Evaluation and Modification of Cepstral Moment Normalization for Speech Recognition in Additibe Babble Ensemble

Roberto Togneri[1], Aik Ming Toh[1], Sven Nordholm[2]

[1]School of Electrical, Electronic and Computer Engineering
The University of Western Australian, Australia
[2]Western Australian Telecommunication Research Institute
roberto@ee.uwa.edu.au, aikming@ee.uwa.edu.au, sven@watri.org.au

Abstract

The statistical properties of a speech feature could differ under the influence of noisy environments. These effects are common in mismatched environments such as additive background noise and reverberant environments. Normalization strategies are employed in speech recognition systems to compensate for the effects of environmental mismatch. This work explores the utilization of cepstral moment normalization for speech recognition in additive noisy environment. The author has evaluated and modified the models used for cepstral moment normalization for improved convergence and performance. The cepstral moment normalization schemes were adopted for additive noisy speech recognition on TI-digit database. Further experimental works on cepstral moment normalization involving dynamic features are also presented. Both odd and even order cepstral moment normalization have shown significant contributions to speech recognition in low signal-to-noise ratio non-stationary noisy environment.

1. Introduction

The statistical properties of a speech feature could differ under the influence of noisy environments. The level of variation depends on the type of noise and also the level of contamination. These effects are common in mismatched environments such as additive background noise and reverberant environments.

Normalization strategies are employed in speech recognition systems to compensate the effects of environmental mismatch. Most of the normalization schemes are applied in the cepstral domain and these techniques are often performed as post-processing scheme on speech features. Normalization techniques 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 such as mean, variance (Jain and Hermansky 2001) and moments (Suk, Choi, and Lee 1999) (Hsu and Lee 2004) to reduce the residual mismatch in feature vectors. Histogram equalization (Torre, Peinado, Segure, J.L.Perez, Bentez, and Rubio 2005), quantile normalization (Hilger and Ney 2001) and feature space normalization (Molau, Hilger, and Ney 2003) fall into the distribution normalization category. These techniques aim at normalizing the database or the distribution to match the reference.

This work explores the utilization of cepstral moment normalization for speech recognition in additive babble noise. The author has evaluated and modified the odd order cepstral moment normalization model used for improved convergence and performance. Evaluation of cepstral moment normalization and dynamic feature will also be addressed.

The paper is organized as follows: The next section explores the statistical properties of a cepstrum and cepstral distributions in additive noisy environments. Section 3 introduces popular normalization schemes adopted for speech recognition. Section 4 specifies the cepstral moment normalization algorithm and modification. The experimental setup is presented in section 5, followed by the results in section 6. The last section comprises the conclusions.

2. Statistical properties

2.1. Moments and statistics

Statistical properties such as mean and variance provide much information on the variations induced by noise on speech cepstra. Theoretically, only the first four statistics have physical definition in the mathematical study. These are the mean, variance, skewness and kurtosis. These properties can be derived from the expectation value of the speech cepstra.

The time averages over a finite time interval can be used to estimate the mean and variance of the speech. The mean, μ is the weighted average of cepstral values across the utterance over a finite interval.

$$\mu = E[X] = \frac{1}{N}\sum_{k=1}^{N} X[k] \tag{1}$$

Variance is the second central moment of a distribution. It is a measure of statistical dispersions about the

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).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

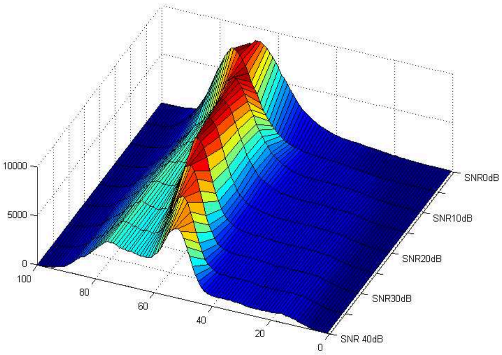

Figure 1: Cepstral distribution of the first cepstra across different SNR levels of additive babble ensemble

