音声言語情報処理　4－10
(1994. 12. 16)

音素認識ネットワーク素子における音素学習区間の検討

矢口　保博　　三輪　譲二

岩手大学　工学部　情報工学科

〒020　岩手県盛岡市上田4－3－5

あらまし　本論文では、音素認識ネットワーク素子において、音素学習区間と音素認識率の関係について検討した。すなわち、ATRデータベースの1名の話者（MAU）の5240単語を用いて、音素標準パターンの作成時に、ATR音素ラベルと異なる学習区間を用いることにより、音素認識率、脱落率、付加率の改善を検討した。このデータベースを用いた特定話者の認識実験の結果、(1) 語頭と語尾の数フレームの無音区間も促音 /Q/ の学習区間と仮定すること、(2) 語頭の /k/,/h/,/g/ や /p/ では、ATRラベルより数フレーム程度学習区間を短く仮定することにより、1．57%音素認識率が改善され、95．25%の認識率が得られると共に、脱落率と付加率も改善された。

和文キーワード　音素学習区間，音素認識，事後確率，ネットワーク素子

On the Learning Intervals of Phoneme in Network Units for Recognition of Phoneme Based on Probability.

Ysuhiro YAGUCHI and Jouji MIWA

Department of Computer and Information Science,
Faculty of Engineering, Iwate University

4-3-5, Ueda, Morioka-shi, Iwate-ken, 020 Japan

Abstract　This paper describes on the relation between a recognition score and the learning intervals of phoneme references in the network units for recognition of phoneme base on probability (NEUROPHONE). The new intervals for the phoneme references are shorter than the original intervals in ATR speech database for phonemes of initial /k/, /h/, /g/ and /p/ in words and are several frame intervals of silence in the word boundary for choked sound phoneme /Q/. In the recognition experiment of the phonemes in 5240 words uttered by one adult male speaker, the recognition score increases to 95.25% and the rates of omission and addition decrease.

英文 key words　learning intervals, phoneme recognition, a *posteriori* probability, network units

1. はじめに

音素を認識単位とする音声認識システムは、認識対象単語の変更の容易さや多数単語への拡張の容易さなどの利点を持っている。しかし、この音素を認識単位とした認識方式では、音素間の調音結合の問題を回避する必要があり、そのため音素のセグメンテーションが非常に難しくなる問題がある。

このため、我々は、これまで、確率に基づいた音素認識ネットワーク素子（Neurophone：Network unit for recognition of phoneme based on probability）を用いた方式[1][2] [3][4]を提案し、種々検討してきた。

この音素認識ネットワーク素子は、条件付き確率層、事後確率層、時間平滑化層および音節構造層の4つの層からなり、並列ネットワーク構造のためハードウェア向きであり、高速化が可能である。また、この素子は、話者の個人差や調音結合に対処するため、事後確率に基づいた音素側抑制の機構を有し、素子より出力される事後確率値は、音素区間でのみ凸形の時間パターンとなり、音素区間の検出とセグメンテーションを、非常に簡単に行うことができる。

しかし、この事後確率に基づいた音素側抑制の機構も、各音素の標準パターン間の微妙な勢力圏争いにより、逆に、目的とする音素の事後確率値が抑制性の働きをしてしまったり、他の音素に対し興奮性の働きをしてしまう恐れがある。この原因により、目的音素が他の音素に誤ってしまう音素の置換誤りや、音素が存在しない部分で音素が認識されてしまう音素の付加誤り、目的の音素が認識されず音素の脱落誤りが生じる。このため、音素の標準パターン作成において、音素の学習区間の決定は非常に重要な問題である。

本論文では、音素認識ネットワーク素子の音素標準パターンを作成する際に、学習する音素区間を変更することにより、各音素標準パターン間の勢力圏を変更させ、音素の置換、付加および脱落による誤りを低減し、音素認識率を改善する方法を検討する。

2. 音素認識システムの概要

2.1 音素認識ネットワーク素子

図1に確率に基づいた音素認識ネットワーク素子の構成を示す。音素認識ネットワーク素子は、条件付き確率層、事後確率層、事後確率平滑化層、音節構造重み付け層の4層から構成されている。また、条件付き確率層には、各音素カテゴリの平均ベクトル、固有値、および固有ベクトル、事後確率層には先験確率、事後確率平滑化層には、事後確率平滑化関数、音節構造重みづけ層には、音節を構成する子音と母音との時間間隔である音節時間係数のパラメータが必要になる。この音素認識ネットワーク素子は、特徴抽出によって得

