Supplementary Material for the Paper: Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection

Xiao-Lei Zhang, *Member, IEEE* and DeLiang Wang, *Fellow, IEEE*

***Abstract*—In this supplementary material, we first give an efficient way of calculating the area under the ROC curve (AUC). Then, we report results in terms of hit rate minus false alarm rate (HIT−FA). HIT−FA evaluates the optimal operating point on a ROC curve. The higher the HIT−FA is, the better the performance is. Furthermore, we show 10 randomly selected AURORA4 utterances, and their manual and automatic voice activity labels. Finally, we provide results with a fixed decision threshold.**

***Index Terms*—Boosting, cochleagram, deep neural network, multi-resolution stacking, voice activity detection.**

APPENDIX

A. AUC Calculation

Algorithm 1 AUC calculation.

Input: Number of training data points n, manual label vector $\mathbf{y} = [y_1, \ldots, y_n]^T$, and predicted soft values $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_n]^T$

Initialization: $a = 0$, $b = 0$, $swapped\_pairs = 0$

Output: AUC A

```
Sort ŷ in descending order, denoted as ŷ*, and reorder y along with ŷ, denoted as y*
for i = 1, ..., n do
    if y_i^* > 0 then
        swapped_pairs ← swapped_pairs + b
        a ← a + 1
    else
        b ← b + 1
    end if
end for
A = 1 − swapped_pairs / (ab)
```

B. Supplementary Results of Noise-dependent Models

1) Main Results: Table I lists the HIT−FA results of 4 machine-learning-based VADs with noise-dependent training, including SVM VAD, Zhang13 VAD, bDNN-based VAD and MRS-based VAD, on 42 noisy environments of AURORA2. Table II lists the HIT−FA results on 8 noisy environments of AURORA4. See the main paper for the analysis of the experimental phenomena.

Xiao-Lei Zhang and DeLiang Wang are with the Department of Computer Science & Engineering and Center for Cognitive & Brain Sciences, The Ohio State University, Columbus, OH, USA (e-mail: xiaolei.zhang9@gmail.com, dwang@cse.ohio-state.edu).

TABLE I
HIT−FA (%) COMPARISON BETWEEN COMPARISON VADS AND PROPOSED BDNN- AND MRS-BASED VADS ON THE AURORA2 CORPUS. THE NUMBERS IN BOLD INDICATE THE BEST RESULTS.

Noise	SNR	SVM	Zhang13	bDNN	MRS
Babble	−5 dB	30.60	33.18	45.14	**49.71**
	0 dB	46.63	51.59	60.96	**64.11**
	5 dB	61.63	64.52	70.20	**71.86**
	10 dB	68.67	72.81	73.31	**74.11**
	15 dB	72.60	76.37	76.36	**76.99**
	20 dB	74.47	77.49	76.42	**77.56**
Car	−5 dB	50.61	51.93	66.78	**69.39**
	0 dB	64.62	68.96	75.22	**77.17**
	5 dB	72.87	73.07	76.66	**78.90**
	10 dB	75.52	76.84	78.86	**80.65**
	15 dB	78.27	78.75	80.62	**82.23**
	20 dB	80.23	81.80	81.65	**83.29**
Restaurant	−5 dB	32.49	35.91	46.13	**50.22**
	0 dB	48.74	48.20	60.72	**64.79**
	5 dB	64.11	66.81	72.42	**74.86**
	10 dB	69.59	72.87	76.72	**78.34**
	15 dB	73.81	76.69	78.28	**80.75**
	20 dB	77.40	78.69	80.55	**82.04**
Street	−5 dB	34.34	35.81	55.42	**57.35**
	0 dB	47.78	47.59	63.19	**65.86**
	5 dB	61.21	64.52	70.67	**72.05**
	10 dB	67.36	69.55	73.91	**75.20**
	15 dB	71.13	74.27	74.14	**75.80**
	20 dB	74.17	76.99	76.33	**76.94**
Airport	−5 dB	35.29	39.49	49.14	**53.01**
	0 dB	50.25	52.62	66.42	**68.59**
	5 dB	62.99	63.92	73.90	**75.89**
	10 dB	70.05	73.78	77.79	**79.32**
	15 dB	76.01	78.18	79.78	**80.92**
	20 dB	77.35	80.45	81.71	**82.95**
Train	−5 dB	37.11	39.96	54.93	**57.34**
	0 dB	52.51	55.05	65.20	**68.99**
	5 dB	65.29	67.10	72.45	**75.40**
	10 dB	70.55	72.75	74.60	**76.90**
	15 dB	75.41	76.66	76.57	**79.62**
	20 dB	75.90	76.67	77.96	**80.46**
Subway	−5 dB	52.82	54.16	69.35	**71.42**
	0 dB	64.11	66.75	74.26	**75.37**
	5 dB	71.37	74.11	77.63	**80.06**
	10 dB	75.27	76.81	78.95	**80.68**
	15 dB	75.92	77.96	79.34	**81.51**
	20 dB	77.21	79.40	80.32	**82.21**

