統計的手法およびトップダウン情報を用いた 男女声対象の自動ラベリング

正員 小森 康弘 (ATR 自動翻訳電話研究所)

非会員 小坂 哲夫 (キャノン)

非会員 大洞 恭則 (キャノン)

正員 新中 新二 (キャノン)

1. はじめに

現在、音声認識の研究は、不特定話者・大語彙を対象にした連続音声認識の実現を目指している。この研究を行うためには、音声のデータベース⁽¹⁾⁻⁽⁷⁾が必要不可欠である。しかし、音声のデータベースを構築するには3つの大きな解決すべき課題がある。

その1：不特定話者・大語彙を対象にした連続音声認識の研究用データベース構築の際のラベルには、どのような単位を用いるか？

その2：不特定話者・大語彙を対象にした連続音声認識の研究用データベース構築の際のラベリングにかかる莫大な手間をどうするか？

その3：ラベリングにかかる莫大な手間を軽減するために自動化するならば、不特定話者や環境による音韻の変動にどう対処するか？

その1のラベル単位に関しては、いくつかの研究機関⁽⁸⁾⁻⁽¹⁰⁾を除いて、ほとんどの研究機関⁽⁸⁾⁻⁽⁹⁾では、音素や音素に準じたものを用いている。しかし、音素は、音声の持つ言語的情報の最小単位にはほぼ等しいが、音響的特徴を十分に表すことができない欠点がある。音声認識研究用の音声データベースのラベルにおいては、ラベルのカテゴリは、音素よりも更に細かい音響的にコンパクトな(物理的に特徴の類似した)単位で、

しかも、言語的情報としても扱いやすいラベルを用いることが望ましい。

その2のラベリングの手間に関しては、幾つかの研究機関⁽⁸⁾⁻⁽¹¹⁾で自動化の試みがあるものの、一般にラベルを人手により入力するという莫大な手間がかかる方法⁽⁵⁾で作成しているのが現状である。この際、人手によりラベリングすることによる個人差の問題も生じる。

その3に関しては、不特定話者や環境による音韻の変動に対してマルチパターンを用いて解決するのか、統計的手法でその分散を吸収するのか、のいずれかが考えられる。

上述の3つの問題点を考慮したラベリングシステムには、電総研田中ら⁽¹⁰⁾による研究がある。このシステムは、標準パターンを音素片という音響的にコンパクトな単位を用いて音素片ラベルの自動ラベリングを実現している。また、不特定話者や環境による音韻の変動に対して、音素片標準パターンのマルチパターン化により対処している。しかし、異なった環境下の音素片が現れたときは、手動ラベリングにより新しい音素片標準パターンを登録する必要がある、この意味で、このシステムは半自動ラベリングシステムである。

我々は、ラベルとして、言語的情報である音素との対応が取りやすく冗長性も少ない音響的にコンパクトな音響境界ラベル (Acoustic-Boundary-Label: ABL) を用いる。更に、統計的手法を用いてあらかじめ設計した標準パターンで特徴抽出を行い、トップダウン情報と組み合わせることにより、男女性の不特定話者に対処した自動ラベリング手法⁽¹²⁾⁽¹³⁾を提案する。

本論文では、まず、提案する自動ラベリング方法について述べ、続いて評価実験を行い本手法の有効性を

An Automatic Labeling Method for Male and Female Utterance Using Statistical Technique and Top-Down Information. By Yasuhiro Komori, Member (ATR Interpreting Telephony Research Laboratories, Osaka), Tetsuo Kosaka, Non-member, Yasunori Ohara, Non-member & Shinji Shinnaka, Member (Canon Inc. Information Systems Research Center Tokyo).

