マイコンによる音声認識システムと中国語の認識

王　之庭*・高浪五男**・谷口　弘**・井上克司**

Construction of a Voice Recognition System with Microcomputer and it's Recognitive Ability of Chinese Speech

Wang Zhi-Ting, Itsuo TAKANAMI, Hiroshi TANIGUCHI and Katushi INOUE

Abstract

Hardware and software of a voice recognition system are constructed. It's hardware consists of a voice recognition board and a host computer FM-7. The former is constructed using the set of LSI's for processing words spoken by specific speakers, and a small number of standard IC's. The set of LSI's are MC4760, μPD7761D and μPD7762G made by NEC. According to the 12 commands reserved by the processing unit μPD7762G for voice recognition, the system software is programmed with the FBASIC. According to the number of registered standard patterns and/or the repetitive number of speeches per word or syllable, the exprimental results of identification accuracy are shown. When the number of standard patterns and the repetitive number are two respectively, the average accuracy of identification is above 94%, which seems to be a comparatively satisfactory result.

1．まえがき

コンピュータに中国語の文字を入力するのは世界の種種の文字の中で最も難しいことの一つであろう．現在，世界で用いられているコンピュータのキーボードはアルファ・ニューメリック・キーとシンボリック・キーを含んでおり，キーの種類はおよそ60～70位である．しかし，中国の常用文字の数量は3515余りである．ほとんどの漢字は偏傍冠で索引できるが，その種類は189種にも達する．しかし，これだけでも含みえない漢字が僅かではあるが存在する．従って，現在中国ではほとんどローマ字による弁音字母（一種のローマ字による記法）で漢字を入力する．この方法は確かに良い入力方法であるが，オペレータに厳しい要求を課すことになる．すなわち，彼らは必ず正確な中国共通語の発音を身に付けなければならない．しかし，国土も広く，人口も多く，言語も複雑な中国においてこの点が短期間に達成されるのはかなり困難と考えられる．

上述したことから，もし直接に音声で中国語を入力できれば，これは画期的な意義をもたらすことになろう．中国における中国語の音声認識に関する研究はまだ初期の段階にあり，特定話者の離散発声についてなされているのが現状である．例えば，1986年7月，北京清華大学で音声で動く漢字ワードプロセッサが製作されている(清華大学音声認識研究グループ研究報告)．このシステムは896個の漢字または単語を入力でき，認識率が95%，平均応答時間が6秒であることが報告さ

*大慶石油学院計算機学科
**山口大学工学部電子工学科

れている．1987年1月には中国の研究者郭栄江氏は東京工業大学において中国語の音声理解システムを完成したことが報告されている（人民日報）．

ここでは，マイクロコンピュータを用いた安価な音声認識装置の製作とそれによる中国語の音声認識の実験結果について報告する．本システムは市販の安価なLSIを用いて製作されており，特定話者の離散発声用である．中国語の単母音6語と全子音21語の計27語について，さらに子音＋母音の全組み合わせ126組について，標準パターンの登録方法をいろいろ試み，比較的良好な認識率が得られている．

2．音声認識装置の概要

2.1　ハードウェア

本装置はFig. 1に示されるように，ホストコンピュータのFM-7と音声認識ボードとから構成されている．

音声認識ボードは，FM-7とのインターフェースを行う8255A，64Kバイトメモリー，及び音声認識用LSIセット（NEC製）から構成されている．このLSIセットはMC4760，μPD7761D，μPD7762Gの三つのユニットからなり，次のような特徴を持っている．