TABLE II
HIT−FA (%) COMPARISON BETWEEN THE COMPARISON VADS AND PROPOSED BDNN-BASED VAD ON THE AURORA4 CORPUS. THE NUMBERS IN BOLD INDICATE THE BEST RESULTS.

Noise	SNR	SVM	Zhang13	bDNN	MRS
Babble	−5 dB	45.69	48.33	56.13	57.92
	0 dB	56.31	60.01	63.94	65.15
	5 dB	67.77	69.94	72.04	73.10
	10 dB	69.75	74.75	75.74	76.03
Factory	−5 dB	42.11	47.42	54.60	55.51
	0 dB	56.93	62.00	66.64	67.18
	5 dB	64.19	70.72	72.25	73.18
	10 dB	73.36	75.66	75.86	76.44

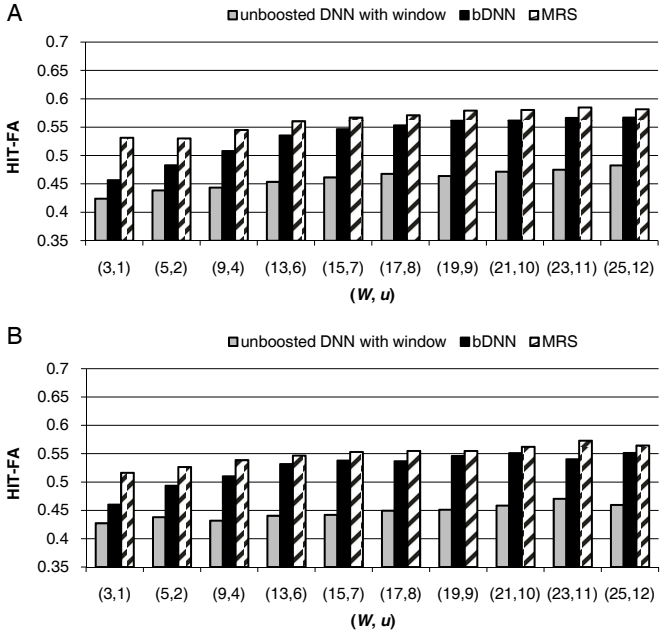

Fig. 1. HIT−FA analysis of the advantage of the boosted algorithm in bDNN-based and MRS-based VADs over the unboosted counterpart that uses the same input $\mathbf{x}'_n$ as bDNN and MRS but the original output y_n as the training target instead of $\mathbf{y}'_n$. (A) Comparison in the babble noise environment with SNR = -5 dB. (B) Comparison in the factory noise environment with SNR = -5 dB. Note that (W, u) are two window parameters of bDNN.