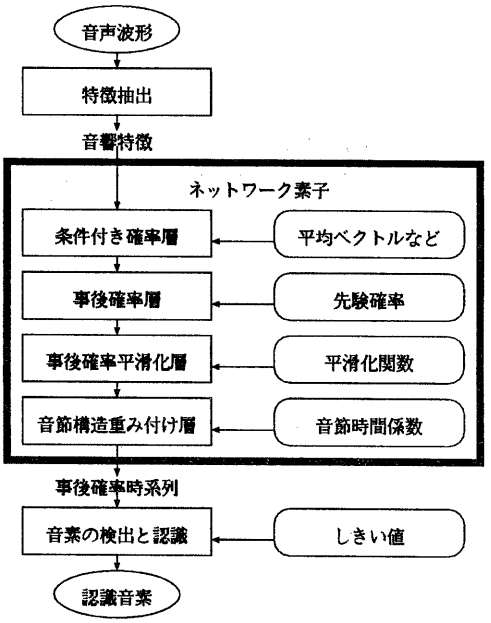

図1. 音素認識ネットワーク素子の構成

られる音響特徴の特徴ベクトルを入力とし、各音素の事後確率を出力する。この素子からは、音素が存在する区間でのみ、凸形の事後確率の時間パターンが出力されるため、音素区間の検出とセグメンテーションが自動的に行われることになる。

2.2 条件付き確率層

音素認識ネットワーク素子の第1層である条件付き確率層は、N次元の特徴ベクトルxを入力とし、各音素カテゴリの条件付き確率を出力する素子である。音素カテゴリω_kの条件付き確率$p(x \mid \omega_k)$は、N次元の特徴ベクトルの分布が、正規分布に従うものと仮定することにより、次のように与えられる。

$$p(x \mid \omega_k) = \frac{\exp(-\frac{1}{2}(x-\mu_k)^t C_k^{-1}(x-\mu_k))}{(2\pi)^{N/2}|C_k|^{1/2}} \quad (1)$$

ここで、μ_kおよびC_kは、音素カテゴリω_kの平均ベクトル（N次元）と共分散行列（$N \times N$）である。

また、共分散行列C_kを固有値分解し、N個の固有値λ_{ik}と固有ベクトルΦ_{ik}で表すと、式(1)は、式(2)のように表される。

$$p(x \mid \omega_k) = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{((x-\mu_k)^t\Phi_{ik})^2}{\lambda_{ik}}\right)}{(2\pi)^{N/2}\left|\prod_{i=1}^{N}\lambda_{ik}\right|^{1/2}} \quad (2)$$

さらに、値の小さい固有値に対応する固有ベクトルは、認識精度を低下させる成分が含まれていることがあるため、次元数 N を M $(M < N)$ で打ち切ると条件付き確率は、式 (3) のようになる。

$$p(x \mid \omega_k) = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{((x-\mu_k)^t\Phi_{ik})^2}{\lambda_{ik}}\right)}{(2\pi)^{M/2}\left|\prod_{i=1}^{M}\lambda_{ik}\right|^{1/2}} \quad (3)$$

この打ち切り次数 M のことを、本論文ではランクということにする。ランク M は、音素認識ネットワーク素子の第1層内のユニットの結合数に対応する。

条件付き確率層で必要となる各音素カテゴリの平均ベクトルのパラメータは、音声データベースに付与されている音素ラベルデータをもとに、各音素区間の加算平均によって求める。また、固有値、固有ベクトルは、各音素区間の加算平均によって求めた共分散行列を、固有値展開することによって求める。

この音素認識ネットワーク素子において最も計算量が多いのは、この第1層で行う2次形式の内積演算であり、次元数 N の2乗に比例することになる。しかし、この層での処理は、各音素カテゴリに関して独立に実行可能であり、ハードウェアで構成する場合は、並列の構成することにより高速化することができる。