小森康弘：正員，ATR 自動翻訳電話研究所（大阪府）

小坂哲夫：非会員，キャノン（株）情報システム研究所，東京都

大洞恭則：非会員，キャノン（株）情報システム研究所，東京都

新中 新二：正員，キャノン（株）情報システム研究所，東京都

示す。

2. 自動ラベリング手法

〈2・1〉 自動ラベリング手法の特徴 本自動ラベリング手法は、以下の特徴を有する。

(1) 音響的レベルと言語的レベルのいずれでも扱いやすい特徴を持ち、音響的にコンパクトな ABL でラベリングを行う。ABL は、音響的特徴の変化点につけ、表1に示す記号を用い、2つの記号を組み合わせることにより1つのラベルを構成する(図1参照)。このことにより、音素と ABL 間の対応も簡単に取れる。すべての音韻環境は、400 余個の ABL で表現される。

(2) 線形判別関数⁽¹⁴⁾⁽¹⁵⁾、KL 変換⁽¹⁶⁾とベイズ判定⁽¹⁶⁾の3種類の統計的手法を用いてあらかじめ設計した標準パターンを利用することにより、不特定話者や環境による音韻の変動に対処する。標準パターンは、男女性10~40名の212音韻バランス単語より作成した。

表1 ABL に用いる記号

Table 1. Symbols used in ABL.

番号	名称	記号
1	母音	A, O, U, I, E
2	無声化母音	a, o, u, i, e
3	撥音	X
4	無音	Q
5	半母音	J, W
6	拗音	Y
7	無声化拗音	y
8	鼻音	M, N, ー(鼻濁音)
9	流音	R
10	有声破裂音	B, D, G, V
11	有声摩擦音	Z
12	無声破裂音	P, T, K
13	無声摩擦音	S, C, H, F
14	バズ音	*
15	音声区間開始	\$
16	音声区間終了	\$

図1 ラベルとラベル位置の例

Fig. 1. Example of labeling points by ABL.

(3) ラベリングは、大まかな特徴からだんだんに細かな特徴へと、トップダウン情報を用いて絞り込みながら進めていく階層構造をしている。トップダウン情報は、5箇所・2通り(特徴の選択・特徴の位置決定)に用いる。

(4) 辞書表記が、音素記号であるため、任意の文節・単語・単音節などのラベリングが可能であり、辞書の追加・変更も容易である。

〈2・2〉 自動ラベリング手法の概要 自動ラベリング手法の概要を述べる(図2参照)。

(1) 音声データベースのインデックスを用いて、音声データおよびそのデータの音素記号列が引き出される。音声は、表2のように分析する。

(2) 音素記号列のモーラ、音素情報とデータ長⁽¹⁷⁾から非常に大まかな各音素の存在可能な区間を決定する(位置のトップダウン情報①)。

(3) 線形判別関数と(2)の音素存在可能区間(特徴選択のトップダウン情報②)を用いてフレームごとに音素レベルの特徴の存在可能性を抽出し音素記号ラティスを求める。個人差の大きい母音では、あらかじめ

図2 自動ラベリング手法の構成

Fig. 2. Structure of the automatic labeling.

表2 音声の分析方法

Table 2. Speech analysis condition.

音声波形	9kHz L. P. F., 24kHz sampling, 12 bit A/D
フィルタ構成	浮動小数点演算 (32 bit) digital-filter
チャネル数	29 channel
中心周波数	250~6,300 Hz
周波数間隔	1/6 octave
フィルタ特性	単峰性 (Q=6.0)
分析周期	10ms/frame

図 3 各種特徴量の出力例「ひとつぶだね」

Fig. 3. Example of parameter features outputs of data /hitocubudane/.