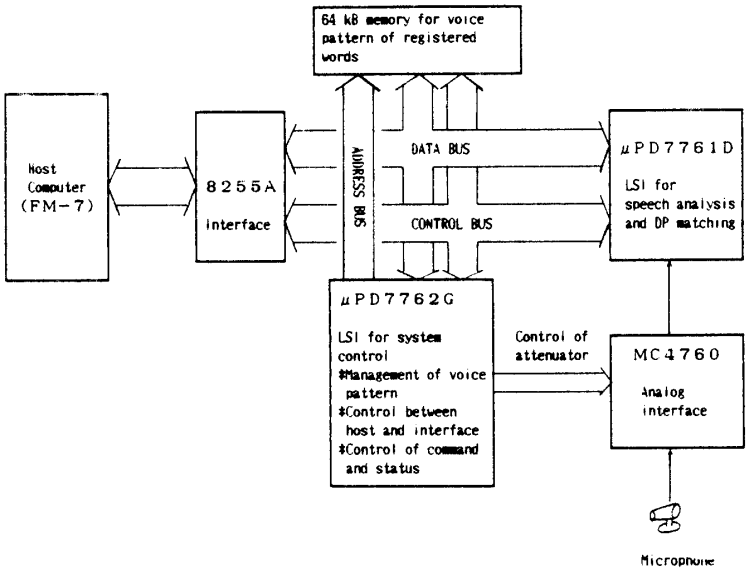


Fig. 1　Block diagram of system

●入力音声

入力単位は離散発声単語（単語間ポーズ時間は最小260mS）

入力音声長は0.2秒から2.0秒

特定話者用

●認識処理方式

特徴抽出　音声周波数分析（8 Ch．デジタル・フィルタリング処理）

認識処理　圧縮DPマッチング

応答時間　発声終了後平均0.5秒

登録単語数　1バンク（16Kバイトのメモリ容量時）当り128語．4バンクまで拡張可能

●システム制御方式

ホストコンピュータからの簡単なコマンド・セット

Table 1(a)
A set of input commands

Command name
Initialize
Level Adjust
Training
Recognition
Second Decision
Hot Start
Down Load
Up Load
Change Reject
Memory Test
Bank Select
Word Reject

Table 1(b)
Meaning of status

Status	Meaning
Acknowledge	Normally ended
Level over	Voice input level is over
Level Under	Voice input level is lower
Time Over	Voice is longer than 2 sec.
No Adjust	Input level of voice for the selected bank is not adjusted
No Syntax	Fitted syntax group is not yet registered
No Pattern	There is no fitted word
No Data	There is no fitted data
No Bank	specified bank is not prepared
Comand Error	Input command is error
Reject Error	Error of word reject value
Time Under	Voice is shorter than 0.2 sec.
I/O Error	Error in I/O

トと音声認識ボードからのステータス情報のやりとりで制御する．Table 1に，これらのコマンドとステータス情報を示す．

Fig．2はシステム中の信号の流れを示す．

ホストコンピュータ(FM-7)と音声認識ボードをつなぐインタフェース8255Aはグループ A をモード 2 (双方向モード) に設定して使用する．このインタフェースの初期化，及びホストコンピュータとのデータの受け渡しの流れをFig．3に示す．

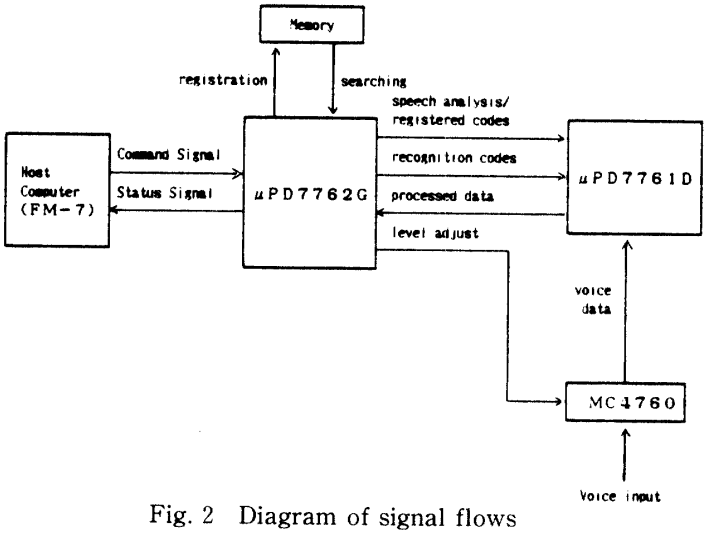


Fig. 2　Diagram of signal flows

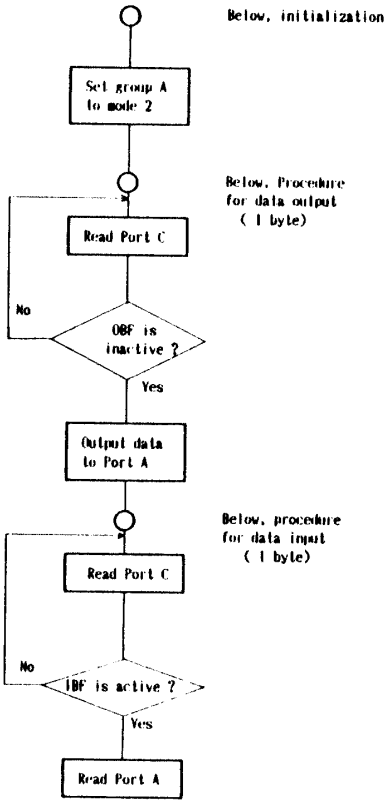


Fig. 3　Initialization of 8255A and it's procedure for data input and/or output with host computer