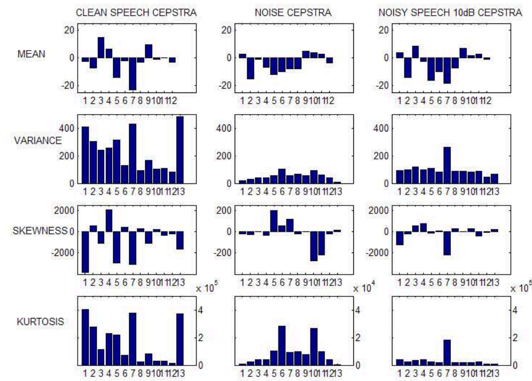

Figure 2: Statistical properties of clean MFCC_0 cepstra, noise cepstra and corrupted speech cepstra

mean of the distribution and is defined by Equation 2 where $E[X]^2 = \mu^2$.

$$\begin{aligned} \sigma^2 &= E[(X - E[X])^2] \\ &= E[X^2] - E[X]^2 \end{aligned} \quad (2)$$

Higher order statistics or moments are also derived about the mean of the distribution. The probability density function of cepstral coefficients of speech signals is usually regarded as a quasi-Gaussian distribution. Under this assumption, the odd order moments should be zero and the even order moments should be some specific constants (Hsu and Lee 2004). The central moments are defined by

$$M_N = E[(X - \mu)^N] = \begin{cases} 0 & \textit{for} \quad N = \textit{odd} \\ k & \textit{for} \quad N = \textit{even} \end{cases} \quad (3)$$

where k is the N-th moment of a normalized Gaussian distribution which is usually a constant such as the unity.

The asymmetrical function of a cepstral distribution is known as the skewness. Skewness characterizes the degree of asymmetry of a distribution around its mean. The indication of skewness is affected by the locus of extreme values or outliers. A distribution is considered to possess negative skewness if the left end of distribution is more pronounced than the right tail. Positive skewness refers to a more prominent right tail distribution. The skewness of a distribution, M_3 is derived from Equation 3 with N=3.

Kurtosis, M_4, characterizes the relative flatness of a distribution in relation to the shape of a normal distribution. Positive kurtosis indicates a relatively peaked distribution while a relatively flat distribution defines a negative kurtosis.

It is often difficult to offer a sensible interpretation to the large values in moment measures. This is because moments can be arbitrarily large due to the high power terms in their expressions and possible existence of one or more large outliers in the data. Sample estimates of moments are likely to be unstable in the presence of outliers. This problem is severe for higher order statistics, such as the kurtosis, which is employed in independent component analysis (ICA) algorithm for speech related researches (Sahguchi, Ozawa, and Kotani 2002).

2.2. Moments and noise

The influence of additive noise ensemble on statistical variations cannot be explained by mere transformation or addition of clean speech and noise cepstra. However, observations from cepstral distributions and statistical variability of cepstral features provide much information on the effects of additive noise.

Figure 1 shows the cepstral distribution of cepstra 1 across several additive noise levels ranging from 40dB to 0dB with an interval of 5dB. 1232 utterances from 16 speakers, both male and female, were analyzed for the cepstral distributions with a bin of 100. Observation from Figure 1 clearly shows a shift in the global mean of the cepstral features. The mean of cepstral features generally increases as the signal-to-noise ratio (SNR) decreases. This can also be supported by Equation 4.

$$\begin{aligned} Y &= X + N \\ log(Y) &= log(X + N) \\ &= log(X) + log\left(1 + \frac{N}{X}\right) \end{aligned} \quad (4)$$

where noisy speech spectra, $Y = |Y[k]^2|$, clean speech spectra, $X = |X[k]^2|$ and noise spectra, $N = |N[k]^2|$.

