TR-I-0050

Duration control methods for HMM phoneme recognition

HMM音韻認識におけるモデル継続時間長の制御手法

Toshiyuki Hanazawa, Takeshi Kawabata and Kiyohiro Shikano

花沢 利行　川端 豪　鹿野清宏

1988.11

Abstract

Two kinds of duration control for HMM (Hidden Markov Model) phoneme recognition are proposed: phoneme-duration control for an HMM phone model and a state-duration control for an HMM state. The phoneme-duration control is carried out by combining an HMM output probability with a phoneme duration penalty. The phoneme duration penalty is calculated using a phoneme duration histogram obtained from training samples. Phoneme duration control is effective in discriminating phonemes with different durations such as /n/ and /N/. State-duration control is realized as a state duration penalty calculated from an HMM state duration distribution of training samples. State- duration control is effective in discriminating phonemes with different event structures such as /s/ and /ts/. Recognition experiments are carried out using Japanese phonemes extracted from an isolated word database uttered by one male speaker. The phoneme recognition rate is improved from 84.8% to 89.8% using these duration control methods.

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

Contents

1. Introduction

HMM is effective in expressing speech data statistically, so it has been used for speech recognition[1][2][3][4]. But duration of the speech data is not expressed accurately in the ordinary HMM as in Fig.1. So in order to express the duration of speech data in the HMM, various duration control methods were proposed[5][6][7][8][9][10].

In this paper two kinds of non-parametric duration control methods for HMM phoneme recognition are proposed, a phoneme-duration control for an HMM phoneme model and state-duration control for an HMM state. These duration control methods are evaluated by recognition experiments of all Japanese phonemes extracted from an isolated word database (5,240 words) uttered by one male speaker.

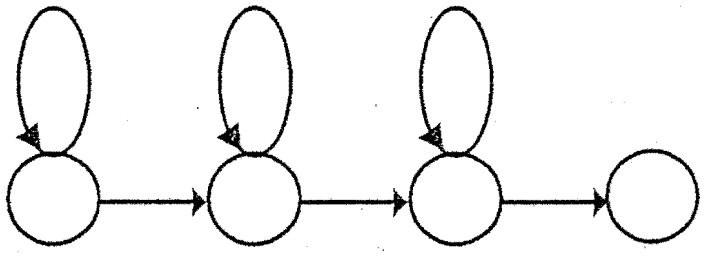


Fig1. Hidden Markov Model

2. Phoneme-Duration Control for an HMM Phoneme Model

The phoneme-duration control is carried out by combining an HMM output probability with a phoneme duration penalty. The phoneme duration penalty is calculated using a phoneme duration histogram obtained from training samples. The HMM output probability is calculated using trellis algorithm. Calculation of the HMM probability controlled by phoneme duration is realized through the following two steps.

Step1. Calculation of the Phoneme Duration Penalty

The phoneme duration penalty $p^{(m)}_{\tau}$ is calculated as follows.

$$p_{\tau}^{(m)} = c_{\tau} / \max_{k}(c_{k})$$

where c_{τ} is the degree of phoneme duration obtained from training samples.

Step 2. Calculation of the HMM Probability Controlled by Phoneme Duration

The HMM probability controlled by the phoneme duration is calculated as follows.

$$P^{(m)} = P * (p_{\tau}^{(m)})^{w_1}$$

where $P^{(m)}$: HMM probability controlled by phoneme duration,
P : HMM probability calculated by the trellis algorithm,
$p^{(m)}{}_{\tau}$: phoneme duration penalty
w_1 : constant.

3. State-Duration Control for an HMM State

The state-duration control is realized as a state duration penalty calculated from an HMM state duration distribution of training samples. The HMM state duration distribution is calculated by modified forward-backward probabilities of training samples. The HMM probability controlled by the state duration is calculated through the following two steps.

Step 1. Calculation of the HMM State Duration Penalty

An HMM state duration distribution $d(i, \tau)$ (i represents the state number, and τ represents duration time) is calculated as follows.

$$d(i, \tau) = \sum_{t} \tilde{\alpha}(i, t) \cdot \prod_{k=t+1}^{t+\tau} (a_{ii} b_{iiv_k}) \cdot \tilde{\beta}(i, t+\tau)$$

where

$$a_{ij} : \textit{transition probability}, \qquad b_{ijv_k} : \textit{output probability}$$

$$\tilde{\alpha}(i, t) = \alpha(i, t) - \alpha(i, t-1)\, a_{ii} b_{iiv_t}$$

$$\tilde{\beta}(i, t) = \beta(i, t) - \beta(i, t+1)\, a_{ii} b_{iiv_{t+1}}$$

$$\alpha(i, t) : \textit{forward probability}, \qquad \beta(i, t) : \textit{backward probability}$$