2.2　ソフトウェア構成

音声認識の基本フローチャートをFig．4に示す．各コマンドはそれが正常に実行されたことをステータス情報によって確認されると，次のコマンドに移る．以下に各ステップを概説する．

a）イニシャライズ

メモリ及びシステム全体を初期化するものである．また，メモリは16Kバイト毎のバンク単位で使うので，バンク・セレクトをし，更にそのメモリのリード・ライト可能かを確認するメモリ・テストも行う．

b）音声レベル調整

音声の入力レベルを調整するものである．このレベル調整は一度だけ行えばよい．これを行った後でないと音声の登録や認識はできない．

c）音声の登録

初めて音声を登録する場合はトレーニング・コマンドを，ホスト側にあらかじめ登録パターンが格納されている場合はダウンロード・コマンドをそれぞれ実行する．

トレーニング・コマンドを実行した場合は，その後入力される音声は自動的に標準パターンとして登録される．

d）音声の認識及び認識応答

リコグニション・コマンドを正常に終了したときはその旨を示すステータス情報に続いて，認識したパター

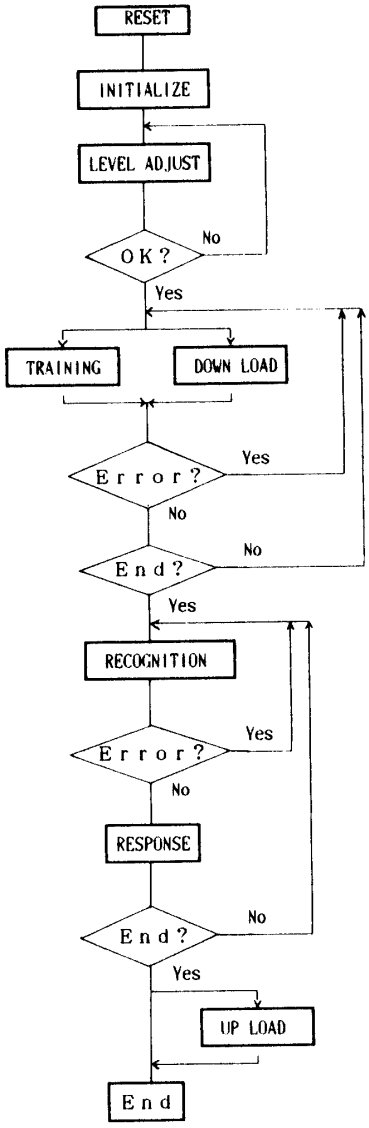


Fig. 4　Flow chart of speech recognition

ンの登録番号と，入力した音声パターンと認識した登録パターンとの距離のデータが返って来る．

セカンド・デシジョン・コマンドの場合は，データが有ればステータス情報に続いて，第2近似パターンの登録番号と距離のデータを返して来る．

e）登録パターンの保存

登録されたパターン・データをホスト側に送り，保存するためのコマンドである．登録単語数，リジェクト・テーブル，登録番号，辞書テーブル，標準パターン・テーブルの各データが音声認識ボード側より送られてくるので，これらのデータをホストコンピュータの外部記憶部に格納する．

以上の基本動作を行うためのソフトウエアはすべてFBASICで書かれており，夫々のコマンドがファンクションキーに割付けてあり，ワンタッチで操作できるようになっている．