Equation 4 shows that a reduction in SNR, $N > X$, would theoretically increases the value of the logarithmic Y and results in incremental mean. However, it should be noted that the influence of additive noise on speech cepstra is more than just a simple additive operation. Non-stationary background noise such as babble noise would affect the cepstra differently from additive white Gaussian noise.

Cepstral distributions of the 13 MFCC_0 vectors also illustrates a reduction in the variance with the slope of the distribution gradually increases. A shift in the skewness is also observed for cepstra 1 distribution.

Observations on the variability of individual cepstra also yield valuable information. Figure 2 illustrates four statistical properties of the clean speech cepstra, noise cepstra and corrupted 10dB speech cepstra. The first row depicts the mean of the 12 cepstra excluding the zeroth cepstral coefficient. The subsequent rows represent the vari-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

ance, skewness and kurtosis of all 13 cepstra respectively. Zeroth cepstral coefficient is illustrated by the 13th cepstra. From observation, one major mismatch that could affect the performance of speech recognition in noisy environment is the large variance and moments of clean speech cepstra. Noisy cepstra tends to have lower variance and moments and this could contribute to a large mismatch in the recognition process. Thus, environmental mismatch may be reduced or compensated if these statistical properties are normalized.

Figure 2 also shows evident and significant reduction in the variance of cepstral features under the influence of additive babble ensemble. 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 kurtosis would have similar variation as the variance since it is the power component of the variance. A large variability in the variance would imply a larger variation in the kurtosis. In general, higher order moments, such as the skewness and kurtosis, are reduced in presence of additive noise.

3. Cepstral normalization

3.1. Cepstral mean normalization

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

The main advantage of using sentence-based normalization methods such as CMN or augmented CMN (Acero and Huang 1995) instead of adaptation method is the additional improvement in recognition accuracy for clean conditions and significant decrease in error rate for mismatched conditions. 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.

However, CMN does not discriminate silence and voice in computing the utterance mean. In addition, CMN cannot be used directly in a real-time system because it requires the complete utterance for the cepstral mean computation.

Nevertheless, CMN has been shown to improve the recognition performance in both additive and reverberant environments (Toh, Togneri, and Nordholm 2005). One explanation is the removal of the mean will eliminate the bias component imposed by various noises.

3.2. Cepstral variance normalization

Cepstral variance normalization (CVN) is a popular supplement technique to CMN. It is also known as the mean and variance normalization (MVN) (Jain and Hermansky 2001) because CVN is 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 and limiting the range of deviation in cepstral features. It scales the deviations to a normalized boundary commonly the unity. The normalized parameters are assumed to be resistant to the degradation imposed by noises. Variance normalization has been a popular technique in speaker recognition in dealing with noise and channel mismatch (Barras and Gauvain 2003) (Melin and Lindberg 1999).

3.3. Cepstral moment normalization

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 kurtosis measures the flatness of a distribution. Skewness has been used for risk analysis in finance (Harvey and Siddique 1999) while kurtosis has been employed in ICA for speech related researches (Sahguchi, Ozawa, and Kotani 2002).

The third order moment or skewness was the first higher order statistic proposed for normalization scheme in speech recognition context (Suk, Choi, and Lee 1999). Exhaustive observations of the statistic of noisy cepstra have revealed that additive noise also affected moments of a cepstrum. The paper reported that moments must be normalized to fully compensate for the statistical variations under noisy conditions. Cepstral third order moment normalization (CTN) has achieved robust performance in additive white Gaussian noise and car noise environments.

C.W. Hsu and L.S. Lee proposed the use of higher order moments for cepstral moment normalization (Hsu and Lee 2004). They reported that optimal cepstral moment normalization was achieved with the use of the fifth and hundredth order moments. The use of the hundredth moment was not necessary since it has been mentioned that higher order statistics greater than four lack of theoretical or physical meaning. Furthermore, the X^N term in the paper (Hsu and Lee 2004) resulted in extremely slow convergence as the order N increased.

