Extension and Further Analysis of Higher Order Cepstral Moment Normalization (HOCMN) for Robust Features in Speech Recognition

Chang-wen Hsu and Lin-shan Lee

Graduate Institute of Communication Engineering
National Taiwan University, Taiwan, Republic of China
poseidons@speech.ee.ntu.edu.tw , lslee@gate.sinica.edu.tw

10.21437/Interspeech.2006-11

Abstract

Cepstral normalization has been popularly used as a powerful approach to produce robust features for speech recognition. Good examples of approaches include the well known Cepstral Mean Subtraction (CMS) and Cepstral Mean and Variance Normalization (CMVN), in which either the first or both the first and the second moments of the Mel-frequency Cepstral Coefficients (MFCCs) are normalized [1, 2]. Such approaches were extended previously to Higher Order Cepstral Moment Normalization (HOCMN) for normalizing moments with orders much higher than two [3]. Here we further extend HOCMN to a more generalized form with the generalized moment with non-integer orders defined in this paper. Extensive experimental results based on a newly defined development set for AURORA 2.0 indicated that not only HOCMN for integer moment orders can perform significantly better than the well-known approach of Histogram Equalization (HEQ), but some further improvements can be consistently obtained for almost all SNR values with non-integer moment orders. The theoretical foundation behind the approaches proposed here which explains why HOCMN can perform well and how the statistical properties of the distributions of the MFCC parameters are adjusted during the normalization processes were also discussed.
Index Terms: Robust speech recognition, cepstral normalization, *N*-th order moment.

1. Introduction

We start with the conventional cepstral moment normalization and introduce the notation to be used here. The *N*-th order moment of a MFCC parameter sequence $X(n)$ is the expectation value of $X^N(n)$, a simplified notation for $[X(n)]^N$ to be used throughout this paper, where N is usually an integer and the expectation value is approximated by the time average over some interval, $\{k = 0, 1, 2, \cdots, T\text{-}1\}$, where k is the time index in the interval,

$$E[X^N(n)] \triangleq \frac{1}{T}\sum_{k=0}^{T-1} X^N(k) \ . \tag{1}$$

In such cases, the purpose of moment normalization of order N is to have

$$E\left[X_{[N]}^N(n)\right] = 0 \tag{2}$$

if N is an odd integer, where the subscript $[N]$ indicates that the sequence $X_{[N]}(n)$ is the normalized version of $X(n)$ whose N-th order moment has been normalized, and

$$E\left[X_{[N]}^N(n)\right] = M_N \tag{3}$$

if N is an even integer, where M_N is the N-th moment of a Gaussian distribution with unit variance, $N(0,1)$, obtained by the moment generating function. Here we assume for simplicity the reference distribution to be used in the normalization is Gaussian with unit variance, although other distribution can also be used, for example that obtained with some training corpus. With the above notation, the well-known CMS is

$$X_{CMS}(n) \triangleq X_{[1]}(n) = X(n) - E\left[X^1(n)\right] \ , \tag{4}$$

where $E[X^1(n)]$ can be obtained by equation (1) with $N = 1$, and the well-known CMVN is

$$X_{CMVN}(n) \triangleq X_{[1,2]}(n) = X_{[1]}(n) \Big/ \sqrt{E\left[X_{[1]}^2(n)\right]} \ , \tag{5}$$

where $X_{[L,N]}(n)$ is the normalized version of $X(n)$ whose L-th and N-th moments have both been normalized as in equations (2) and (3), and so on.

Previously, we proposed an extension referred to as Higher Order Cepstral Moment Normalization (HOCMN) which is developed by extending the concept of CMS and CMVN to moment orders much higher than two [3]. We showed that the recognition accuracy was significantly improved if the normalized even moment order was extended from 2 to 100 using a carefully chosen scaling factor and the normalized odd moment order was extended from 1 to 3 or 5 using an iterative procedure [3]. Here in this paper we further extend HOCMN to a more generalized form with the generalized moment with non-integer orders being normalized, and offer further analysis and discussions regarding the fundamental principles behind these approaches based on the statistical properties of the distributions of the MFCC parameters.

In the following, the extended HOCMN and statistical principles are first formulated in section 2. The experimental setup based on AURORA 2.0 testing environment is described in section 3, and some experimental results and discussions are presented in section 4. Finally, we make concluding remarks in section 5.

