長さの異なるセグメントを組み合わせた 音声認識のための特徴ベクトルの構成法*

二宮 和則[†]・大槻 恭士[†]・大友 照彦[†]

Construction of Segmental Unit Input with Dual-Width Window for Speech Recognition*

Kazunori NINOMIYA[†], Takashi OTSUKI[†] and Teruhiko OHTOMO[†]

In this paper, we propose the dual-width windowed segment (DWWS) which consists of a short cepstrum segment and a long Δ cepstrum segment, as a more effective segmental input vector than the conventional one. First, using a criterion for cluster analysis, we evaluate various compositions of feature vectors based on their ability of phoneme separation to show the effectiveness of DWWS. Then we carry out discrete HMM speech recognition experiments to verify the result of evaluation. As a result, it is shown that the DWWS brings on high recognition performance when a categories-dependent codebook is used for vector quantization.

1. はじめに

高度な音声認識システムの実現において、音声の時間的変化特性を導入することが重要とされ、これまでに様々な検討がなされてきている。

時間的特性を含んだ特徴パラメータとして、静的特徴量であるケプストラムの時間軸方向の回帰係数である Δ ケプストラム [1] や、聴覚モデルに基づく動的ケプストラム [2] などが提案されている。また、当該フレームの入力として前後数フレームを同時に使用するセグメント単位入力も検討されている。これは、井出らによる TSP 法 [3] に準じたものであり、Waibel らによる TDNN[4] といった脳の神経情報処理をモデル化したニューラルネットワークや、McDermott らによる Shift-Tolerant LVQ[5] などに適用され、7 フレーム前後のセグメント単位入力が音声認識の特徴ベクトルとして有効であることが示されている。

一方、認識手法としての HMM は時系列パターンをモデル化するものであり、ニューラルネットワークや LVQ などと比べると時間的特性をうまく表現した効果的な方法である。しかし、同状態の中では観測値の順序が入

れ代わっても出力確率値に影響がないなど、その時間表現能力は十分なものではない。そこで、モデルの状態数を増加させる方法 [6]、状態の継続時間長確率密度を組み込む方法 [7]、条件付き HMM[8] など時間特性を表現した様々なモデルが検討されている。また、連続する音素コンテキストを考慮した音素環境依存型 HMM[9] など大きな意味で一種の時間モデルと考えられる。セグメント単位入力を用いた HMM では、中川らが連続分布型 HMM[10] においてその有効性を示し、離散分布型 HMM においても SQ-HMM[11] として耐雑音性への効果が検討されている。

ところで、これまでセグメント単位入力を構成する場合、ケプストラムおよび Δ ケプストラムには同じ幅の時間窓をかけるか、あるいは Δ ケプストラムを補足的な特徴量としてセグメント単位のケプストラムに付加して使用するのが一般的であった。しかし、セグメント内における二つの特徴量は、それぞれ性質が異なり、互いの特性を活かした適切な併用方法が望まれていた。また、様々な特徴ベクトルを評価するためには膨大な認識実験を行う必要があり、簡単な計算のみで各特徴ベクトルの有効性を評価できる評価尺度の定義が望まれていた。

本論文では、セグメント単位入力におけるケプストラムと Δ ケプストラムの有効な併用方法に着目し、音声認識に適したセグメント単位入力の構成法について検討する。まず、クラスタ分析における分離評価基準を用いて、様々なセグメント長の組合せからなる特徴ベクトル

* 原稿受付 2000 年 8 月 31 日

[†] 山形大学 工学部 Faculty of Engineering, Yamagata University; 4-3-16, Jonan, Yonezawa city, Yamagata 992-8510, JAPAN

Key Words: cepstrum, Δ cepstrum, cluster analysis, discrete HMM, dual-width windowed segment.