One interesting observation could be contemplated from cepstral moment normalization (CMtN) schemes. A subsequent CMtN scheme would remove the normalization effects of previously normalized moments. This is applicable to both the odd and even order CMtN schemes. It is not possible to have two even or two odd moments normalized at a particular instant. In addition, both odd and even order CMtN affect their lower order counterparts. Nevertheless, it is still possible to normalize an odd and an even order moment simultaneously with a hybrid normalization scheme.

4. Cepstral moment normalization scheme

Cepstral mean normalization is performed prior to all cepstral moment normalization to ensure $\mu = 0$.

4.1. Even cepstral moment normalization

The even order cepstral moment normalization can be normalized with the scaling of the first order moment nor-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

malized coefficients by a constant (Hsu and Lee 2004).

$$\begin{aligned} X_{N=even} &= bX_{cmn} \\ E[X^N_{N=even}] &= E[(bX_{cmn})^N] \\ &= b^N E[X^N_{cmn}] \\ &= M_N \end{aligned} \tag{5}$$

X_{cmn} is the mean normalized vector, μ is zero due to mean normalized sequence 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 6.

$$b = \left[\frac{M_N}{E[X^N_{cmn}]}\right]^{1/N} = \left[\frac{1}{E[X^N_{cmn}]}\right]^{1/N} \tag{6}$$

4.2. Odd cepstral moment normalization

CTN was first proposed by Y.H. Suk, S.H. Choi and H.S. Lee (Suk, Choi, and Lee 1999) using a non-linear transform defined by Equation 7.

$$\begin{aligned} X_{ctn} &= aX^2_{cvn} + bX_{cvn} + c \\ &= \frac{a}{b}X^2_{cvn} + X_{cvn} + \frac{c}{b} \end{aligned} \tag{7}$$

$$\begin{aligned} E[X_{ctn}] &= E\left[\frac{a}{b}X^2_{cvn} + X_{cvn} + \frac{c}{b}\right] \\ &= \left[\frac{a}{b}\cdot 1 + 0 + \frac{c}{b}\right] \\ &= 0 \end{aligned} \tag{8}$$

X_{cvn} is the variance normalized vector which can be defined by Equation 5 with $N = 2$. The constants a, b and c were determined so that the resultant coefficient should have zero mean, unity variance and zero third-order moment. The constant b was set to unity with variance normalization. The relationship between a and c was derived from Equation 8, giving $c = -a$. The value of a is selected from the root nearest to the average value of the one real-valued root (Suk, Choi, and Lee 1999).

We used the non-linear model described in Equation 7 and evaluated it for higher order odd cepstral moment normalization. The variance can be set to unity with CVN and thus, we removed constant b from the derivation.

$$X_N = aX^2_{cvn} + X_{cvn} + c \tag{9}$$

$$c = -aE[X^2_{cvn}] = -a$$

For the third order cepstral moment normalization,

$$\begin{aligned} E[X^3] &= E[(aX^2 - a + X)^3] \\ &= E[a^3X^6 + 2a^2X^5 - 2a^3X4 + aX^4 + \cdots \\ &= -2a^2X^3 + a^3X^2 + a^2X^5 + 2aX4 + \cdots \\ &= -2a^2X^3 + X^3 - 2aX^2 + a^2X + \cdots \\ &= -a^3X^4 - 2a^2X^3 + 2a^3X^2 + \cdots \\ &= -aX^2 + 2a^2X - a^3] \\ &= 0 \end{aligned} \tag{10}$$

Instead of solving and choosing the root for a as in CTN, we derived an approximation for the value of a similar to paper (Hsu and Lee 2004). When a was small, higher order terms in Equation 10 were removed and the last three terms as shown in Equation 11 were retained. The solution for a in the third order CMtN is defined by Equation 12.