2. Extension and further analysis of HOCMN

2.1. Generalized moments for a non-integer order *u*

In the initial development of HOCMN [3], the values of N, L or the orders of the moments being normalized are always taken as even and odd integers respectively. However, these orders do not have to be integers but can be any positive real number u. With such considerations, we can further define two types of generalized moments as given below. In the first case, the sign of each parameter sample $X(n)$ in equation (1) is retained first and only the absolute value of the sample $X(k)$ is raised to a non-integer power order u. So equation (1) is generalized to

$$E_1\left[X^u(n)\right] \triangleq \frac{1}{T}\sum_{k=0}^{T-1} \mathrm{sgn}\left[X(k)\right] \cdot \left(abs\left[X(k)\right]\right)^u \ . \tag{6}$$

$E_1[X^u(n)]$ in equation (6) is referred to as the generalized moment of the first type with order u here. In the second case, in evaluating the moments the sign of each parameter sample $X(k)$ in equation (1) is simply removed,

$$E_2\left[X^u(n)\right] \triangleq \frac{1}{T}\sum_{k=0}^{T-1} abs\left[X(k)\right]^u \ . \tag{7}$$

$E_2[X^u(n)]$ in equation (7) is referred to as the generalized moment of the second type with order u here. With the above

definitions, $E_1[X^u(n)]$ in equation (6) reduces to equation (1) if u is an odd integer, and $E_2[X^u(n)]$ in equation (7) reduces to equation (1) if u is an even integer. So the conventional definition of moments in equation (1) remains valid here for integer orders, in which case the two types in equations (6) and (7) converge into one in equation (1).

2.2. HOCMN with non-integer moment orders

With the set of generalized moments defined in equations (6) and (7), the HOCMN proposed previously can be directly extended to non-integer moment orders. When the generalized moment of the second type $E_2[X^u(n)]$ in equation (7) is used, HOCMN for an even integer N using a carefully chosen scaling factor developed previously [3] can be directly applied for an arbitrary real number u, except N is replaced by u here. Similar to the even integer, a parameter sequence can be normalized with respect to the generalized moment of the second type $E_2[X^u(n)]$ for a single real number u in addition to CMS. Similarly, when the generalized moment of the first type $E_1[X^u(n)]$ in equation (6) is used, HOCMN for an odd number L using an iterative procedure developed previously [3] can be directly applied for an arbitrary real number u, except L is replaced by u. Similar to the odd integer, a parameter sequence can be normalized with respect to the generalized moment of the first type $E_1[X^u(n)]$ for a single real number u in addition to CMS. These two types of normalization with respect to the two types of generalized moments can also be cascaded as developed previously [3] to produce a $HOCMN_{[1,u_1,u_2]}$ process, where u_1 and u_2 are the two real number orders for the generalized moments of the first and the second types being normalized respectively.

2.3. Fundamental principles behind HOCMN

In statistics the "third moment about the mean", normalized to the standard deviation, is referred to as the "skewness" of a distribution, or its departure from symmetry,

$$S' = E\left[\left(X-\mu_X\right)^3\right]\Big/\sigma_X^3 \tag{8}$$

Where μ_X and σ_X are the mean and standard deviation of the random variable X whose distribution is being considered. A positive or negative value of S' in equation (8) indicates the distribution is skewed to the right or left, and S' is zero if the distribution is symmetric [4]. On the other hand, the "fourth moment about the mean", normalized to the standard deviation, is referred to as the "kurtosis" of a distribution, or whether the distribution is "peaked" or "flat with tails of larger size",

$$K' = \left(E\left[\left(X-\mu_X\right)^4\right]\Big/\sigma_X^4\right) - 3 \tag{9}$$

where 3 is the value for a standard normal distribution. A positive value of K' in equation (9) indicates the distribution is flatter with tails of larger size than a standard normal distribution, and a negative value of K' indicates it is "more peaked" with smaller tails than a standard normal distribution.