の有効性を評価する [12]。つぎに、その分析結果に基づき、静的な特徴量であるケプストラムには時間窓幅を小さく、動的な特徴量である Δ ケプストラムには時間窓幅を大きくした異なる長さのセグメントから構成される特徴ベクトルを提案する。さらに、この特徴ベクトルを、離散分布型 HMM 音声認識システムに適用し、種々の認識実験によりその有効性の検証を行い、セグメント単位入力における各特徴空間の特性を明らかにする。

2. 特徴ベクトルとその評価法

2.1 特徴ベクトル

音声認識における特徴パラメータは、スペクトル包絡パラメータを基本としており、本研究では LPC メルケプストラム係数を採用する。また、スペクトル包絡の時間変化パラメータとして、LPC メルケプストラム係数の 1 次の線形回帰係数である LPC Δ メルケプストラム係数も同時に使用する。時刻 t におけるケプストラムを X_t とすると、 Δ ケプストラム Y_t は (1) 式によって求められる。

$$Y_t = \frac{\sum_{a=-n}^{+n} a \cdot X_{t+a}}{\sum_{a=-n}^{+n} a^2} \quad (1)$$

これは、当該時刻の前後 n フレームに対し、重み付き時間窓をかけて分析を行うため、静的な特徴パラメータであるケプストラムに対して、時間特性を含んだ動的な特徴パラメータとなる。

一方、セグメント単位入力は、ケプストラムあるいは Δ ケプストラムに対して、当該フレームの前後数フレームに時間窓をかけて切り出した時間特性を持つ特徴ベクトルである。本研究では、各時刻の入力となる特徴ベクトルを、Fig. 1 に示すように、当該フレームを中心とする幅 Cw のケプストラムセグメントと幅 Dw の Δ ケプストラムセグメントをあわせたものとする。よって、構成される特徴ベクトルは、 Cw と Dw を構成要素とする様々な組合せが考えられる。そこで、様々なセグメント長の組合せからなる特徴ベクトルを比較するためのいくつかの評価尺度を定義する。

2.2 分離度

音素を認識単位とする音声認識において、ベクトル空間上で各音素が形成するクラスを音素クラスとして定義し、その分離の度合いが大きい特徴ベクトルが有効であると考え、特徴ベクトルの評価に、クラス分離の度合いに基づく評価基準を用いる。

まず、 M 個の音素クラスを $P_1, P_2, \dots, P_M$ 、各音素クラスに属する特徴ベクトルの数を $N_1, N_2, \dots, N_M$ 、そして全特徴ベクトルの数を $N = N_1 + N_2 + \dots + N_M$ とする。つぎに、音素クラス P_c に属する特徴ベクトルを

Fig. 1 Feature vector using segmental unit input

$x_{cn}(n=1,2,\dots,N_c)$ 、全特徴ベクトルの平均を $\bar{x}$ 、そして音素クラス P_c に属する特徴ベクトルの平均を $\bar{x}_c$ とすると、全特徴ベクトルに関する共分散行列 T および各音素クラス内での共分散行列の平均 W はそれぞれ次式で与えられる。

$$T = \frac{1}{N} \sum_{c=1}^M \sum_{n=1}^{N_c} (x_{cn} - \bar{x})(x_{cn} - \bar{x})' \quad (2)$$

$$W = \frac{1}{N} \sum_{c=1}^M \sum_{n=1}^{N_c} (x_{cn} - \bar{x}_c)(x_{cn} - \bar{x}_c)' \quad (3)$$

全データのばらつきに比べ、各クラス内のばらつきが小さいほどクラス間の分離が良いと考えられることから、分離度 S を次式で定義する [13]。

$$S = \frac{|T|}{|W|} \quad (4)$$

この計算は、音声データに目視で付けられた音素ラベルに基づき全フレームについてクラス分けして行う。

2.3 等価セグメント長

先に求めた分離度の値だけでは意味が分かりにくいので、ケプストラムセグメントのみで構成される特徴ベクトルの分離度とセグメント長の関係の近似式により、分離度 S をセグメント長に変換する。これにより各特徴ベクトルと同じ音素分離能力を持つケプストラムセグメントのフレーム数が分かる。これを、等価セグメント長 E と呼び次の近似式を用いて計算した [12]。

$$E = 2.0 \times 10^{-10} S^3 + 5.0 \times 10^{-7} S^2 + 0.006 S \quad (5)$$

2.4 分離効率

等価セグメント長 E と実使用フレーム数 $Cw + Dw$ の比で、1 フレームあたりの分離能力を表すことができる。これを分離効率 ρ として (6) 式で定義する。