3．認識結果

本システムの中国語の認識性能を調べるため，中国語の単語と音節（音節の四声を含む）の認識率について夫々調べた．

(1)　単語の認識特性

音声認識ボードに用いられているLSIセットは本来単語認識用であり，中国語の単語の認識率はかなり高い．例えば次のとおりである．

57個の中国の省と市名について，夫々20回づつ発声したところ，その平均正当率は94.5%であった．

中国語の数字0から100まで，及び1千と1万を含む合わせて103個の数字について，上と同じ方法で行った結果，その平均正当率は92.8%以上であった．

以上のように，単語の認識については比較的良い結果を得たので，以下に，情報量の少ない単音節の場合について詳しく調べることにする．

(2)　単音節の認識特性

単音節は単語に比し，発声長も短く，従って情報量も少ないので，認識が困難になることが予想される．実際，実験してみるとそのとおりであり，良い認識率を得るには，先ず次のことが大切であることが分かった．

a)標準パターンを作るとき，工夫する必要がある．最初に各単音節について発声を繰り返し練習し，なるべく正確に発声し，また発声速度もほぼ固定（ほぼ60音節／分程度）して標準パターンの候補を登録し，その中で良い認識率が得られるものを標準パターンとして登録する．

b）認識を行うときは，なるべく高い認識率が得られるように発声し，それを再現性良く行う．

Table 2は上記のことに注意しながら，6個の単母音と21個の子音について，それぞれ一つの単音節に一つの標準パターンを登録し，その認識結果を調べたものである．その結果，この方法に対する認識率は60%程度と低く，この方法ではほとんど実用性がないと思われる．この原因はb(o)，p(o)，m(o)，f(o) などの音節では，どの音も最後は"o"の音になるので，これらの音を区別する鍵は最初の部分の短い発声長の子音部になる．単音節de, te, ne, leなどについても同様であり，音声特徴が似通ったものになるためである．

本システムのハードウエア上の性能は音声認識用のLSIによって決まってしまっているので，さらに認識率

を改善するにはソフトウエア上で工夫するしかない．ここでは，標準パターンの登録について以下の三つの方法を試みた．

1）同一単音節を異なる登録番号で複数登録する方法．

Table 2 The case when the duplication No. of registered patterns is 1 and the repetition No. of speech is 1.

A : single vowel, B : recognition rate of A
C : consonant , D : recognition rate of B

A	B	C	D
a	70	b(o)	50
o	70	p(o)	70
e	60	m(o)	50
i	70	f(o)	50
u	50	d(e)	60
ü	70	t(e)	70
		n(e)	70
		l(e)	70
		g(e)	50
		k(e)	70
		h(e)	70
		j(i)	50
		q(i)	60
		x(i)	70
		z h(i)	50
		c h(i)	70
		s h(i)	60
		r(i)	70
		z(i)	50
		c(i)	60
		s(i)	70

average recognition rate is 62.2%

Table 3は各単音節に対し3個づつ登録した場合の認識結果を示す．認識率が83%程度になることが分かった．この様に，この方法はある程度有効な方法であると思われるが，同一単音節を複数個登録するため1バンク当りの登録単語数がその分減少し，それだけ同一メモリ量に対し登録単語数が減ることになる．

2）各単音節を繰り返し数回連続して発声する方法．例えば，"aaa", "ooo", shishishi"…のように発声する．この場合は3回連続して発声したことになる．

Table 4は例のように3回連続して発声した場合の認識結果を示している．単音節の認識率が94%程度になっていることが分かる．この方法もかなり有効な方法であるが，単語の発声時間が3倍長くなること，さらに，3回も繰り返すことは比較的煩わしいことなどの欠点がある．

Table 3 The case when the duplication No. of registered patterns is 3 and the repetition No. of speech is 1.

A : single vowel, B : recognition rate of A
C : consonant , D : recognition rate of B

A	B	C	D
a	80	b(o)	70
o	80	p(o)	90
e	80	m(o)	70
i	90	f(o)	70
u	80	d(e)	80
ü	90	t(e)	90
		n(e)	90
		l(e)	90
		g(e)	80
		k(e)	90
		h(e)	90
		j(i)	80
		q(i)	80
		x(i)	90
		z h(i)	80
		c h(i)	90
		s h(i)	90
		r(i)	90
		z(i)	70
		c(i)	80
		s(i)	90

average recognition rate is 83.3%

Table 4 The case when the duplication No. of registered patterns is 1 and the repetition No. of speech is 3.

A : single vowel, B : recognition rate of A
C : consonant, D : recognition rate of C

A	B	C	D
a	100	b(o)	80
o	100	p(o)	100
e	80	m(o)	100
i	100	f(o)	80
u	90	d(e)	80
U	100	t(e)	100
		n(e)	100
		l(e)	100
		g(e)	90
		k(e)	100
		h(e)	100
		j(i)	80
		q(i)	90
		x(i)	100
		z h(i)	80
		c h(i)	90
		s h(i)	90
		r(i)	90
		z(i)	80
		c(i)	90
		s(i)	100

Average recognition rate is 94.8%

Table 5 The case when the duplication No. of registered pattern is 3 and the repetition No. of speech is 3.