And , the HMM state duration penalty is calculated as follows.,

$$p^{(s)}(i,\tau) = d(i,\tau) / \max_{k}(d(i,k))$$

where $p^{(s)}(i,t)$: HMM state duration penalty
$d(i,\tau)$: HMM state duration distribution

Step 2. Calculation of the HMM Probability Controlled by the State Duration

A forward probability controlled by the state duration is calculated as follows,

$$\alpha^{(s)}(j,t) = \sum_{i}\sum_{\tau}\alpha^{(s)}(i,t-\tau-1)a_{ij}b_{ijv_{t-\tau}} \cdot \prod_{k=t-\tau+1}^{t}(a_{jj}b_{jjv_k})\cdot(p^{(s)}(j,\tau))^{w_2}$$

where $\alpha^{(s)}(i,t)$: forward probability controlled by the state duration,
$p^{(s)}(i,t)$: HMM state duration penalty,
w_2 : constant.

And the HMM probability controlled by the state duration is calculated as follows,

$$P^{(s)} = \alpha^{(s)}(J,T)$$

where $P^{(s)}$: HMM probability controlled by the state duration
$\alpha^{(s)}(i,t)$: forward probability controlled by the state duration
J : final state
T : final frame of input speech

4. Phoneme Recognition Experiments

4.1 Speech Data

The recognition experiments were carried out using all Japanese phonemes extracted from an isolated word database[11] (5,240 words) uttered by one male speaker. The speech data was sampled at 12kHz, pre-emphasized by (1 - $0.97z^{-1}$) and windowed using a 256-point Hamming window every 3 msec. Then a 12-order LPC analysis was carried out. A codebook of 256 LPC spectrum envelopes was generated from 216 phonetically balanced words. The weighed Likelihood ratio augmented with power value (PWLR)[12] was used as an LPC distance measure for vector quantization. Phoneme tokens for training were extracted from the even numbered entries of 5,240 words, and tokens for testing were extracted from

the odd numbered entries. All phonemes were extracted manually using acoustic-phonetic labels[11] provided with the database.

4.2 Hidden Markov Models

HMMs with 4 states and 3 loops shown in Fig.1 were used for each phoneme. Two models were trained for each affricate (/ch/, /ts/) and stop (/p/ /t/ /k/ /b/ /d/), where one HMM was for consonants extracted from the word-initial position and the other was for word-middle consonants. For the other phonemes only one context-independent model was trained. The number of tokens for training and testing of each phoneme are shown in Table 1.

4.3 Phoneme Recognition Using Phoneme-Duration Control

The phoneme recognition experiment using the phoneme duration control described in section 2 was carried out. Improvement of the phoneme recognition rates for each phoneme is shown in Fig.2. The total recognition rate was improved from 84.8% (without duration control) to 88.2%. The phoneme duration control is especially effective in discriminating phonemes with different durations such as /n/ and /N/.

4.4 Phoneme Recognition Using State-Duration Control

The phoneme recognition experiment used the state duration control described in section 3. Improvement of the phoneme recognition rates for each phoneme is shown in Fig.3. The total recognition rate was improved from 84.8% (without duration control) to 88.3%. The state duration control is especially effective in discriminating phonemes with different event structures such as /s/ and /ts/.

4.5 Phoneme Recognition Using Phoneme-Duration Control and State-Duration Control

Recognition results described in section 4.3, 4.4 show that the phoneme-duration control is effective in discriminating phonemes which have different durations and the state-duration control is effective in discriminating phonemes with different event structures. More accurate recognition is accomplished by using both duration control methods simultaneously. In this section the phoneme recognition experiment using the phoneme-duration control and the state-

duration control was carried out. Improvement of the phoneme recognition rates for each phoneme is shown in Fig.4. Compared with Fig.2, Fig.3, it is proved that this composite duration control method has both advantages in phoneme-duration control and state-duration control. The total phoneme recognition rate was improved from 84.8% (without duration control) to 89.8% using this composite duration control method.