$$\rho = \frac{E}{Cw + Dw} \quad (6)$$

ここで、 $\rho > 1$ の場合、同じ次元数のケプストラムセグメントより分離能力が大きく、逆に $\rho < 1$ の場合、同じ

次元数のケプストラムセグメントより分離能力が小さいことが分かる。

3. 特徴ベクトルの評価実験

ここでは、前述の評価基準に基づき、実際の音声データを用いて様々なセグメント長の組合せからなる特徴ベクトルの評価実験を行う。

3.1 音声データ

音声資料として、東北大・松下単語音声データベースより男性話者10人の発声した212単語データを用いた。音素カテゴリは25音素とし、これらすべての音声データには目視による音素ラベルが付与されている。音声の分析条件をTable 1に示す。

Table 1 Condition of speech analysis

Sampling frequency	12 [kHz]
Pre-emphasis	$1 - 0.97z^{-1}$
Hamming window	25 [ms]
Frame period P_f	10 [ms]
Feature parameters	mel cepstrum 16 order Δ mel cepstrum 16 order

3.2 分離度による評価

ケプストラムおよび Δ ケプストラムの単一特徴量において各セグメント長と分離度の関係をFig. 2に示す。その結果、ケプストラムはセグメント長が大きくなると分離度も大きくなる右上がりの曲線となった。一方、 Δ ケプストラムの分離度はケプストラムに比べてきわめて低い値を示し、セグメント長が大きくなってもほとんど増加していない。

Fig. 2 Relationship between the length of segment and separation ($P_f = 10$ [ms], $N_s = 7$)

3.3 等価セグメント長による評価

等価セグメント長の様子をFig. 3に示す。この結果は、分離度が特徴ベクトル中の線形独立なフレームの数に依存すると考えるとある程度説明ができる。以下、Fig. 4 (簡単のために Δ ケプストラム分析窓幅 N_s を3として

Fig. 3 Relationship among cepstrum segment length, delta cepstrum segment length and equivalent segment length ($P_f = 10$ [ms], $N_s = 7$)

Fig. 4 Three case of feature vector ($N_s = 3$)

いる)に示す三つの場合に分けて説明する。

case 1: ケプストラムより Δ ケプストラムがすべて求まる、つまり Δ ケプストラムは線形従属な要素である。この場合、全く情報を落とさずに Δ ケプストラム部をすべて除くことが可能である。(Fig. 3の斜線部)

case 2: ケプストラムと Δ ケプストラムから、新たなケプストラムが導ける。ただし、 Δ ケプストラムセグメントの中央($Cw - N_s + 1$)フレームが線形従属で除去可能となる。(Fig. 3の網かけ部)

case 3: Cw が $N_s - 1$ よりも小さく、 Δ ケプストラムセグメントからケプストラムセグメントを導くことはできない。この場合、すべてのフレームは線形独立である。(Fig. 3の網かけなしの部分)

3.4 分離効率による評価

分離効率 ρ の様子をFig. 5に示す。case 1, 2では不要なフレームを含むため効率が1以下となる。case 3ではケプストラムセグメントの長さが3のところに稜線があり、ケプストラムセグメント3フレームと Δ ケプストラムセグメントからなる特徴ベクトルの効率が良いことが分かる。したがって、効率よく高い分離度を得るためには、長さ3のケプストラムセグメントと長い Δ ケプストラムセグメントからなる特徴ベクトルを用いる必要がある。

Fig. 5 Relationship among cepstrum segment length, delta cepstrum segment length and efficiency ρ ($P_f = 10[\text{ms}]$, $N_s = 7$)

4. 音声認識のための特徴ベクトル構成法

4.1 DWWSの構成

3.で明らかにしたように、ケプストラムと Δ ケプストラムの分離度ではケプストラムの方がきわめて高く、静的な特徴量であるケプストラムは当該フレームの音素性を強く表現したパラメータであり、逆に、動的な特徴量である Δ ケプストラムは、音素性そのものは小さい代わりに、当該フレーム周辺の時間変化特性を表現したパラメータであるといえる。よって、ケプストラムに対する時間窓幅を適度に小さく、 Δ ケプストラムに対する時間窓幅を大きく取することは、局所的な当該時刻の音素性と、大局的な時間変化特性を無駄なく取り入れた、音声認識に適した特徴ベクトルの構成法であると考えられる。以後、このような構成の特徴ベクトルを、DWWS (Dual-Width Windowed Segment) と呼ぶことにする。

