TR－A－0075

HMM Speech Recognition using DFT and Auditory Spectrograms (Part 2)

DFTと聴覚スペクトログラムを用いた
HMM音声認識（PART 2)

Tatsuya HIRAHARA

平原 達也

1990. 2. 21

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷 ☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

1. Introduction

In our previous report (Patterson and Hirahara, 1989, ATR Technical Report TR-A-0063), we showed some results for HMM /b,d,g/ phoneme recognition using DFT and SAS, auditory spectrograms. As it was a very preliminary report, many tests remained:

1. recognition tests with added independent pink noise.
2. recognition tests with different S/N ratios.
3. recognition tests with a larger phoneme set.
4. recognition tests training HMM on both clean and noisy tokens.
5. recognition tests with different reference vector sizes.
6. optimization of the auditory model parameters.

In this report, we will focus on the first three problems. The 4th and 5th are the problems concerning the HMM phoneme recognition system itself. As they are topics which are interesting from a practical viewpoint, we will leave them for the speech recognition people. The 6th is a fundamental problem. However, all experiments should be repeated when the auditory model is returned. Therefore, we decided not to do it at the moment.

2. Method

Experiment methods and procedures are the same as those described in the previous report.

Speech samples are CV-syllables, including /b/, /d/ or /g/, extracted from Japanese words which were uttered in isolation by a male speaker. Each category of /b/, /d/ and /g/ includes 180 to 260 tokens for training and test sets, respectively. Input vectors for HMM are reduced 16 by 15 spectra based on DFT and SAS. K-means clustering was used to make a codebook, and an HMM with four states and six transitions was used. The recognition task is to classify input tokens into /b/, /d/ or /g/ regardless of the following vowel.

3. Results for independent pink noise

In the frozen noise condition, only one pink noise data set, which was previously sampled and stored on the disk, was added to each token. On the other hand, in the random noise (independent noise) condition, a pink noise generator (B&K 1049) output was sampled (12kHz, 16bit) just prior to the analysis for each token. The level of the pink noise was fixed so that the overall S/N ratio was approximately 0dB. In the HMM recognition system, the reference vector size was 16 channels by 7 frames.

The results for the DFT system under frozen and random noise conditions are shown in Figures 1(a) and 1(b), respectively. The results for the SAS system under frozen and random noise conditions are shown in Figures 2(a) and 2(b), respectively. In each figures, the abscissa represents the codebook size and the ordinate represents the total performance, the percent correct identification of the individual categories for a particular combination of training and test conditions. Outlined and solid symbols indicate results for the open and closed recognition experiments, respectively.

First, when the DFT system is trained on clean speech and tested on noisy speech, there is no significant difference in the closed experiment noise condition performance curves. In the open experiments, on the other hand, the performance improvement compared to the code book size increase in the random noise condition is smaller than that in the frozen noise condition. Second, when the DFT system is trained and tested on noisy speech, the performance saturates around 80% for the large code book in the open experiment under the random noise condition. Third, when the DFT system is trained on noisy speech and tested on clean speech, performance in the frozen noise condition is 5 to 10% better than in the random noise condition for the small code book. However, the difference for the largest code book (= 85) is not significant.

The noise condition affects the performance in the SAS system, as well as in the DFT system. One exception is that the performance in the random noise condition is a few percent better than that in the frozen noise condition for the small codebook, when the SAS system is trained on noisy speech and tested on clean speech.

For open experiments, when the system is tested on noisy speech and the codebook is large (=85), the performance of the two front ends in the random noise condition is worse than that in the frozen noise condition. The performance drops about 9% when the system is trained on clean speech (84.7% --> 75.3% [DFT], 81.6% --> 73.1% [SAS]) and about 7% when the system is trained on noisy speech (87.4% --> 79.9% [DFT], 84.1% --> 78.1% [SAS]). It should be noted that such performances under random noise conditions when the S/N = 0dB is nearly equal to that under frozen noise conditions when the S/N = -6dB. Namely, adding independent pink noise to each token (random noise condition) is equivalent to decreasing the frozen noise condition S/N ratio -6dB.

The advantage of the SAS system compared with the DFT system for noisy speech is observed where the codebook size is small (5 or 10). This is the same result observed in the frozen noise condition when the S/N = -6dB. When the codebook is small, the HMM learns noisy speech features better in the SAS system than in the DFT system.

4. Result for several S/N ratio conditions

Recognition experiments were repeated in the frozen noise condition with three overall S/N ratios, +6dB, 0dB and -6dB. The S/N ratio is infinite when clean speech is used. In the HMM recognition system, the reference vector size was 16 channels by 7 frames.

The results for the DFT system when the S/N = -6dB and +6dB are shown in Figures 3(a) and 3(b), respectively. The results for the SAS system when the S/N = -6dB and +6dB are shown in figures 4(a) and 4(b), respectively. The results for the DFT and SAS system when the S/N = 0dB are shown in Figures 1(a) and 2(a) in section 3. Figures 5(a), 5(b) and 5(c) are reproduced from Figures 1(a), 2(a), 3(a)(b) and 4(a)(b) to show the performance curves versus S/N ratio in open experiments. Figure 5(a) shows the performance vs. S/N ratio when the system was trained on clean speech and tested on noisy speech. Figure 5(b) shows the result when the system was trained on noisy speech and tested on clean speech. Figure 5(c) shows the result when the system was trained and tested on noisy speech. In these figures, the abscissa represents the overall S/N ratio and the ordinate represents the total performance of open experiment. Squares with a dashed line and circles with a solid line indicate the performance of the DFT system and the SAS system, respectively.

As seen in Figure 5(a), when the system is trained on clean speech, performance improves as the S/N ratio increases. It is also clear that the SAS system gives better performance than the DFT system only when the codebook size is less than 20. Figure 5(b) and (c) show that a system trained on noisy speech gives better performance than the one trained on clean speech. It is amazing that the performance dropped only 5% when system was trained on noisy speech (S/N = -6dB) and tested on clean speech with codebook size = 85, while it dropped 20% when trained on clean speech and tested on noisy speech.

5. Result for a larger phoneme set

In this section, the performance of the DFT system and the SAS system is compared with a larger phoneme set rather than the small /b,d,g/ phoneme set.

Speech data used in the experiments are CV-syllables extracted from a large database of 5,240 common Japanese words, which were uttered in isolation by a native male Japanese speaker (MAU). All CV-syllables including /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/, /h/, /z/, /ch/, /r/, /w/ and /y/ were extracted using manually selected acoustic-phonetic labels provided with the database. Table 1 shows the number of tokens for each category in a training set and in a test set.

Token	Training set	Test set
b	218	227
d	202	179
g	260	252
p	32	15
t	425	440
k	1152	1164
m	471	481
n	260	265
N	503	488
s	475	538
sh	186	177
h	214	207
z	115	115
ch	79	71
ts	212	177
r	754	722
w	71	81
y	159	174
TOTAL	6940	6937

Table 1 Number of tokens

Only one frozen noise condition (PN-12) was used in the experiment. The S/N ratio was about 6 dB for the /b/, /d/ and /g/ tokens. The S/N ratio of other tokens were not measured precisely. By looking at wave forms of several tokens, it can be seen that the S/N ratio for the voiceless stops, in particular /p/, is lower than 6dB.

In the HMM system, 20 codebooks are chosen for each category. Since the number of tokens included in each category is different, the ratio of the codebook size to the number of tokens ratio varies considerably among the categories. It is about 1 for /p/ and 1/500 for /k/.

The reference vector size was fixed at 16 channels by 7 frames.

The total performance of the DFT and the SAS system for each training and testing condition are compared in Figure 6. The confusion matrices for all DFT results and SAS results are shown in Table 2 and Table 3, respectively.

From these Figure and Tables, it is hard to find any SAS system advantage. SAS system performance is 5 to 20% poorer than that of the DFT system. In particular, performance differences are large when HMM systems are trained on noisy speech and tested on clean speech. These results conflict with the results for the /b, d, g/ task, in which SAS gives better performance than DFT when 20 codebooks are prepared for each category.

6. Conclusion

(1) Differences between the DFT and SAS systems were small under random noise conditions as well as frozen noise conditions.

(2) Under the random noise condition, both DFT and SAS system performance improvement compared to the codebook size increase are small when the HMM is trained on noisy tokens.