A : single vowel, B : recognition rate of A
C : consonant, D : recognition rate of C

A	B	C	D
a	100	b(o)	90
o	100	p(o)	100
e	90	m(o)	100
i	100	f(o)	90
u	100	d(e)	100
ü	100	t(e)	100
		n(e)	100
		l(e)	100
		g(e)	100
		k(e)	100
		h(e)	100
		j(i)	90
		q(i)	100
		x(i)	100
		zh(i)	100
		ch(i)	100
		sh(i)	100
		r(i)	100
		z(i)	90
		c(i)	90
		s(i)	100

average recognition rate is 97.7%

Table 6 The case when the duplication No. of registered pattern is 1 and the repetition No. of speech is 2.

A : single vowel, B : recognition rate of A
C : consonant, D : recognition rate of C

A	B	C	D
a	100	b(o)	80
o	100	p(o)	100
e	90	m(o)	90
i	90	f(o)	90
u	100	d(e)	80
ü	100	t(e)	90
		n(e)	100
		l(e)	100
		g(e)	90
		k(e)	100
		h(e)	100
		j(i)	90
		q(i)	90
		x(i)	90
		zh(i)	80
		ch(i)	100
		sh(i)	90
		r(i)	100
		z(i)	80
		c(i)	80
		s(i)	100

Average recognition rate is 92.6%

3）上記の1）と2）を組み合わせる方法.

Table 5は(標準パターンを)三個登録の三回発声の場合，Table 6は一個登録の2回連続発声の場合，Table 7は2個登録の2回連続発声の場合の認識結果である．3個登録の3回連続発声の場合と2個登録の2回連続発声の場合の認識率はいずれも96%程度であり，これらの認識率はほとんど等しいことがわかる．これらの結果から，2回登録の2回連続発声法が最も良い方法と考えられる．

Table 7 The case when the duplication No. of registered pattern is 2 and the repetition No. of speech is 2.

A : single vowel, B : recognition rate of A
C : consonant, D : recognition rate of C

A	B	C	D
a	100	b(o)	90
o	100	p(o)	90
e	90	m(o)	100
i	100	f(o)	100
u	100	d(e)	90
ü	100	t(e)	100
		n(e)	100
		l(e)	100
		g(e)	90
		k(e)	100
		h(e)	100
		j(i)	90
		q(i)	100
		x(i)	100
		zh(i)	90
		ch(i)	100
		sh(i)	100
		r(i)	100
		z(i)	90
		c(i)	90
		s(i)	100

Average recognition rate is 96.6%

Table 8とTable 9は子音の場合の結果である．子音の認識率は後続の母音に依存しており，特に，後続母音がaをもつ単音節の認識率が高く，u，üの場合は低い．これは前者では口の開きが大きく，後者では小さいなど口の開きの大小による差と考えられる．このほか，音声のエネルギーも小さく，また発声の度に子音の雑音成分のスペクトルや強度の変化し易い無気子音b，d，g，i，zh，zなどの認識率が比較的低くなっているのが見られる．

子音間の誤認の傾向を見ると，bとp，jとg，zhとchとsh，zとc，hとk，mとnの間で顕著である．

Table 8　The recognition rate of consonants in the case when the duplication No. of registered patterns is 1 and the repetition No. of speech is 2