め話者ごとに線形判別関数の学習⁽¹⁸⁾を行う。

(4) 音素記号列 (位置のトップダウン情報③) と音素記号ラティスの記号レベルの DP を行い音素レベルの特徴の区間を決定する。

(5) KL 変換とベイズ判定の2種類の統計的手法により表現した時間周波数パターン (TSP) の標準パターンを用いることによりセグメント特徴をフレームごとに抽出し、セグメント特徴ラティスを求める。

(6) セグメント特徴ラティスに対して(4)の音素レベルの特徴を大まかな位置情報 (特徴選択のトップダウン情報④) として用い、セグメント特徴の選択・正規化を行う。

(7) 選択・正規化されたセグメント特徴ラティスと音素記号列から求めたセグメント特徴記号列 (位置のトップダウン情報⑥) の DP を行い、セグメント特徴レベルの特徴の区間を決定する。

(8) ABL 位置決定規則により(7)のセグメント特徴レベルの特徴の区間から ABL に最も適当な位置を決定し、データベースに格納する。

以下、各手順の詳細とここで用いた統計的手法について述べる。

〈2.3〉 音素存在区間の推定 一般的に、一単語長程度の発声におけるモーラ継続時間⁽¹⁷⁾は、語頭で短く語尾で長くなる傾向がある。このモーラ継続時間の伸縮が線形に変化すると仮定すると、一発声の継続時間とモーラ数より各モーラの大まかな存在区間を推定できる。これより音素の存在可能な区間が大まかではあるが推定できる。

〈2.4〉 音素レベルの特徴の検出 音素レベルでの特徴検出には、基本的に線形判別関数を用いる。線

形判別関数は、少ない演算で、精度良く特徴を抽出するという利点があり、よく用いられる。線形判別関数は、ある対立する2つのクラスを設定しその両者を最も良く分離できる特徴量を見付け出す方法である。ここでは10種類の線形判別関数⁽¹⁴⁾⁽¹⁵⁾(5母音性・有聲/無声性・鼻音性・摩擦性・有聲破裂性・無聲破裂性)と対数パワー、高域と低域のパワー比を用いる。音素の存在可能な区間内において、その音素の有すべき特徴が存在するか否かにより、その音素の存在をフレームごとに判定し、音素記号ラティスを出力する。図3-(6)に、音素のフレームごとの判定の例を示す。また図3-(7)に、判別関数の出力例を示す。

〈2.5〉 母音の学習 話者ごとに特徴の変動が大きい母音は、線形判別関数の学習を行い不特定話者に対処する。学習用のテキスト (表3参照) は、いずれも1種類の母音しか含まれていないため、学習用サンプルを自動検出できる。学習方法は、パーセプトロン学習⁽¹⁸⁾と呼ばれる誤り訂正学習法を用いる。本学習により、母音定常部のコンフュージョンおよび、母音定常部以外での母音特徴の付加を大幅に減少できる (図4参照)。日本語においては母音区間が決定できれば、多くの場合、子音区間は標準パターンがなくてもほぼ決定できる。このため、本学習により、後段での記号レベルの DP で音素レベルの特徴区間の決定を可能とする。

表3 母音学習用リスト
Table 3. List of speaker adaptation.

/akasatanahamajarawa/	/ekesetenehemereee/
/ikisicinihimirii/	/okosotonohomojorowo/
/ukusucunuhumujuruu/

図4 母音線形判別関数の学習例「ひとつぶだね」

Fig. 4. Output example of vowel linear discriminant function before and after speaker adaptation /hitocubudane/.

(2.6) 音素レベルの特徴の区間決定 音素記号列と音素ラティスの記号レベルでの DP を行い、音素レベルの特徴の区間を決定する。DP の計算は次の漸化式で行う。

$$g(i, j) = d(i, j) + \max[\{g(i-1, j-1) + p_1\}, \{g(i-1, j) + p_2\}] \dots (1)$$

$d(i, j)$: i フレーム目の音素記号と j 番目の音素記号との得点 (音素記号間で決定する)

$g(i, j)$: 累積得点

p_1, p_2 : 音素と時間長に応じて決まるペナルティ p_1 は、音素の平均的な継続時間長より長い時加点、

表 4 セグメント特徴の種類
Table 4. Segment type features.

