



Higher Order Cepstral Moment Normalization for Improved Robust Speech Recognition

Author : Chang-Wen Hsu, Lin-Shan Lee

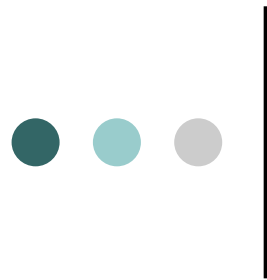
Reporter : 邱聖權

Professor : 陳嘉平



Introduction

- CMS and CMVN are popular approach to reduce mismatch between training and testing by normalizing the first and second moments, mean and variance, of the feature distribution.
- The higher order moments are more dominated by samples with larger values, which very likely include the harmful parts of environmental disturbances and constitute a major source of mismatch.



Fundamental principles behind HOCMN

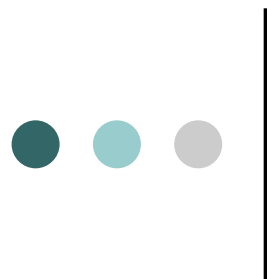
- The third moment about the mean, normalized to the standard deviation, is referred to as the “skewness” of a distribution

$$S' = \frac{E[(X - \mu)^3]}{\sigma^3}$$

- The fourth moment about the mean, normalized to the standard deviation, is referred to as the “kurtosis” of a distribution

$$K' = \left(\frac{E[(X - \mu)^4]}{\sigma^4} \right) - 3$$

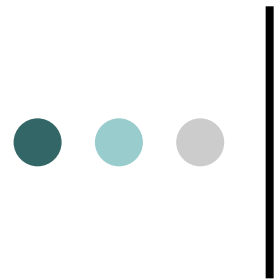
where 3 is the 4th moment for a standard normal distribution.



Conventional cepstral moment normalization

- The Nth-order moment of a MFCC parameter sequence $X(n)$ is the expectation value of $X^N(n)$, a simplified notation for $[X(n)]^N$
- The expectation value is approximated by the time over some interval $\{k = 0, 1, 2, \dots, T - 1\}$

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



Normalization of odd and even order moments

- When order N is odd, the purpose of order N moment normalization is to have

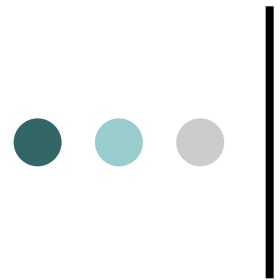
$$E[X_{[N]}^N(n)] = 0$$

where $X_{[N]}(n)$ is the N th order moment normalized $X(n)$.

- When N is even,

$$E[X_{[N]}^N(n)] = M_N$$

where M_N is the N th moment of a Gaussian distribution with unit variance $N(0,1)$ obtained by the moment generating function.



CMS and CMVN

- CMS

$$X_{CMS}(n) = X_{[1]}(n) = X(n) - E[X^1(n)]$$

- CMVN

$$X_{CMVN}(n) = X_{[1,2]}(n) = \frac{X_{[1]}(n)}{\sqrt{E[X_{[1]}^2(n)]}}$$

● ● ● | HOCCMN for a larger even integer N

- The high even order moment normalization can be extended from first-order normalization

$$X_{[1,N]}(n) \approx bX_{[1]}(n) = bX_{CMS}(n)$$

$$b = \left[\frac{M_N}{E[X_{[1]}^N(n)]} \right]^{1/N}$$



Derivation of parameter

- When N is relatively large, b can be approximated as

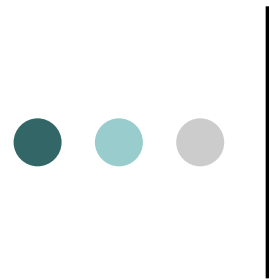
$$b \approx \frac{(M_N)^{1/N}}{\exp\left\{\frac{1}{N} \log\left[\frac{1}{T} \sum_{k=0}^p (X_{[1](k)})^N\right]\right\}}$$

$$= \frac{(M_N)^{1/N}}{\exp\left\{\log(X_{[1](0)}) + \frac{1}{N} \left[1 + \sum_{k=1}^p \left(\frac{X_{[1](k)}}{X_{[1](0)}}\right)^N\right] - \frac{1}{N} \log(T)\right\}}$$

, $X_{[1](0)} \geq X_{1} \geq \dots \geq X_{[1](T-1)} \geq 0$,

p is the number used in the approximation,

when $k > 4$ the term $\frac{X_{[1](k)}}{X_{[1](0)}}$ is very close to zero.