In the definition of N-th order moment in equation (1), however, for simplicity the mean is not subtracted and the normalization with respect to standard deviation is not performed either as were done in equations (8) and (9). So the N-th order moment in equation (1) for $N = 3$ or 4 are simply "un-normalized third or fourth moments about the origin". Although slightly different from those in equations (8) and (9), the statistical properties they carry are very similar. So they may be referred to as "modified skewness or kurtosis about the origin", for $N = 3$ or 4 respectively,

$$S = E\left[X^3\right],\ K = E\left[X^4\right]. \tag{10}$$

The above concept of "modified skewness or kurtosis" as defined in equation (10) can be further extended to other moment orders different from 3 and 4. For an odd integer L (which is not necessarily 3), the L-th moment can be considered as the "generalized skewness of order L about the origin", while for an even integer N (which is not necessarily 4), the N-th moment can be considered as the "generalized kurtosis of order N about the origin",

$$S^{(L)} = E\left[X^L\right],\ L\text{: an odd integer}, \tag{11}$$

$$K^{(N)} = E\left[X^N\right],\ N\text{: an even integer}. \tag{12}$$

Both of $S^{(L)}$ and $K^{(N)}$ as defined above have very similar interpretation as the "skewness" and "kurtosis" in equations (8) and (9), except here the distance of the parameter values from the origin are emphasized by different orders.

The normalization to have a moment with odd integer order L being zero in equation (2) is then to constrain the distribution to be symmetric about the origin in the sense of "generalized skewness of order L about the origin". The normalization to have a moment with even integer order N being that of a standard normal distribution in equation (3) is to constrain the distribution to be "equally flat with tails of equal size" as compared to a standard normal distribution in the sense of the "generalized kurtosis of order N about the origin". The above interpretation can then be extended to the generalized moments of non-integer orders discussed in section 2.1 as well. So the generalized moments of the first and second types in equation (6) and (7) have to do with the "generalized skewness" and "generalized kurtosis" of the distribution, and so on.

3. Experimental setup

The above approaches were evaluated by the AURORA 2.0 testing environment with an English connected-digit string corpus. Two training conditions (clean condition / multi-condition) and three testing sets (sets A/B/C) were defined by AURORA 2.0 [5]. In clean-condition training the acoustic models are trained by clean speech only, while in multi-condition training the models are trained by a corpus with both clean and noisy speech. The testing set A included four different types of noise which were used in the multi-condition training (subway, babble, car and exhibition), while the testing set B included another four different types of noise not used in the multi-condition training (restaurant, street, airport and train station). The testing set C then included two noise types respectively from sets A and B (subway and street), plus additional convolutional noise. Five different SNR values, ranging from 20dB to 0dB, were tested in each case. Whole-word HMM models were used as specified by AURORA 2.0. Each word had 16 states and 3 Gaussian mixtures per state. The speech features were extracted by the AURORA WI007 Front-end, which converted each signal frame into 13 cepstral coefficients (MFCCs, C0~C12), on which all the normalization techniques proposed above were performed. The first and second derivatives were then computed from the normalized cepstral coefficients and used as well in the tests. The implementation of the normalization approaches proposed here was based on the statistics of progressively moving segments. In other words, the summation in equation (1) was performed over a progressively moving segment with length l+1, including the preceding l/2 frames and following l/2 frames.

3.1. Development set and object function based on the clean-condition training set of AURORA 2.0

In addition to the testing environment as summarized above, we also defined a development set based on this testing environment, to be used for the selection of the various moment orders L, N or u, u_1, u_2 as mentioned in section 2.1-2.2. For this purpose, we divided all the 8440 utterances in the clean training corpus of AURORA 2.0 into two subsets, 7544 utterances for training and the rest 896 for testing. We added the eight types of noise used in AURORA 2.0 as summarized above on the second

subset of 896 utterances (now defined as the testing data of the development set, originally in the clean training corpus of AURORA 2.0) with SNR ranging from 20dB to -5dB respectively as the testing data for the development set. The first subset of 7544 utterances was then used for clean-condition training for the development set. So the testing conditions for the development set is very similar to those with clean-condition training and testing sets A and B defined in AURORA 2.0. The averaged word accuracy for all these forty conditions (eight types of noise and five SNR values) was then used as the object function for parameter selection.

4. Preliminary experimental results

4.1. HOCMN for integer orders with parameters selected by the development set

With the tests as reported above, it is clear that the performance of $HOCMN_{[1,L,N]}$ depends on many parameters, the odd and even moment orders L and N, the best lengths of the processing segments to be used in estimating these moments of orders L and N, referred to as l_L and l_N, also the number of iterations used for the odd order moment normalization. It is thus reasonable to use the development set and object function as defined in section 3 to find a sub-optimal set of these parameters. This makes sense for practical applications, because it is always possible to obtain a set of parameters in this way using a development set with conditions similar to the application task.