2.3 事後確率層

第2層の事後確率層は、第1層で求めた各音素カテゴリの条件付き確率 $p(x \mid \omega_k)$ から、各音素の事後確率を求める層である。音素カテゴリ ω_k の事後確率 $p(\omega_k \mid x)$ は、$p(\omega_k)$ を先験確率、$p(x)$ を分布確率とすると、ベイズ則に従い式 (4) で与えられる。

$$p(\omega_k \mid x) = \frac{p(\omega_k)p(x \mid \omega_k)}{p(x)} \quad (4)$$

認識音素全体が全事象をカバーしていると仮定すると、全カテゴリの事後確率の和は1となることから、全認識音素に対して式 (4) の両辺の和を取ると、以下の式が得られる。

$$\begin{aligned}\sum_{k}^{All} p(\omega_k \mid x) &= \sum_{k}^{All}\frac{p(\omega_k)p(x \mid \omega_k)}{p(x)} \\ &= \frac{\sum_{k}^{All} p(\omega_k)p(x \mid \omega_k)}{p(x)} \\ &= 1 \end{aligned} \quad (5)$$

式 (4) と式 (5) より、式 (6) が得られる。

$$p(\omega_k \mid x) = \frac{p(\omega_k)p(x \mid \omega_k)}{\sum_i p(\omega_i)p(x \mid \omega_i)} \quad (6)$$

この式は、分母項の他の音素カテゴリの条件付き確率により、目的の音素カテゴリの事後確率を抑制する機構が組み込まれることを示している。この式において、目的の音素カテゴリの先験確率が興奮性の重みとなり、他の音素カテゴリの先験確率が抑制性の重みとなる。事後確率層で必要なパラメータは各音素カテゴリの先験確率であるが、先験確率は音声認識のタスクに依存するため、ここではすべて同一の値とし、$p(\omega_k) = 1/$全音素数 として与える。

2.4 事後確率平滑化層

第3層の事後確率平滑化層は、第2層で求めた事後確率を平滑化する層である。この事後確率平滑化層は、特徴ベクトルの微妙な時間変動によって生ずる事後確率の変動を除去する役割をもっている。平滑化前の事後確率を $p(t)$、平滑化後の事後確率を $p_s(t)$ とすると、事後確率平滑化は、式 (7) のように行われる。

$$p_s(t) = \sum_{\tau} w(\tau)p(t-\tau) \quad (7)$$

ここで $w(\tau)$ は事後確率平滑化関数である。

2.5 音節構造重み付け層

第4層の音節構造重み付け層は、実際に音節を構成可能な事後確率に重み付けを行う層である。例えば、日本語の音声の場合、子音には必ず母音が後続し、半母音 /w/ には母音 /a/ のみが後続する。このような音節構造の重み付けを、重み付けする前の音素カテゴリ k の事後確率を $p_k(t)$、重み付け後の事後確率を $p_{w_k}(t)$ として、式 (8) のように行う。

$$p_{w_k} = p_k(t)\sum_i p_i(t+\tau_i) \quad (8)$$

ここで、τ_i は、音素カテゴリ ω_i に後続する母音との時間間隔であり、音節時間係数と呼ぶ。

2.6 認識音素決定

音素認識ネットワーク素子から出力された事後確率は、音素区間でのみ凸形の時間パターンとなる。このため、事後確率の時間パターンの凸区間を音素として検出し、凸形の時間パターンの変曲点から変曲点までの区間を、音素区間として自動的にセグメンテーションを行う。自動検出された音素区間の事後確率の最大値と音素の持続時間長の両方が、各音素カテゴリ毎に設定したしきい値を越えた場合にのみ、認識音素と決定する。

3. 音素認識実験

3.1 音声資料と認識対象音素

音素認識実験のための音声資料としては、ＡＴＲ翻訳電話研究所作成の音声データベース [5] の男性話者１名（ＭＡＵ）による重要語５２４０単語を用いた。

そのうち偶数番の単語番号の２６２０単語を音素標準パターンの作成に用い、奇数番の２６２０単語を音素認識実験の評価に用いた。

認識対象音素は、表 1に示す 23 音素とした。

表 1. 認識対象音素

母音	/a/, /o/, /u/, /i/, /e/
半母音	/j/, /w/
鼻音	/m/, /n/, /η/
有声破裂音	/b/, /d/, /g/
流音	/r/
有声摩擦音	/z/
喉頭摩擦音	/h/
無声摩擦音	/s/
無声破擦音	/c/
無声破裂音	/p/, /t/, /k/
促音	/Q/
撥音	/N/

3.2 音響特徴量と標準パターン

音響特徴量の抽出のための音声分析方法を、表 2に示す。

表 2. 音声分析方法

サンプリング	10kHz, 16 ビット
分析フレーム周期	10ms
分析窓	25.6ms ハミング窓
分析方法	256 点 FFT
特徴量	16 チャネルメルスケール パワースペクトル