HOCMN for a larger odd integer L

- HOCMN for an high odd order can be extended from the third-order cepstral moment normalization.
- It also coexist with first-order normalization.

$$X_{[1,L]}(n) \approx aX_{[1,L-1]}^{L-1}(n) + X_{[1,L-1]}(n) + c$$

$$X_{[1,L]}(n) \approx a\left(X_{[1,L-1]}^{L-1}(n) - M_{L-1}\right) + X_{[1,L-1]}(n)$$

Derivation of parameter

$$\begin{aligned}
 E[X_{[1,L-1]}^L(n)] &= E\left\{ \left[a(X_{[1,L-1]}^{L-1}(n) - M_{L-1}) + X_{[1,L-1]}(n) \right]^L \right\} \\
 &= a^L \times E\left\{ \left[X_{[1,L-1]}^{L-1}(n) - M_{L-1} \right]^L \right\} \\
 &+ L \times a^{L-1} \times E\left\{ \left[X_{[1,L-1]}^{L-1}(n) - M_{L-1} \right]^{L-1} X_{[1,L-1]}(n) \right\} \\
 &+ \dots + L \times a \times E\left\{ \left[X_{[1,L-1]}^{L-1}(n) - M_{L-1} \right] X_{[1,L-1]}^{L-1}(n) \right\} \\
 &+ E[X_{[1,L-1]}^L(n)] = 0
 \end{aligned}$$

When a is small (usually on $[-0.1, 0.1]$ practically), we can delete the higher order terms and keep only the last two terms

$$a \approx \frac{-E[X_{[1,L-1]}^L(n)]}{L \times E[X_{[1,L-1]}^{2(L-1)}(n) - M_{L-1} X_{[1,L-1]}^{L-1}(n)]}$$

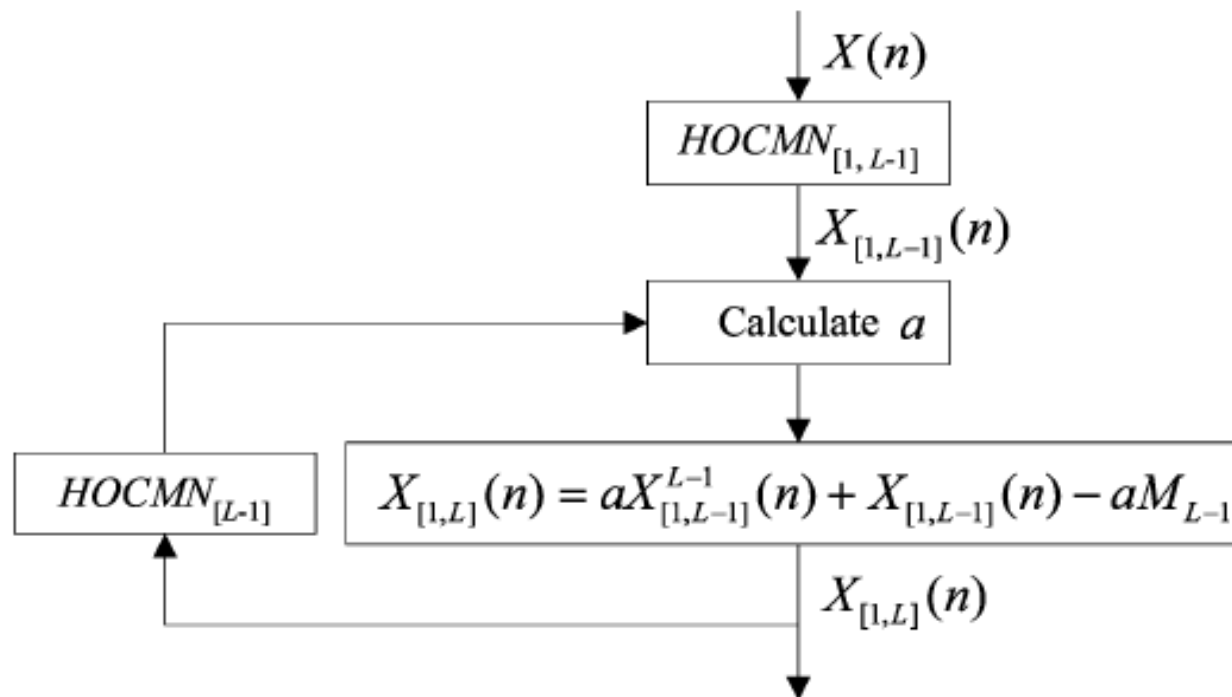
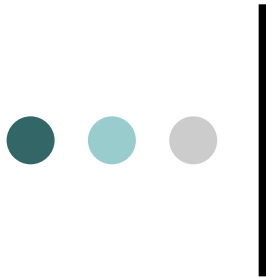
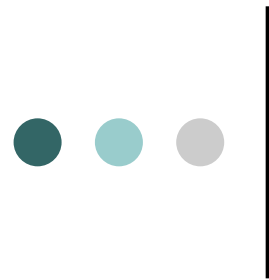