$$a3E[X^4] + E[X^3] - a3E[X^2] = 0 \tag{11}$$

$$a = \frac{-E[X^3]}{3E[X^4 - X^2]} \quad \textit{for} \quad N = 3 \tag{12}$$

We extend the extensive but straightforward calculation to N=5. Similar expression was obtained for odd moment of order 5. The last three terms of interest for higher order fifth moment is described by Equation 13 and the solution for a is given by Equation 14.

$$a3E[X^6] + E[X^5] - a3E[X^4] = 0 \tag{13}$$

$$a = \frac{-E[X^5]}{3E[X^6 - X^4]} \quad \textit{for} \quad N = 5 \tag{14}$$

Thus, the value of a for higher order odd N-th moments can be approximated with Equation 15.

$$a = \frac{-E[X^N]}{3E[X^{N+1} - X^{N-1}]} \tag{15}$$

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, $M_{odd} = 0$.

This work expanded the second order non-linear transform used for CTN to higher order cepstral moment normalization. The non-linear transform in Equation 9 was used instead of the $aX^{N-1}_{[1,N]} + X_{[1,N]} + c$ term in paper (Hsu and Lee 2004). This reduced the convergence rate significantly since the solution converged faster from the second order X^2 rather than the high power term, X^{N-1} in (Hsu and Lee 2004). This implementation also reduced the complexity and computation time significantly. Furthermore, it achieved the similar aim in reducing or converging the corresponding odd moment to zero.

The odd order CMtN in this work is a hybrid technique that embedded the CVN or the second order CMtN 2 scheme in the odd order CMtN algorithm. This is necessary given the assumption of the second order non-linear transformation used for odd order CMtN. The incorporation of CVN in the algorithm allows fast convergence and robust performance.

5. Experimental setup

The adult portion of the full TI-digit database was used to evaluate the recognition performance of different normalization schemes. The database comprised both isolated and connected digit utterances. The training data contained 8624 utterances from 112 speakers and there were 3850 utterances from 50 speakers for the testing data.

An ensemble of non-stationary babble noise was extracted from the NOISEX-92 database to corrupt utterances

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9

Accepted after full paper review

Table 1: Speech recognition with cepstral moment normalization (CMtN) in additive babble environments

SNR (dB)	clean	40	35	30	25	20	15	10	5	0
CMN	98.56	98.35	98.20	97.82	96.49	93.17	82.03	57.40	27.78	12.92
CMtN 2 (CVN)	98.47	98.34	98.09	98.00	97.64	96.88	94.91	89.72	75.44	46.46
CMtN 4	98.39	98.33	98.11	97.99	97.51	96.81	94.76	90.49	77.49	48.49
CMtN 6	98.39	98.32	98.02	97.93	97.47	96.80	94.69	90.55	76.97	50.29
CMtN 8	98.41	98.28	98.05	97.90	97.39	96.69	94.67	90.57	77.08	51.09
CMtN 10	98.38	98.24	98.00	97.84	97.32	96.69	95.01	90.53	77.02	51.70
CMtN 25	98.31	98.09	97.92	97.65	97.28	96.48	94.79	89.80	76.63	49.13
CMtN 100	98.34	98.05	97.94	97.65	97.28	96.46	94.73	89.79	76.51	48.98
CMtN 3	98.36	98.21	98.08	98.00	97.27	97.11	94.95	90.70	77.74	49.81
CMtN 5	98.35	98.30	98.19	98.05	97.71	96.91	94.69	90.11	76.45	48.98

for speech recognition in additive babble noise. The testing data was corrupted with additive babble noise at different signal-to-noise ratio (SNR) from -5dB to 40dB at an interval of 5dB.

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. The segmented signal was transformed into power spectrum using the short-time Fourier transform. A Mel-scale triangular filterbank with 26 filterbank channels was used to generate the Mel-frequency cepstral coefficients (MFCC) features. The MFCC_0 coefficients constituted 12 static MFCC vectors and the zeroth cepstral coefficients. The HMM model used 15 states and 5 mixtures for the connected digit recognition.