音素認識ネットワーク素子へ入力する音響特徴量の次元数 N は、16 チャネル帯域スペクトル×６フレームの 96 次元とした。ここで、６フレーム分の特徴量は、図 2に示すように、フレーム周期 10ms としたとき、11 フレーム中から１フレーム毎に間引くことで得られる。また、式 (3) におけるランク M の値は、従来の研究結果 [3] より、音素 /g/ を 70、音素 /p/ を 65、その他の音素を 80 とした。

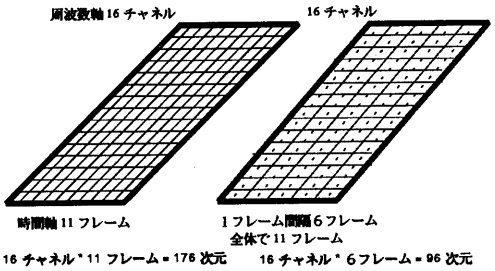

図 2. 音響特徴量の周波数・時間表現

3.3 音素学習区間

音素の標準パターンの学習では、従来は、ＡＴＲの音素ラベルと全く同一の開始フレームから終了フレームまでの区間を音素区間として、音素の標準パターンを学習した。本実験では、図 3に示すように、ＡＴＲの音素ラベルと相対的に異なる開始フレームと終了フレームを音素区間と仮定して、音素の標準パターンを学習し、音素認識実験によって評価を行った。ただし、フレーム周期を、ＡＴＲの音素ラベルでは 5ms としているが、本研究では 10ms としているため、例えば、1235ms は 1240ms として取り扱った。

ＡＴＲの音素ラベルと異なる学習区間を用いる音素とコンテキストは、従来のＡＴＲ音素ラベルを用いて学習した音素標準パターンによる音素認識実験を行い、実験結果の３音素組コンテキストの音素混同確率行列の情報より、付加、脱落、および置換の誤りの多い音素やコンテキストを選択した。

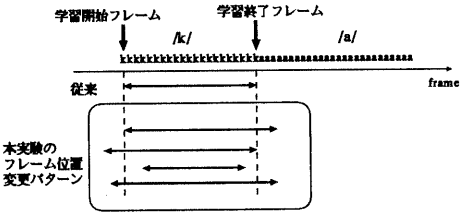

図 3. 音素 /k/ の標準パターン学習音声区間の変更例

3.4 音素認識結果

新しい音素標準パターンを用いる音素認識実験を、下記の音素とコンテキスト条件に従って、以下に示す番号の順に行った。ここで、＃は語頭と語尾の単語境界を表すものとする。

(1) /Q/ ＋ ＃ または ＃ ＋ /Q/ の場合

語頭や語尾の単語境界のコンテキストでは、他のコンテキストと比較して、音素の付加が生じることが多い。このため、促音 /Q/ の音素の学習区間を、本来のＡＴＲの音素ラベルの促音区間とは別に、語頭の数フレームの無音部分

と語尾の数フレームの無音部分を、促音 /Q/ の音素区間と仮定して、促音の標準パターンを学習した。

実験では、最初に、語尾の数フレームの無音区間を促音と仮定して、学習フレーム位置の検討を行い、音素認識率が最大となるフレーム区間を決定した。その後、最適な語尾のフレーム区間に加えて、語頭の数フレームの無音区間を、促音と仮定して、学習フレーム位置の検討を行った。

実験結果を表3に示す。この結果、語頭と語尾の4フレーム程度の無音区間を、促音 /Q/ の音素区間に加えることにより音素認識率が、約1.2%改善され、音素脱落率を1%程度減少させることができた。

表3. 音素 /Q/ の学習区間変更による結果

語頭学習フレーム位置	0(従来)	-13 ~ -10
語尾学習フレーム位置	0(従来)	+3 ~ +6
音素認識率 [%]	93.68	94.87
音素付加率 [%]	14.44	14.30
音素脱落率 [%]	3.15	2.16

(2) # + (/k/,/h/,/g/) の場合

語頭の /k/、/h/、/g/ は、他の音素と比較して認識率が悪く、特に、付加率が高い音素とコンテキストである。このため、語頭の /k/、/h/、/g/ の標準パターンの検討を行った。検討は、/k/、/h/、/g/ の順に行うが、/k/ は、(1) の /Q/ の実験で音素認識率が最も良かった学習区間を用いて行う。同様に、/h/ は /k/、/g/ は /h/ の実験で最も音素認識率が良かった学習区間を引き継いで用いる。