Table 1 : The number of tokens for training and testing of each phoneme

phoneme	number of tokens (for training)	number of tokens (for testing)
/s/	475	269
/sh/	187	177
/h/	266	262
/z/	225	227
/ch/ (initial / middle)	12/67	71
/ts/ (initial / middle)	36/176	177
/p/ (initial / middle)	8/25	15
/t/ (initial / middle)	196/229	220
/k/ (initial / middle)	417/736	233
/b/ (initial / middle)	59/159	227
/d/ (initial / middle)	68/135	179
/g/	68	67
/ng/	192	185
/m/	471	241
/n/	260	265
/N/	503	244
/r/	753	241
/w/	72	81
/y/	159	174
/a/	1715	220
/i/	1324	215
/u/	1498	209
/e/	664	226
/o/	877	214

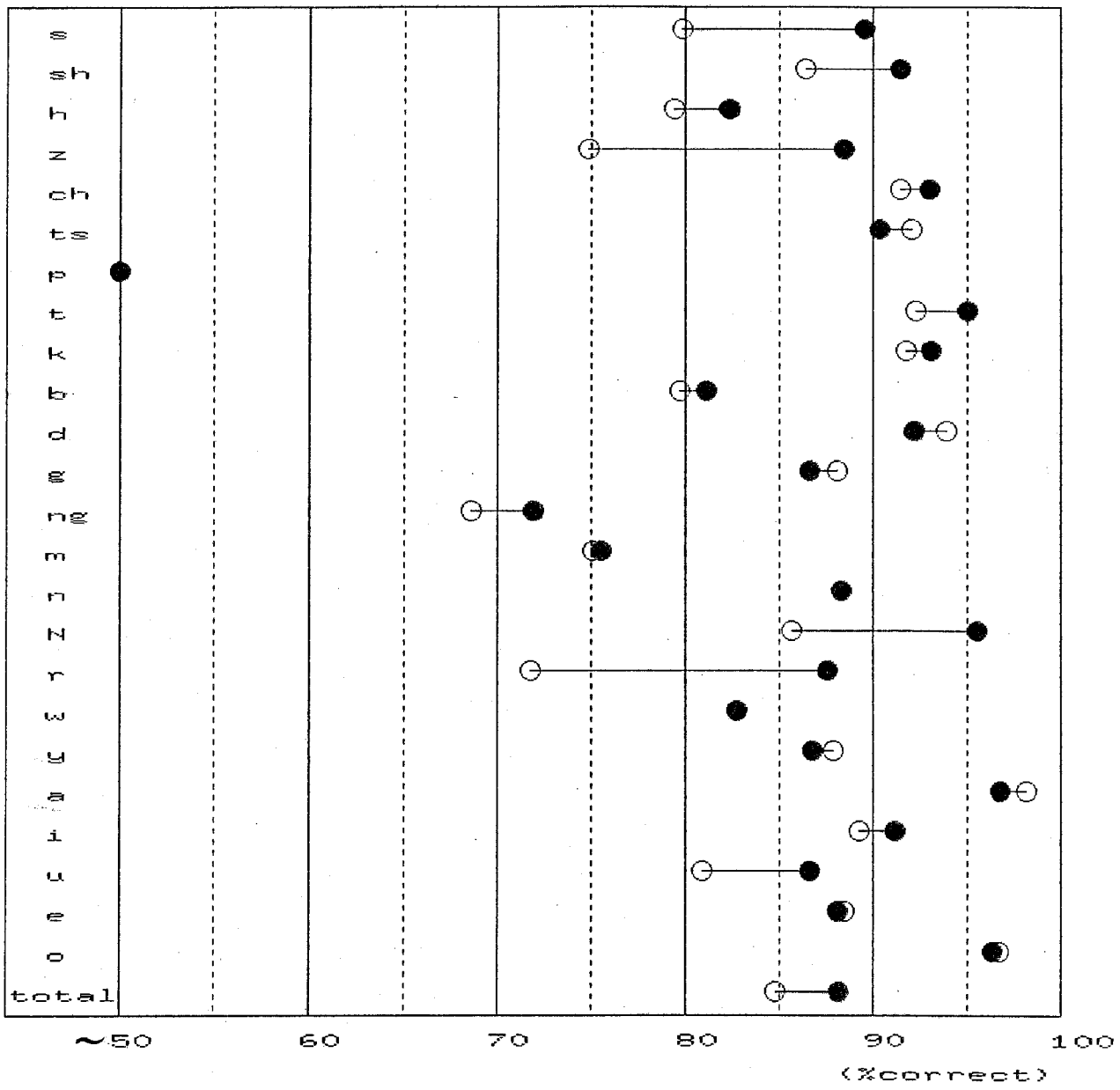


Fig.2. Improvement of phoneme recognition rates using the phoneme-duration control.