Fig. 1. Flow chart of HOCMN with a larger odd integer L .



Generalized moments for a noninteger-order U

- For odd order

$$E_1[X^u(n)] = \frac{1}{T} \sum_{k=0}^{T-1} \text{sign}(x(k)) \times (\text{abs}(X(k)))^u$$

- For even order

$$E_2[X^u(n)] = \frac{1}{T} \sum_{k=0}^{T-1} (\text{abs}(X(k)))^u$$



Experimental setup

- The cepstral normalization approaches were tested in two different ways, by whole-utterances and by progressively moving segments.
- The segment length is $L+1$, moment is evaluated by present frame and preceding $L/2$ frames and following $L/2$ frames.

Initial results for HOCMN with even-integer orders

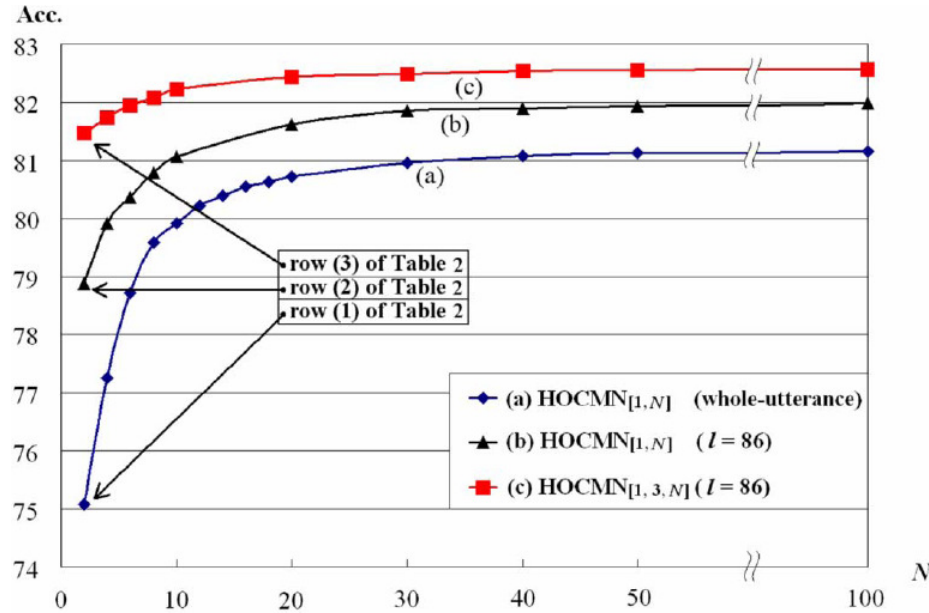


Fig. 4. Recognition accuracy for HOCMN averaged for all noise types and all SNR values in all the three test sets A, B, and C with different even orders N . (a) $\text{HOCMN}_{[1,N]}$ (whole-utterance). (b) $\text{HOCMN}_{[1,N]} (l = 86)$. (c) $\text{HOCMN}_{[1,3,N]} (l = 86)$.

TABLE II
AVERAGED RECOGNITION ACCURACY FOR THE BASELINE EXPERIMENTS
WITH CLEAN-CONDITION TRAINING, OR THE LEFT-MOST POINTS
IN CURVES (a) (b) (c) WITH $N = 2$ IN FIG. 4

Baseline Experiments	Clean-condition Training			
	Set A	Set B	Set C	Avg.
(1) CMVN (whole-utterance)	74.32	75.48	75.79	75.08
(2) CMVN ($l = 86$)	78.03	79.77	78.75	78.87
(3) $\text{HOCMN}_{[1,3,2]} (l = 86)$	80.23	82.70	81.48	81.47

Experiment results on Aurora2 (clean training)