2) Effects of Boosted DNN and MRS on the Performance: To investigate how bDNN and MRS improve the performance, we ran DNN, bDNN, and MRS with MRCG as the input feature on AURORA4, where the model "DNN" used the same input as bDNN, i.e. $[\mathbf{x}_{m-W}^T, \ldots, \mathbf{x}_m^T, \ldots, \mathbf{x}_{m+W}^T]^T$, but y_m as the target instead of $[y_{m-W}, \ldots, y_m, \ldots, y_{m+W}]^T$.

Fig. 1 shows the HIT−FA result with respect to the window length. The result is consistent with the AUC result in the main paper.

3) Effects of the MRCG Feature on the Performance: To evaluate how the MRCG feature affects the performance, we compared it with the combination (COMB) of 10 conventional acoustic features in Zhang13 VAD [1], on AURORA4 with DNN, bDNN, or MRS as the classifier.

Table III lists the HIT−FA result between MRCG and COMB. The result is consistent with the AUC result in the main paper.

TABLE III
HIT−FA (%) ANALYSIS OF THE RELATIVE CONTRIBUTIONS OF BDNN, MRS, AND MRCG. "COMB" DENOTES A COMBINATION OF 10 ACOUSTIC FEATURES IN [1].

Noise	SNR	DNN		bDNN		MRS	
		COMB	MRCG	COMB	MRCG	COMB	MRCG
Babble	−5 dB	49.15	46.14	53.58	56.13	55.75	57.92
	0 dB	57.60	55.79	61.81	63.94	63.44	65.15
	5 dB	66.40	65.06	70.58	72.04	72.47	73.10
	10 dB	71.83	70.28	74.22	75.74	75.21	76.03
Factory	−5 dB	45.82	42.32	51.30	54.60	55.29	55.51
	0 dB	54.54	55.82	62.36	66.64	65.64	67.18
	5 dB	64.46	63.88	70.50	72.25	72.45	73.18
	10 dB	69.17	68.21	73.83	75.86	76.09	76.44

C. Supplementary Results of Noise-independent Models

Table IV lists the HIT−FA comparison of the NI and ND models of the DNN-, bDNN-, and MRS-based VADs, and 3 referenced VADs on AURORA2. Table V lists the comparison on AURORA4. The experimental conclusion is consistent with that in the main paper.

D. Visualizing 10 Randomly Selected Utterances of AURORA4

Ten randomly selected utterances of AURORA4 with their automatic labels and manual labels are shown in Figs. 2 and 3.

E. Results with Fixed Decision Threshold

Table VI lists the HIT−FA comparison between the NI models with the decision thresholds fixed to 0.8 (i.e. Thres=0.8) and the NI models with the optimal decision thresholds (i.e. Thres=*Opt*) for AURORA2. Table VII lists the comparison for AURORA4. The FA rates are also listed in the tables. The experimental results show that the decision threshold of the proposed method is not sensitive to the test environments.

REFERENCES

[1] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.

TABLE IV
HIT−FA (%) COMPARISON BETWEEN THE NOISE-INDEPENDENT (NI) MODELS, NOISE-DEPENDENT (ND) MODELS, AND 3 REFERENCED VADS AT AURORA2.