語頭の /k/ の音素区間変更前と変更後の付加音素 /k/ の音素混同行列を、それぞれ表4と表5に示す。この表から語頭において /k/ の付加を大幅に減らすことができたことがわかる。

次に、語頭の /k/、/h/、/g/ のそれぞれの実験結果を、表6、7、8に示す。この結果から、ＡＴＲの音素ラベルより数フレーム短い区間から、音素標準パターンを学習した方が、音素認識率を高くすることができ、また、付加数や脱落数も少なくすることが出来ることが分かった。

表4. 付加音素 /k/ の音素混同行列 (音素 /k/ の学習区間変更前)　単位: 個数

		後続音素																						
		a	o	u	i	e	j	w	m	n	η	b	d	g	r	z	h	s	c	p	t	k	Q	=	#	all
先後音素	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	3
	i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	e	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	h	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	s	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3
	c	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	k	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3
	=	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	2
	#	1	8	5	38	1	40	1	13	8	0	5	2	3	8	5	0	5	2	0	0	2	0	0	0	147
		音素 /k/ の付加数 = 166　音素 /k/ の付加率 12 %																								

表5. 付加音素 /k/ の音素混同 (音素 /k/ の学習区間変更後)　単位: 個数

		後続音素																						
		a	o	u	i	e	j	w	m	n	η	b	d	g	r	z	h	s	c	p	t	k	Q	=	#	all
先後音素	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	3
	i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	e	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	h	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	s	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	3
	c	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	k	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	=	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	#	1	5	3	27	0	31	0	7	5	0	2	1	2	6	2	0	0	1	0	0	2	0	0	0	95
		音素 /k/ の付加数 = 111　音素 /k/ の付加率 8 %																								

表6．音素 /k/ の学習区間変更による結果

注目音素 /k/		開始フレーム位置 0	1	2
終了フレーム位置	0	94.88	94.83	94.77
		14.30	14.35	14.45
		2.16	2.21	2.12
		166	98	86
	-1	94.90	94.96	94.83
		14.33	14.31	14.39
		2.16	2.10	2.10
		180	111	94
	-2	94.92	94.91	94.91
		14.36	14.30	14.38
		2.15	2.09	2.11
		180	113	103

表記法
音素認識率[%]
音素付加率[%]
音素脱落率[%]
注目音素付加数[個]

表7．音素 /h/ の学習区間変更による結果

注目音素 /h/		終了フレーム位置 0	1	2
終了フレーム位置	-2	95.07	94.97	94.93
		13.94	13.49	13.36
		2.00	2.06	2.11
		251	122	57
	-3	95.04	95.09	95.09
		13.95	13.47	13.33
		2.03	2.02	2.03
		201	106	70
	-4	95.01	95.02	95.00
		13.75	13.38	13.37
		2.06	2.05	2.06
		138	89	113

表8．音素 /g/ の学習区間変更による結果

注目音素 /g/		開始フレーム位置 -1	0	1
終了フレーム位置	0	94.90	95.09	95.05
		13.59	13.47	13.40
		2.01	2.02	2.01
		267	106	75
	-1	95.00	95.11	95.09
		13.66	13.56	13.33
		2.02	2.01	2.01
		239	158	81
	-2	95.03	95.07	95.07
		13.71	13.44	13.31
		2.00	2.02	2.01
		202	127	79

(3) /p/ の場合

無声破裂音 /p/ は、持続時間が短く出現頻度も少ないため、一般に、認識が困難である。このため、/p/ の音素区間を変更して音素標準パターンを作成し、認識実験を行った。この実験でも、(2) と同様に、他の音素の学習区間は、音素認識率が最も良かった学習区間を用いた。

実験結果を表9に示す。この結果からも、ＡＴＲの音素ラベルと異なる区間から音素標準パターンを学習した方が、音素認識率を高くすることができることが分かった。

表9．音素 /p/ の学習区間変更による結果

注目音素 /p/		開始フレーム位置 0	1	2
終了フレーム位置	0	95.11	95.20	95.23
		13.56	13.45	13.52
		2.01	2.02	2.03
		190	26	58
	-1	95.20	95.25	95.15
		13.45	13.50	13.59
		2.01	2.02	2.03
		66	106	170
	-2	95.23	95.16	95.16
		13.51	13.57	13.53
		2.02	2.00	2.00
		89	182	171