定常的特徴			
番 号	セグメント	対応する音素	セグメント特徴
1	母 音	/a/	A
2		/o/	O
3		/u/	U
4		/i/	I
5		/e/	E
6	無声化母音	無声化の/i//u/	+
7	無 音	/Q/	Q
8	バ ズ	/b//d//g/のバズ	*
9	鼻 音	/N/	X
10		/m/	N
11		/n/	M
12	流 音	/r/	R
13	摩 擦	/z/	Z
14		/h/	H
15		/s/	S
16		/c/	C
非定常的特徴			
17	半 母 音	/j/	j
18		/w/	w
19	有声破裂	/b/	b
20		/d/	d
21		/g/	g
22	破裂流音	/r/	r
23	鼻音の出わたり	/g/ (鼻濁音)	=
24		/m/	m
25		/n/	n
26	無 声 破 裂	/p/	p
27		/t/	t
28		/k/	k
29	摩擦破裂部	/z/	z
30		/c/	c
31	語 頭 母 音	/a/	a
32		/o/	o
33		/u/	u
34		/i/	i
35		/e/	e

図 5 セグメント特徴の検出

Fig. 5. Detection of segment type feature.

短い時減点。 p_2 は、その逆。

(2.7) セグメント特徴とその検出方法 セグメント特徴⁽²⁰⁾は、G. Fant らが提案した音声スペクトログラフを記述するための最小単位である。ここでは定常時特徴 16 種類・過渡的特徴 19 種類の計 35 種類 (表 4 参照) のセグメント特徴を設定する。セグメント特徴の検出に周波数情報と時間情報とを同時に表現できる時間周波数パターン⁽¹⁹⁾ (TSP) を用いる。TSP の音韻性に寄与する変動は、50 ms 以内に起こり、また、個人性・発話速度・音韻環境による変動は、50ms 以上に起こると仮定し、すべての特徴を 50ms に設定する。

図 5 に、セグメント特徴の検出法の構成図を示す。

図 3-(8) に、セグメント特徴の検出例を示す。

音声データは、帯域通過フィルタでまず周波数の特徴抽出を行う。続いて KL 変換を用いて 2 段階の次元圧縮を行う。KL 変換を適用することにより、1 段階ではスペクトル空間の冗長性を低減し、2 段階では計算量の削減および特徴の圧縮をし、冗長性も除去される。このことにより、突発的な変動に対しても影響を受けにくい安定した特徴を抽出でき、不特定話者や環境による音韻の変動に対しても高い識別率が期待できる。

次に、事後確率による各セグメント特徴の正規化を行う。音声の特徴を表す入力ベクトル x 、セグメント特徴 i の事後確率はベイズ則により以下のように求める。

$$P(i|x) = P(x|i) \cdot P(i) / \sum_{k=1}^N \{P(x|k) \cdot P(k)\} \dots (2)$$

$P(x|i)$: 確率関数, $P(i)$: 事前生起確率,

N : クラス数

セグメント特徴の出現確率が等しいと仮定すると,

$$P(x|i) = P(x|i) \prod_{k=1}^N P(x|k) \dots\dots\dots (3)$$

となり, 多数サンプルを用いれば $P(x|i)$ は, 多次元の正規分布であると仮定でき,

$$P(x|i) = \exp(D^2) / \{(2\pi)^{n/2} \cdot |S_i|^{1/2}\} \dots\dots\dots (4)$$

$$D^2 = -1/2 \cdot \{(x - m_i)^t \cdot S_i^{-1} \cdot (x - m_i)\} \dots\dots\dots (5)$$

n : 次元数, S_i : 共分散行列, m_i : 平均ベクトル

D^2 : マハラノビス距離

となる。

この方法により, セグメント特徴の検出をフレームごとに行う。本方法は, 各パタン内の分布を考慮しているため, 不特定話者や環境による音韻の変動をかなり吸収できる。

〈2・8〉 セグメント特徴の選択・正規化 上述のような方法を用いても, フレームごとに検出するセグメント特徴は, 図3-(8)のように, 多くの特徴の付加が存在する。音素レベルの特徴抽出の結果を用いれば, セグメント特徴の存在し得る区間を大まかに推定できる。これをトップダウン情報として用い, セグメント特徴の選択・正規化を行いラベリングに必要な特徴の検出を行う。その手順を以下に示す。

(1) i フレームから $\pm N$ フレーム以内の音素レベルの特徴を用いて, i フレームにおけるセグメント特徴の選択を行う。

(2) i フレームの音素も含めて前方向・後方向にそれぞれ最大3音素記号(計5音素記号)をサーチし,

表5 音素-セグメント特徴対応表
Table 5. Relation of phoneme and segment type feature.

