論　文

三角形の面積を用いた音素対のセグメンテーション*

安居院　猛**　細　村　宏**　中嶋正之**

Abstract

It is suggested to recognize connected speech through the decomposition of input sequences into phoneme-dyads (for example VC, CV continuum) as recognition units.

Feature vector sequences whose elements are spectra are thought to segment the phoneme-dyad from the connected speech.

The feasibility of the segmentation of the phoneme-dyads by using the area of a triangle whose vertices are these three feature vector sequences is discussed.

As consequences of the experiments, the following result became clear.

The area of the triangle is useful for the segmentation of phoneme-dyads from the connected speech.

1. はじめに

音声の認識においては，認識単位の選び方によって認識方法を大きく，2つの手法に分類することができる．一方は単語単位に認識する方法[1]であり，他方は音素単位に認識する方法[2]である．認識対象の単語数を，100 程度に限定した場合には，ともにかなり高い認識率を得ることができるが，単語数をさらに増した場合には，単語単位で認識する場合，用意すべき標準パタンがかなり増加し，認識の実現が困難になる．したがって，多くの単語を認識するためには，音素単位に認識するほうが有利と考えられる．しかしながら，連続音声中の音素は前後の音素の影響を受けて，単独で発声した場合の音素と比較すると，発声レベルでも音響レベルでも，その物理的性質が変ってくるという，いわゆる調音結合がある．そこで，調音結合を考慮した方法として，VCV 音節を認識単位に用い，VCV 音節ごとに認識を行う方法もかなり行われており[3]，また，2個の音素の組み合わせである音素対を用いる方法も提案されている[4]．上記の2種の方法については，後で詳しく比較検討するが，我々は，主として，音素対のほうが用意すべき標準パタンの数が少なくてすむことから，音素対単位に連続音声を認識する方法について，検討を進めている．連続音声を音素対単位に認識するためには，音素対をセグメンテーションしなければならない．そのためには，音素境界を抽出する必要がある．音素境界には大きく分けて，子音から母音 (C→V)，母音から子音 (V→C)，母音から母音 (V→V) の場合があるが，日本語においては (V→C)，(C→V) の音素境界の出現確率が 92% 近くを占める[5]ことから，本研究では (C→V)，(V→C) の音素境界を抽出することにする．そのための一方法として，時間的に連続した3個の特徴ベクトル列よりなる三角形の面積を定義し，これを用いることによって，音素境界の抽出が可能であることを示した．

2. 音素対の概略

2.1 音素対とは

音素対の種類およびその数についての詳しい報告は，以前に行なった[4]ので，本節では代表的な音素対とその数を **Table 1** (次頁参照) に示す．Table 1 では，音素対の数は，CV，VC，XC 型に属するものが多いが，日本語における出現率を調べた[5]ところ，CV，VC 型のものが全体の 92% 近くを占めることが明ら

* A Study on the Segmentation of Phoneme-Dyads through the use of the Area of a Triangle by Takeshi AGUI, Tsukasa HOSOMURA and Masayuki NAKAZIMA (Imaging Science and Engineering Laboratory, Tokyo Institute of Technology).

** 東京工業大学像情報工学研究施設

Table 1 The number of each phoneme-dyad in Japanese.

V: vowel　C: consonant
X: mora phonemes or semi-vowels
Each entry in the table expresses the number of phoneme-dyads.

	V	C	X
V	VV (25)	VC (60)	VX (29)
C	CV (60)		CX (11)
X	XV (39)	XC (88)	XX (31)
		Total	343

かとなった．したがって，音素境界の出現確率も (C→V), (V→C) の場合が92%近くを占めることになる．それ故，本研究では，(C→V), (V→C) の音素境界を抽出することについてのみ解析を行った．

2.2 音素対と VCV 音節との概略的な比較

音素対単位での認識確認を行っていないため，音素対単位での認識，あるいは VCV 音節を用いた場合との比較について詳しく論ずることはできない．しかし，概略的に VCV 音節と音素対とを比較すると，VCV 音節の場合は，子音が前後の母音の影響で変形されていたとしても，VCV 音節自体が，その変形をも含んでいる．したがって，一個の子音に対して，前後の母音の違いに対応したすべての変形パタンが用意されていることになる．これに対して，音素対の場合には，例えば，CV 音素対では，Cとして後のVの違いによる変形パタンが用意され，VC 型音素対では，Cとして前のVの違いによる変形パタンが用意されていると考えることができる．