Also in the initial experiments of HOCMN, we observed that in the AURORA 2.0 corpus, many utterances were actually very short for precise estimate of the odd order moments, but still acceptable for normalization of even order moments. Therefore in the following experiments $HOCMN_{[1,L,N]}$ with processing segment length l_L and l_N was performed in a slightly different way, in which the odd order moment normalization was not performed as long as the utterance length was less than l_L, but the even order normalization was always performed regardless of the utterance length.

With all the above, the best set of parameters obtained here with the help of the development set was found to be $L = 3$, $N = 100$, $l_3 = 120$, $l_{100} = 160$, with complete results for the testing sets A, B and C and overall average listed in row (b) of Table 1. The overall average is 84.73%, representing a relative error rate reduction of 12.43% as compared to the baseline of $HOCMN_{[1,3,100]}$, $l = 86$ (row (a) of Table 1) proposed in [3].

4.2. Comparison of HOCMN with the well known Histogram Equalization (HEQ)

Also shown in row (c) of Table 1 for comparison is the well known approach of Histogram Equalization (HEQ) [6, 7] (in the case of equalizing into a standard Gaussian), also processed with a moving segment with length $l = 98$ which was similarly selected with the development set and the object function. It can be found that the best case of $HOCMN_{[1,3,100]}$ ($l_3 = 120$, $l_{100} = 160$) in row (b) outperformed HEQ in row (c) in all testing sets, with overall averaged accuracy of 84.73% as compared to that of 83.38% with HEQ.

It is important to discuss why HOCMN can perform better than HEQ. HEQ essentially equalized (or normalized) the entire distribution of the MFCC parameters, so all moments of all orders are normalized to a good degree. As a result the parameters may be over-fitted to a given distribution and thus slightly different from their original nature. HOCMN, however, only normalized three moments of orders 1, 3 and 100, may therefore preserve more original nature of the parameters. On the other hand, here the normalization was performed only with short-term statistics for either HEQ or HOCMN. Equalizing the entire distribution based on only the limited quantity of data over a short period of time is inherently difficult. In the case of

Clean Condition Training	Clean Condition Training (HOCMN, $N = 100$)				Relative Error Rate Reduction
	Set A	Set B	Set C	Avg.	
(a) $HOCMN_{[1,3,100]}$ (l=86) [3]	81.24	83.95	82.46	82.57	–
(b) $HOCMN_{[1,3,100]}$ (l_3=120, l_{100}=160)	83.78	86.12	83.87	84.73	12.43%
(c) HEQ (l=98)	82.44	84.45	83.11	83.38	–

Table 1. *Complete data for the best results obtained previously [3] in row (a), the best results obtained here in row (b), and for HEQ in row (c).*

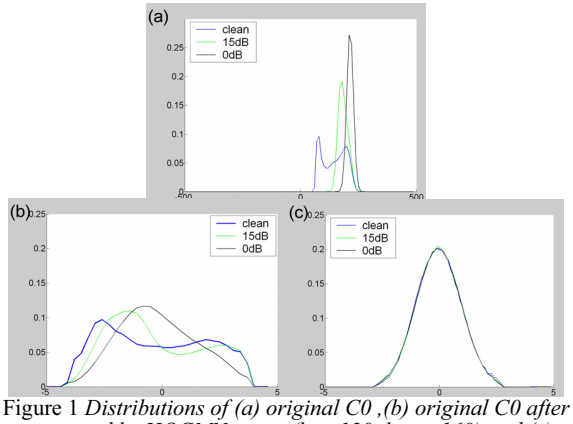

Figure 1 *Distributions of (a) original C0 ,(b) original C0 after processed by $HOCMN_{[1,3,100]}$ ($l_3 = 120$, $l_{100} = 160$) and (c) original C0 after processed by HEQ ($l = 98$).*

equalizing with respect to a standard Gaussian distribution, the limited quantity of data is not necessarily best approximated by a standard Gaussian distribution. For HOCMN, however, it may be more reasonable to try to normalize only the three moments of orders 1, 3 and 100, rather than the entire distribution, with the limited quantity of data. Although the limited quantity of data is also not adequate to estimate the few moments including the 100-th moment precisely, the normalization for the moment of even order 100 only needs to estimate a scaling factor which depends on the 100-th root of the 100-th moment [3]. Therefore, the estimation error of this scaling factor can be significantly reduced. All these are probably why HOCMN can perform better than HEQ.

