



# Feature and Distribution Normalization Schemes for Statistical Mismatch Reduction in Reverberant Speech Recognition

A.M. Toh<sup>1</sup>, R. Togneri<sup>1</sup>, S. Nordholm<sup>2</sup>

<sup>1</sup>School of Electrical, Electronic and Computer Engineering  
The University of Western Australia, Australia

<sup>2</sup>Western Australian Telecommunications Research Institute

aikming@ee.uwa.edu.au, roberto@ee.uwa.edu.au, sven@watri.org.au

## Abstract

Reverberant noise has been a major concern in speech recognition systems. Many speech recognition systems, even with state-of-art features, fail to respond to reverberant effects and the recognition rate deteriorates. This paper explores the significance of normalization strategies in reducing statistical mismatches for robust speech recognition in reverberant environment. Most normalization works focused only on ambient noise and have yet been experimented on reverberant noise. In addition, we propose a new approach for the odd order cepstral moment normalization which is computationally more efficient and reduces the convergence rate in the algorithm. The proposed method is experimentally justified and corroborated by the performance of other normalization schemes. The results emphasize the significance of reducing statistical mismatches in feature space for reverberant speech recognition.

verberant environments. It is essential to explore the performance of normalization strategies for reverberant speech recognition. Two main normalization approaches, cepstral moment normalization and histogram equalization, are investigated for robustness in reverberant environments. In addition, we propose a computationally efficient odd order moment normalization scheme and evaluate it with other normalization schemes on the full TI-digit database. Robust results for MFCC\_0 post-processed with normalization schemes in reverberant environments illustrate the contribution of the proposed method.

The paper is organized as follows: Section 2 discusses statistical mismatches due to reverberation. Section 3 explores the proposed cepstral moment normalization methods and section 4 introduces the distribution normalization schemes in particular histogram equalization. The experimental setup is presented in section 5, followed by the results in section 6. The final section comprises the conclusions.

## 1. Introduction

The statistical properties of a speech feature differ under the influence of noisy environments such as reverberant condition [1]. Human ears are robust to reverberation time such as RT 0.4s which is both clearly and audibly perceived. However, for speech recognition system, this is the reverberation level where the recognition performance starts to deteriorate significantly. Reverberation time of 0.5s onwards can be regarded as severe reverberation and the recognition accuracy degrades rapidly in these regions.

Normalization strategies are often employed as a post-processing scheme in speech recognition systems to compensate for the effects of environmental mismatch. These schemes are preferred because a priori knowledge and adaptation are not required under any environment. Normalization methods can be classified into feature normalization and distribution normalization. Feature normalization attempts to normalize certain statistical property of speech vectors such as the mean, variance and moments [2] [3] to reduce the residual mismatch in feature vectors. Histogram equalization [4], and feature space normalization [5] belong to the distribution normalization category which aim at normalizing the database or the distribution to match the reference distribution.

Normalization schemes mentioned above have been evaluated only in additive and convolutional noise but not in re-

## 2. Statistical Mismatches

In addition to temporal smearing and spectral flattening effects observed in [1], reverberation also affects statistical properties of speech features such as the first order (mean), second order (variance) and higher order moments. The analysis of statistical properties is important because these parameters play a major role in the mismatch between the clean and reverberant speech features. Figure 1 illustrates the mean-variance plot of the MFCC cepstra affected by reverberation. The blue region (o) refers to the parameters of the clean vectors, the red (+) and black (\*) clusters denote the mean-variance of the cepstral vector in reverberation time RT 0.2s and RT 0.6s respectively.

Figure 1 shows evident reduction in the parameter variance. The reduction in the variance is validated since the statistical dispersion or the deviation of a cepstra about the mean is smaller with the inclusion of noise. The increase in the cepstral mean is expected. Similar observations are obtained through an analysis on the skewness and kurtosis of the cepstral vector. It is evident that the reverberation affects statistical properties significantly and contributes to mismatches in speech parameters. This corroborates the need to reduce statistical mismatches in reverberant environment. Thus, normalization schemes are employed for normalizing statistical properties in mismatch conditions.