All cepstral features were mean normalized and all normalization schemes were performed on the full utterance. The odd order cepstral moment normalization algorithm required the odd moment in interest to converge to at least less than 0.0001. Apart from that, all even moments were normalized to unity in this work.

6. Experimental results

6.1. Baseline performances

The mean normalized MFCC_0_Z feature or the CMN scheme was used as the baseline to evaluate the performance of cepstral moment normalization (CMtN) schemes for speech recognition in non-stationary additive babble noise.

Table 1 records the recognition accuracy results for even order CMtN in additive babble ensemble. CMN gave the best recognition performance in high SNR regions such as 40dB and 35dB. CMtN 2 or CVN yielded best results for middle SNR range such as 30dB to 20dB. Even order (CMtN) has demonstrated favourable and robust recognition performance for low SNR levels such as 15dB and below. Table 1 depicts that even order CMtN schemes achieved at least 12% improvement in the recognition accuracy compared to the CMN only scheme in SNR of 15dB. Recognition accuracy of more than 90% was maintained at SNR of 10dB where CMN only gave 57.40%. Higher even order CMtN showed comparable performance to CMtN 2 or CVN. In addition, it was observed that there was no need for extremely high order even moments such as the hundredth moment proposed in (Hsu and Lee 2004).

The last 2 rows in Table 1 record the results for odd order CMtN schemes using Equation 9 and a computed from Equation 15. The odd order CMtN scheme was performed up to the fifth moment only, since higher order moments such as 7 and 9 were extremely slow in convergence. Preliminary experiments with a smaller database and seventh order CMtN showed degradation in the recognition performance.

Similar improvements were achieved for CMtN 3 and CMtN 5 in SNR less than 20dB environments. It could be speculated that these improvements were contributed by the hybrid normalization of both variance and odd order moment. Optimal cepstral moment normalization may not be evident in the additive babble ensemble experiments but Table 1 has demonstrated the effectiveness of both odd and even order cepstral moment normalization for speech recognition, especially in the low SNR region such as 15dB and below. In addition, Table 1 shows that CMtN 2 and CMtN 3 are sufficient for robust speech recognition in noisy environments.

6.2. Moment normalization and dynamic features

Delta (D) and double delta (A) features are time regression features or dynamic features computed from the time trajectories of MFCC_0 features. CMN does not affect the computation of D and A because CMN only removes the mean bias or shift introduced by background noises. Thus, the time trajectories of cepstral features remain the same. However, cepstral variance normalization and moment normalization would alter the trajectories of the cepstra. The normalization of variance and moments would apply a transformation, not just a shift onto cepstral features. Thus, it is essential to explore the effects of CMtN on time regression features for speech recognition.

Table 2 records the recognition performance of dynamic features and moment normalized cepstral features. D refers to the time regression of the moment normalized cepstral features while D_{MFCC_0} is the dynamic features from the MFCC_0 features.

The use of CMtN schemes did not improve the recognition accuracy for high SNR so Table 2 only presented the recognition performance in low SNR region. Recog-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

Table 2: Speech recognition with cepstral moment normalization (CMtN) and dynamic features (D)

SNR (dB)	clean	25	20	15	10	5	0
CMN + D	99.68	98.99	98.37	95.81	87.18	62.43	27.35
CMtN 2 + D	99.51	99.03	98.62	97.37	92.52	77.63	41.45
CMtN 2 + DMFCC0	99.56	99.15	98.86	98.09	95.21	86.12	58.76
CMtN 3 + D	99.49	99.00	98.71	97.74	94.06	82.95	52.28
CMtN 3 + DMFCC0	99.55	99.16	98.91	98.05	95.36	86.39	60.57
CMtN 4 + D	99.60	99.00	98.57	97.42	92.88	79.93	45.06
CMtN 4 + DMFCC0	99.60	99.06	98.77	97.96	94.81	85.94	58.90

nition accuracy of above 92% could be maintained in SNR of 10dB with the use of CMtN and delta features. The improvements over the state-of-the-art MFCC_0_Z_D or CMN + D scheme were more than 5%. In addition, recognition rate greater than 80% was achieved in SNR of 5dB compared to 62.43% by CTM + D.