(3) The SAS system's advantage was not observed even under lower S/N ratio conditions. When the codebook size was 85 per category, the DFT system gave better performances under any S/N ratio and any training/testing conditions. However, when the codebook size was 5 per category, the SAS system always outperformed the DFT system.

(4) The difference between the DFT and SAS systems become clear in an eighteen-phoneme set experiment. The DFT system gave 5 to 20% better performance than the SAS system under any training/testing conditions.

These conclusions suggest that the 16X15 DFT representation is more suitable for the HMM pattern classifier than 16X15 SAS representation. We find this result to be very odd. We expected the SAS system outperform the DFT system, because the SAS spectrum seemed to grasp phoneme cues, such as formant structure, better than DFT.

One possible reason is that the SAS systems we used were not optimized. Another possible reason is that the SAS representations are reasonable for our spectrum reading knowledge but are NOT reasonable for HMM to classify input tokens. This may be because, the SAS representations were formed either by emphasizing certain components or eliminating certain components based on psychophysical and physiological knowledge. This elimination might lose some important information for HMM. On the other hand, the DFT system does not emphasize any component and preserves all power spectrum information.

Further, it is clear that HMM is not a model of the way human beings classify patterns.

It is a hypothesis that a well designed auditory front end overcomes the traditional DFT and/or LPC based frontend. However, this report failed to prove this hypothesis.

What went wrong?

(a) *Was it because the SAS is not a well-designed auditory model?*
The recognition tests should be repeated with a well-tuned SAS front end and other auditory front ends.

(b) *Was it because the affinity between the SAS frontend and HMM is weak?*
It is difficult to say at present. The recognition tests should be repeated with other pattern classifiers, such as DTW, LVQ, NN and/or feature based pattern classifiers.

(c) *Is that because the hypothesis is wrong?*
I still believe the hypothesis is true.

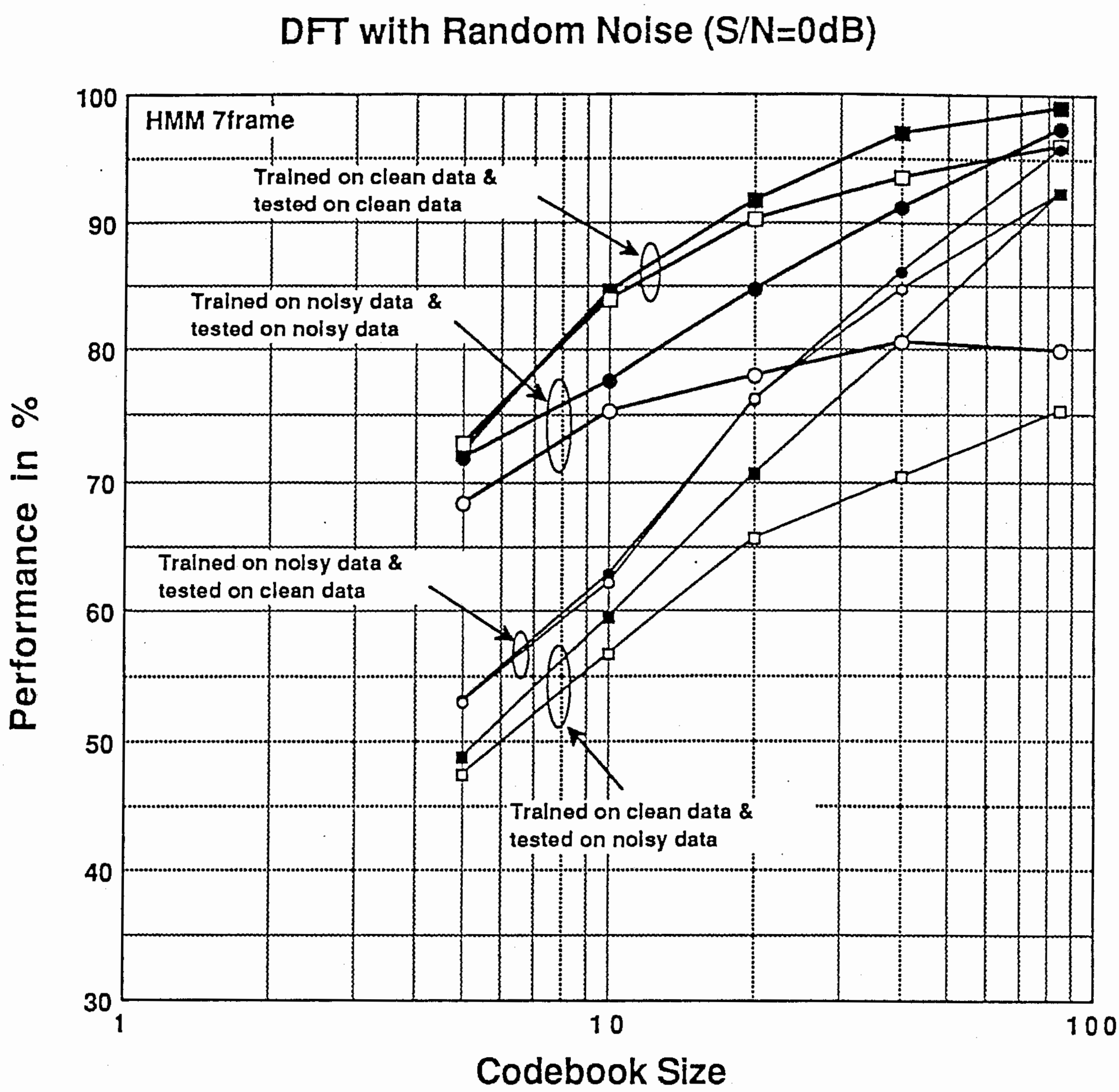

Figure 1(b) The /b, d, g/ phoneme recognition results for the DFT system under random noise conditions. Outlined and solid symbols indicate results for the open and closed recognition experiments.