Consider the distributions of parameters C0 shown in Figure 1 (a) for all utterances in the testing set with all types of noise but separately under three SNR values: clean, 15dB and 0dB. All these parameters were respectively normalized by the best cases of $HOCMN_{[1,3,100]}$ ($l_3 = 120$, $l_{100} = 160$) and HEQ ($l = 98$) as represented by rows (b) (c) of Table 1, and respectively shown in Figure 1 (b) and (c). Comparing Figure 1 (b) with Figure 1 (a), it can be found that with HOCMN the parameters were normalized to have excellent symmetry about the origin in the sense of "generalized skewness of order 1 and 3", and excellent flatness with tail size equal to a standard normal distribution in the sense of "generalized kurtosis of order 100", although the original shapes of the distributions were more or less preserved. For those processed by HEQ in Figure 1 (c), on the other hand, the distributions for clean, 15dB and 0dB were very close to each other because they were all equalized into a standard Gaussian. Such a "forced matched" condition is not necessarily the best for the reasons mentioned above, because the characteristic nature of the original parameter distributions may be smeared out by the equalization process.

The above difference between HOCMN and HEQ can also be observed with some averaged distance measure,

$$d = E\left[\frac{\|\bar{y} - \bar{x}\|}{\|\bar{x}\|}\right], \qquad (13)$$

where $\bar{x}$ is the 13-dimensional vector of MFCC parameters for

clean speech and $\overline{y}$ the corresponding noisy speech version but processed by HEQ ($l = 98$) or by $HOCMN_{[1,3,100]}$ ($l_3 = 120$, $l_{100} = 160$) respectively, $\|\cdot\|$ is the Euclidean distance, and the average $E[\cdot]$ is performed over all utterances in the AURORA 2.0 testing set, including all different types of noise but separated for different SNR values. So the distance measure d here reflects how the normalized feature vectors are "individually" matched to their clean speech versions, as compared to Figure 1 which shows how the feature parameters are normalized "collectively".

The results of the distance measure d evaluated by equation (13) for feature vectors $\overline{y}$ processed by HEQ and $HOCMN_{[1,3,100]}$ are listed in rows (1) and (2) of Table 2 respectively. It can be found that the averaged distance d for HOCMN-processed features is consistently smaller than that for HEQ-processed features across all SNR values. The "relative distance reduction" for $HOCMN_{[1,3,100]}$ as compared to HEQ is also listed in row (3) of the table. Also listed in rows (4) and (5) in the lower half of Table 2 are respectively the accuracies achieved by HEQ and by $HOCMN_{[1,3,100]}$, averaged over all different types of noise but separated for different SNR values, with the error rate reduction ratio achieved by $HOCMN_{[1,3,100]}$ as compared to HEQ listed in row (6) of the table. It can be found that not only higher accuracies were consistently obtained by $HOCMN_{[1,3,100]}$ for all SNR values and the error rate reduction were significant except for 0dB of SNR, but the error rate reduction in row (6) were in fact in parallel with the "relative distance reduction" in row (3) to a good extent, i.e., more significant error rate reduction or distance reduction for higher SNR, but minor difference for lower SNR. This verified that when processed by HOCMN the feature vectors $\overline{y}$ were in fact better matched "individually" to its corresponding clean speech versions $\overline{x}$ than HEQ-processed vectors, although the "complete distributions" of HEQ-processed feature vectors $\overline{y}$ looked better matched to those of the corresponding clean speech versions $\overline{x}$ (e.g. those in Figure 1 (b) vs. those in Figure 1(c)).