It was observed that the hybrid combination of CMtN and dynamic features gave more robust performance when dynamic features were derived from baseline MFCC_0. Even though the inclusion of temporal information from cepstral moment normalized features improved the recognition performance in low SNR significantly, optimal recognition performances were achieved with DMFCC_0 features derived from baseline MFCC_0_Z. This demonstrated that the time trajectories of the cepstral features had been affected by the transformation of CMtN.

Nevertheless, we have shown the robustness of cepstral moment normalization for speech recognition in non-stationary additive babble ensemble environments.

7. Conclusion

Cepstral moment normalization schemes (CMtN) have shown potentials in low signal-to-noise ratio non-stationary additive environments. Significant improvements were achieved and the recognition performance surpassed those of the state-of-the-art features in low SNR environments.

Odd order CMtN schemes were slow in convergence but the incorporation of cepstral variance normalization in the algorithm has reduced the convergence time and improved the robustness of cepstral features. The performance of CMtN could be further enhanced with the inclusion of temporal information through dynamic features. The hybrid combination of moment normalized cepstral features and dynamic features derived from baseline MFCC_0_Z features contributed to optimal recognition performance in this work.

Nevertheless, it has been shown that CMtN 2 and CMtN 3 offer significant robustness in speech recognition. There was not evident need for higher order moments such as 5 and above. Higher order CMtN may give slight improvements in certain conditions but the computation time and complexity increased as the order increased especially for higher odd order moment normalization.

This paper has presented a concise work on the limitations and contributions of cepstral moment normalization schemes for speech recognition in additive noisy environments

References

Acero, A. and X. Huang (1995, Dec.). Augmented cepstral normalization for robust speech recognition. In *Proc. IEEE Workshop on ASR*.

Barras, C. and J. Gauvain (2003). Feature and score normalization for speaker verification of cellular data. In *Proc. ICASSP*.

Harvey, C. and A. Siddique (1999). Autoregressive conditional skewness. *Journal of Financial and Quantitative Analysis 34*, 465–487.

Hilger, F. and H. Ney (2001, Sept.). Quantile based histogram equalization for noise robust speech recognition. In *Proc. 7th European Conference on Speech Communication and Technology*.

Hsu, C. and L. Lee (2004). Higher order cepstral moment normalization (hocmn) for robust speech recognition. In *Proc. ICASSP*, pp. 197–200.

Jain, P. and H. Hermansky (2001). Improved mean and variance normalization for robust speech recognition. In *Proc. ICASSP*.

Melin, H. and J. Lindberg (1999). Variance flooring, scaling and tying for text-dependent speaker verification. In *Proc. EUROSPEECH*, pp. 1975–1978.

Molau, S., F. Hilger, and H. Ney (2003). Feature space normalization in adverse conditions. In *Proc. ICASSP*, pp. 656–659.

Sahguchi, Y., S. Ozawa, and M. Kotani (2002). Feature extraction using supervised independent component analysis by maximizing class distance. In *Proc. ICONIP*.

Suk, Y., S. Choi, and H. Lee (1999, April). Cepstrum third-order normalization method for noisy speech recognition. In *IEEE Electronic Letters*, Volume 35, no.7, pp. 527–528.

Toh, A., R. Togneri, and S. Nordholm (2005). Investigation of robust features for speech recognition in hostile environments. In *Proc. APCC*, pp. 956–960.

Torre, A., A. M. Peinado, J. Segure, J.L.Perez, C. Bentez, and A. J. Rubio (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions SAP 13*, 355–366.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review