音 素	セグメント特徴	音 素	セグメント特徴
/Q/	(Q)	/b/	{*}-b
/a/	A	/d/	{*}-d
/o/	O	/g/	{*}-[g/=]
/u/	U	/r/	{*}-[r/R]
/i/	I	/z/	{*}-Z
/e/	E	/s/	(+)-S-+
/j/	j	/h/	(+)-H-+
/w/	w	/c/	(Q)-C-+
/N/	(X)	/p/	(Q)-p-+
/m/	M-m	/t/	(Q)-t-+
/n/	N-n	/k/	(Q)-k-+

(), { } は, 語頭において省略可能, [/] は, いずれかを選択, 各記号は, 表4に準ずる。

i フレームにおけるセグメント特徴の選択情報とする。ただし, 音素記号に無音・無声音・有声音のグループの変わり目が存在すれば, その方向のサーチを打ち切る。

(3) 音素-セグメント特徴対応表(表5参照)を用いて(2)で求めた音素の情報を, セグメント特徴の情報に置き換える。

(4) i フレームにおけるセグメント特徴を(3)で求めたセグメント特徴の情報を基に, 選択・正規化を行う。具体的には, セグメント特徴の情報で選ばれない特徴には0を代入し, 選択された特徴はその中の事後確率が最大のものが1になるように正規化を行う。

図3-(13)のように, セグメント特徴を選択・正規化することにより, その他の無関係な競合する特徴を除去し, 各特徴が明確にセグメンテーションされる。

〈2・9〉 DP によるセグメント特徴の区間の決定

選択・正規化されたセグメント特徴と音素記号列から音素-セグメント特徴対応表(表5参照)を用いて求まるセグメント特徴記号列のDPを行い, 各セグメント特徴の存在し得る区間を決定する。DPの計算は, 以下の漸近化式で行う。

$$g(i, j) = d(i, j) + \max \left\{ \begin{array}{ll} g(i-1, j), & g(i-i, j-1), \\ \text{(継続)} & \text{(遷移)} \\ \{g(i-1, j-2) + p\} \end{array} \right\} \dots\dots\dots (6)$$

(脱落)

$d(i, j)$: i フレームにおける j 番目のセグメント特徴記号の正規化された得点 ($(0, 1)$)

$g(i, j)$: 累積得点

表6 ABL 位置決定規則

Table 6. Rules for defining positions of ABL.

無 音	パワーがある値より下がった位置をラベル位置とする。
母 音 抑 音 摩 擦 音 バズ 始 端 バズ 終 端	連続して出現するセグメント特徴から, その特徴のマハラノビス距離が最大となる値の K 分の1の値を求め, その値より大きい値を取る区間の始端と終端をラベル位置とする。ここで K は特徴グループにより異なる値をもつ。ただし, 母音の区間において, 無声音と判定された区間は無声化のラベルに変更し, 特徴が変わる位置をラベル位置とする。十分な定常区間が現れない場合, ラベルは脱落する。
半 母 音 始 音	セグメント特徴のマハラノビス距離が最大となる位置をラベル位置とする。ただし, 始音の区間において, 無声音と判定された区間は無声化のラベルに変更し, その特徴が変わる位置をラベル位置とする。
鼻音定常始端 鼻音出わたり	鼻音のセグメント特徴が出現した位置を鼻音定常の始端のラベル位置とし, 鼻音の出わりのセグメント特徴のマハラノビス距離が最大となる位置
流 音	パワー値が最小になる位置をラベル位置とする。
有聲破裂音	有聲破裂音のセグメント特徴のマハラノビス距離が最大となる位置をラベル位置とする。
無聲破裂音	無音から有音に移った位置をラベル位置とする。

表 7 話者ごとの代表的なラベルのラベリング率 (±30 ms 以内)

Table 7. Automatic labeling ratio of each speaker within ±30 ms deviation.