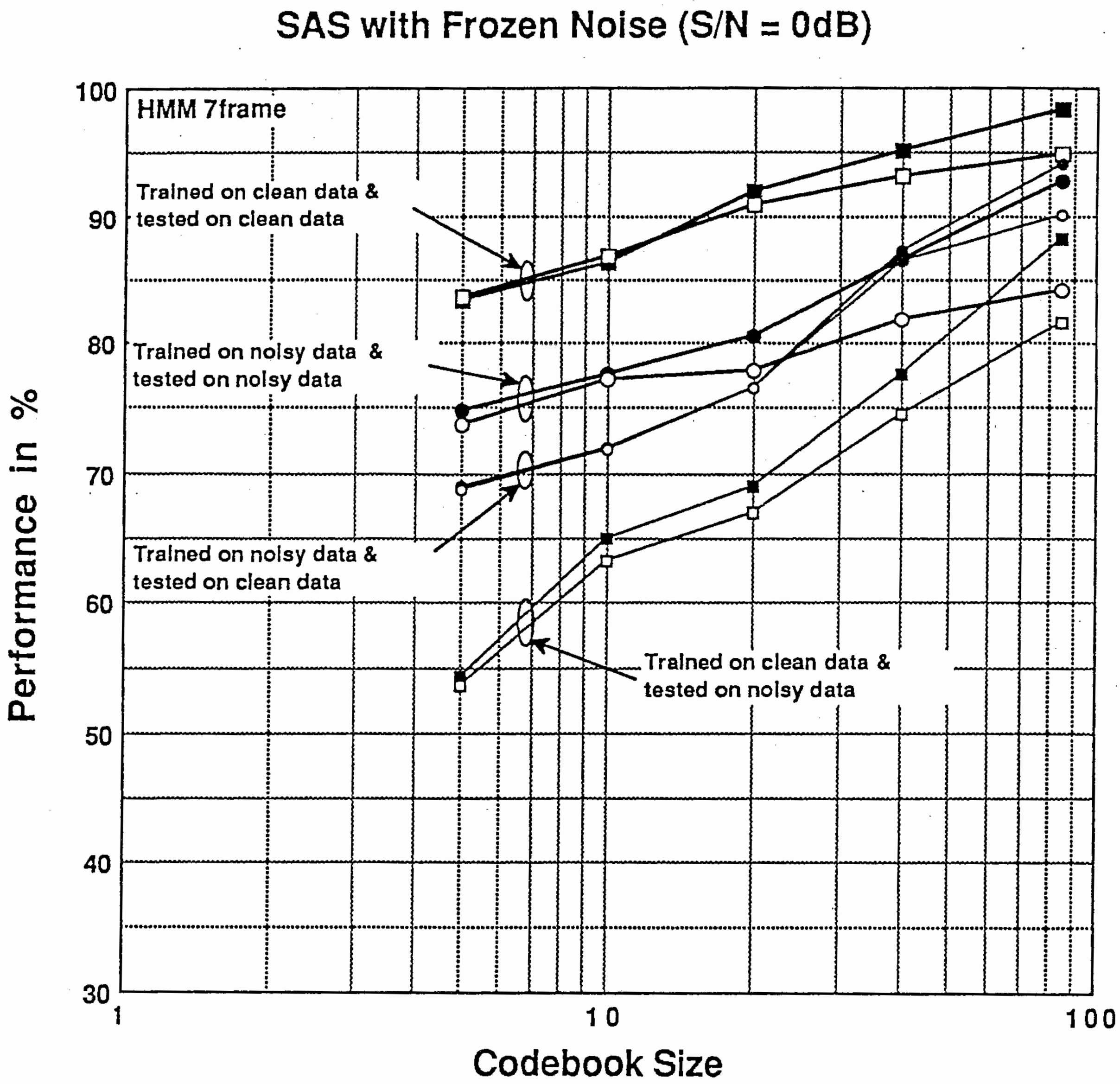

Figure 2(a) The /b, d, g/ phoneme recognition results for the SAS system under frozen noise condition. Outlined and solid symbols indicate results for the open and closed recognition experiments.

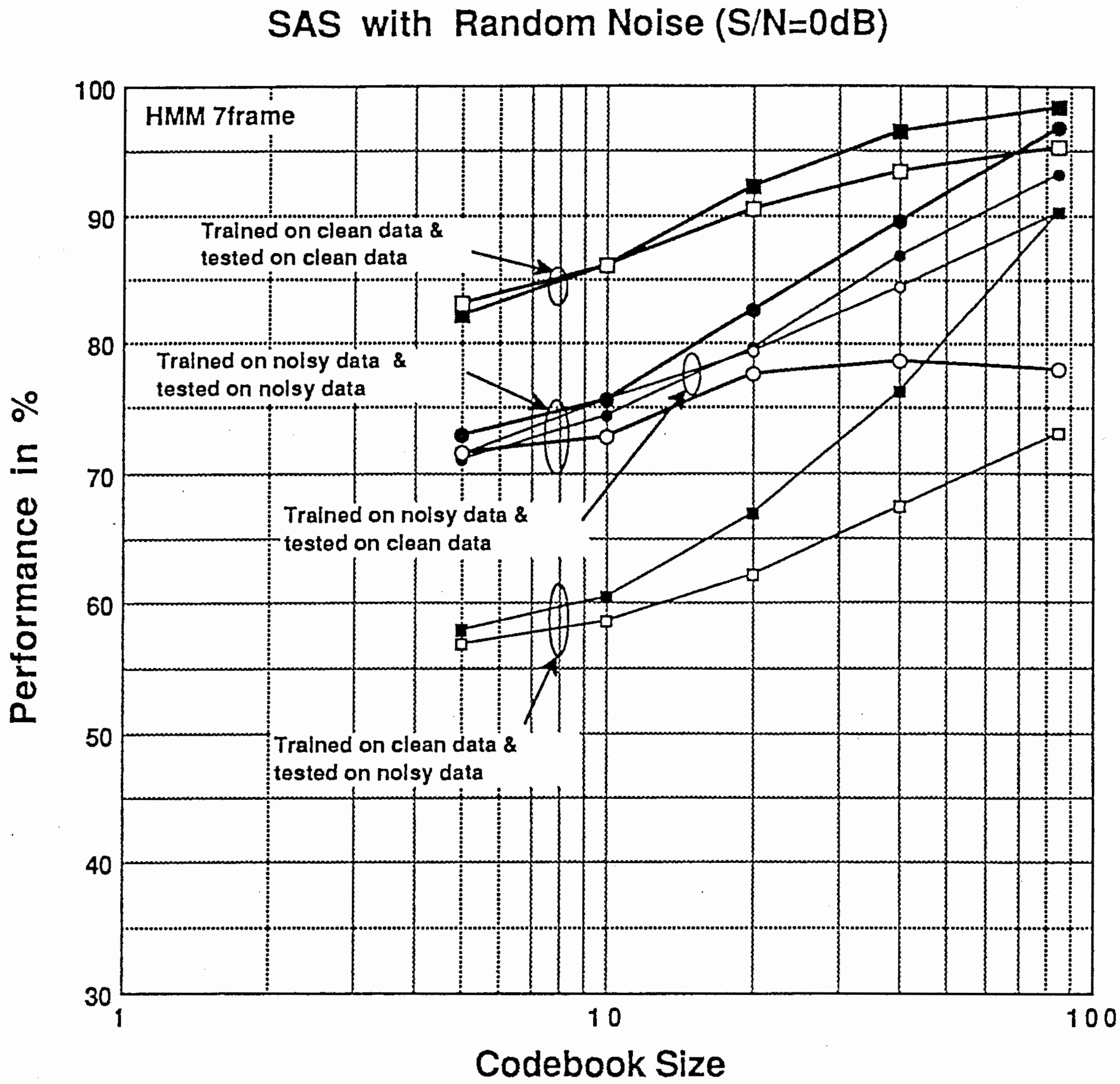

Figure 2(b) The /b, d, g/ phoneme recognition results for the SAS system under random noise conditions. Outlined and solid symbols indicate results for the open and closed recognition experiments.

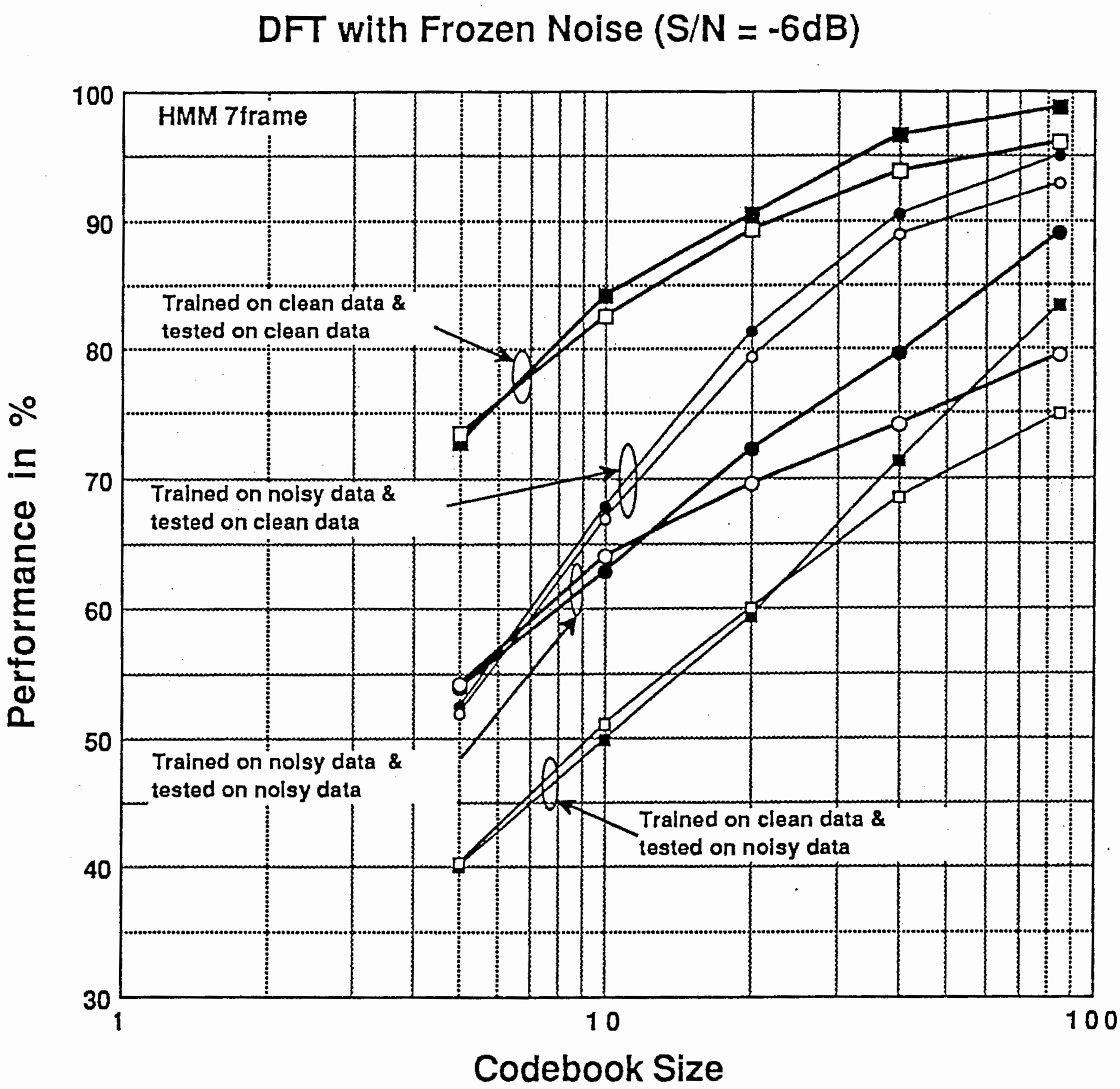

Figure 3(a) The /b, d, g/ phoneme recognition results for the DFT system under frozen noise conditions where the S/N=-6dB. Outlined and solid symbols indicate results for the open and closed recognition experiments.