## 3. Feature Normalization

### 3.1. Cepstral moment normalization

Cepstral mean normalization (CMN) or cepstral mean subtraction has always been applied to state-of-the-art features such

This research is partly funded by the National ICT Australia (NICTA). National ICT Australia is funded through the Australian's Government Backing Australia's Ability initiative, in part through the Australian Research Council (ARC)

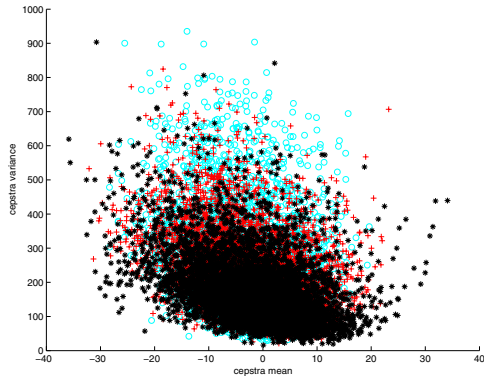


Figure 1: MFCC cepstral mean variance distribution

as the Mel-frequency cepstral coefficients (MFCC\_0). It is a de facto standard for most large vocabulary speech recognition systems. The algorithm computes a long-term mean value of the feature vectors and subtracts the mean value from the cepstral vectors of that utterance. CMN has been shown to be effective in alleviating the effects of linear filtering or convolutional distortion caused by characteristics of different communication channels or recording devices. In addition, it was also shown to improve the recognition performance in both additive and reverberant environments [1].

Cepstral variance normalization (CVN) is a popular technique often used in conjunction with CMN. The mean and variance of cepstral coefficients are assumed to be invariant in the CVN analysis. Therefore the exclusion of these properties would result in the removal of irrelevant information such as the effects of mismatched environments. CVN contributes to robustness by scaling the cepstral deviations to a normalized boundary such as the unity.

Moments and cumulants have been employed in diversity of disciplines that deal with data, random variables and stochastic processes. They can be used to quantify statistical properties of a distribution such as its location and scale. Two popular higher order statistics commonly used are the skewness and kurtosis. The skewness measures the asymmetry while the kurtosis measures the flatness of a distribution. The third order moment or skewness was the first higher order statistic proposed for normalization scheme in the speech recognition context [2]. Hsu and Lee later proposed the use of higher order moments for cepstral moment normalization [3]. The  $N$ -th order cepstral moment normalization is denoted by CMtN  $N$  in this work.

### 3.2. Even cepstral moment normalization

Cepstral mean normalization is performed prior to all cepstral moment normalization (CMtN) to ensure zero mean. The even order cepstral moment normalization can be normalized with the scaling of the first order moment normalized coefficients by a constant [3].

$$\begin{aligned}
 X_{N=\text{even}} &= bX_{\text{CMN}} \\
 E[X_{N=\text{even}}^N] &= E[(bX_{\text{CMN}})^N] \\
 &= b^N E[X_{\text{CMN}}^N] \\
 &= M_N
 \end{aligned} \tag{1}$$

$X_{\text{CMN}}$  is the mean normalized vector and  $M_N$  is the  $N$ -th order even moment of a normalized Gaussian distribution

where unity has been adopted,  $M_N = 1$ , in this work. The solution for  $b$  can be obtained from equation (2).

$$b = \left[ \frac{M_N}{E[X_{\text{CMN}}^N]} \right]^{1/N} = \left[ \frac{1}{E[X_{\text{CMN}}^N]} \right]^{1/N} \tag{2}$$

### 3.3. Odd cepstral moment normalization

We used the non-linear model proposed for the cepstral third order normalization (CTN) [2] and extended it for higher order odd cepstral moment normalization. The variance can be effectively set to unity with CVN.

$$X_N = \frac{a}{b} X_{\text{CVN}}^2 + X_{\text{CVN}} + \frac{c}{b} \tag{3}$$

$$\begin{aligned}
 E[X_N] &= E \left[ \frac{a}{b} X_{\text{CVN}}^2 + X_{\text{CVN}} + \frac{c}{b} \right] \\
 &= 0
 \end{aligned} \tag{4}$$

$X_{\text{CVN}}$  or CMtN 2 is the variance normalized vector which can be computed from equation (1) with  $N = 2$ . The relationship between  $a$  and  $c$  can be derived from equation (4), giving  $c = -aE[X_{\text{CVN}}^2]$  or simply  $c = -a$ . Instead of solving and choosing the root for  $a$  as in CTN, we derived an approximation for the value of  $a$  similar to [3].