Noise	**SNR**	Sohn	Ramirez05	Ying	DNN		bDNN		MRS	
					NI	ND	NI	ND	NI	ND
Babble	−5 dB	15.61	17.14	13.91	33.71	41.48	39.48	45.14	41.25	49.71
	0 dB	27.18	30.24	23.38	56.58	55.22	62.71	60.96	64.70	64.11
	5 dB	45.88	50.29	40.79	70.40	68.97	73.69	70.20	74.03	71.86
	10 dB	59.93	63.23	55.34	74.77	72.06	77.44	73.31	77.21	74.11
Car	−5 dB	14.22	15.96	17.96	56.98	62.49	61.74	66.78	63.02	69.39
	0 dB	29.83	34.58	30.38	70.63	71.68	74.06	75.22	74.08	77.17
	5 dB	47.02	51.23	43.82	75.67	74.48	78.66	76.66	78.82	78.90
	10 dB	61.20	64.88	55.17	77.40	76.71	80.01	78.86	79.65	80.65
Restaurant	−5 dB	8.46	8.39	10.89	31.43	41.41	34.76	46.13	37.91	50.22
	0 dB	23.81	22.19	18.23	51.48	54.33	57.51	60.72	59.90	64.79
	5 dB	32.94	36.01	32.87	68.21	67.93	73.30	72.42	74.35	74.86
	10 dB	47.39	49.82	44.92	75.16	73.45	78.18	76.72	78.49	78.34
Street	−5 dB	5.50	8.30	9.42	44.13	51.05	47.41	55.42	48.99	57.35
	0 dB	16.28	16.39	16.19	62.95	60.96	66.21	63.19	66.34	65.86
	5 dB	28.96	34.05	30.96	71.30	68.87	73.95	70.67	73.49	72.05
	10 dB	40.20	43.62	39.92	74.86	72.83	77.85	73.91	76.81	75.20
Airport	−5 dB	8.80	15.35	13.83	41.37	44.61	45.51	49.14	48.92	53.01
	0 dB	22.93	25.86	26.02	60.63	61.69	64.61	66.42	64.75	68.59
	5 dB	36.73	42.92	37.73	71.32	70.00	74.46	73.90	74.83	75.89
	10 dB	52.96	60.35	54.72	76.08	75.00	79.29	77.79	79.10	79.32
Train	−5 dB	9.75	10.89	15.98	45.77	50.31	51.00	54.93	53.26	57.34
	0 dB	16.50	19.31	27.07	62.81	60.95	66.54	65.20	67.59	68.99
	5 dB	36.20	45.08	41.79	72.77	68.84	76.77	72.45	76.73	75.40
	10 dB	54.10	58.47	50.87	74.58	72.07	78.08	74.60	77.63	76.90
Subway	−5 dB	8.23	9.69	11.19	33.29	65.09	38.14	69.35	40.92	71.42
	0 dB	19.69	18.99	16.85	55.85	71.52	60.47	74.26	60.94	75.37
	5 dB	30.66	41.90	25.76	69.61	75.09	72.91	77.63	73.40	80.06
	10 dB	45.53	49.98	39.32	76.03	75.93	78.64	78.95	79.25	80.68

TABLE V
HIT−FA (%) COMPARISON BETWEEN THE NOISE-INDEPENDENT (NI) MODELS, NOISE-DEPENDENT (ND) MODELS, AND 3 REFERENCED VADS AT AURORA4.

Noise	**SNR**	Sohn	Ramirez05	Ying	DNN		bDNN		MRS	
					NI	ND	NI	ND	NI	ND
Babble	−5 dB	29.44	38.45	21.03	42.44	46.14	47.14	56.13	51.51	57.92
	0 dB	40.64	52.09	29.76	51.04	55.79	55.36	63.94	58.81	65.15
	5 dB	54.42	65.23	42.70	62.10	65.06	64.75	72.04	66.24	73.10
	10 dB	67.50	70.89	56.12	69.26	70.28	71.90	75.74	72.48	76.03
Factory	−5 dB	12.00	13.43	19.50	42.41	42.32	48.13	54.60	51.77	55.51
	0 dB	21.04	25.63	28.42	51.83	55.82	56.61	66.64	59.88	67.18
	5 dB	33.40	40.11	38.83	63.23	63.88	66.68	72.25	67.11	73.18
	10 dB	47.33	55.39	50.47	69.81	68.21	73.32	75.86	73.40	76.44