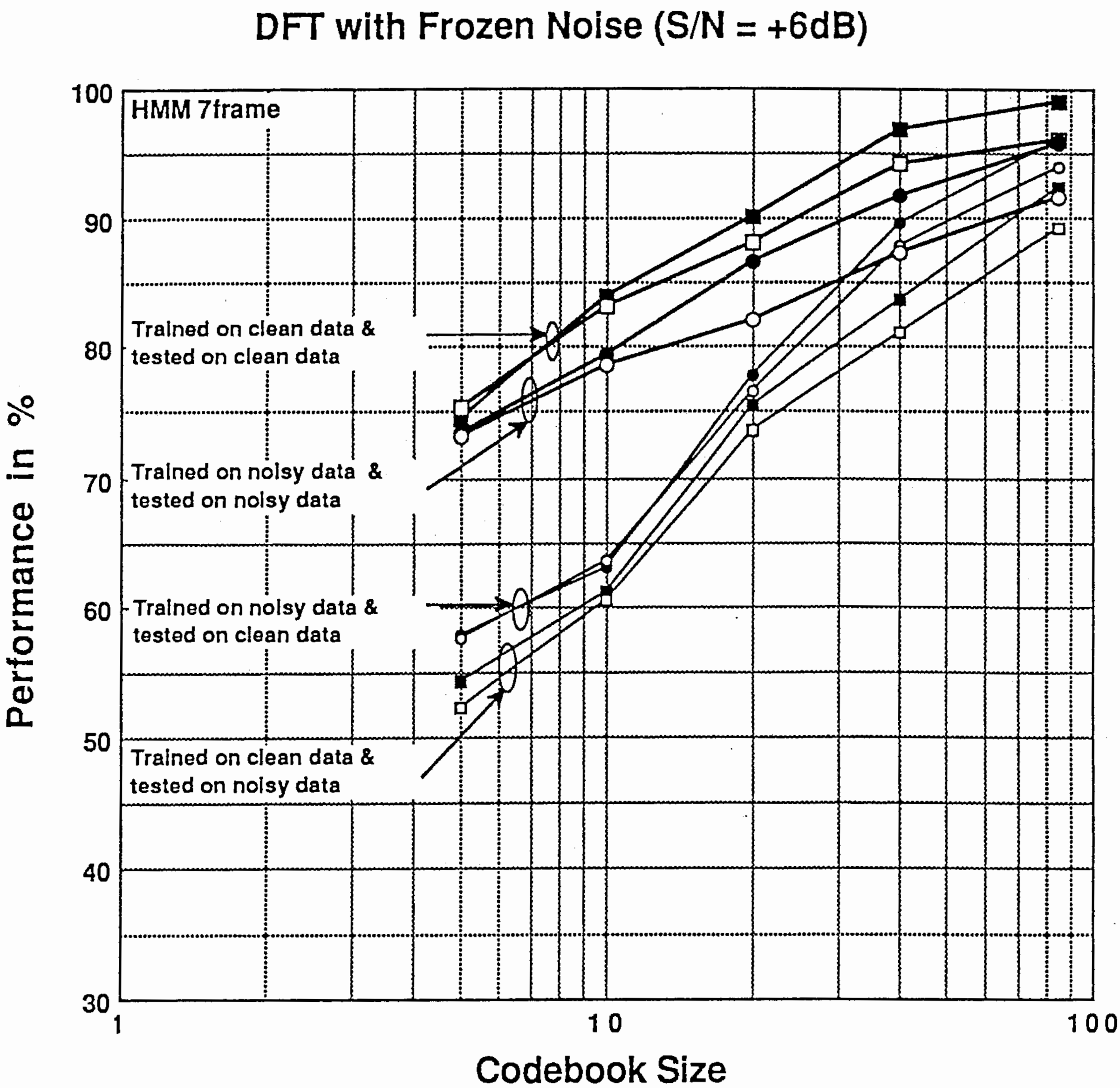

Figure 3(b) The /b, d, g/ phoneme recognition results for the DFT system under frozen noise conditions where the S/N=+6dB. Outlined and solid symbols indicate results for the open and closed recognition experiments.

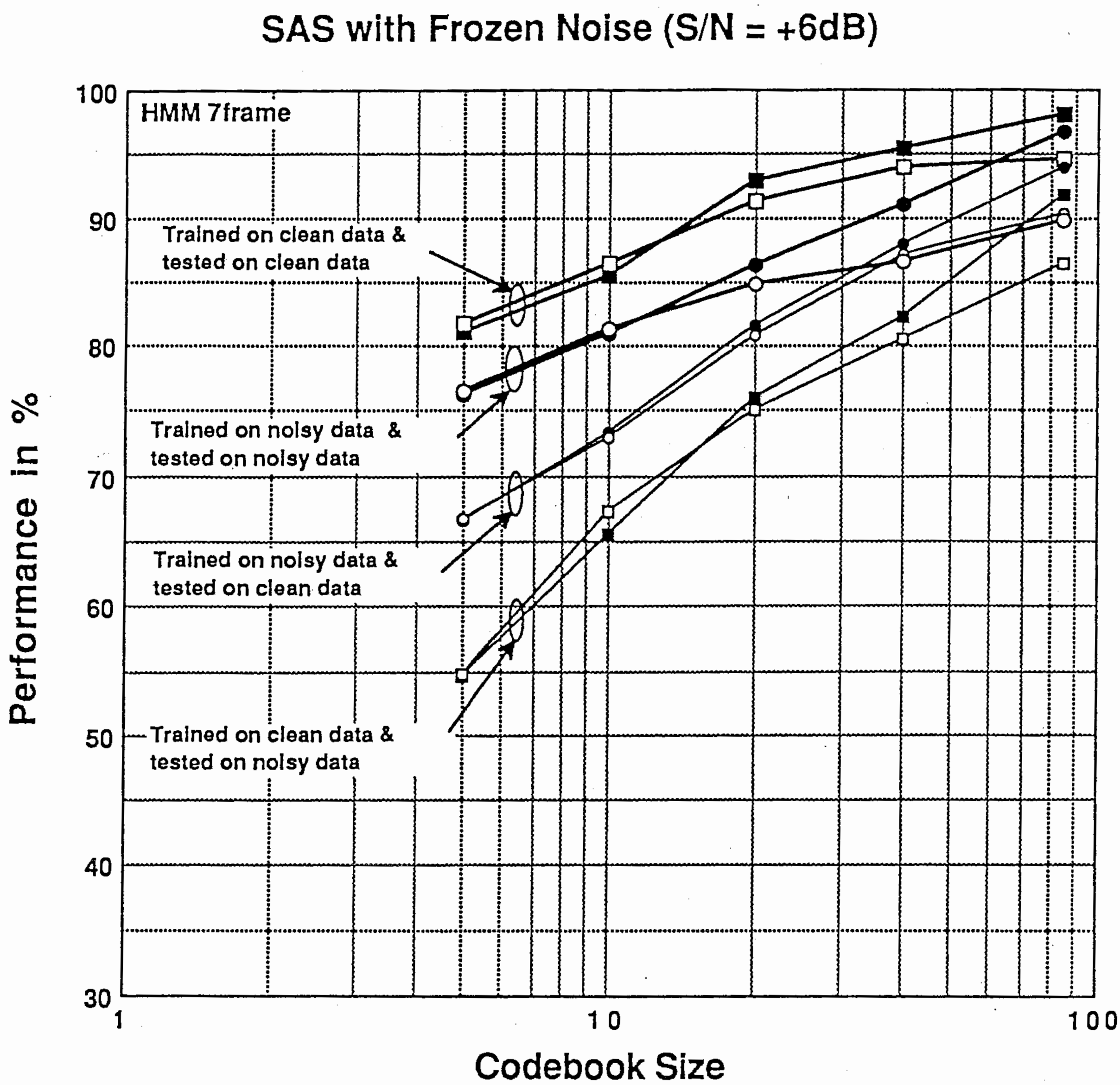

Figure 4(b) The /b, d, g/ phoneme recognition results for the SAS system under frozen noise conditions where S/N=+6dB.