次に、総合音素認識実験結果を表10に示す。この結果、ＡＴＲラベルより数フレーム程度学習区間を短く仮定することで、1.57% 音素認識率が改善され、95.95% の認識率が得られると共に、脱落率は 1.13%、付加率は 0.94% 改善された。

表10．総合音素認識実験結果

	従来	本実験				
音素		/Q/	/k/	/h/	/g/	/p/
学習開始フレーム位置	0	-13~-10	+1	+1	0	+1
学習終了フレーム位置	0	+3~+6	-1	-3	-1	-1
音素認識率[%]	93.68	95.25				
音素付加率[%]	14.44	13.50				
音素脱落率[%]	3.15	2.02				

3.5 考察

本論文の方法による音素認識の改善例を図4、図5および図6に示す。これらは、入力音声「気持ち (/kimoci/)」に対しての音素認識結果を表している。

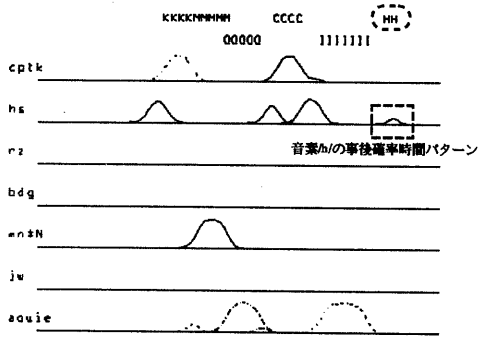

図4. 従来の音素認識結果

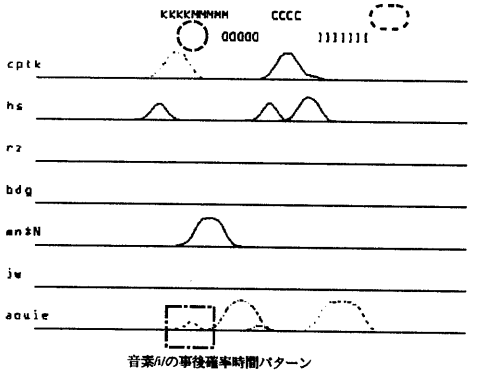

図5. /Q/ の音素学習区間変更後の音素認識結果

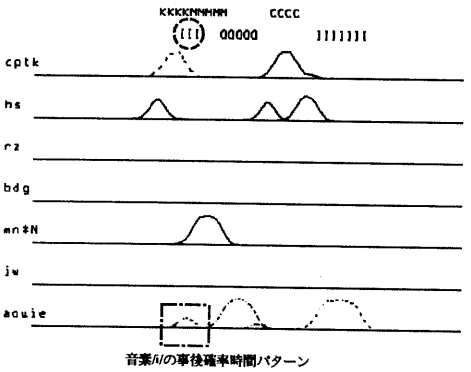

図6. /k/ の音素学習区間変更後の音素認識結果

従来の音素認識結果 (図4) と /Q/ の音素学習区間を変更した後の音素認識結果 (図5) を比較してみると、従来の図4では、語尾部分に /h/ が付加しているが、/Q/ の音素学習区間を増した図5では、それが改善されている。

また、/Q/ の音素学習区間を変更した後の音素認識結果 (図5) と、/k/ の音素学習区間を変更した後の音素認識結果 (図6) を比較すると、図6では、/i/ の事後確率値が大きくなり、正しく認識されていることがわかる。

これらの実験結果より、音素の標準パターン学習のための音素区間を変更することにより、音素認識率を改善できることがわかった。また、NEUROPHON方式は、音素認識結果を図式表現できるため、このような音素認識の改善の戦略を検討するために有用であることがわかる。