したがって，認識方法として VCV 音節単位の場合には，まず子音の前後の母音を認識し，その後 VCV 音節の標準パタンとのマッチングをとることによって認識が可能である．これに対して，音素対単位の場合，まず，子音の前後の母音を認識し，その後，VC 型音素対あるいは，CV 型音素対の標準パタンとのマッチングをとるわけであるが，このとき，例えば CV 型音素対の標準パタン中の子音に注目すると，子音の先頭部は他の母音の影響を受けていないのに対して，未知パタン中の子音が前の母音の影響を強く受けている場合には，標準パタンと未知パタンとのマッチング誤差が大きくなる．したがって，この場合にはCからVへの「わたり」の部分に重点を置いてマッチングを行うべきである．

以上のことをまとめると，VCV 音節単位に認識する場合には，子音に注目すると，前後の母音の違いに対応した変形パタンがあるため，高い認識率を期待できる反面，変形パタンが増えるだけ，標準パタンも多く用意しなければならない．これに対して，音素対単位に認識する場合，VCV 音節単位の場合と同程度の認識率を得るためには，未知パタンとのマッチングにおいて，工夫をこらさなければならないが，用意すべき標準パタンをあまり多く必要としないという利点がある．本研究では，用意すべき標準パタンが少なくてすむという利点を持つ音素対を用いて，連続音声の認識を行うことを検討している．

3. 特徴ベクトルおよび分析資料

3.1 特徴ベクトル

特徴ベクトルの算出には，音声の分析に最も良く使われている短時間スペクトルを用いた．一般に短時間スペクトルは次式で定義される．

$$F(\omega,\tau)=\int_{-\infty}^{\infty} f(t)w(\tau-t)\exp\{-j\omega(\tau-t)\}\,dt \tag{1}$$

ここで，$f(t)$ は音声波形，$w(t)$ は時間窓，$F(\omega,\tau)$ は時刻 τ における周波数スペクトルを表わす．

いま，音声波形 $f(t)$ を標本化したとき，時刻 t_i における $f(t_i)$ の値に，時間窓をかけたものを $x_i(i=0,\cdots,N-1)$ とすると，FFT 演算を用いて，周波数スペクトル $X(k)$ を求めることができる．

$$X(k)=\sum_{p=0}^{N-1} x_p \exp(-j2\pi pk/N). \tag{2}$$
$$(k=0,1,\cdots,N/2)$$

このとき，時刻 t_i における特徴ベクトル $\boldsymbol{F}_i$ を，次式で定義する．

$$\boldsymbol{F}_i=(F_i(0),F_i(1),\cdots,F_i(n-1))^T. \tag{3}$$

ここで，$F_i(j)=\log|X_i(j)|$, $n=N/2+1$

T: 転置を表わす．

ただし，$\|\boldsymbol{F}_i\|=1$ となるように正規化しておくものとする．そうすると，特徴ベクトル $\boldsymbol{F}_i$ は時刻とともに，n 次元空間内の単位球面上に，ある軌跡を描くことになる．

3.2 分析資料

(C→V), (V→C) の音声境界を求めるための資料として，成人男性一名の発声した対称型 VCV 音節を用いた．各 VCV 音節は，長さ 810 msec であり，10 kHz で標本化し，8 bit で量子化した．Vとして母音5個，Cとして子音12個 (/r/, /b/, /d/, /g/, /p/, /t/, /k/, /z/, /s/, /h/, /m/, /n/) を選んだので，資料の数は

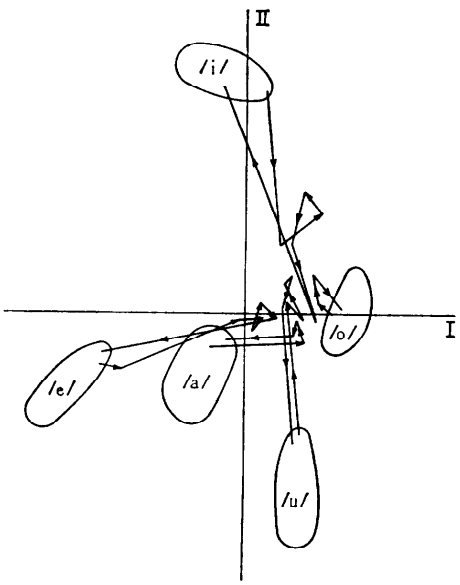

Fig. 1 Graph (I, II) of feature vector sequences of VbV syllables.
Where I: First principal axis.
II: Second principal axis.