Outlined and solid symbols indicate results for the open and closed recognition experiments.

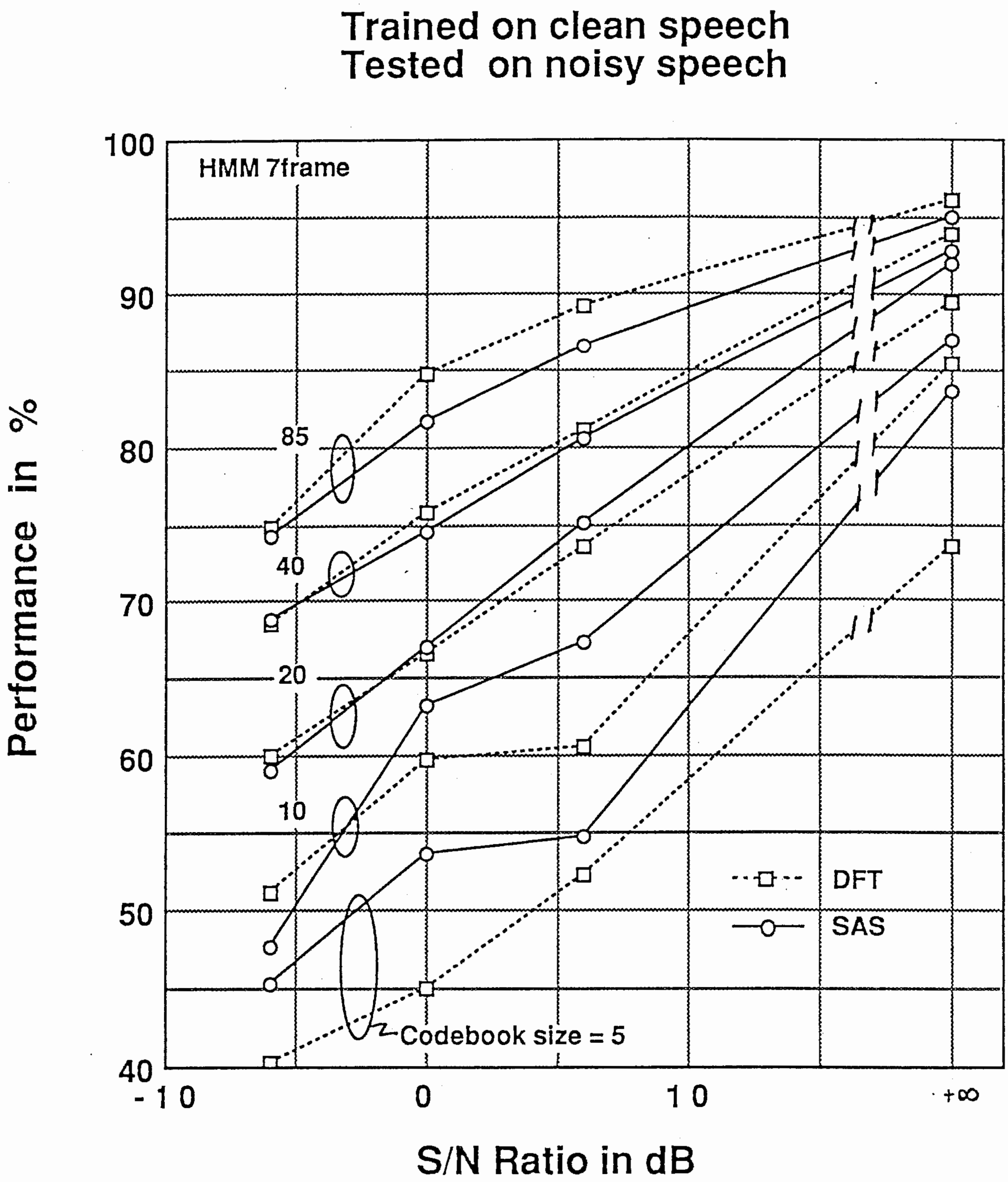

Figure 5(a) performance curves for the /b, d, g/ phoneme recognition versus S/N ratio in open experiments. HMM was trained on clean speech and tested on noisy speech.

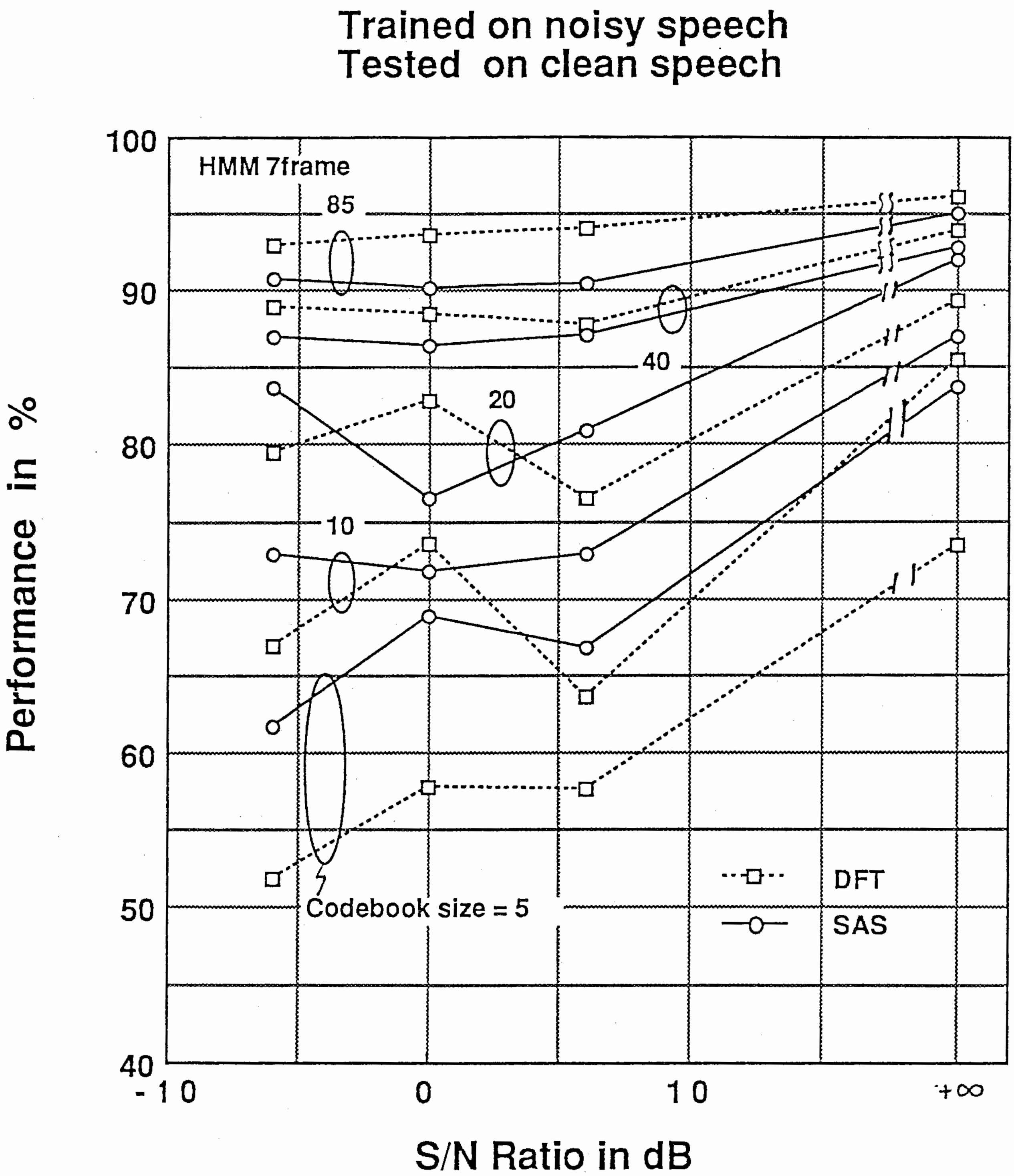

Figure 5(b) performance curves for the /b, d, g/ phoneme recognition versus S/N ratio in open experiments. HMM was trained on noisy speech and tested on clean speech.