For the  $N$ -th order odd cepstral moment normalization,

$$\begin{aligned}
 E[X^N] &= E[(aX^2 - a + X)^N] \\
 &= a^N E[X^2 - 1]^N + \dots \\
 &+ aN E[X^{N+1}] + E[X^N] - aN E[X^{N-1}] \\
 &= 0
 \end{aligned} \tag{5}$$

The expansion and approximation of equation (5) has been documented in [6]. When  $a$  was small, higher order terms in equation (5) were removed and the last three terms were retained. Thus, the value of  $a$  for higher order odd  $N$ -th moments can be approximated with equation (6). The expression for  $a$  has been verified with  $N$  of 3, 5 and 7 in equation (5).

$$a = \frac{-E[X^N]}{N \cdot E[X^{N+1} - X^{N-1}]} \tag{6}$$

Instead of the high power terms,  $X^{2(N-1)}$  for expression  $a$  in [3], we derived a simpler approximation for  $a$  shown in equation (6). This implementation reduced the complexity and computation time significantly. Furthermore, it achieved similar aim in reducing or converging the corresponding odd moment to zero. A recursive loop was embedded in the algorithm to fine tune the approximation for  $a$  and to converge the odd moment to zero. The algorithm has been verified for zero odd moments convergence,  $M_{\text{odd}} = 0$ .

## 4. Distribution Normalization

Distribution normalization schemes target at mapping the database or the distribution of the testing data to match the reference. We concentrate on histogram equalization (HEQ) for its effectiveness in reverberant speech recognition. HEQ, a popular method used for digital image processing, maps the testing data such that its distribution matches the reference distribution.

The purpose of histogram equalization is to derive a transformation  $x(y)$  which transforms the probability distribution of

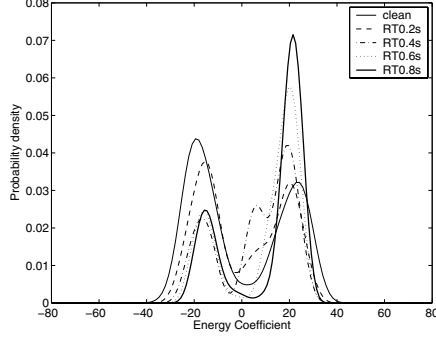


Figure 2: Histogram for the first cepstral coefficient in clean and reverberant environments.

the noisy speech  $p_y$  into the reference distribution  $p_{ref}$  [4]. The probability distribution can be represented by a finite number of observations or the histogram in the implementation. The transformation is established from the noisy and the reference cumulative distributions, given by equation (7).

$$x(y) = C_{ref}^{-1}[C_V(y)] \quad (7)$$

$x(y)$  is the transformation that converts the testing data distribution,  $p_y$  to the reference distribution,  $p_{ref}$ .  $C_{ref}^{-1}$  is the inverse function of the cumulative probability function,  $C_{ref}$  and  $C_V$  is cumulative histogram of the testing data.

Histogram equalization can be effectively implemented with interpolation or table look-up approach. It is applied to each feature vector on a sentence-by-sentence basis. The normalized cumulative histogram of the training and testing data or the probability distribution function of a Gaussian distribution are computed.

The histogram is estimated by considering a 100 uniform interval between  $min(y)_i - \sigma_i$  and  $max(y)_i + \sigma_i$  due to large variances of the cepstral features.  $\sigma_i$  is the standard deviation for the  $i$ -th component of the feature vector. The normalized cumulative histogram can then be computed from the histogram shown by equation (8).

$$C_V(y) = \int_{-\infty}^y p_V(y') dy' \quad (8)$$

The reference cumulative histogram and the target cumulative histogram are used to compute the required transformation,  $x(y)$ . The transformation are then used to interpolate the original cepstral features to the normalized features.

Figure 2 shows the probability distribution of the first cepstral coefficient for the utterance “1-1-1” in reverberant environments. The figure clearly depicts a shift in the distribution towards higher energy coefficients as the reverberant level increases. This could be explained by the temporal smearing effect introduced by reverberant noise [1]. Thus, the primary aim of histogram equalization is to normalize the reverberant distribution to match the clean or the reference distribution.