4.2 DWWSのHMMへの適用

セグメント単位入力をHMMに適用する場合、入力ベクトルの次元数が問題となる。DWWSは、ケプストラムに対する時間窓幅は小さいが、 Δ ケプストラムに対する時間窓幅は前後の音素区間を含む程度に大きなものを想定するので、入力ベクトルの次元数はきわめて大きくなることが予想される。特徴ベクトル空間における連続確率密度を用いる連続分布型HMMでは、入力ベクトルの次元数が大きくなると、共分散行列の推定精度が低下するという問題がある。この対策として、K-L展開による次元圧縮を行う方法[10]があるが、大幅な次元圧縮は不適切である。

一方、学習サンプルをベクトル量子化し、その離散シンボル系列をモデル化する離散分布型HMMでは、量子化誤差や学習サンプルに強く依存するという問題もあるが、入力ベクトルの次元数はさほど問題とならない。したがって、DWWSのベクトル特性を劣化させることなく観測するためには、離散分布型HMMを用いた音声認識システムを使用するのが適当であり、その場合、DWWSで切り出された入力ベクトルを、同じDWWS

を使って求めたコードブックによりベクトル量子化することで実現できる。

コードブックの作成方法としては、分類するカテゴリへの依存性によって以下の二つの方法が考えられる。カテゴリ非依存コードブックは、全学習データを一括してLBG法などでクラスタリングし、 K 個の代表ベクトルを求めるものであり、カテゴリ依存コードブックは、分離したい各音素ごとに K_s 個の代表ベクトルを求め、最後にそれらをまとめて一つのコードブック (サイズ= $K_s \times$ 音素数) とするものである。とくに後者は、音声のラベリングデータを必要とするが、各音素ごとに求めた代表ベクトルをLVQ[14]などで再度修正学習を行うことで、より有効なコードブックが得られる。

5. 認識実験による検証

ここでは、認識実験によりDWWSの有効性を検討する。まず、不特定 (学習に使用していない) 話者・特定 (学習に使用した) タスクに対する認識実験として、前述の二つのコードブック作成方式ごとに、各特徴パラメータに対する時間窓幅と認識率の関係を調べ、セグメント単位入力における各特徴空間の特性を考察する。つぎに、不特定タスクにおける特定話者および不特定話者を対象とした認識実験を行い、認識対象と各特徴量に対する時間窓幅の関係を考察する。

5.1 実験条件

音声資料およびその分析条件は3.と同様とする。各特徴パラメータは、各次元ごとに平均と分散により-10~10の範囲に正規化して使用した。この正規化は、ケプストラム軸と Δ ケプストラム軸を含む単一の空間で距離計算を行う (単一コードブック) ためのものである。

5.2 不特定話者・特定タスク連続音素認識実験

Cw, Dw を0,1,3,5,...と変化させ、それぞれの組合せに対して、10名の212単語データから、カテゴリ非依存・依存の二つのコードブックを作成した。コードブックサイズは、カテゴリ非依存が512、カテゴリ依存が各音素16の計400とする。

HMMは各コードブックおよびサンプルデータを用いて初期モデルを学習し、さらに15名の212単語を追加して連結学習を行った。HMMの状態数は4とする。また、特徴ベクトルの時間的表現能力を見るために、HMMモデルおよび音素認識システムには、継続時間などの時間制御は一切使用しないものとした。ここでは、評価用データとして、不特定話者・特定単語である5名の212単語を使用し、(7)式で定義される音素Accuracyを用いて評価する。

$$\text{Accuracy} = \frac{\text{総音素数} - \text{挿入} - \text{置換} - \text{脱落}}{\text{総音素数}} \quad (7)$$