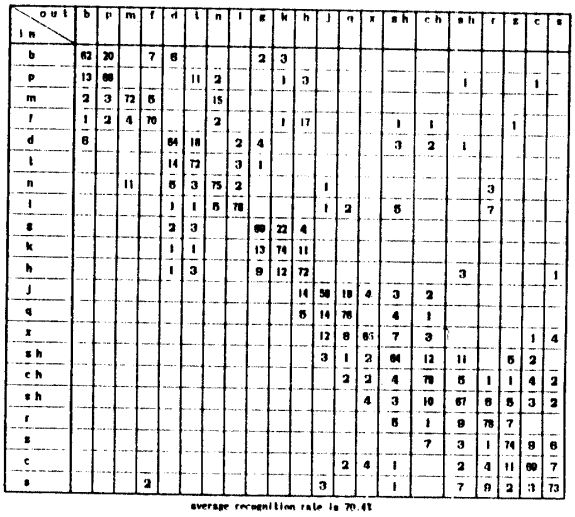
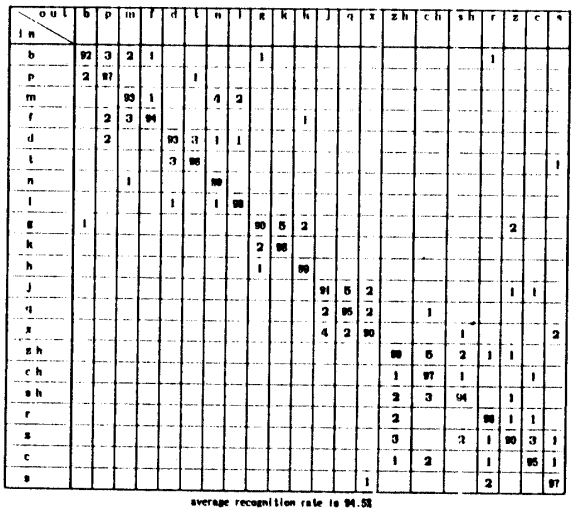


Table 9　The recognition rate of consonants in the case when the duplication No. of registered patterns is 2 and the repetition No. of speech is 2



これは，これらの間の音声が酷似しているためと考えられる．

(3)　中国語の声調の識別結果

中国語には一つ一つの音節に平，高，下げ上げ，低など四種類の声調，いわゆる四声がある．これは音を区別し，したがって言葉の意味そのものを区別するための重要な役割を果たす．Table 10はこの四声の認識結果をまとめたものである．この結果から，第三声と第四声で認識率が高く，第一声と第二声で低いことが分かった．

Table 10　Recognition rate of the four tones

vowel \ tone	1st tone	2nd tone	3rd tone	4th tone
a	86.4	83.6	97.1	92.3
o	84.6	85.7	95.8	92.1
i	83.5	78.2	92.5	90.7
e	80.3	77.8	93.4	91.4
u	82.6	83.2	93.9	92.2
ü	81.5	84.7	92.3	91.5
average	83.2	82.2	94.3	91.7

4．あとがき

音声で言葉を入力するとき，日本語の場合と中国語の場合について有利な点と不利な点を考えてみると，ちょうど反対になっているようである．すなわち，日本語の単音節の種類は比較的少なく，大体68種程度であるが，一方，一字多音である．例えば「生」という字は20種類の読み方がある．これに対し，中国語ではほとんどの場合一つの漢字には一つの読み方しかない．二つ以上の音を持った漢字はまれである．このことは一つの音節から決まる漢字の数が少ないので，中国の言葉の入力にとって有利になる．しかし，音節の種類からみると407種，四声を含めば1357種にも達し，もし，これらの千余りの音節をすべて音声で入力すれば，登録パターンもそれだけ多くなり，従って認識率も低くなり，認識の応答時間も長くなるなど中国語の入力にとって問題点となる．

これらのことから，漢字の入力音声の標準パターンの作り方の工夫や，認識率の低い単音節の認識率をさらに向上させるなどの研究が残されている．

参　考　文　献

1）磯山　隆：LSIセットを活用—音声認識で動くワード・プロセッサ，トランジスタ技術，CQ出版社，第22巻，第3号，427-434，同　第4号，433-440 (1985)

2）伊福部達：音声タイプライタの設計，CQ出版社 (1983)

3）上田次郎：池田雄一郎，音声認識とその活用技術，センサ・インターフェーシング，No.，，CQ出版社，6-28 (1983)

4）Reddy. D. R(ed)："Speech Recognition" Acadedmic Press, New York, (1979)

5) 森下, 小畑：信号処理, 計測自動制御学会 (1982)
6) 迫江博昭：音声認識における動的計画法の応用, bit Vol. **15**, No. 8, 131-142 (1983)
7) 川嶋弘尚：音声・画像信号の圧縮と音声合成, bit Vol. **15**, No. 8, 143-160 (1983)
8) H. Sakoe and S. Chiba："Dynamic programming algorithm optimization for spoken word recognition", IEEE Trans. on ASSP vol. **26**, 1., 78-82 (1978)
9) 鐘ヶ江信光：中国語の学習 (1972)

(昭和62年4月15日受理)