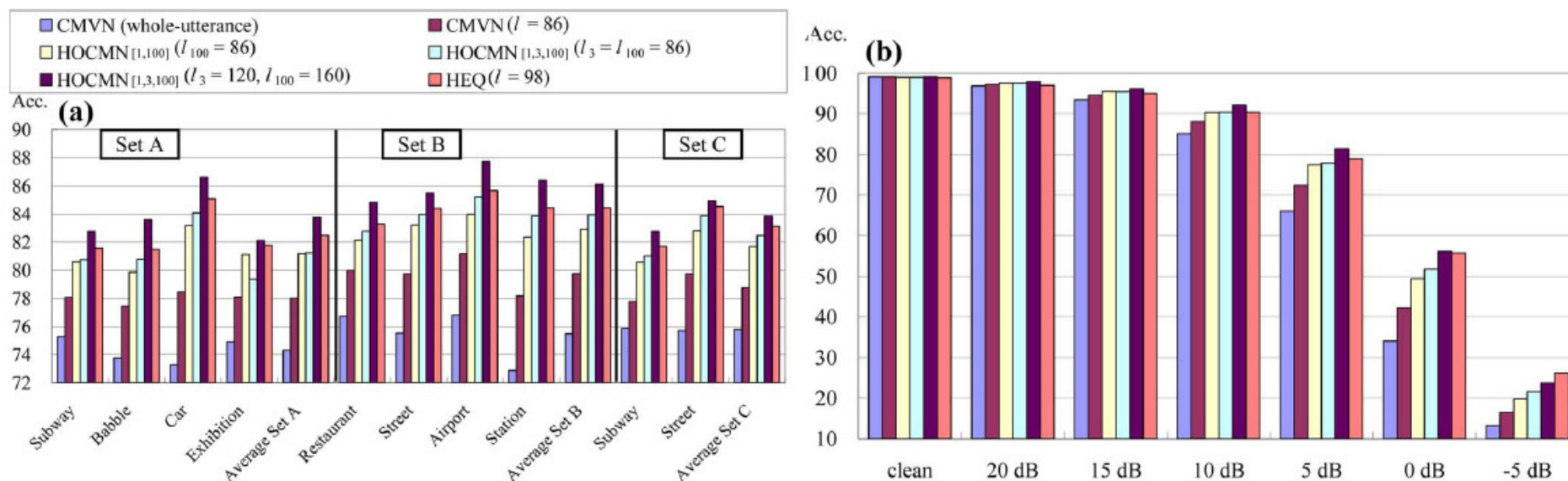


Fig. 5. Comparison of several representative cases tested here for HOCMN with integer moment orders for (a) different types of noise in different testing sets, averaged over all SNR values and (b) different SNR values, averaged over all types of noise in all different testing sets A, B, and C.