5.2.1 コードブック作成方法の比較

カテゴリ非依存コードブックにおける各時間窓幅の組

Table 2 Relationship between width of time window and phoneme recognition rate using a categories-independent codebook ($Ns=7$)

cepstrum segment length(Cw)	Δ cepstrum segment length(Dw)									
	0	1	3	5	7	9	11	13	15	17
0	—	60.8	55.8	52.0	48.3	45.2	23.8	13.7	10.0	12.9
1	46.7	73.2	70.3	75.1	76.2	78.0	71.1	76.9	76.1	75.3
3	46.7	66.5	70.0	76.0	71.7	71.7	74.0	72.6	77.2	70.9
5	62.0	70.1	74.0	69.8	77.2	78.3	77.7	78.3	78.1	77.6
7	69.3	70.9	72.8	74.7	77.2	77.4	78.0	77.6	78.8	77.8
9	72.2	68.0	74.1	73.3	72.2	77.5	71.3	74.5	78.2	77.7
11	72.3	70.7	75.2	76.0	75.7	74.4	77.9	77.2	76.8	77.1
13	68.4	69.8	72.2	74.4	72.5	71.3	71.9	73.6	76.2	76.6

※ The recognition rate 78.0% or above is written in Gothic type, and the underline is used at the highest value.

Table 3 Relationship between width of time window and phoneme recognition rate using a categories-dependent codebook ($Ns=7$)

cepstrum segment length(Cw)	Δ cepstrum segment length(Dw)									
	0	1	3	5	7	9	11	13	15	17
0	—	53.3	72.4	76.6	80.5	80.3	83.2	83.5	84.1	84.5
1	29.8	74.4	77.9	83.1	85.4	85.6	85.6	86.8	<u>87.5</u>	87.2
3	54.4	67.7	75.2	78.9	80.7	85.6	85.2	85.9	86.8	87.1
5	70.4	68.4	78.7	77.5	83.0	83.8	84.8	84.5	86.7	86.2
7	75.5	74.7	77.8	80.3	82.5	83.4	84.7	85.5	84.4	86.1
9	78.4	79.8	79.4	79.7	82.5	84.1	84.4	84.9	85.2	84.9
11	79.8	79.8	81.2	82.0	82.8	83.5	83.7	84.5	85.0	85.1
13	77.9	81.1	81.4	81.2	83.1	83.4	84.3	85.1	84.9	84.8

※ The recognition rate 86.0% or above is written in Gothic type, and the underline is used at the highest value.

合せと認識率の関係を Table 2 に示す。 $Cw=5\sim9$ に対して、 $Dw=9\sim17$ 程度にピークが見られるものの、 DWWS の効果は明確に現れていない。ただし、 $Cw > Dw$ の領域に比べると、 $Cw = Dw$ および $Cw < Dw$ の領域は良好な結果を示している。

ケプストラムのみのセグメント単位入力では $Cw=9, 11$ 程度までは良好に働いており、 HMM 内に時間特性を導入する要素となっていると考えられる。一方、 Δ ケプストラムのみのセグメント単位入力では、時間窓幅の増加に伴い認識率は大きく低下する。とくに Dw が 9 と 11 の間で 20% 以上の差があるが、1 フレームのケプストラムを加えることで、認識率およびその減少傾向は大きく改善されている。

つぎに、カテゴリ依存コードブックにおける各時間窓幅の組合せと認識率の関係を Table 3 に示す。認識率は、カテゴリ非依存方式に比べかなり高く、 $Cw=1$ 、 $Dw=15$ において最高認識率 87.5% を得た。傾向を見る

と、ケプストラムの時間窓幅を小さく、 Δ ケプストラムの時間窓幅を大きく取った DWWS の領域で高い認識率が得られており、とくに Dw を 7 以上に取った場合に良好な認識性能が得られることが分かる。

ここで、この結果と等価セグメント長の結果 (Fig. 3) との相関を取ったところ 0.55 と、また特徴ベクトルの空間特性をじかに反映する LVQ 音声認識システム [15] における同様の認識結果との相関を取ったところ 0.82 となり高い値を示した。これより、分離度を用いた評価基準により各特徴ベクトルの有効性を評価することは妥当であり、一連の実験が正当であるといえる。ただし、HMM および LVQ の各認識システムに対して相関値に差が出たのは、前者がシンボル時系列をモデル化したものであるのに対し、後者はベクトル空間そのものをモデル化している違いであり、LVQ の方がより統計計算の考え方に則しているためであると考えられる。