○ : without duration control

● : using the phoneme-duration control

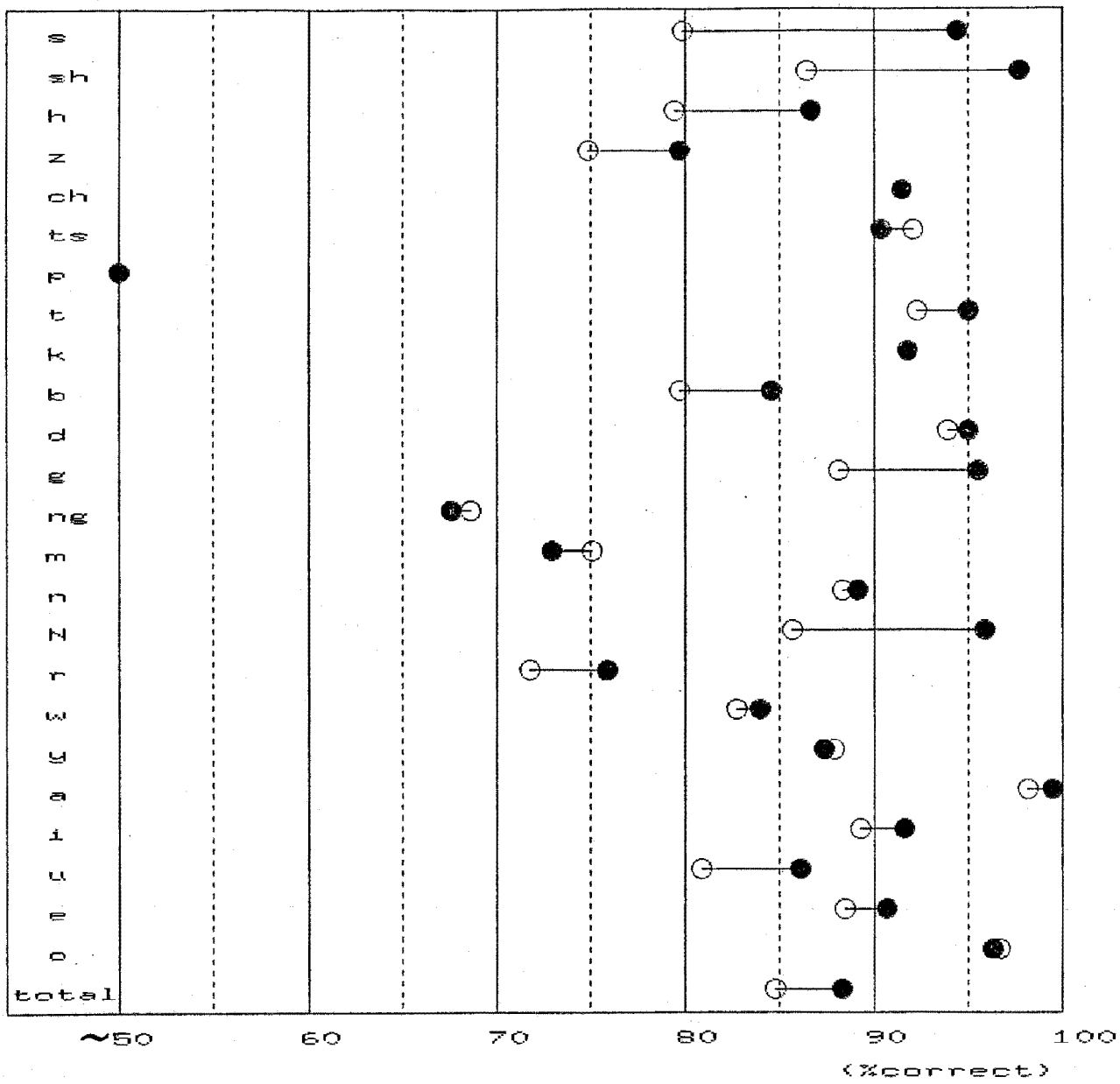


Fig.3. Improvement of phoneme recognition rates using the state-duration control.