合計 12×5=60 個である.

本研究で対象としている,(C→V),あるいは (V→C) の音素境界のように,時間的に変化の速い信号のスペクトルを求める場合には,時間分解能を良くしなければならない.ところが,時間分解能を良くすると,周波数分解能が悪くなる[6],そこで,本研究では,これらのかねあいより,1フレームの長さを 12.8 msec に選んだ,また,フレーム間隔は 12.8 msec とし,時間窓はハミングウィンドウを用いた.このようにすると,各パラメータは,$N=128$, $n=65$ となる.

3.3 特徴ベクトルの軌跡

(C→V),(V→C) の音素境界を抽出するのに先だち,音素境界付近でのスペクトルの動きを調べておく必要がある.そのために,ここでは VCV 音節で,中心の子音が /b/ である場合について,主成分分析を行い,第1,第2主軸のなす平面上に,特徴ベクトルの軌跡を描いた.これを **Fig. 1** に示す.母音とみなされる各領域を実線で囲んである.母音から子音に推移するにつれて,母音の領域をぬけ出し,$\boldsymbol{F}_i$ の動きが大きくなる.このとき,特徴ベクトルの動く方向は,母音の各領域の中心を結んだ5角形の内部の一点を目ざしているようにみえる.

4. 音素境界の抽出

4.1 距離を用いた音素境界の抽出

連続音声から音素対をセグメンテーションするためには,まず,音素境界を抽出しなければならない.$\boldsymbol{F}_i$ は,子音部で動きが大きいことから,$\boldsymbol{F}_i$ の差分ベクトルのノルムを距離 D_i とすると,D_i を利用することによって,音素境界を抽出できる可能性がある.

$$D_i=\|\boldsymbol{F}_{i+1}-\boldsymbol{F}_i\|. \qquad (4)$$

そこで,60個の VCV 音節について距離 D_i を求めた.その結果の一例を **Fig. 2** に示す.(a)が /aba/,(b)が /iri/,(c)が /ama/,(d)が /asa/ の場合であり,縦軸が D_i,横軸が時間である.

Fig. 2(a)に示す /aba/ の場合のように,破裂音を含む音節の場合には,子音に対する時点における距離が母音部と比較してかなり大きいが,(b)に示す /iri/ の場合には,母音部にも子音部における距離と同程度のピークが現われている.さらに,(c)に示す /ama/ や(d)に示す /asa/ の場合にいたっては,子音部でのピークが母音部におけるピークと見分けがつかない程度なので,音素境界の抽出を行うことはほとんど不可能である.

母音部でピークの現われる原因としては,距離 D_i が $\boldsymbol{F}_i$ と $\boldsymbol{F}_{i+1}$ の2個の特徴ベクトルの差のノルムにより定義されているため,$\boldsymbol{F}_{i-1}=\boldsymbol{F}_{i+1}$ で $\boldsymbol{F}_i$ が $\boldsymbol{F}_{i-1}$, $\boldsymbol{F}_{i+1}$ と離れていれば,D_{i-1}, D_i は大きな値になるということが考えられる.したがって,定常な状態にお

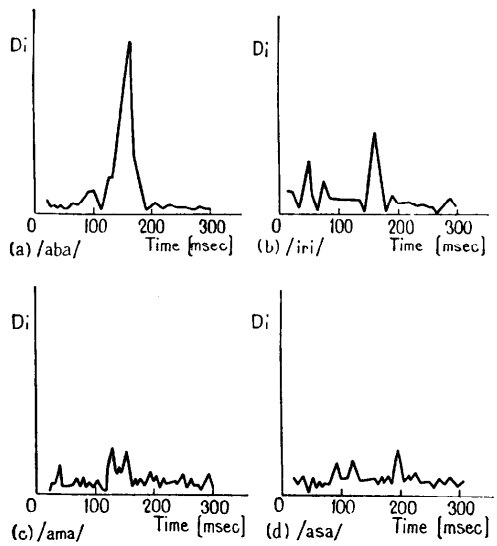

Fig. 2 D_i expressed relative to VCV syllables.