カテゴリ依存コードブックにおいて注目すべきことは、

Δ ケプストラムのみのセグメント単位入力でも非常に高い認識率を示し、時間窓幅の増加に伴い認識率がさらに向上している点である。これは、カテゴリ非依存コードブックにおける結果と全く逆の傾向である。一方、ケプストラムセグメントのみでは、共に $Cw=9$ 程度でほぼ収束している。以上のようにコードブック作成方法の違いにおいて、ケプストラムおよび Δ ケプストラムのみでの認識性能に大きな差が見られた。これについて、次節で詳しく説明する。

5.2.2 単一特徴量における検討

Figs. 6, 7 に、カテゴリ非依存コードブックおよびカテゴリ依存コードブックを用いた場合の、ケプストラムまたは Δ ケプストラムのみで構成された特徴ベクトルの時間窓幅と誤り率の関係を示す。ただし、図における del., ins., rep. は順に脱落誤り率、挿入誤り率、置換誤り率である。

Fig. 6 Relationship between width of time window and error rate using a categories-independent codebook ($N_s=7$)

Fig. 7 Relationship between width of time window and error rate using a categories-dependent codebook ($N_s=7$)

ケプストラムセグメントは、どちらのコードブックにおいても、 $Cw=9$ 程度までの時間窓幅の増加に伴い、置換・挿入誤り率が減少している。これは、ベクトル次元数の増加による分離空間の拡大と、入力ベクトル内への時間特性の導入によるものと考えられる。

一方、 Δ ケプストラムセグメントは、カテゴリ非依存コードブックにおいて、時間窓幅の増加に伴う置換・挿

入誤り率の増加が顕著で、とくに、 $Dw=9$ から 11 の間で 20% 以上の挿入誤り率の増加を示す。これに対し、カテゴリ依存コードブックでは、時間窓幅の増加に対し、置換・挿入誤り率は比例的に減少する。

以上の実験から、各特徴量およびそのセグメントに関して以下のように考察される。

Δ ケプストラムは、音声の時間情報を取り入れた有効な特徴パラメータであるが、セグメントを構成して利用する場合、時間窓幅の増加に伴って、現時刻の音素と同時に隣接する他の音素情報も多く含んだ特徴ベクトルとなる。このために、ベクトル空間内では各カテゴリ境界が不明確となり、線形分離型のカテゴリ非依存コードブックでは分類が困難で、認識率の急激な低下を引き起こすものと考えられる。

一方、各セグメントベクトルに対して教師を与え、前後の音素情報を含んだうで現時刻の音素を特徴付けることができれば、一種の音素環境依存型の入力ベクトルとして効果的に働くことが期待される。カテゴリ依存コードブックは、このような教師付きクラスタ化を実現しており、 Δ ケプストラムセグメントで得られた高認識率を説明することができる。このようなベクトル空間の教師付き非線形分離は、LVQ やニューラルネットワークなどの得意な分野であり、これらの手法とセグメントベクトルを組み合わせることが効果的で、とくに、音素性が小さい Δ ケプストラムで有効に働くものと考えられる。

ケプストラムは、当該時刻の音素性を強く表した特徴ベクトルであり、ベクトル空間内では各音素クラスがある程度固有の位置に存在し、その分布もさほど複雑ではなく、線形的なクラスタリングでも十分分類できるものと考えられる。したがって、ケプストラムにおける 7~9 フレーム程度までのセグメントは、時間的特性を導入するために効果的に機能するが、それ以上時間窓幅を広げると、セグメント内に含まれる他の音素の影響を強く受けて逆効果となる。また、統計的な分析からも明らかに、大きな時間窓幅から構成されるケプストラムセグメントに Δ ケプストラムを付加しても、ケプストラムセグメント内に重複した区間情報が存在し、さほどの効果は期待できない。以上のことから、ケプストラムに対する時間窓幅は小さく取るほうが適していると考えられる。

5.3 不特定タスク連続音素認識実験

DWWS は Δ ケプストラムに対して非常に大きな時間窓をかけているため、一種の音素環境依存型の入力ベクトルであると述べた。また、ベクトル量子化 (VQ) ベースのシステムでは、学習データに強く依存しすぎるという問題もあり、DWWS は学習コンテキストおよび学習話者へ強く依存した入力ベクトルであるとも考えられる。