○ : without duration control

● : using the state-duration control

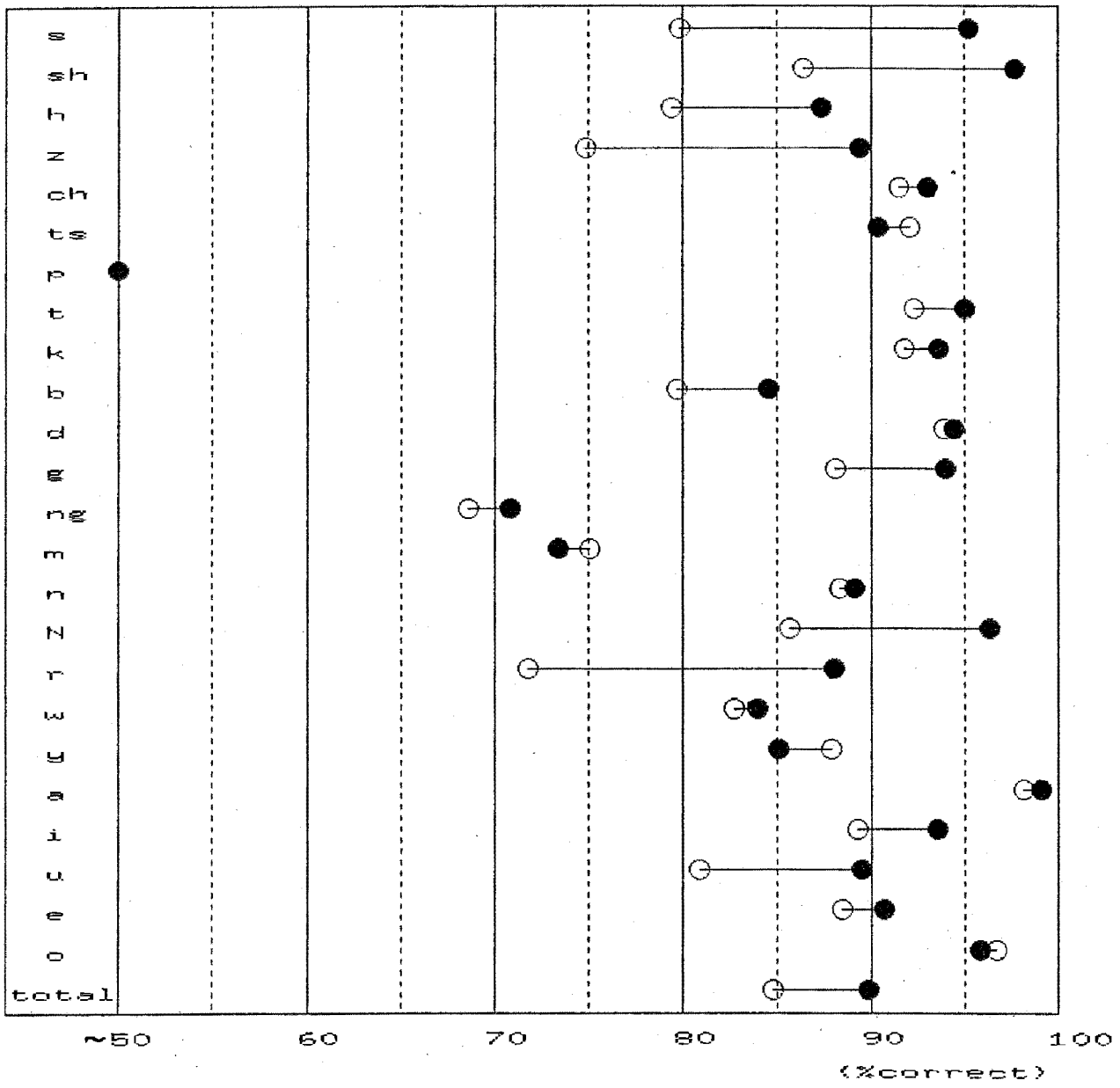


Fig.4. Improvement of phoneme recognition rates using the phoneme-duration control and the state-duration control.

○ : without duration control

● : phoneme-duration and state-duration control

5. Summary and Conclusion

In this paper two kinds of non-parametric duration control methods for HMM phoneme recognition were presented: phoneme duration control for an HMM phoneme model and a state duration control for an HMM state. The phoneme duration control is carried out by combining an HMM output probability with a phoneme duration penalty. The phoneme duration penalty is calculated using a phoneme duration histogram obtained from training samples. State duration control is realized as a state duration penalty calculated from an HMM state duration distribution of training samples.

These duration control methods were evaluated by recognition experiments of all Japanese phonemes extracted from an isolated word database (5,240 words) uttered by one male speaker. Phoneme duration control is effective in discriminating phonemes with different durations such as /n/ and /N/. The state duration control is also effective in discriminating phonemes with different event structures such as /s/ and /ts/. The best recognition result was obtained using the two duration control methods simultaneously, and the total phoneme recognition rate was improved from 84.8% (without duration control) to 89.8%.

Acknowledgment

The authors would like to thank Dr. Akira Kurematsu, president of ATR Interpreting Telephony Research Laboratories for his continuous support of this research. We are also grateful to Dr. Kai-Fu Lee of Carnegie-Mellon University for his many valuable suggestions. We are also indebted to the member of the Speech Processing Department at ATR.