いて，ノイズなどが原因で一個の特徴ベクトル $\boldsymbol{F}_i$ が前後の特徴ベクトルと，かなり違った値をもったとき D_{i-1} および D_i が大きな値をとることになる．

そこで，子音部のピークを強調し，かつ母音内でのピークを抑えるような尺度として，距離 D_i のかわりに，時間的に連続した3個の特徴ベクトル列の囲む面積を考え，これによって音素境界を求める方法について，つぎに検討する．

4.2 面積を用いた音素境界の抽出

3個の特徴ベクトル $\boldsymbol{F}_{i-1}$, $\boldsymbol{F}_i$, $\boldsymbol{F}_{i+1}$ よりなる三角形の面積を S_i とし

$$S_i=\frac{1}{2}\|(\boldsymbol{F}_i-\boldsymbol{F}_{i-1})\times(\boldsymbol{F}_{i+1}-\boldsymbol{F}_{i-1})\|.$$

$$=\frac{1}{2}\sqrt{\sum_{j=1}^{n}\begin{vmatrix} f_{ij}-f_{i-1j} & f_{ij+1}-f_{i-1j+1} \\ f_{i+1j}-f_{i-1j} & f_{i+1j+1}-f_{i-1j+1}\end{vmatrix}^2}. \quad (5)$$

ここで，$\boldsymbol{F}_i=(f_{i1}, f_{i2}, \cdots, f_{in})$, $f_{in+1}=f_{i1}$ と定義する．(5)式のように S_i を定義すると，スペクトルの変化の大きな所では，その大きさを強調できるとともに，$\boldsymbol{F}_{i+1}=\boldsymbol{F}_{i-1}$ のとき $\boldsymbol{F}_i\neq\boldsymbol{F}_{i-1}$ であっても，$S_i=0$ となり，定常状態において，$\boldsymbol{F}_i$ が前後の特徴ベクトルと異なった値をとっても，その影響を除去できる．

したがって，S_i は音素境界を抽出するための，非常に有効なパラメータと考えられる．このことを確かめるため，60個の VCV 音節に対して，実際に S_i の値を求めた．距離の場合と比較するため，/aba/, /iri/, /ama/, /asa/ の場合をそれぞれ **Fig. 3** の(a), (b), (c), (d) に示す．縦軸が S_i，横軸が時間である．距離の場合の Fig. 2 と比較するといずれの音節の場合も，面積の場合のほうが母音部に比較して，子音部が大きな値を持っていることがわかる．

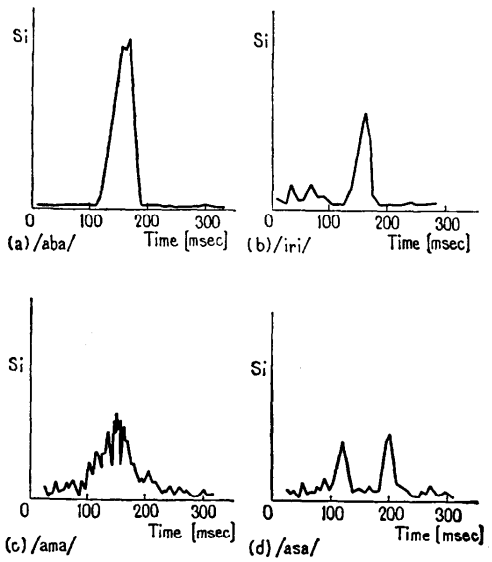

Fig. 3 S_i expressed relative to VCV syllables.

S_i の値の時間的な変化の形状は，中心の子音の違いに依存し，母音の違いにはほとんど依存しない．中心の子音が破裂音の場合は，Fig. 3(a) に示すように子音部で大きな値をとり，(b), (c) に示す流音や鼻子音の場合には，子音部において，破裂音のときほど大きな値はとらない．また(d)の /asa/ の場合には，音素境界付近でピークを持ち子音部では S_i は母音部と同程度の値となっている．このように，中心の子音が /s/ の場合には2個の大きなピークを持つ傾向がある．