4.3. HOCMN with non-integer moment orders

As mentioned previously in sections 2.2, HOCMN can also be performed with generalized moments with non-integer orders. The previous work [3] indicated that the performance of $HOCMN_{[1,L,N]}$ for even moment order N was saturated with larger integer of N such as $N = 100$. So replacing $N = 100$ by a non-integer u_2 cannot help. From the experiments described in section 4.1 leading to the best set of parameters $HOCMN_{[1,3,100]}$ ($l_3 = 120$, $l_{100} = 160$) in row (b) of Table 2, however, it seemed the odd moment order $L = 3$ was not necessarily the best, thus replacing it by a non-integer u_1 to have $HOCMN_{[1,u_1,100]}$ may help, though in that case the parameters such as the processing segment length l_{u_1}, l_{100} and number of iterations in normalizing the generalized moment of first type may all need to be adjusted as well. In addition, the orders of the moments being normalized can be different for the 13 MFCC parameters, which were assumed always the same for all the experiments reported above. All these were done with $HOCMN_{[u_1,u_1,100]}$ here, with all parameters selected for each MFCC parameter based on the development set and the object function defined in section 3.1.

The results averaged over all different types of noise but separated for different SNR values are listed as row (2) in Table 3, as compared to the case in row (1) which is the same as the case in row (b) of Table 1, or the best results for HOCMN with the same integer moments for all MFCC parameters. It can be found from Table 3 that the improvements obtainable are very limited except for the 0dB and -5dB cases (57.21 vs. 56.07% and 25.72 vs. 23.71) respectively. Although the improvements by the non-integer orders are limited when averaged over many conditions as shown in Table 3, they may make the difference under specific conditions, for example for a specific type of noise, which can be found by a development set describing such conditions.

SNR		20dB	15dB	10dB	5dB	0dB
(1)	d(HEQ)	0.8764	0.9621	1.0483	1.1376	1.2291
(2)	d(HOCMN)	0.8314	0.9168	1.0056	1.1010	1.2013
(3)	$\frac{d(HEQ)-d(HOCMN)}{d(HEQ)}$	5.13%	4.71%	4.07%	3.22%	2.26%
(4)	Accuracies(HEQ)	96.98	94.99	90.27	78.93	55.71
(5)	Accuracies(HOCMN)	97.90	96.15	92.15	81.40	56.07
(6)	Error Rate Reduction (HOCMN vs. HEQ)	30.46%	23.15%	19.32%	11.72%	0.81%

Table 2. *Comparison of distance measure d and accuracies obtained by HEQ and by $HOCMN_{[1,3,100]}$ respectively, together with distance reduction and error reduction ratios.*

Clean Condition		20dB	15dB	10dB	5dB	0dB	-5dB	Avg.
(1)	$HOCMN_{[1,3,100]}$ (l_3=120, l_{100}=160)	97.90	96.15	92.15	81.40	56.07	23.71	84.73
(2)	$HOCMN_{[1,u_1,100]}$ (u_1, l_{u_1}, l_{100} selected for each MFCC parameter)	97.93	96.23	92.17	81.41	57.21	25.72	84.99

Table 3. *Performance of best cases of HOCMN when the moment orders being normalized can be integers or non-integer values and different for different MFCC parameters.*

5. Conclusions

In this paper, we extended the previously proposed concept of HOCMN to generalized moments with non-integer orders. Extensive experimental results based on a newly defined development set for AURORA 2.0 indicated that not only HOCMN for integer moment orders can perform significantly better than the well-known approach of Histogram Equalization (HEQ), but some further improvements can be consistently obtained for almost all SNR values with non-integer moment orders. The theoretical foundation behind the approaches proposed here which explains why HOCMN can perform well and how the statistical properties of the distributions of the MFCC parameters are adjusted during the normalization processes were also discussed.

6. References

[1] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on ASSP, 1981.

[2] O. Viikki, K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition", Speech Communication, Vol. 25, pp. 133-147, August 1998.

[3] Chang-wen. Hsu, Lin-shan Lee, "Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition", *ICASSP'04*, 2004.

[4] David J. Krus, *Visual Statistics*, Aug. 2003. http://www.visualstatistics.net/index.htm

[5] H. G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000, Paris, September 2000.

[6] F. Hilger and H.Ney, "Quantile Based Histogram Equalization for Noise Robust Speech Recognition", Proceedings of Eurospeech, 1135-1138, 2001.

[7] Á. de la Torre, J. C. Segura, C. Benítez, A. M. Peinado, and A. J. Rubio, "Non-linear Transformations of the Feature Space for Robust Speech Recognition", *ICASSP'02*, 2002.