References

[1] J.K.Barker, : "The DRAGON System -An Overview", IEEE Trans., ASSP-23, (1975-02)

[2] F. Jelinek, : " Continuous Speech Recognition by Statistical Methods", IEEE. Proc. 64., (1976)

[3] Y. L. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J.Makoul, P. J. Price, S. Roucos and R. M. Schwartz, : " BYBLOS: The BBN Continuous Speech Recognition System", ICASSP87 (1987)

[4] K-F. Lee and H-W. Hon, : " Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM ", ICASSP88 (1988)

[5] J. D. Ferguson, : " Variable Duration Models for Speech", in Poc., Symp. on Application of hidden Markov models to Text and Speech, ed. J.D. Ferguson, Princeton, (1980)

[6] M. Nishimura and M. Okochi, : " A Speech Recognition Method using HMM with Duration Distribution ", In Conference of Acoustical Society of Japan, March 1986

[7] L. R. Rabiner and S. E. Levinson, : " A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based on Hidden Markov Models and Level Building", IEEE. Trans. Acoust., Speech & Signal Proc. ASSP-32, 3, (1985)

[8] L. R. Rabiner, B-H. Juang, S. E. Levinson and M. M. Sondhi, : " Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities ", AT&T. Tech. J., 64, 6 (1985)

[9] M. J. Russell and R. K. Moore, : " Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition", ICASSP85 (1985)

[10] S. E. Levinson, : " Continuous Variable Duration Hidden Markov Models for Speech Recognition ", Computer Speech and Language, 1 (1986)

[11] K. Takeda, Y. Sagisaka and S. Katagiri, : " Acoustic-Phonetic Labels in a Japanese Speech Database", Proc., of Euro., Conf., on Sp. Tech., Vol.2. (1987)

[12] K. Aikawa, M. Sugiyama, K. Shikano, : " Spoken Word recognition Using Vector Quantization with Power-Weighted WLR Measure ", Trans. of the Comm., on Speech Research The Acoustical Society of Japan, S82-61 (December 1982)

Appendix

Fig-A-1. Confusion Matrix of All Japanese Phonemes
(without duration control)

Fig-A-2. Confusion Matrix of All Japanese Phonemes
(using the phoneme duration control)

Fig-A-3. Confusion Matrix of All Japanese Phonemes
(using the state duration control)

Fig-A-4. Confusion Matrix of All Japanese Phonemes
(using the phoneme duration control and the state duration control)

phon	s	sh	h	z	ch	ts	p	t	k	b	d	g	ng	m	n	N	r	w	y	a	i	u	e	o	dat	err	zero	%
s	215			10	4	40																			269	54	0	79.9
sh		153		1	23																				177	24	0	86.4
h		1	208			1		9	34											5			4		262	54	0	79.4
z	20	12		170	12	12															1				227	57	0	74.9
ch		3		3	65																				71	6	0	91.5
ts	9				5	163																			177	14	0	92.1
p			1				6	4	2													1	1		15	9	0	40.0
t	1		5				4	203	5								1						1		220	17	0	92.3
k		5	6		1		1	5	214												1				233	19	0	91.8
b										181	13	2		7	4		9	7			2	1	1		227	46	0	79.7
d								2		1	168			1	1		5			1					179	11	0	93.9
g										5	1	59			1						1				67	8	0	88.1
ng										2	1		127	23	5		4		2		9	3	8	1	185	58	0	68.6
m			1							6		1	13	181	32	3	2	1			1				241	60	0	75.1
n											3	1	3	19	234	1	3				1				265	31	0	88.3
N													9	4	15	209			1		5		1		244	35	0	85.7
r			1	1						4	5	2	7	3	12		173	7	2	1	6	1	15	1	241	68	0	71.8
w										4				1			1	67						8	81	14	0	82.7
y			1												1		5		153		3		11		174	21	0	87.9
a			1															1	2	216					220	4	0	98.2
i		1		1	3				2					1	1				10		192		4		215	23	0	89.3
u	3			1		3	1		1	3	2		2	1		2	3		2		3	169	9	4	209	40	0	80.9
e													2				1		7	2	13	1	200		226	26	0	88.5
o														2			2	1				2		207	214	7	0	96.7
sum	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	4639	706	0	84.8
ave	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---				83.5