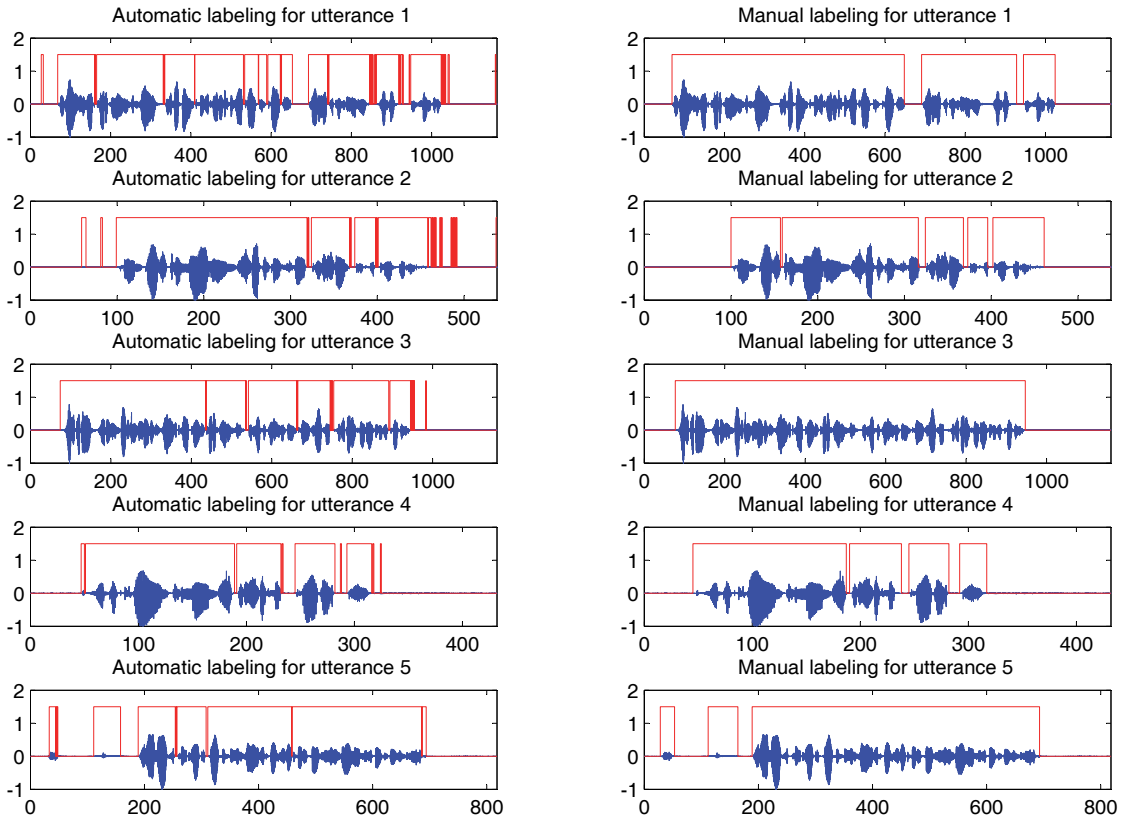

Fig. 2. Comparison between automatic labelling (by Sohn VAD) and manual labelling for the first 5 utterances of 10 randomly selected clean utterances from AURORA4.