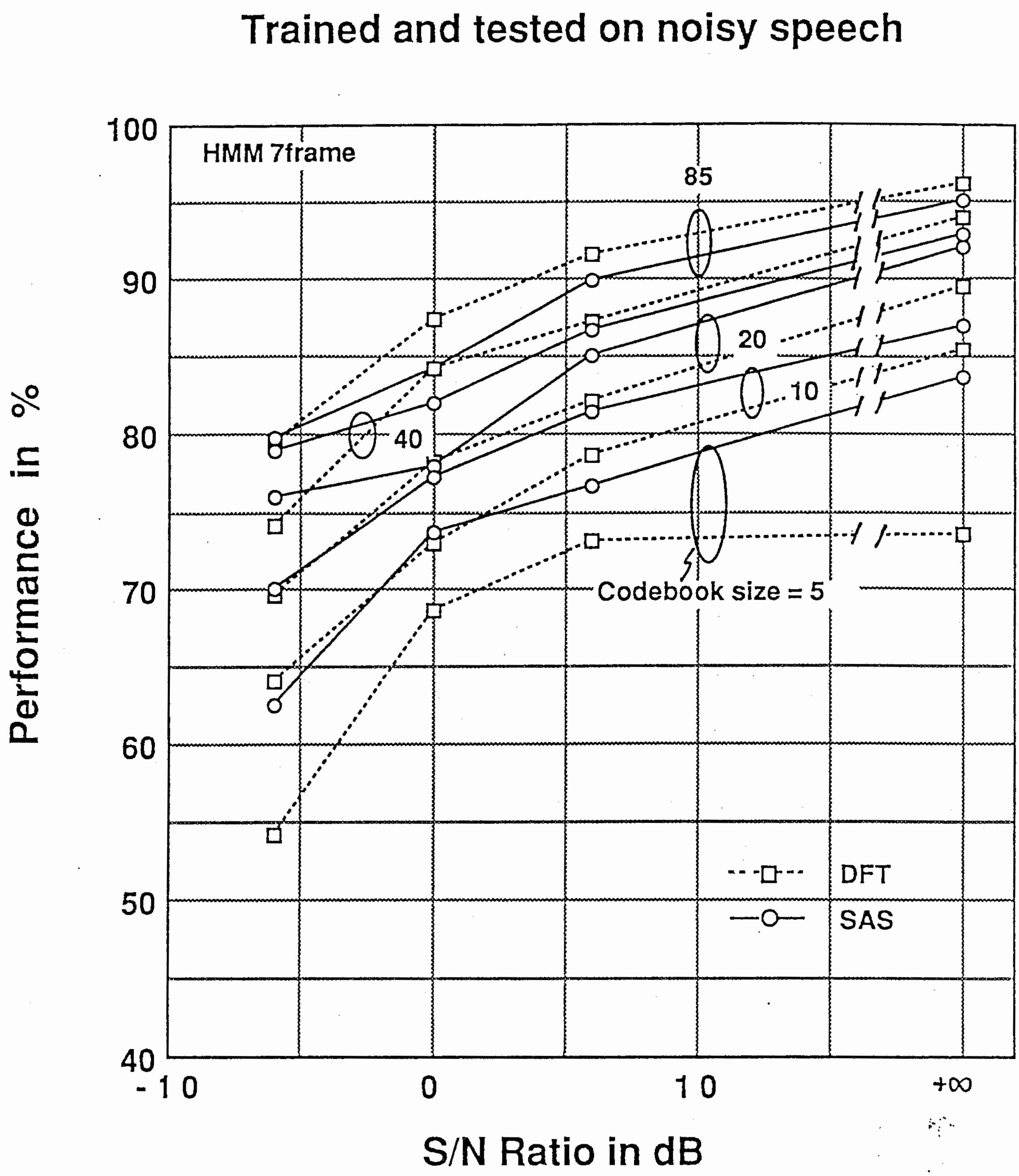

Figure 5(c) performance curves for the /b, d, g/ phoneme recognition versus S/N ratio in open experiments. HMM was trained and tested on noisy speech.

Results for 18 phonemes

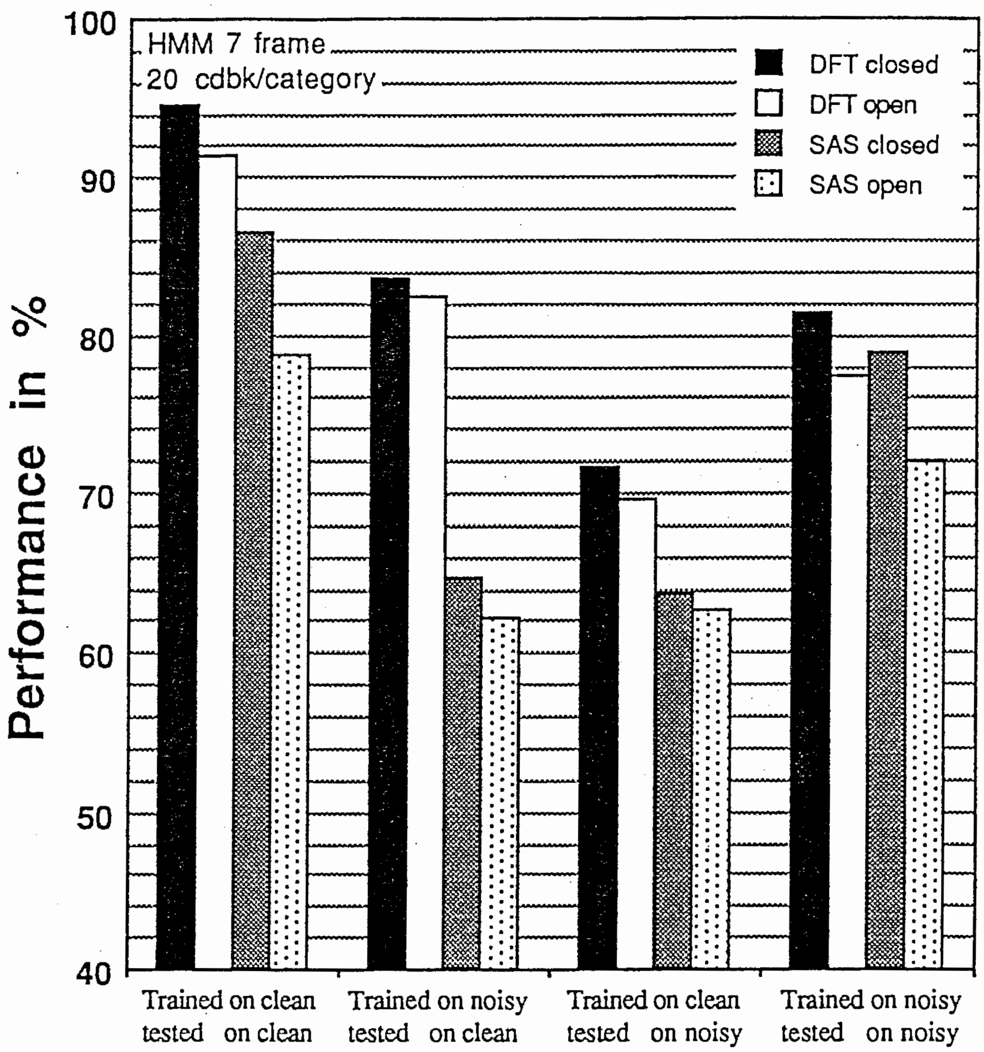

Figure 6 Performances for a larger phoneme set.