Comparison of HOCMN with CMVN

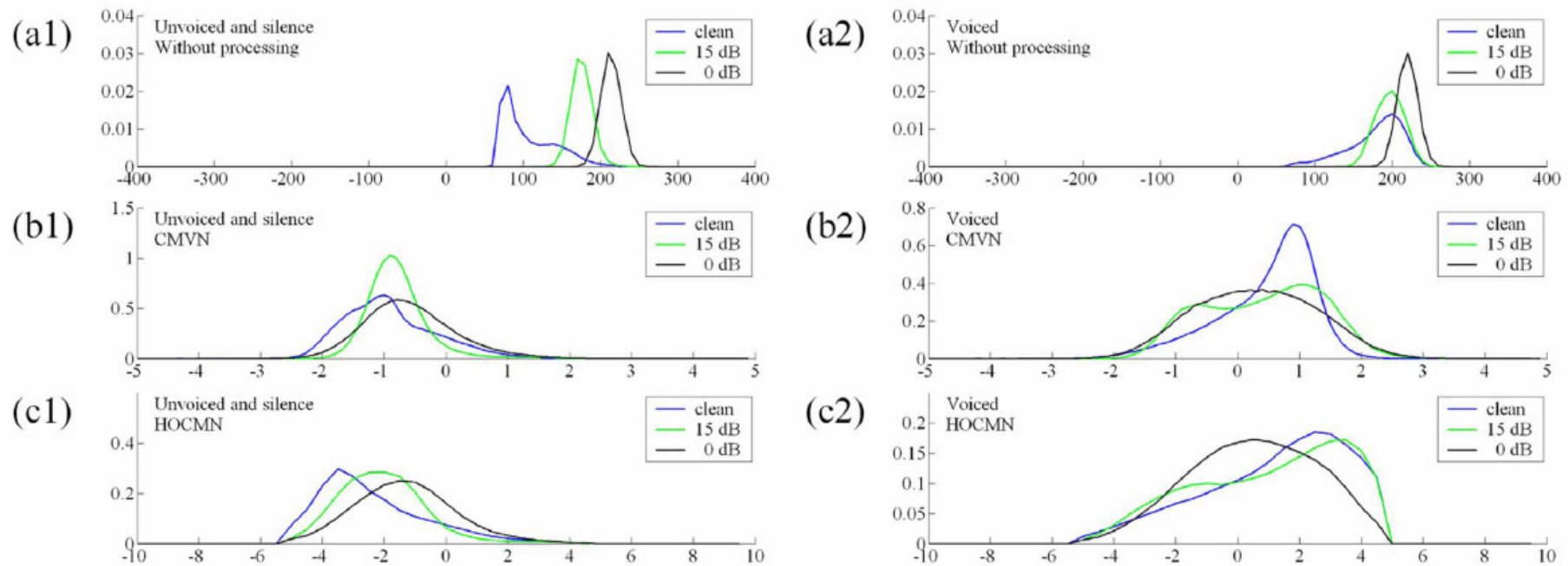


Fig. 6. Distributions of C0 parameters divided into two classes: unvoiced and silence parts in (a1), (b1), (c1) and voiced parts in (a2), (b2), (c2), obtained from all utterances in the testing sets of AURORA 2, including all types of noise (in sets A, B, and C) separately under three SNR values: clean, 15 dB, and 0 dB, without processing in (a1), (a2), with CMVN ($l = 86$) in (b1), (b2), and with HOCMN_[1,3,100] ($l_3 = 120$, $l_{100} = 160$) in (c1), (c2).

Comparison of HOCMN with the Well-Known HEQ

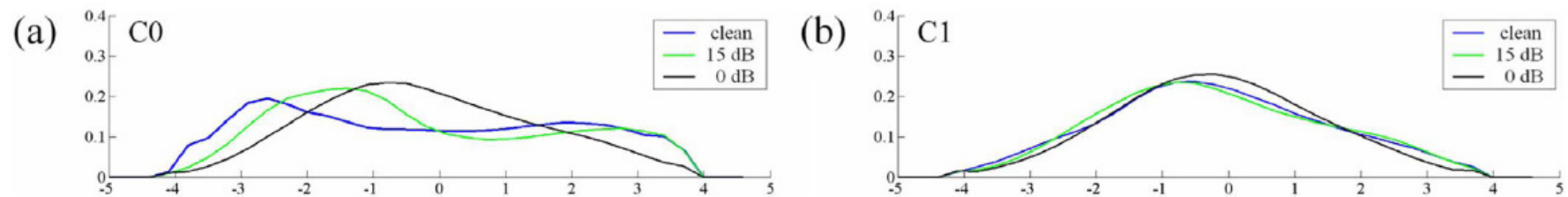


Fig. 8. Distributions of (a) C0 and (b) C1, obtained from all utterances in the testing sets of AURORA 2, including all different types of noise (4 in set A, 4 in set B, 2 in set C) separately under three SNR values: clean, 15 dB, and 0 dB after processing by HOCMN_[1,3,100] ($l_3 = 120$, $l_{100} = 160$).

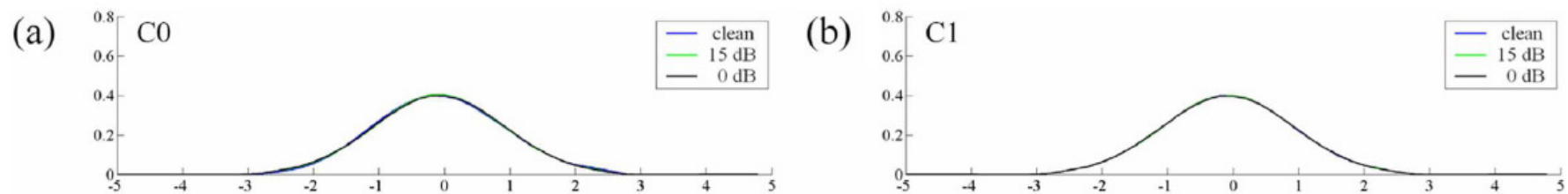


Fig. 9. Distributions of (a) C0 and (b) C1, obtained from all utterances in the testing sets of AURORA 2, including all different types of noise (4 in set A, 4 in set B, 2 in set C) separately under three SNR values: clean, 15 dB, and 0 dB after processing by HEQ ($l = 98$).