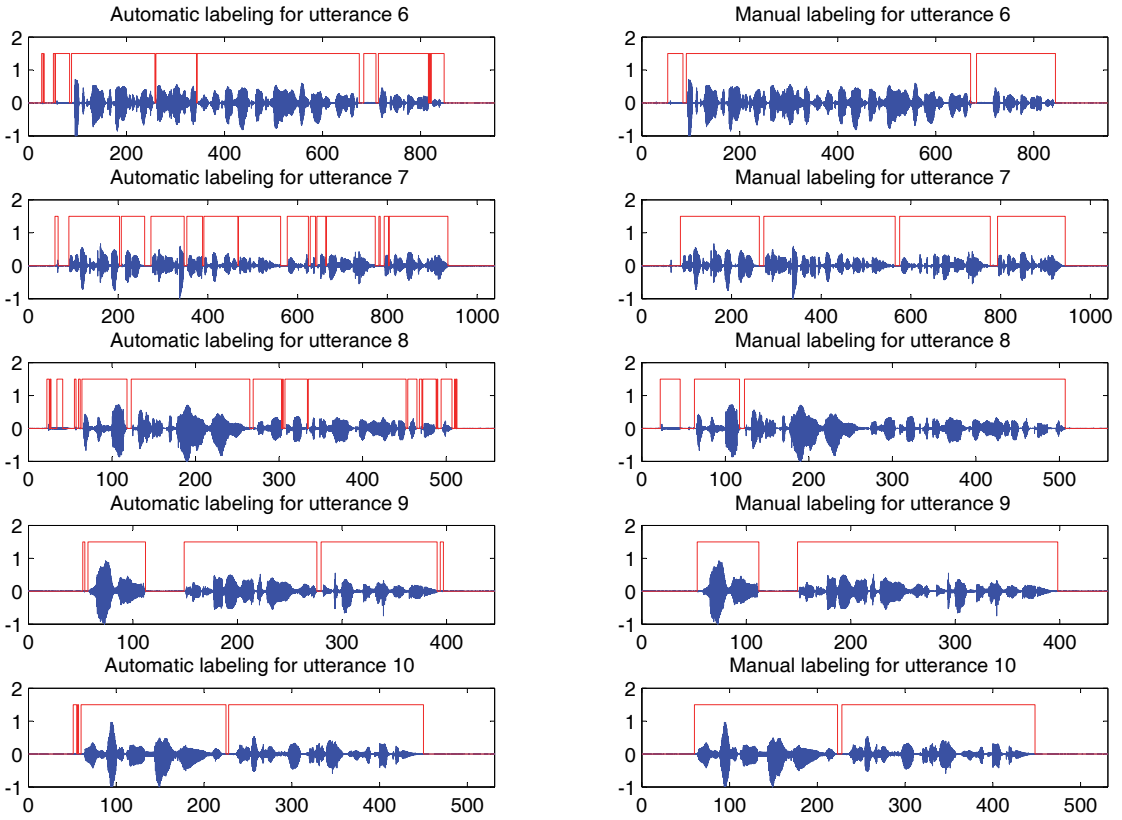

Fig. 3. Comparison between automatic labelling (by Sohn VAD) and manual labelling for the last 5 utterances of 10 randomly selected clean utterances from AURORA4.

TABLE VI

HIT−FA (%) COMPARISON BETWEEN THE NI MODELS WITH THE DECISION THRESHOLDS FIXED TO 0.8 (I.E. THRES=0.8) AND THE NI MODELS WITH THE OPTIMAL DECISION THRESHOLDS (I.E. THRES=*Opt*) FOR AURORA2. THE NUMBERS IN PARENTHESES ARE FA RATES.