Closed

I\0	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	199	15	4																91.3
d	5	193	3					1											95.5
g	7	16	227				4					1				1		4	87.3
p				31	1														96.9
t				14	405	6													95.3
k				1	50	1078						7		12	4				93.6
m			4				427	30	6							3	1		90.7
n							15	242	1							2			93.1
N			6				3	1	491									2	97.6
s										458					17				96.4
sh											183			3					98.4
h			1								1	210		1		1			98.1
z													115						100.0
ch														79					100.0
ts										5					207				97.6
r	2		9	2	4	9	10	5	5							703	2	3	93.2
w																	71		100.0
y																		159	100.0
																			94.6

open

I\0	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	202	16	7				1									1			89.0
d	14	163	1										1						91.1
g	11	18	202		1		10	4	1							1	2	2	80.2
p			1	7	4	3													46.7
t				18	411	9			2										93.4
k					54	1071		1				14		18	3	3			92.0
m			5				422	44	5							3	2		87.7
n			1				31	224	3							6			84.5
N			4	2			4	2	472							2		2	96.7
s										515					23				95.7
sh											171			6					96.6
h			1			1	1				1	202		1					97.6
z						1			1				113						98.3
ch											3			68					95.8
ts										4					173				97.7
r	4	1	11	3	5	10	17	5	4		2					651	3	6	90.2
w	1		2				1									5	72		88.9
y		1	4													3	1	165	94.8
																			91.9

Table 2(a) Confusion matrix of DFT result. The HMM system is trained and tested on clean speech.

I\0	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	179	19	6		1	7	1									5			82.1
d	20	175	3			3												1	86.6
g	22	19	171	1		9	19	5	2							7		5	65.8
p	1			16	11	2										2			50.0
t				7	401	12										4		1	94.4
k	16		11	20	177	861				3	2	19		21	4	15	1	2	74.7
m	7		20			2	345	69	13				1			10	3	1	73.2
n	3		3			2	37	207	3							5			79.6
N			16				11	5	463							7		1	92.0
s										447			1		27				94.1
sh											185			1					99.5
h	2		1			5	3				3	190	1		1	8			88.8
z						1							114						99.1
ch						1					7			71					89.9
ts										61				2	149				70.3
r	29	5	11	7	7	15	8	11	10							642	2	7	85.1
w	1		1	1												1	67		94.4
y			2													2		155	97.5
																			83.6

I\0	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	194	16	5	1	2	7		2											85.5
d	8	159	3		1	3		1					1			3			88.8
g	20	13	165		1	9	20	7	2							7		8	65.5
p	1			4	5	3										1		1	26.7
t				6	421	7										5		1	95.7
k	14		9	38	174	850				4	5	17		25	5	19		4	73.0
m	5	1	16			2	346	86	16							3	4	2	71.9
n		1	4				46	205	3							6			77.4
N			14				10	5	446							11	1	1	91.4
s										507		1	3	1	26				94.2
sh											174			3					98.3
h	4		1			4	6				5	182				5			87.9
z								2				1	111			1			96.5
ch											9			62					87.3
ts										61				2	114				64.4
r	20	10	26	3	11	15	10	11	5		1	1				597	2	10	82.7
w	2		3			1	1									7	66	1	81.5
y		2	7			1	1	1								3		159	91.4
																			82.5

Table 2(b) Confusion matrix of DFT result. The HMM system is trained on noisy speech and tested on clean speech.

Closed

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	127	20	26		1	34	1	1	1			3	3			1			58.3
d	13	168	9			9						1				1		1	83.2
g	5	9	154	1	1	16	19	6	24			13	1			9	1	1	59.2
p	3		1	10	3	7			3							5			31.3
t	11	12	3	21	323	51			1							3			76.0
k	34	15	97	6	70	823	2	4	4			74	2	3	1	9	2	6	71.4
m			23			5	199	89	46			102	4			3			42.3
n	1		8			5	53	150	9	1		28	1		1	3			57.7
N			3			40	13		440				7						87.5
s						11				288	2	21		1	152				60.6
sh						1				1	177			7					95.2
h						11			1		1	196	2		3				91.6
z						8	1		2	2		3	86		13				74.8
ch						2					2			71	4				89.9
ts						21				5		9			177				83.5
r			21	14	5	33	16	9	25		2	44	6			575		4	76.3
w	1						4					2					64		90.1
y	1	2	10			6	2	5	1			7	1			4		120	75.5
																			71.7

open

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	125	27	22		3	35	2	2	3			3	1			3		1	55.1
d	16	145	9			4						1				4			81.0
g	4	11	156		1	20	17	5	18			9	3			6		2	61.9
p	2			6	1	3							1			1		1	40.0
t	12	10	2	26	314	71			2							2		1	71.4
k	22	25	101	5	77	794	3	6	6			84	1	9	3	17	1	10	68.2
m			31			10	227	85	33			88	4		1	2			47.2
n			10			5	48	135	21	1		32	4		1	7		1	50.9
N			3			32	9		434				8			2			88.9
s						24				311	5	15		2	181				57.8
sh						1				3	162			11					91.5
h		1		1		9	1				1	189	2		3				91.3
z						7			4	7		4	77		16				67.0
ch						1					5			62	3				87.3
ts						14				4		8			151				85.3
r		2	12	10	6	40	18	11	20	1	2	44	12		1	540	2	1	74.8
w	2					1	6		1				1			2	68		84.0
y		2	21			7	1	2				5	2		1	9		124	71.3
																			69.6

Table 2(c) Confusion matrix of DFT result. The HMM system is trained on clean speech and tested on noisy speech.

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	183	12	1		1	10	3	1	2			2	3						83.9
d	8	186				5						2						1	92.1
g	5	16	184			12	6	5	19			3	3		1	2	1	3	70.8
p				26	4							1	1						81.3
t	6			9	389	18										3			91.5
k	64	14	6	7	137	746	9	7	7	1		111	7	2	7	1	2	24	64.8
m			1			2	304	83	26			41	9		2			3	64.5
n						3	28	204	11	1		6	1		1	1		4	78.5
N			8				7	1	453			13	20			1			90.1
s										419	3	18	3	1	31				88.2
sh										2	180			4					96.8
h						1					1	207	1		4				96.7
z									1	3			106		5				92.2
ch											3			72	4				91.1
ts										9		3	10		190				89.6
r			2	4	9	5	16	9	16	1		15	16		4	650	2	5	86.2
w								1				1					69		97.2
y						1	5	2				1						150	94.3
																			81.5

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	167	24	1		1	17	3	3	2			2	4					3	73.6
d	14	156	2		3	2						1						1	87.2
g	16	18	171		1	11	8	3	13			3	4			1	1	2	67.9
p			1	8	4	2													53.3
t	2			11	393	29		1				1	1			2			89.3
k	60	21	16	6	144	698	8	7	3	2		129	7	10	17	5	3	28	60.0
m	1			1		3	305	86	25			43	10		1		4	2	63.4
n			1			1	49	173	16			11	5			3		6	65.3
N			5				6	3	440			10	23			1			90.2
s										476	5	15	5		37				88.5
sh										3	166			8					93.8
h	1	1				1	1				1	195	4		3				94.2
z								2		5		1	96		7	4			83.5
ch											8			60	3				84.5
ts										12		3	3	1	158				89.3
r			8	5	9	6	26	13	11	1		18	19		4	598	2	2	82.8
w	3	1	1				3						2			5	66		81.5
y		3	3			2	4	5	1			4				3		149	85.6
																			77.5

Table 2(d) Confusion matrix of DFT result. The HMM system is trained and tested on noisy speech.

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	197	7	3		1	1	4	1					1		1		2		90.4
d		200	1															1	99.0
g	11	3	218			3	8	4	8							1		4	83.8
p				31														1	96.9
t		3	3	6	392	13		1								5	1	1	92.2
k	30	4	16	26	49	897	13	7		14	20	11		24	6	9		26	77.9
m	3		35				358	44	11		1	3				4		12	76.0
n		1	3				11	228	11				3			2		1	87.7
N	1		20				6	4	460			1	4	1		4		2	91.5
s						3				420			2		50				88.4
sh											173			13					93.0
h	1					1	2		2		7	199		2					93.0
z										2			112		1				97.4
ch											6			73					92.4
ts										9			2		201				94.8
r	4	4	9	8	1	11	17	15	7	6	3	6	12		1	624		26	82.8
w	1																70		98.6
y			2		1													156	98.1
																			86.5

Closed

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	191	12	2			6	5	3		2							4	2	84.1
d	4	166			2			3		2						2			92.7
g	15	3	173		1	8	20	2	14					1	1	6		8	68.7
p	2			5	2	3	1									1		1	33.3
t	5	12	2	4	395	16										4	1	1	89.8
k	30	2	24	38	50	840	9	14		13	27	20	4	23	17	12	1	40	72.2
m	5		32			4	311	75	19		1	7	1			12		14	64.7
n	2	2	4		1		22	185	20		2	2	5			19		1	69.8
N	3		23	2		1	9	11	422		3	1	6	1		5		1	86.5
s		5		1		2				466			12	1	51				86.6
sh			1			1				2	145	4		24					81.9
h	1	2		1	1	8	3		1	1	11	175		1	1	1			84.5
z	1	10						3	1	5	1		84		5	5			73.0
ch											20			51					71.8
ts						1				26			2		148				83.6
r	3	5	10	4	8	19	21	22	1	6	3	4	16		2	573		25	79.4
w	1					3	4	1				1				7	64		79.0
y			2		5	2	4	2					1			5		153	87.9
																			78.8

Open

Table 3(a) Confusion matrix of SAS result. The HMM system is trained and tested on clean speech.