そこで、評価用データとして不特定単語である 1000 単

Table 4 Relationship between width of time window and phoneme recognition rate for open task ($Ns=7$)

cepstrum segment length(Cw)	Δ cepstrum segment length(Dw)									
	0	1	3	5	7	9	11	13	15	17
0	—	63.8	75.5	79.0	82.9	83.3	84.5	83.8	83.8	82.4
	—	64.7	77.1	79.7	82.3	83.4	83.7	83.8	83.2	82.0
1	38.3	76.3	82.3	84.1	85.4	85.3	86.5	85.5	86.4	86.2
	39.3	76.5	81.6	83.3	84.7	85.1	84.9	84.5	<u>85.9</u>	84.4
3	60.1	74.4	83.4	83.8	86.2	<u>87.1</u>	86.2	86.2	86.2	85.7
	61.3	73.5	82.3	83.1	84.8	85.0	85.0	84.7	84.3	84.6
5	74.7	75.7	81.8	83.7	85.5	84.7	85.0	86.0	86.7	85.6
	72.9	73.6	79.9	81.9	83.7	83.1	83.6	84.5	84.2	83.8
7	78.5	78.3	82.1	84.8	84.4	85.9	85.4	86.3	84.9	86.0
	76.0	75.7	81.0	83.8	83.2	84.5	83.3	84.4	83.7	84.0
9	79.2	81.0	81.3	84.1	84.9	83.8	85.2	85.9	85.9	84.4
	76.7	80.0	79.9	82.1	83.2	82.2	83.2	85.0	84.7	83.1
11	80.1	80.5	82.2	81.0	83.7	84.9	85.1	85.1	85.5	85.1
	78.6	79.6	81.2	81.5	82.5	83.2	82.7	84.0	83.7	83.4
13	80.1	81.7	82.6	83.0	84.3	84.4	83.9	85.7	85.1	84.2
	78.3	81.1	81.5	81.0	83.6	83.2	81.7	83.8	83.4	82.7

※ Upper section : known speaker (2 male), Lower section : unknown speaker (3 male)

※ The recognition rate above 5 in rank is written in Gothic type, and the underline is used at the highest value.

語について、特定話者2名および不特定話者3名の発声データを用いて検討を行った。その結果を Table 4 に示す。ただし、コードブックはカテゴリ依存とし、HMM は状態数を5として、前実験と同様に作成した。Table 4 は、上段が特定話者、下段が不特定話者に対する各時間窓幅の組合せと認識率の関係を示す。

不特定単語・特定話者における認識率では、 $Cw=3$ 、 $Dw=9$ において最高の87.1%を示すが、大きめの Cw までのかなり広いDWWSの領域で飽和的な値を示す。これに対して、不特定単語・不特定話者における認識率は、 $Cw=1$ 、 $Dw=15$ において最高の85.9%を示す。この場合、 Cw は小さく取るほうが望ましく、 Dw を大きく取りすぎると認識率は低下する。

5.4 認識対象とDWWSに関する考察

これまでの認識実験により、DWWSは認識対象によらず非常に有効で、学習データへの依存性も小さい入力ベクトルであることが確かめられた。認識対象と各特徴パラメータに対する最適時間窓幅の関係をまとめると以下ようになる。

- (1) ケプストラムに対する最適時間窓幅は発声話者に関係し、話者が未知の場合小さな値をとる。したがって、 Cw を小さくすると話者性が減少する。
- (2) Δ ケプストラムに対する最適時間窓幅は発声内容に関係し、タスクが未知の場合若干小さな値をとる。ただし、発声話者が既知の場合大きくとって

もさほど問題がない。

以上、DWWSはセグメント単位入力の本質である時間的特性をより有効に表現した不特定話者音声認識に適した入力パターンであると考えられる。不特定話者・不特定単語の認識率をさらに向上させるためには、より多くの話者および音素環境を学習する必要がある。

6. おわりに

音声の時間的变化を表現するためのセグメント単位入力において、クラスタ分析における分離度を用いた評価尺度を定義し、様々な組合せからなる特徴ベクトルを評価した。その結果を基に、ケプストラムおよび Δ ケプストラムの有効な併用方法として、ケプストラムには時間窓幅を小さく、 Δ ケプストラムには時間窓幅を大きくとるDWWSを提案した。