	速度	AVR.	vwIS	vwIE	NS	NE	nsIS	nsIE	ufrS	ufrE	vfrS	vfrE	buzS	buzE	QS	QE	v-pl	u-pl	Rdip
男性 1	7.79	93.3	93.0	92.6	87.5	84.0	83.3	80.0	98.6	98.4	100.	83.3	93.8	93.8	98.5	98.2	87.0	95.9	93.9
男性 2	8.05	95.0	97.8	93.0	87.0	73.1	83.3	95.9	97.3	100.	100.	100.	100.	95.8	100.	94.9	91.7	98.6	87.9
男性 3	6.47	94.9	95.7	91.3	86.4	84.0	95.8	95.3	100.	95.2	100.	83.3	82.5	97.5	100.	97.7	93.5	98.6	100.
男性 4	6.60	93.8	96.2	85.4	91.7	88.0	91.5	93.8	89.7	100.	83.3	100.	93.8	93.8	98.6	99.4	88.0	98.6	97.0
男性 5	6.74	95.1	95.6	92.7	76.0	80.0	87.9	85.4	100.	100.	100.	85.7	93.9	97.1	100.	98.3	96.0	100.	100.
女性 1	6.73	94.1	96.6	89.7	92.0	88.5	92.5	91.7	93.3	98.4	100.	100.	97.3	100.	98.7	96.6	100.	95.9	90.9
女性 2	5.69	94.7	94.5	92.5	88.0	88.5	88.9	90.9	95.9	96.7	100.	100.	91.7	97.3	100.	96.6	100.	98.6	97.0
女性 3	6.50	95.3	97.1	90.9	95.7	80.8	88.9	100.	98.7	100.	100.	100.	95.0	94.9	97.6	97.3	96.4	100.	87.9
女性 4	6.47	94.1	96.6	90.6	100.	72.0	92.2	90.9	98.6	93.4	100.	100.	94.6	97.3	94.1	94.4	96.6	98.6	87.9
女性 5	7.54	95.8	97.4	93.8	87.5	92.3	93.0	97.9	100.	96.7	100.	83.3	90.9	97.0	98.7	96.1	100.	96.7	88.2

速度: モーラ/秒, AVR.: 平均ラベリング率, vwIS: 母音定常開始, vwIE: 母音定常終了, NS: 撥音定常開始, NE: 撥音出わり, nsIS: 鼻音定常開始, nsIE: 鼻音出わり, ufrS: 無声摩擦開始, ufrE: 無声摩擦終了, vfrS: 有声摩擦開始, vfrE: 有声摩擦終了, buzS: バズ開始, buzE: バズ終了, QS: 無音開始, QE: 無音終了, v-pl: 有声破裂, u-pl: 無声破裂, Rdip: 流音
(注) 半母音・拗音は、除く

表 8 代表的なラベルの分布とラベリング率 (1 フレーム: 10 ms)

Table 8. Distribution and labeling ratio of automatic labeled ABL (1 frame: 10 ms).