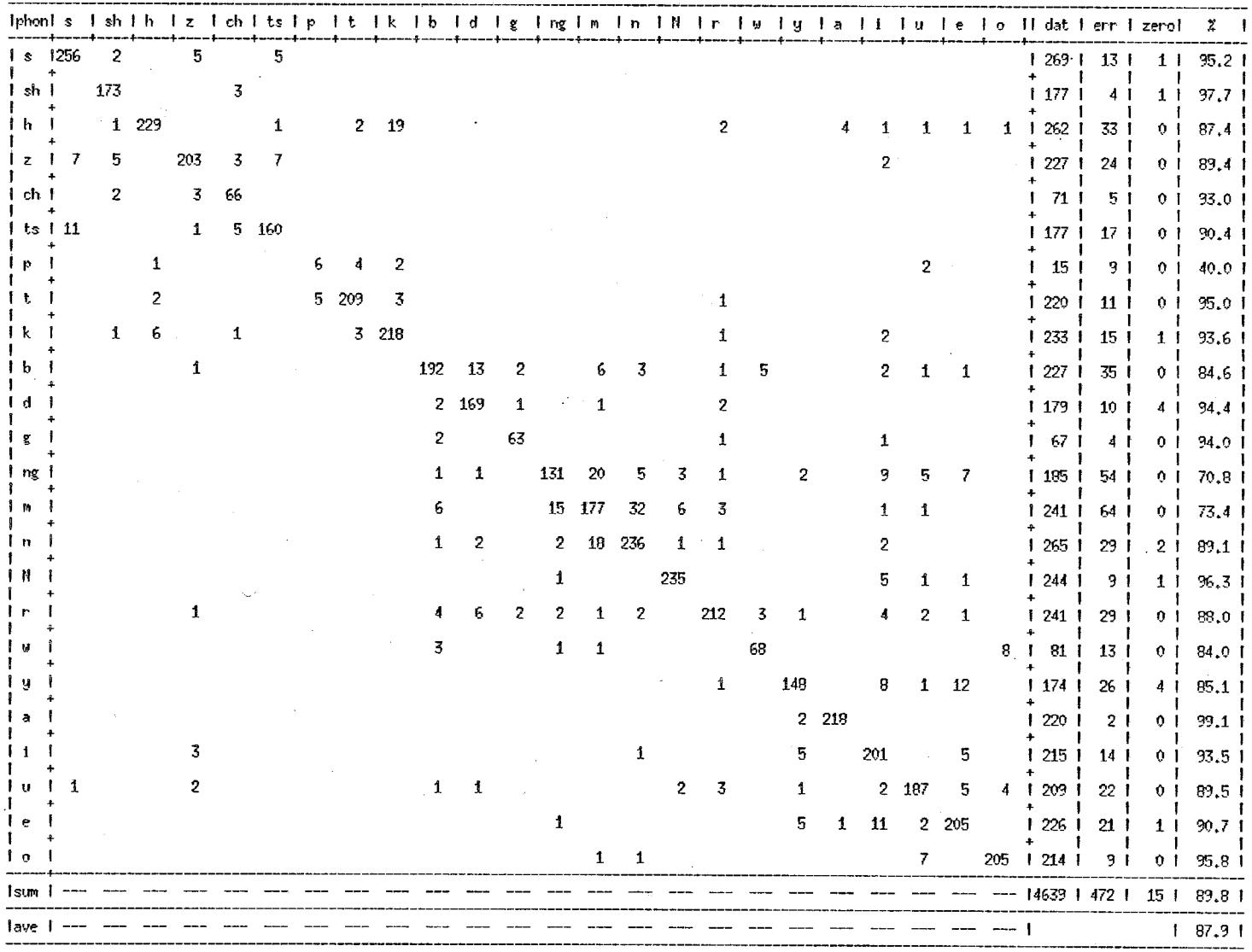
Fig-A-1. Confusion Matrix of All Japanese Phonemes
(without duration control)

phon	s	sh	h	z	ch	ts	p	t	k	b	d	g	ng	m	n	N	r	w	y	a	i	u	e	o	dat	err	zero	%
s	241	1		5	3	18																			269	28	1	89.6
sh		162		1	13																				177	15	1	91.5
h		1	216			1		4	29								2			6	1		2		262	46	0	82.4
z	6	4		201	5	10															1				227	26	0	88.5
ch		2		3	66																				71	5	0	93.0
ts	12				5	160																			177	17	0	90.4
p			1				6	4	2														2		15	9	0	40.0
t			4				5	209	1								1								220	11	0	95.0
k		1	5		1		1	5	217								1				1				233	16	1	93.1
b										184	15	1		8	4		3	8			2	1	1		227	43	0	81.1
d										2	165			1	1		6								179	14	4	92.2
g										5	1	58			1		1				1				67	9	0	86.6
ng										1	1		133	22	4		2		2		10	3	7		185	52	0	71.9
m										6			14	182	31	2	3	1			1	1			241	59	0	75.5
n											2	1	3	17	234	1	4				1				265	31	2	88.3
N													1			233					7	1	1		244	11	1	95.5
r				1						3	8	2	2	1	2		211	4	1		4	1	1		241	30	0	87.6
w										4			1	1				67						8	81	14	0	82.7
y			1											1	1		2		151		3		11		174	23	4	86.8
a			2															3	2	213					220	7	0	96.8
i				4	1										1		2		7		196		4		215	19	0	91.2
u	1			1		1			1	3	1				1	2	5		1		2	181	5	4	209	28	0	86.6
e													2						7	2	13	2	199		226	27	1	88.1
o										1				2			1					4		206	214	8	0	96.3
sum	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	4639	548	15	88.2
ave	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---				86.3