DWWSを離散分布型HMMに適用し、種々の認識実験を行った結果、カテゴリ依存型のコードブックを用いることで、その有効性が確かめられた。また、この結果は分離度による評価結果にほぼ則したのとなっており、音声認識における特徴ベクトルの評価方法として分離度を用いることで、おおよその見当をたてて実験を行うことが可能であることが分かった。

さらに、DWWSは認識対象にかかわらず有効であり、学習データに強く依存するというVQベースシステムの問題も大幅に改善した非常に簡単に構成できる音声認識

に適した特徴ベクトルであるといえる。

また、DWWSは前後の音素環境をふまえた一種の音素環境依存型の入力ベクトルであり、Triphoneのような効果を実現しているとも考えられる。ただし、認識システムや認識対象などの条件により、最適な時間窓幅は若干異なり、様々な環境に対応する特定の最適値を決定することはできない。

今後、異なる発話様式に対する検討を行うと共に、DWWSの雑音耐性、音素環境依存モデルとの比較、またセグメント長を入力に合わせて自由に変化することのできる可変長セグメント単位入力への拡張についても検討する予定である。

参考文献

- [1] S. Furui: Speaker - independent isolated word recognition using dynamic features of speech spectrum; *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-34, No. 1, pp. 52-59 (1986)
- [2] 相川, 河原, 東倉: 順行マスキングの時間周波数特性を模擬した動的ケプストラムを用いた音声認識; 電子情報通信学会論文誌 A, Vol. J76-A, No. 11, pp. 1514-1521 (1993)
- [3] 井出, 牧野, 城戸: 時間周波数パターンを用いた子音の認識; 音響学会音声研資, S82-23, pp. 177-184 (1982)
- [4] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang: Phoneme recognition using time-delay neural networks; *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-37, pp. 328-339 (1989)
- [5] E. McDermott and S. Katagiri: Shift-invariant multi-category phoneme recognition using Kohonen's LVQ2; *Proc. of the Int. Conf. on Acoustics, Speech & Signal Processing*, S3. 1, pp. 81-84 (1989)
- [6] 中川, 平田, 橋本: 連続分布型 HMM による日本語音韻認識; 日本音響学会誌, Vol. 46, No. 6, pp. 486-496 (1990)
- [7] M. J. Ressel, and A. E. Cook: Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition; *Proc. of the Int. Conf. on Acoustics, Speech & Signal Processing*, pp. 5-8 (1985)
- [8] F. J. Smith, J. Ming, P. O'Boyle and A. D. Irvine: A hidden Markov model with optimized inter-frame dependence; *Proc. of the Int. Conf. on Acoustics, Speech & Signal Processing*, Vol. I, pp. 209-212 (1995)
- [9] 本間, 嵯峨山: 音素環境クラスタリングに基づいた HMM による音素認識; 音響学会秋期全大, 1-1-18, pp. 35-36 (1989)
- [10] 中川, 山本: セグメント統計量を用いた隠れマルコフモデルによる音声認識; 電子情報通信学会論文誌, Vol. J79-D-II, No. 12, pp. 2032-2038 (1996)
- [11] 大倉, 杉山: セグメント統計量を用いた HMM 音声認識; 信学技報, SP91-55, pp. 1-6 (1991)
- [12] 大槻, 大友: 音声認識に適した特徴ベクトルの構成の検討; 信学技報, SP97-61, pp. 27-34 (1997)
- [13] H. P. Frieman and J. Rubin: On some invariant criteria for grouping data; *J. American Statistical Association*, Vol. 63, pp. 1159-1178 (1967)
- [14] S. Makino, M. Endo and K. Kido: Recognition of phonemes in continuous speech using a modified LVQ2 method; *The Journal of the Acoustical Society of Japan (E)*, Vol. 13, No. 6, pp. 351-360 (1992)
- [15] 二宮, 大槻, 大友: 2 段窓セグメント単位入力を用いた LVQ 音声認識システムの検討; 信学技報, SP96-27, pp. 53-60 (1996)