HOCMN with noninteger moment orders

TABLE V
PERFORMANCE OF THE BEST CASES OF HOCMN (1) WHEN ONLY INTEGRAL MOMENT ORDERS ARE NORMALIZED, AND (2) NONINTEGER VALUES AND MAY VARY FOR EACH MFCC PARAMETER AND SNR VALUE, ASSUMING SNR VALUE IS PRECISELY KNOWN

Clean Condition	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	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 VI
OPTIMAL NONINTEGER MOMENT ORDERS FOR HOCMN FOR EACH MFCC PARAMETER AND SNR VALUE, ASSUMING SNR VALUES ARE KNOWN

	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
C1	3.0	3.2	3.1	3.0	3.2	3.1
C2	3.0	3.0	3.1	3.0	3.1	3.0
C3	3.0	3.2	3.1	3.0	2.9	3.1
C4	3.0	3.0	3.1	3.0	2.9	3.1
C5	3.0	3.1	3.1	3.0	3.1	3.1
C6	3.0	3.1	3.0	3.0	3.1	3.1
C7	3.0	3.0	3.1	3.0	3.1	3.1
C8	3.0	3.0	3.1	3.0	3.1	3.1
C9	3.0	3.1	3.2	3.0	3.1	3.1
C10	3.0	3.0	3.1	3.0	3.1	3.1
C11	3.0	3.1	3.1	3.0	3.2	3.1
C12	3.0	3.1	3.1	3.0	3.1	3.1
C0	3.0	3.0	3.1	3.0	3.1	3.4



HOCMN with multicondition training

TABLE VII

COMPARISON OF ACCURACIES OBTAINED WITH MULTICONDITION TRAINING BY (1) CMVN, (2) HEQ, AND (3) HOCMN, AVERAGED OVER ALL DIFFERENT TYPES OF NOISE BUT SEPARATED FOR DIFFERENT SNR VALUES, TOGETHER WITH (4), (5) RELATIVE ERROR RATE REDUCTIONS FOR HOCMN COMPARED WITH CMVN AND HEQ, RESPECTIVELY

Multi-condition		20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg
(1)	CMVN ($l=200$)	98.36	97.53	95.61	88.92	69.83	33.78	90.05
(2)	HEQ ($l=240$)	98.07	97.24	95.59	90.34	74.03	40.22	91.05
(3)	$HOCMN_{[1,u_1,100]}$ (u_1, l_{u_1}, l_{100} selected for each MFCC parameter)	98.42	97.57	96.10	90.32	73.45	39.78	91.17
Error Rate	(4) HOCMN vs. CMVN	3.48%	1.50%	11.21%	12.64%	12.01%	9.07%	11.28%
Reduction	(5) HOCMN vs. HEQ	18.15%	11.78%	11.71%	-0.22%	-2.22%	-0.73%	1.33%



Experimental results on Aurora 3

Danish	WM	MM	HM	Avg
baseline	78.61	55.37	38.09	60.35
CMVN ($l = 200$)	81.98	64.12	57.99	69.73
HEQ ($l = 240$)	77.04	58.47	59.17	66.07
HOCMN ($L = 3$)	82.99	61.30	64.38	70.75
HOCMN ($u_1 = 6.0$)	84.87	65.68	65.24	73.25

German	WM	MM	HM	Avg
baseline	91.00	79.87	75.95	83.34
CMVN ($l = 200$)	90.90	80.75	80.43	84.73
HEQ ($l = 240$)	92.31	81.48	82.38	86.04
HOCMN ($L = 3$)	92.31	85.80	85.43	88.31
HOCMN ($u_1 = 5.5$)	93.07	85.21	85.48	88.42

Spanish	WM	MM	HM	Avg
baseline	86.94	73.52	37.53	69.89
CMVN ($l = 200$)	92.02	86.64	80.90	87.36
HEQ ($l = 240$)	92.09	83.31	81.41	86.35
HOCMN ($L = 3$)	93.52	89.94	82.71	89.56
HOCMN ($u_1 = 4.5$)	93.38	90.69	84.27	90.16

Finnish	WM	MM	HM	Avg
baseline	93.82	72.16	42.08	73.30
CMVN ($l = 200$)	93.19	85.98	72.01	85.37
HEQ ($l = 200$)	90.44	82.69	76.18	84.16
HOCMN ($L = 3$)	94.05	87.48	80.88	88.46
HOCMN ($u_1 = 3.0$)	94.05	87.48	80.88	88.46