Two forms of histogram equalization are investigated in this work, HEQ I and HEQ II. HEQ I normalizes both the training and testing data to the Gaussian probability distribution with zero mean and unity variance. In HEQ II, the probability distribution for the clean training data was computed from the full adult portion of the TI-digit database and used as the reference. The testing data were then normalized based on clean training data distribution.

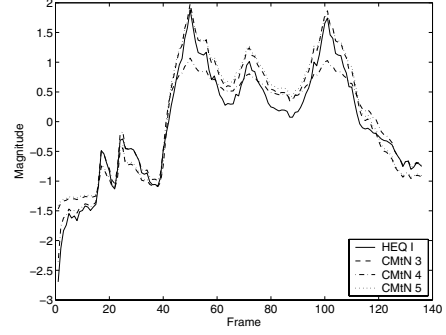


Figure 3: Normalized zeroth cepstral coefficient in reverberant level RT 0.4s

## 5. Experimental Setup

The adult portion of the full TI-Digit database was used in this work. The training data contained utterances of 55 male and 57 female speakers. There were also 23 male and 27 female speakers for the testing data. Each speaker subset composed of 77 digit utterances.

Simulated reverberant data has been commonly used in many reverberant speech recognition researches [7], [8]. In this work, reverberant effects were captured by estimating the impulse response of the room environment from long segments of speech. The experiment used the room impulse response designed to match the characteristic of a 2.2m high, 3.1m wide and 3.5m long room. The microphone and the speakers were localized 0.5m from the wall at opposite ends. The speech was then convolved with the RT60 room impulse response. RT60 is the time interval in which the reverberation level decays by 60dB.

All the speech files were pre-emphasized and windowed with a Hamming window. The speech signal was analyzed every 10ms with a frame width of 25ms. A Mel-scale triangular filterbank with 26 filterbank channels was used to generate the MFCC\_0 coefficients (MFCC) features. The MFCC 0 coefficients constituted 12 static MFCC vectors and the zeroth cepstral coefficients. The Hidden Markov Model used 15 states and 5 mixtures for the connected digit recognition.

## 6. Experimental Results

Figure 3 shows the normalized cepstral sequence using normalization schemes HEQ I and higher order moment normalization CMtN 3, CMtN 4 and CMtN 5 respectively. The normalized sequences for HEQ I match the higher order moment normalized sequences CMtN 3 and CMtN 5 considerably. This demonstrates that HEQ I also attempts to normalize statistical properties such as the variance and moments of cepstral features. This is justified by the mapping of the testing data to the zero mean and unity variance Gaussian distribution. Although the HEQ I and HEQ II would not have exactly the same statistical properties as their references after the mapping, they would have properties close to their respective reference distributions.

The static MFCC\_0 features was used as the baseline to evaluate the performance of cepstral moment normalization (CMtN) counterparts and histogram equalization schemes (HEQ) for speech recognition in reverberant environments. Table 1 records the recognition results for several order cepstral moment normalization and histogram equalization schemes in

Table 1: *Speech recognition with cepstral moment normalization (CMtN) in reverberant environments*

RT (s)	clean	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
MFCC_0	97.67	94.99	91.97	87.80	78.27	67.02	54.11	43.64	37.48
CMtN 1 (CMN)	98.56	96.93	95.42	92.48	84.65	74.23	59.18	47.35	38.34
CMtN 2 (CVN)	98.47	97.15	96.02	93.95	87.07	77.30	63.10	49.73	39.57
CMtN 3	98.36	97.07	95.89	94.15	86.84	78.26	66.27	56.47	45.28
CMtN 4	98.35	97.12	96.00	94.17	86.80	77.66	66.20	55.44	45.48
CMtN 5	98.35	97.27	96.26	94.18	87.69	78.89	67.15	56.32	46.29
CMtN 6	98.39	97.05	95.97	93.92	86.28	76.22	60.95	48.98	38.70
HEQ I	98.17	96.77	95.79	93.46	87.08	78.48	68.55	56.24	47.44
HEQ II	97.53	96.50	95.35	92.94	87.03	76.94	65.61	54.88	45.42

reverberant environments. These results have been assessed with a signed-rank significance test and gave a confidence level of at least 98%. Moment normalization schemes such as cepstral variance normalization denoted by CMtN 2 and higher order moment normalization, CMtN 3, CMtN 4 and CMtN 5 demonstrated favourable and robust recognition performance both in low and severe reverberant levels.