4.3 音素境界抽出アルゴリズムおよび抽出結果

S_i をもとに，音素境界を決定することが可能であるかを確かめるため，種々の VCV 音節に対する S_i の描く曲線の大局的な形状を考慮して，つぎのような音素境界の自動抽出アルゴリズムを作成した．

i) ある域値 θ を設定する．

ii) $S_i\geqq\theta\Rightarrow\phi_i=1$
$S_i<\theta\Rightarrow\phi_i=0$ とする．

iii) $\phi_{i-1}=0, \phi_i=1, \phi_{i+1}=0\Rightarrow\phi_i=0$ とする．

iv) $\phi_I=1, \phi_{I+1}=0, \cdots, \phi_{J-1}=0, \phi_J=1$ のとき
$J-I\geqq10\Rightarrow I<k<J$ を満たすすべての k に対して $\phi_k=0$ とする．
$J-I<10\Rightarrow I<k<J$ を満たすすべての k に対して $\phi_k=1$ とする．

v) ϕ_i が $0\rightarrow1$ あるいは $1\rightarrow0$ となる時点を音素境界とする．

ここで，θ としては，子音部では連続した2個以上の時点，すなわち 25.6 msec 以上は $S_i\geqq\theta$ となり，かつ母音部では連続した2個以上の点で $S_i\geqq\theta$ とならないように決めた．また iv) において，$J-I\geqq10$ と $J-I<10$ との場合に分けたのは，連続音声中の母音の継続時間は，150 msec 以上であるのに対して子音の場合，特に，VSV 音節の2個のピークに囲まれた区間は，110 msec 以下であり，連続した10個の時点はちょうど，128 msec に相当するからである．

上記のアルゴリズムを，例を用いて図示したのが，**Fig. 4** (次頁参照) である．このアルゴリズムを実際に60個の VCV 音節に用いたところ，すべての VCV 音

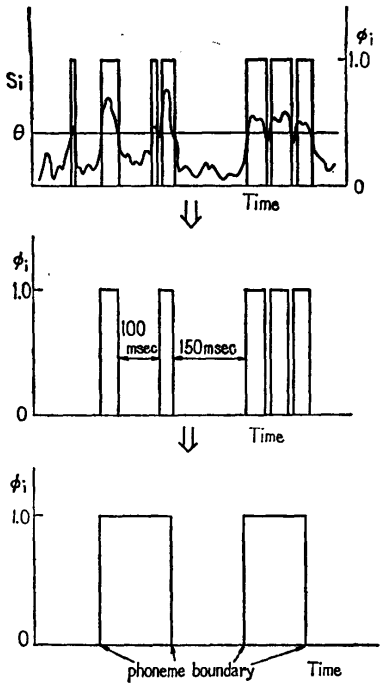

Fig. 4 Procedure to seek the phoneme boundary between consonant and vowel.

節に対して，妥当な音素境界を確定することができた．

5. あとがき

連続音声を音素対単位に認識することによって，調和音結合による認識の際の困難さを改善することが期待されるが，この場合，音素対をセグメンテーションするためには音素境界を抽出することが不可欠となる．そこで，本研究では，日本語において出現確率の高い (C→V), (V→C) の音素境界を抽出するために，時間的に連続した3個の特徴ベクトル列によって囲まれる三角形の面積を用いた．つぎに，この面積を利用して，妥当な音素境界を定めることができることを示した．

今回は，S_i を求める際に，特徴ベクトルの成分として，すべての周波数の値を用いているが，特徴を表わすための有効な成分は，そのうちの一部と考えられるので，これを抽出することによって，次元数を減らすことができるため，S_i を求める時間を，より短縮することができる．その他，1フレームの長さや，フレーム間隔などのパラメータを，子音部と母音部によって切り変えることによって，より精度の良い分析が可能になると考えられる．これらのことについては，現在検討中である．

参 考 文 献

1) F. Itakura: Minimum prediction residual principle applied to speech recognition, IEEE ASSP, Vol. 23, No. 1, pp. 67～72 (1975)
2) 長島他：音韻の標準パターンを用いた単語音声の認識，音響学会春季大会，p. 521 (1976)
3) 中津他：VCV 音節の端点フリー DP マッチングを用いた連続単語音声の認識，信学全大，p. 172 (1976)
4) 細村他：音素対の時間的な可逆性に関する研究，音学誌，Vol. 31, No. 9, pp. 521～528 (1975)
5) 細村他：日本語における CV, VC 音節の出現確率，行動計量学会総会，pp. 132～133 (1976)
6) C. R. Patisaul: Time-frequency resolution experiment in speech analysis and Synthesis, J. A. S. A., Vol. 58, No. 6, pp. 1296～1397 (1975)

(昭和51年7月5日受付)
(昭和52年4月6日再受付)