音素の標準パターン勢力圏の例を図7に示す。

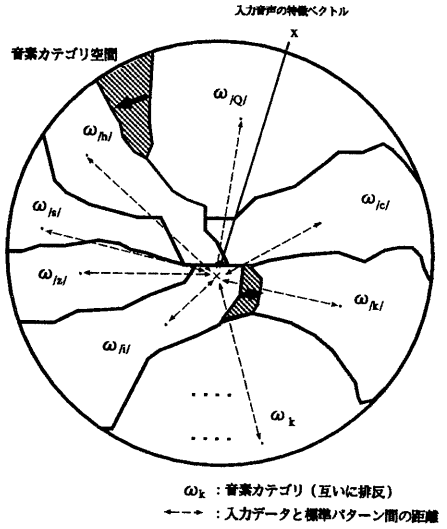

図7. 各音素の標準パターン勢力圏の例

図5の例では、/Q/ の音素学習区間変更により、/Q/ の標準パターンの勢力が強まり、逆に /h/ の標準パターンの勢力を抑制したため、/h/ の付加が改善されたと考えられる。すなわち、図7の上方斜線部に示す様に、/h/ の領域が狭くなり /Q/ の領域が広くなったものと考えられる。

また、図6の例では、/k/ の音素学習区間を狭めることにより、/k/ の標準パターンの勢力が弱まり、逆に /i/ の標準パターンの勢力が強まったため、/i/ が認識されたと考えられる。すなわち、図7の中央斜線部に示す様に、/k/ の領域が狭くなり /i/ の領域が広くなったものと考えられる。

ここで、話者ＭＡＵと異なる話者ＭＨＴに対して、ＭＡＵと同じ条件で音素 /k/ の学習区間の変更による検討を行った。その実験結果を、表 11に示す。

この結果、話者ＭＨＴにおいても音素 /k/ の学習区間変更により、音素 /k/ の付加数は、193 増加しているが、音素認識率は、0.33% 改善し 94.39% の認識率を得られた。それと共に、付加率は 0.13%、脱落率は 0.06% 改善された。

しかし、話者ＭＡＵでの学習区間は、実際の長さより狭くすると良い結果が得られたが、話者ＭＨＴでは、学習区間の長さは変っておらず、前後２フレームづつ移動しただけになっている。このことより、音素学習区間は話者ごとに異なることがわかった。

表 11. 音素 /k/ の学習区間変更による結果 (ＭＨＴ)

注目音素 /k/		開始フレーム位置			
		-3	-2	-1	0
終了フレーム位置	0		94.32	94.18	94.06
			22.23	21.99	22.10
			1.38	1.38	1.42
			479	281	246
	-1	94.34	94.35	94.20	94.11
		22.37	22.03	21.85	22.00
		1.32	1.37	1.37	1.39
		605	464	261	235
	-2	94.35	94.39	94.23	94.18
		22.18	21.97	21.76	21.81
		1.37	1.36	1.39	1.40
		582	439	241	221
	-3	94.36	94.35	94.21	
		22.03	21.93	21.82	
		1.37	1.39	1.46	
		565	436	257	

表記法
音素認識率 [%]
音素付加率 [%]
音素脱落率 [%]
注目音素付加数 [個]

4. むすび

本論文では、音素認識ネットワーク素子において、音素学習区間と音素認識率の関係について検討した。すなわち、ＡＴＲデータベースの１名の話者（ＭＡＵ）の５２４０単語を用いて、音素標準パターンの作成時に、ＡＴＲ音素ラベルと異なる学習区間を用いることにより、音素認識率、脱落率、付加率の改善を検討した。その結果、ＡＴＲラベルより数フレーム程度学習区間を短く仮定することにより、1.57% 音素認識率が改善され、95.25% の認識率が得られると共に、脱落率は 1.13%、付加率は 0.94% 改善された。

このとき、音素学習区間を決める開始フレーム位置と終了フレーム位置は、話者により、また、音素により最適な位置があり、音素の学習区間の変更が音素認識において有効であることがわかった。

しかし、本報告の方法では、最適なフレーム位置を決定するためにかなりの時間を要している。今後、最適な音素学習区間を自動的に求める方法等について検討する必要がある。

参考文献

[1] 斎藤、三輪: “事後確率の時間特徴を利用した音素認識”, 信学技報,SP85-56(June 1985).

[2] 三輪、山崎: “事後確率に基づいたネットワークモデルを用いた音素認識”, 信学技報,SP89-27 (June 1989).

[3] 三輪、今: “確率に基づいた音素認識ネットワーク素子の次元数の検討”, 信学技報, PRU91-78 (Nov. 1991).

[4] 今、三輪: “音素事後確率の時間パターンを特徴とした音素認識”, 信学技報,SP92-124(June 1993).

[5] 武田、匂坂、片桐、阿部、桑原: “研究用日本語音声データベース利用解説書”、(株) エイ・ティ・アール自動翻訳電話研究所、(1990).