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	else	脱落	±1	±3	総数
vwI S	15	42	70	439	867	1,642	120	115	67	35	19	1	172	76.6	96.7	3,432
vwI E	37	45	111	176	269	1,892	465	210	110	29	26	183	0	71.9	91.2	3,653
N S	3	3	3	17	31	97	36	19	11	5	4	10	14	68.6	89.5	239
N E	2	3	15	16	13	141	15	9	3	4	2	32	0	66.3	83.1	255
nsI S	11	18	18	50	150	238	34	6	1	1	3	26	0	75.9	89.4	556
nsI E	4	3	3	5	21	324	45	25	7	8	6	16	0	83.7	92.1	466
ufr S	0	2	4	16	108	388	147	53	0	3	2	7	0	86.8	98.1	741
ufr E	0	0	0	0	5	549	26	15	0	1	0	12	0	94.9	97.9	611
vfr S	0	1	0	2	2	27	13	7	2	0	0	0	2	78.9	98.2	57
vfr E	0	0	3	9	9	33	1	0	0	0	0	4	0	73.8	93.4	61
buz S	0	0	2	2	11	174	60	48	22	10	3	11	0	71.4	93.0	343
buz E	0	2	1	2	10	269	34	13	3	0	1	9	0	91.0	96.5	344
Q S	0	0	0	0	1	772	1	0	1	0	0	12	0	98.3	98.5	787
Q E	0	1	1	0	2	1,704	4	1	2	4	3	47	0	96.7	96.9	1,769
v-pls	0	3	2	7	22	197	18	5	3	2	1	7	0	88.8	95.1	267
u-pls	0	0	1	1	1	711	4	1	0	2	0	9	0	98.1	98.5	730
R dip	0	1	3	2	3	287	3	5	5	2	2	16	0	88.5	93.1	331

vwI S: 母音定常開始, vwI E: 母音定常終了, NS: 撥音定常開始, NE: 撥音出わり, nsI S: 鼻音定常開始, nsI E: 鼻音出わり, ufr S: 無声摩擦開始, ufr E: 無声摩擦終了, vfr S: 有声摩擦開始, vfr E: 有声摩擦終了, buz S: バズ開始, buz E: バズ終了, QS: 無音開始, QE: 無音終了, v-pls: 有声破裂, u-pls: 無声破裂, R dip: 流音 (注) 半母音・拗音は除く

p: 無音・バズは小さい負の値を与え脱落を認め、
その他の特徴は大きい負の値を与え脱落を認めない。

〈2・10〉 ABL のラベル位置の決定 ABL の位置は、求めたセグメント特徴の区間より ABL 位置決定規則(表 6 参照)を用いて、各特徴に応じて決定する。

3. 実験

〈3・1〉 実験データおよび評価方法 実験データの単語リストは、母音学習用 5 単語(表 3 参照)および電総研速水らが提案した VCV バランス単語⁽²¹⁾の初めの 95 単語(計 100 単語)を使用した。本実験データは、標準ボタン設計用とは異なる男女性各 5 名(計 10 名)が、1 秒あたり 6 から 8 モーラで防音室

内にて発声したものである。本手法の評価には、自動ラベリングされた ABL とその ABL を視察により訂正したものとのズレで行った。

〈3・2〉 評価および考察 表 7 に、話者ごとの代表的な ABL のラベリング率(±3 フレーム以内: 1 フレームは 10 ms)を示す。表 8 に、全データに対する ABL のズレを示し、±3 と±1 フレーム以内のラベリング率を示す。ABL の平均自動ラベリング率として、±3 フレーム以内においては 94.8% [14656/15461], ±1 フレーム以内においては 82.3% [12732/15461] を得た。ラベル位置の修正がされなかった ABL は、64.4% [9951/15461], であった。音声データベースを構築する際、1 フレームのズレが大きく影響するラベル(例: 破裂・鼻音出わり・流音)や、

あまり大きく影響されないラベル（例：母音・撥音・鼻音の定常）が存在する。前者の自動ラベリング率は、 ± 1 フレーム以内にて 91.2%，後者は、 ± 3 フレーム以内にて、93.2% であった。男女性各 5 名のどの話者に対しても 93% から 96% 程度の自動ラベリング率（ ± 3 フレーム以内）が得られ、男女性を含めた不特定話者対象に、比較的良好なラベリング率で安定な結果が得られた。

