parametrization of the speech signal using C0 instead of the logarithmic energy. First and second order regressions are augmented to 13 MFCC vectors, yielding a final 39 component feature vector. CMS is performed by sentence-by-sentence subtraction of the mean values of each cepstral coefficient. The HEQ reference distribution for overall cepstra have been obtained by averaging over the whole clean training set of utterances. The reference distribution for LP and HP equalization are obtained by averaging LP and HP filtered clean training unequalized utterances. Both training and test utterances have been then equalized as discussed in the paper. Cepstral coefficients are equalized before the computation of the regressions. HMMTool Kit (HTK) software is used for Training and Recognition.

4.2. Discussion

Fig. 4 shows the results for different combinations of equalization. To understand the effect of noise on different sub-bands

	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avrg.	Rel. Impro.
Baseline	87.61	75.42	53.3	53.17	46.95	56.57	45.4	76.89	64.21	45.28	41.98	36.26	47.51	36.45	54.79	0%
HEQ	88.21	76.59	62.77	61.29	61.5	64.41	60.55	79.06	68.39	53.48	50.83	49.28	54.42	49.52	62.88	14.8%
S-HEQ	88.51	81.56	68.22	64.08	65.08	69.64	64.69	82.42	76.69	62.1	57.89	54.94	63.27	58.88	68.43	24.9%

Table 1: Aurora-4 recognition accuracy results for different additive noise types (T-02 to T-07) and convolutive and additive noise type (T-08 to T-14). T-01 shows the recognition accuracy for clean speech

the following experiments were done on Aurora-2 database with above experimental setup. Filtering technique used was as explained in section 3, Eqn. 2 and 3. Expt. 2 uses only low pass filtered components of conventional MFCC features. Small improvement was observed over baseline, indicating the domination of noise in high frequency region. In Expt. 3, equalized LPF cepstra is added to unequalized HPF cepstra to get the combined set of features. In contrast, in Expt. 4 unequalized LPF cepstra is added with equalized HPF cepstra. Expt. 4 shows significantly less WER (19%) compared to Expt. 3 (29%), indicating the low SNR's in high frequency sub-band. Also, surprisingly, Expt. 4 WER was slightly less than Expt. 7 (HEQ, WER=19.5%), showing the effectiveness of sub-band equalization. Equalizing both LPF and HPF cepstra (Expt. 5) showed slightly poor results compared to Expt. 4, which was also reflected in Expt. 6, where another level of equalization was applied on features used in Expt. 5. Conventional HEQ on plain features is performed in Expt. 7 and shows WER of 19.5%. Expt. 8 shows WER for S-HEQ method, which shows least WER of 17.1%. Table 1 shows the results for Aurora-4 database. S-HEQ performs very well in all the test cases, and gives a relative improvement of 15% in WER over HEQ. Table 2 shows the recognition results for Aurora-2 database for different SNR conditions. Baseline results are with CMS. S-HEQ outperforms HEQ at all SNR levels and also for all noise conditions (not shown here) with relative improvement of 12% over HEQ in WER.

5. Conclusion

In this paper, we present a novel efficient approach for noise normalization referred to as S-HEQ. An independent equalization is done on LPF and HPF cepstra along with an overall equalization. An optimal approach to equalize sub-bands is suggested as shown in Fig. 3. S-HEQ has shown significant improvement in results, with a very small computational overhead. Like HEQ, S-HEQ is also not model based, nor does it make any assumption about the underlying density functions. S-HEQ is around 3 times slower than HEQ (3 equalizations are done), yet is very fast compared to VTS [3, 4]. As HEQ in itself is very fast, computational overhead of S-HEQ is negligible and can be used in real time applications, to achieve greater accuracy.

6. Acknowledgements

This work was supported under the Indo-Spanish Joint Program of Co-operation in Science and Technology. The Indian group is supported under project DST/INT/SPAIN/P-5 of Ministry of Science and Technology. The Spanish Group is supported under project ACI2009-0892 by the Ministry of Science and Innovation.

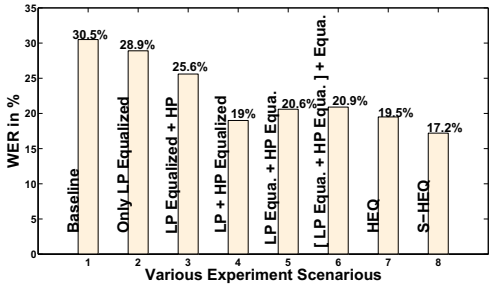


Figure 4: Word error rate in % for the various experiments performed on Aurora-2 database

	Baseline	HEQ	S-HEQ
Clean	99.23	99.07	99.17
20dB	97.35	97.57	97.76
15dB	93.43	95.38	95.75
10dB	80.62	89.73	90.83
5dB	51.87	75.26	78.07
0dB	24.30	44.63	51.79
-5dB	12.03	16.33	21.66
Average	69.51	80.51	82.84
Rel. impro.	0%	15.82%	19.2%

Table 2: Aurora-2 recognition results (Avg. over Set A, B, C)

7. References

[1] A. de la Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 355 – 366, May 2005.

[2] J. Segura, C. Benitez, A. de la Torre, A. Rubio, and J. Ramirez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *Signal Processing Letters, IEEE*, vol. 11, no. 5, pp. 517 – 520, May 2004.

[3] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. ICASSP-96*, 1996, pp. 733–736.

[4] Y. Obuchi and R. Stern, "Normalization of time-derivative parameters using histogram equalization," in *Proc. of EUROSPEECH 2003, Geneva, Switzerland. 2003*, 2003.

[5] F. Hilger, S. Molau, and H. Ney, "Quantile based histogram equalization for online applications," in *Interspeech*, 2002.

[6] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *ASRU*, 2001.

[7] J. W. Hung and H. T. Fan, "Subband feature statistics normalization techniques based on a discrete wavelet transform for robust speech recognition," *Signal Processing Letters, IEEE*, vol. 16, no. 9, pp. 806 –809, sept. 2009.