Noise	**SNR**	DNN		bDNN		MRS	
		Thres=0.8	Thres=*Opt*	Thres=0.8	Thres=*Opt*	Thres=0.8	Thres=*Opt*
Babble	−5 dB	33.37 (28.86)	33.71	37.80 (19.27)	39.48	41.07 (32.35)	41.25
	0 dB	55.95 (26.04)	56.58	62.60 (17.79)	62.71	62.87 (25.73)	64.70
	5 dB	68.70 (21.97)	70.40	73.37 (16.51)	73.69	72.18 (21.66)	74.03
	10 dB	73.80 (19.31)	74.77	76.95 (15.75)	77.44	75.54 (19.66)	77.21
Car	−5 dB	55.92 (16.39)	56.98	58.77 (10.52)	61.74	62.99 (20.25)	63.02
	0 dB	70.36 (18.74)	70.63	74.04 (14.84)	74.06	72.72 (20.85)	74.08
	5 dB	74.62 (18.73)	75.67	78.37 (14.88)	78.66	76.26 (19.11)	78.82
	10 dB	76.06 (18.93)	77.40	79.04 (15.45)	80.01	76.95 (19.20)	79.65
Restaurant	−5 dB	30.98 (34.85)	31.43	33.46 (27.73)	34.76	37.63 (37.47)	37.91
	0 dB	50.79 (29.28)	51.48	57.35 (20.32)	57.51	57.70 (28.07)	59.90
	5 dB	67.23 (19.76)	68.21	73.27 (12.40)	73.30	73.19 (17.05)	74.35
	10 dB	74.63 (14.27)	75.16	78.03 (9.71)	78.18	78.08 (12.83)	78.49
Street	−5 dB	44.05 (27.94)	44.13	46.97 (20.70)	47.41	48.37 (31.02)	48.99
	0 dB	61.42 (25.68)	62.95	65.83 (19.82)	66.21	64.34 (27.06)	66.34
	5 dB	69.39 (22.45)	71.30	72.64 (18.30)	73.95	70.40 (23.67)	73.49
	10 dB	73.07 (20.25)	74.86	76.88 (15.75)	77.85	74.95 (20.21)	76.81
Airport	−5 dB	41.33 (31.64)	41.37	44.93 (23.44)	45.51	47.50 (33.75)	48.92
	0 dB	58.45 (29.13)	60.63	63.83 (22.03)	64.61	61.91 (29.65)	64.75
	5 dB	69.43 (22.77)	71.32	74.13 (17.07)	74.46	72.62 (21.77)	74.83
	10 dB	74.39 (19.96)	76.08	78.56 (15.25)	79.29	76.39 (19.30)	79.10
Train	−5 dB	45.44 (27.62)	45.77	50.74 (19.17)	51.00	52.44 (29.75)	53.26
	0 dB	60.47 (26.45)	62.81	66.27 (19.53)	66.54	65.16 (26.08)	67.59
	5 dB	71.08 (20.04)	72.77	76.31 (14.27)	76.77	74.78 (18.64)	76.73
	10 dB	72.83 (19.32)	74.58	77.28 (14.51)	78.08	75.61 (18.43)	77.63
Subway	−5 dB	33.20 (27.10)	33.29	38.09 (22.79)	38.14	40.62 (32.54)	40.92
	0 dB	54.51 (23.00)	55.85	59.85 (18.08)	60.47	59.00 (25.23)	60.94
	5 dB	68.85 (14.17)	69.61	72.89 (9.39)	72.91	72.91 (13.94)	73.40
	10 dB	75.95 (8.58)	76.03	78.10 (5.27)	78.64	78.99 (7.61)	79.25

TABLE VII

HIT−FA (%) COMPARISON BETWEEN THE NI MODELS WITH DECISION THRESHOLDS FIXED TO 0.8 (I.E. THRES=0.8) AND THE NI MODELS WITH OPTIMAL DECISION THRESHOLDS (I.E. THRES=*Opt*) FOR AURORA4. THE NUMBERS IN PARENTHESES ARE FA RATES.

Noise	**SNR**	DNN		bDNN		MRS	
		Thres=0.8	Thres=*Opt*	Thres=0.8	Thres=*Opt*	Thres=0.8	Thres=*Opt*
Babble	−5 dB	41.22 (45.59)	42.44	46.66 (39.18)	47.14	51.23 (37.35)	51.51
	0 dB	48.30 (37.81)	51.04	54.06 (30.81)	55.36	58.29 (29.52)	58.81
	5 dB	59.71 (24.48)	62.10	64.41 (18.05)	64.75	66.18 (19.85)	66.24
	10 dB	68.67 (13.95)	69.26	71.66 (8.54)	71.90	71.65 (10.95)	72.48
Factory	−5 dB	42.35 (39.24)	42.41	48.12 (34.38)	48.13	51.44 (35.70)	51.77
	0 dB	51.02 (31.96)	51.83	56.32 (27.15)	56.61	59.66 (27.12)	59.88
	5 dB	62.61 (19.47)	63.23	66.59 (15.24)	66.68	66.93 (16.88)	67.11
	10 dB	69.78 (11.74)	69.81	72.57 (7.59)	73.32	72.42 (10.59)	73.40