4. む す び

本論文では、音響的にコンパクトな ABL というラベルを用い、統計的手法とトップダウン情報を組み合わせた自動ラベリング手法について述べた。更に、男女性各 5 名（計 10 名）のデータを用いて自動ラベリング手法の評価を行い、本手法が男女性を含めた不特定話者に対して有効であることを示した。このことから、自動ラベリングのみでラベリング作業は完了しないものの、簡単な確認と修正作業だけで、ラベリングが可能となる。本手法を導入すれば、従来の手動ラベリングと比較して、高精度なラベルデータを短期間で作成することを可能にする。今後は、本システムを用いて音声データベースを充実させ、音声認識の研究を進める予定である。

謝辞 本研究に東北大学応用情報学研究センター城戸研究室の成果を利用させていただきました。記して、城戸教授に感謝申し上げます。

（昭和 63 年 4 月 4 日受付）

文 献

- (1) 板橋秀一：「単語音声データベース」，日本音響学会誌，41，10，723（昭 60）
- (2) S. Itahashi: "A Japanese Language Speech Database", *proc. ICASSP* 86, 7, 4, 321 (1986)
- (3) 中島・鈴木・大村：「音声研究用データファイル制御システム」，音声研究会資料，S 73-07，1（昭 48）
- (4) 秋場・入間野・金指・真船：「日本語音声認識研究用音声データベース」，音響学会講義集，1-4-22，43（昭 57-3）
- (5) 武田・勾坂・片桐・桑原：「音韻ラベルを持つ日本語音声データベースの構築」，信学技報，SP 87-19，25（昭 62-6）
- (6) 城風・牧野・城戸：「光ディスク上の大規模音声データベースを中核とした GP-IB ネットワーク」，信学技報，SP 87-42，47（昭 62-7）
- (7) 前田・浜口・芥子・溝口・角所：「音声データベース SPEECH-DB」，音声研究会資料，S 851-79，633（昭 57）
- (8) 相川・杉山・鹿野：「音声の自動ラベリング」，音響学会講義集，2-1-3，89（昭 58-10）
- (9) 盛合・牧野・城戸：「音素判別フィルタを用いた孤立発声単語の自動ラベリング」，音声研究会資料，S 85-54，413（昭 60-12）
- (10) 田中・速水・太田：「音声の音素片ネットワーク表現と時系列のセグメント化法を用いた自動ラベリング手法」，日本音響学会誌，42，11，860（昭 61）
- (11) 松尾・小坂・岡田・城戸：「セグメント特徴に基いた連続音声に対する自動ラベリング手法の検討」，音響学会講義集，1-4-22（昭 60-9）
- (12) 小森・小坂・大洞・内山・城戸：「連続音声の自動ラベリングについて」，音響学会講義集，2-5-8（昭 62-3）
- (13) 小森・小坂・大洞・麻生・新中：「ABL の自動ラベリング」，音響学会講義集，3-3-9，107（昭 62-10）
- (14) 川端・牧野・城戸：「判別分析による音素の特徴抽出」，信学論(A)，J 65-A，12（昭 57-12）
- (15) S. Makino, S. Homma & K. Kido: "Speaker Independent Word Recognition System Based on Phoneme Recognition for a Large Size (212 words) Vocabulary", *J. A. S. J. (E)*, 6, 3, 171 (1985)
- (16) 小坂・岡田・松尾・城戸：「連続音声の認識を目的としたセグメント特徴の抽出」，音声研究会資料，S 85-54，405（昭 60）
- (17) 勾坂・東倉：「規則による音声合成のための音韻時間制御」，信学論(A)，J 67-A，7，629（昭 59-7）
- (18) 川端・牧野・城戸：「パーセプトロン学習」，音声研究会資料，S 82-24，185（昭 57-6）
- (19) 井出・石森・岡田・牧野・城戸：「時間周波数ボタンを用いた子音認識」，第 8 回応用情報研究学センター，シンポジウム予稿集，43（昭 58）
- (20) G. Fant: "Speech Sound and Features" the MIT Press, ch. 2 (1973)
- (21) 速水・田中・横山・太田：「研究用音声データベースのためのバランス単語セットの作成」，電総研彙報，49，10，65（昭 60）