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	151	10	7	3	3	22	6	1	1			5	1			3	2	3	69.3
d	12	148	5		9	11	1	1					1			13		1	73.3
g	15	2	139	2	4	36	23	5	11			1				16	1	5	53.5
p	8			15	1	1	2	1								2		2	46.9
t	5	8	5	5	359	27	1	1				1				12	1		84.5
k	118	20	39	15	87	656	14	11	2	11	10	48	5	16	5	68	4	23	56.9
m	33		50	2		18	250	50	14		1	8				31	5	9	53.1
n	2	4	6			22	46	98	11			2	13	1		50		5	37.7
N	10	1	34	15		5	56	19	327			1	10	8	2	13		2	65.0
s	9	3				11	3	7		329		14	31	1	58	9			69.3
sh											153	9		23	1				82.3
h	1	2		1		7	13	2	5	1	2	178		1		1			83.2
z	9	2	2			6	2	9	1	9		1	62			9		3	53.9
ch			1								8	2		67	1				84.8
ts	2	1				3	2	3		50	4	9	7	4	124	3			58.5
r	14	9	17	10	20	33	40	14	5	3	1	11	17	1		509	2	48	67.5
w	3					3	2									2	61		85.9
y			2			12	1				1	2				16		125	78.6
																			64.8

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	141	15	7		6	26	6	5	3			2	1			8	5	2	62.1
d	11	133	2		7	10		3		3			2			7		1	74.3
g	15	3	138		2	39	20	6	12		1	1	1			8	2	4	54.8
p	3			4	2	2	2	1							1				26.7
t	3	7	6	8	373	22						1				17	1	2	84.8
k	121	14	45	15	87	643	18	12	1	17	16	47	9	22	6	57	2	32	55.2
m	32		50	4		30	244	61	14		3	2				25	3	13	50.7
n	5	1	6		2	22	39	99	12		1	4	9			61		4	37.4
N	9	1	40	13		6	63	17	304	1		2	8	6	2	15		1	62.3
s	12	4				12	4	14		356	4	21	40	3	57	11			66.2
sh						2	1			1	136	14		21	2				76.8
h	6	1			1	16	14	3		3	4	154		1		3		1	74.4
z	10	8	2			7	6	11	1	5	2	1	48		2	11		1	41.7
ch			1			2			1		25	2	2	38					53.5
ts	1		1			5	3	4		56	3	3	6	4	89	1	1		50.3
r	21	7	11	14	19	40	29	16	6	7	2	4	11	1		495	1	38	68.6
w	3	1	1			3	3					1				9	60		74.1
y					2	12	4	1			1	2				15		137	78.7
																			62.2

Table 3(b) Confusion matrix of SAS result. The HMM system is trained on noisy speech and tested on clean speech.

closed

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	112	9	12	3	2	42	14	3	3	3			2			10	3		51.4
d	7	152	2		4	13		4		5		4	2			8		1	75.2
g	10	15	128		1	35	32	8	10	5		1	1		1	8		5	49.2
p	4			15		7	2					1				3			46.9
t	4	18	4	1	317	61	1	4								11	1	3	74.6
k	30	39	20	5	49	797	13	2	2	65	4	17		34	31	33	2	9	69.2
m	17		30			93	212	21	39	11		16				30	2		45.0
n	8		11			23	46	88	18	40			1		2	23			33.8
N	8	1	27	1		19	45	23	369	2	1					7			73.4
s						36				340	3	19	12	2	62	1			71.6
sh						2					160	1		23					86.0
h	2	2	1			76	1			17	8	96		6	1	4			44.9
z	1	1				6	1	7	5	19		1	62	1	4	7			53.9
ch						5				3	6			64	1				81.0
ts						38				61		3		2	108				50.9
r	13	8	15	4	10	46	19	20	24	18		5	7		6	540	1	18	71.6
w	3					15	2									4	47		66.2
y		1	5		1	26	6	2		1			1		1	31		84	52.8
																			63.8

open

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	95	10	14	2	4	62	16	1	3	1		1			1	11	5	1	41.9
d	6	134	1		4	17	1	4		4						8			74.9
g	10	17	121	1	1	43	25	5	12	3		2		1		7		4	48.0
p	1	1	1	5	3	2						1				1			33.3
t	3	19	9	3	321	55	1	2		3		2				19	1	2	73.0
k	18	30	24	5	42	839	15	3		64	6	19	1	26	26	30	1	15	72.1
m	15		43		1	92	212	22	27	11		17	1		1	38	1		44.1
n	14	2	12			26	57	73	16	34			3		1	27			27.5
N	8	2	22			24	34	14	377	2						5			77.3
s	1					28		1		401	2	21	9	4	70	1			74.5
sh						5					140			31	1				79.1
h	2	1		1		76			1	26	3	80		7	2	7		1	38.6
z	4	12				1	1	11	4	24		1	43		6	8			37.4
ch						7				1	7			56					78.9
ts						29				60	1	2	2		82	1			46.3
r	10	8	13	7	16	39	22	26	27	27		4	4		3	498		18	69.0
w	3				1	15	4	2								6	50		61.7
y	3	1	12		3	21	5	1		1		2		1	3	28		93	53.4
																			62.7

Table 3(c) Confusion matrix of SAS result. The HMM system is trained on clean speech and tested on noisy speech.

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	182	6	3			11	3	1	1	1			5		2			3	83.5
d	2	190				7				1		2							94.1
g	13	13	180			7	17	6	14			3	4			1		2	69.2
p				29		1			2										90.6
t	2	8	1	3	391	14										5		1	92.0
k	31	31	9	12	80	761	4	6	2	14	3	29	1	38	97	4		30	66.1
m	1		9			4	323	28	31	1		57	4	2	8	1	1	1	68.6
n			2			3	23	195	14	8		3	4		5	1		2	75.0
N			11				9	3	446	6	1		21		5	1			88.7
s										354	2	10	2		107				74.5
sh											169			16	1				90.9
h						2	1			2	10	185		7	6	1			86.4
z								1	1	13			94		5	1			81.7
ch											5	1		68	5				86.1
ts										8		1	2	2	199				93.9
r	1	3	9	5	6	20	19	19	12	16		6	17		6	584		31	77.5
w	1																70		98.6
y			1			1	4	1	1	1		1	1		1			147	92.5
																			78.9

I\O	b	d	g	p	t	k	m	n	N	s	sh	h	z	ch	ts	r	w	y	
b	154	12	4			31	4	4	3	2		1	3		2	2	1	4	67.8
d	4	162	4			5						2				2			90.5
g	26	6	159			8	18	2	18			1	1		2	5		6	63.1
p	1			6	4	2						1				1			40.0
t	3	10	1	3	394	25										4			89.5
k	36	19	27	10	94	729	7	5	1	18	6	33	1	39	96	6		37	62.6
m	5		16			5	293	49	23	7		51	6	2	11	6	4	3	60.9
n	4		1			2	44	162	14	8		1	11		3	13		2	61.1
N			14				11	3	419	5			25		10	1			85.9
s						2	1	2		371	3	18	5	3	133				69.0
sh											142			32	3				80.2
h	1	1	1			3	4	1		15	3	149	1	15	6	6		1	72.0
z			3				3	8	3	15			75		6	2			65.2
ch						1					10			55	5				77.5
ts										12	1	4			160				90.4
r	1	5	14	6	16	20	30	19	10	25		1	14		8	530	1	22	73.4
w	2					1	4	1					1			4	68		84.0
y	5		4		3	2	9	1		2		2	2		2	6		136	78.2
																			72.1

Table 3(d) Confusion matrix of SAS result. The HMM system is trained and tested on noisy speech.