The proposed odd order cepstral moment normalization schemes, CMtN 3 and CMtN 5, yield best performances compared to other moment normalization schemes. This was mainly due to the hybrid contribution of variance normalization and odd order moment normalization embedded in the algorithm. The algorithm ensured that at least three moments, the mean, variance and higher order odd moment of interest, were normalized. The effectiveness of odd and even order CMtN for speech recognition was apparent from reverberation time of 0.3s or more where an improvement of 6% or better resulted.

Table 1 also shows the results for histogram equalization scheme, HEQ I and HEQ II. Both HEQ schemes showed comparable results to the CMtN schemes. However, histogram equalization using the Gaussian probability function as the reference probability distribution, HEQ I, yielded better recognition performance compared to HEQ II with probability distribution generated from the clean training data. This is because HEQ I normalizes both the training and testing data to the Gaussian reference where statistical properties such as mean and variance are normalized. HEQ II maps the testing data to the clean training data which is non-Gaussian distributed and has high variability in statistical parameters. The recognition results further corroborate the observations in Figure 3 where HEQ I attempts to perform moment normalization on cepstral features. The proposed CMtN and HEQ I results reveal the significance of reducing statistical mismatches in feature vectors.

## 7. Conclusion

We have explored the contribution of feature normalization and distribution normalization for speech recognition in reverberant environments. The proposed odd order cepstral moment normalization scheme have demonstrated robust performances in reverberant speech recognition. The computational efficiency and simplicity of the proposed scheme highlights its contribution in moment normalization. The scheme ensured at least three moments, the mean, variance and a higher order odd moment of interest are normalized. The use of the Gaussian probability distribution as the reference for HEQ scheme is more effective than its counterpart, HEQ II, which used the distribution of the training data as reference. The CMtN and HEQ I results

corroborate the significance of reducing statistical mismatches in feature vectors.

The dynamic and acceleration features are not included in this work because this paper focuses on the performance of normalization strategies compared to the performance of the baseline static feature. In addition, it was shown in [6] that regression features derived from the static cepstra are more robust than those derived from normalized cepstra. The scope on regression features and normalization schemes remains for further investigation and justification.

Optimal normalization schemes may not be evident in reverberant environments but the results have depicted robust improvements for speech recognition in reverberant environments. The normalization schemes explored in this work could be performed as post-processing methods for a diversity of speech features. In addition, they could also be combined with other noise compensation methods for improved performance.

## 8. References

- [1] A.M Toh, R. Togneri, and S. Nordholm, "Investigation of robust features for speech recognition in hostile environments," in *Proc. APCC*, 2005, pp. 956–960.
- [2] Y.H. Suk, S.H. Choi, and H.S. Lee, "Cepstrum third-order normalization method for noisy speech recognition," in *IEEE Electronic Letters*, 1999, vol. 35, no.7, pp. 527–528.
- [3] C.W. Hsu and L.S. Lee, "Higher order cepstral moment normalization (hocmn) for robust speech recognition," in *Proc. ICASSP*, 2004, pp. 197–200.
- [4] A.d.I. Torre, A. M. Peinado, J.C. Segure, J.L.Perez, C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions SAP*, vol. 13, pp. 355–366, 2005.
- [5] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse conditions," in *Proc. ICASSP*, 2003, pp. 656–659.
- [6] R. Togneri, A.M Toh, and S. Nordholm, "Evaluation and modification of cepstral moment normalization for speech recognition in additive babble ensemble," in *Proc. 11th Australasian International Conference on Speech Science and Technology (SST)*, 2005.
- [7] L. Couvreur, C. Couvreur, and C. Ris, "A corpus-based approach for robust ASR in reverberant environments," in *Proc. ICSLP*, 2000, pp. 397–400.
- [8] K.J. Palomki, G.J. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *Proc. ICASSP*, 2002.