Fig-A-2. Confusion Matrix of All Japanese Phonemes
(using phoneme duration control)

phon	s	sh	h	z	ch	ts	p	t	k	b	d	g	ng	m	n	N	r	w	y	a	i	u	e	o	dat	err	zero	%
s	254	2		5	1	7																			269	15	0	94.4
sh		173		1	3																				177	4	0	97.7
h		2	227			2		3	19												4	1	1	3	262	35	0	86.6
z	13	14		181	7	11																1			227	46	0	79.7
ch		3		3	65																				71	6	0	91.5
ts	12				5	160																			177	17	0	90.4
p			1				6	4	2														1	1	15	9	0	40.0
t			3				5	209	2								1								220	11	0	95.0
k		5	6		5			1	214													2			233	19	0	91.8
b				1						192	12	2		6	3		2	5				2	1	1	227	35	0	84.6
d								2		3	170	1		1			1				1				179	9	0	95.0
g										2		64										1			67	3	0	95.5
ng										3	1		125	24	4	3	2		2			9	3	8	1	185	60	0	67.6
m			1							6			15	176	34	6	1					1	1			241	65	0	73.0
n										1	1		2	17	236	5	2					1				265	29	0	89.1
N													2			234			1			5	1	1		244	10	0	95.9
r			1	1						5	6	2	4	2	9		183	4	3	4	6	2	8	1	241	58	0	75.9
w										3			1	1				68							8	81	13	0	84.0
y																	1		152			8	1	12		174	22	0	87.4
a																			1	219						220	1	0	99.5
i		1		2					1						1				8		197			5		215	18	0	91.6
u	4			2					1	1	1		1			2			2			2	180	9	4	209	29	0	86.1
e													2						5	1	11	2	205			226	21	0	90.7
o														1	1								6		206	214	8	0	96.3
sum	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	14639	543	0	88.3
ave	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---				86.6

Fig-A-3. Confusion Matrix of All Japanese Phonemes
(using state duration control)



phon	s	sh	h	z	ch	ts	p	t	k	b	d	g	ng	m	n	N	r	w	y	a	i	u	e	o	dat	err	zero	%
s	1256	2		5		5																			269	13	1	95.2
sh		173			3																				177	4	1	97.7
h		1	229			1		2	19								2			4	1	1	1	1	262	33	0	87.4
z	7	5		203	3	7															2				227	24	0	89.4
ch		2		3	66																				71	5	0	93.0
ts	11			1	5	160																			177	17	0	90.4
p			1				6	4	2													2			15	9	0	40.0
t			2				5	209	3								1								220	11	0	95.0
k		1	6		1			3	218								1				2				233	15	1	93.6
b				1						192	13	2		6	3		1	5			2	1	1		227	35	0	84.6
d										2	169	1		1			2								179	10	4	94.4
g										2		63					1				1				67	4	0	94.0
ng										1	1		131	20	5	3	1		2		9	5	7		185	54	0	70.8
m										6			15	177	32	6	3				1	1			241	64	0	73.4
n										1	2		2	18	236	1	1				2				265	29	2	89.1
N													1			235					5	1	1		244	9	1	96.3
r				1						4	6	2	2	1	2		212	3	1		4	2	1		241	29	0	88.0
w										3			1	1				68						8	81	13	0	84.0
y																	1		148		8	1	12		174	26	4	85.1
a																			2	218					220	2	0	99.1
i				3											1				5		201		5		215	14	0	93.5
u	1			2						1	1					2	3		1		2	187	5	4	209	22	0	89.5
e													1						5	1	11	2	205		226	21	1	90.7
o														1	1							7		205	214	9	0	95.8
sum	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	4639	472	15	89.8
ave	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---				87.9

Fig-A-4. Confusion Matrix of All Japanese Phonemes
(using phoneme duration and state duration control)