音楽情報科学 24－17
(1998. 2. 20)

# 音声認識における２段窓セグメント単位入力の有効性

二宮 和則　　大槻 恭士　　大友 照彦

山形大学工学部

〒992－0038 山形県米沢市城南 4－3－16
TEL : 0238-26-3382

E-mail : (ninomiya,otsuki,ohtomo) @banana.yz.yamagata-u.ac.jp

あらまし これまで我々は、音声の時間特性を効果的に表現できるセグメント単位入力として、ケプストラムの時間窓を小さく、Δケプストラムの時間窓を大きくした２段窓セグメント単位入力(DWWS)を提案し、LVQ音声認識システムにおいてその有効性を示してきた。このDWWSを離散分布型HMMに適用し、種々の認識実験よりセグメント単位入力における各特徴空間の特性を考察、音声認識全般における有効的な特徴ベクトルであることを明らかにする。その結果、カテゴリー依存型コードブックを用いることで、前後の音素を含んだΔケプストラムセグメント空間を非線形にクラスタリングすることが可能となり、２段窓セグメント単位入力により高い認識率が得られることが分かった。

キーワード ケプストラム，Δケプストラム，離散分布型HMM,２段窓セグメント単位入力，クラスタリング

# The Efficiency of the Dual-Width Windowed Segment on Speech Recognition

Kazunori NINOMIYA , Takashi OTSUKI and Teruhiko OHTOMO

Faculty of Engineering ,Yamgata University

4-3-16 Jounan,Yonezawa-shi,992-0038 Japan
TEL : 0238-26-3382

E-mail : (ninomiya,otsuki,ohtomo) @banana.yz.yamagata-u.ac.jp

Abstract In previous work, for representing the dynamic feature of speech, we proposed the dual-width windowed segment (DWWS) which consists of short segment of cepstrum and long segment of delta cepstrum, and showed the efficiency of DWWS on LVQ speech recognition system. In this report, combining DWWS with the segment quantizing discrete HMM (DWWSQ-HMM), we carry out experiments of speech recognition to investigate the character of the cepstral segment vector and the delta cepstral segment vector. As a result, it is shown that the delta cepstral segment well represents contexts instead of difficulty in being divided linearly into phoneme classes, and also shown that combined with category dependent codebook (e. g. LVQ) which has nonlinear discrimination, DWWS brings high performance of speech recognition.

key words cepstrum, Δcepstrum, discrete HMM, dual-width windowed segment, clustering

## 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取り入れることが重要とされ様々な検討がなされている。

時間的・動的特性を含んだ特徴量として、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、聴覚の順向マスキングに基づく動的ケプストラム[2]などが提案されており、特に前者は様々なシステムで広く利用されている。また、当該フレームの入力に加えその前後数フレームを使用するセグメント単位入力[3][4]も有効的な手法として検討されている。

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、もしくはΔケプストラムは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されるのが一般的であった。これに対して我々は、両者の特性・相関等を考慮して、ケプストラムに対する時間窓幅は小さく、Δケプストラムに対する時間窓幅は大きく取る２段窓セグメント単位入力(Dual-Width Windowed Segment：以降DWWS)を提案し、LVQ音声認識システムに適用し、その有効性を示してきた[5]。

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、LVQ等と比べると時間的特性を上手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替わっても出力確率値に影響がないなど、その表現能力は十分なものではない。よって、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]などの時間特性を表現したモデルが提案・検討されている。

またセグメント単位入力を用いた検討として、文献[9]では、連続分布型HMMにおいてその有効性を示している。さらに、離散分布型HMMにおいてもSQ-HMM[10]として対雑音性への効果が示されている。

本報告では、DWWSを離散分布型HMMに適用したDWWSQ-HMM音声認識システムを構築し[11]、種々の認識実験により、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し、音声認識におけるDWWSの有効性を示す。

## 2. DWWSの構成

ケプストラムの１次の回帰係数であるΔケプストラムは、当該フレームの前後nフレームに対して重み付き時間窓をかけて分析を行う。従って、このΔケプストラムは、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であると言える。

一方、セグメント単位入力は、(Δ)ケプストラムに対して、当該フレームの前後wフレームに時間窓をかけて切り出した(Δ)ケプストラム方向と時間方向の２次元的な特徴ベクトルである。

ここで、セグメント単位入力におけるケプストラムとΔケプストラムの併用を考える。２つの特徴量は、ベクトル空間上で、それぞれ位置や大きさ・性質の異なるものである。また、前述のようにΔケプストラムはケプストラムに対して時間窓をかけて求めた特徴パラメータであり、セグメント単位入力作成時の時間窓内には重複した情報が取り入れられることになる。よって、両者の特性を生かしたセグメントを構成する必要があり、各特徴パラメータに対してそれぞれ異なる時間窓をかけ１つのセグメント単位入力を作成することが意味を持つと考えられる。

まず、各特徴パラメータの特性で考えると、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、Δケプストラムは音素性そのものは小さく、時間的変化特性を表現したパラメータであると言える。

次に、情報量の観点から考えると、Δケプストラム１フレームには、ケプストラムのNs(=2n+1)フレーム分の情報が含まれており、この区間内でのケプストラムセグメントは線形従属な情報となる。逆に１フレームのケプストラムに対して、Δケプストラムセグメントの増加は、線形独立なフレーム数の増加につながると言える。文献[12]では、特徴ベクトル中の線形独立なフレーム数の増加が、音素クラスの分離に貢献することを統計的手法により分析している。

これらの観点から、ケプストラムに対する時間窓幅(Cw)を小さく、Δケプストラムに対する時間窓幅(Dw)を大きく取ることで、局所的な当該時刻の音素性をふまえた上で、大局的な時間的変化特性を無駄なく取り入れることができると考えられる。このよ

うな窓を用いて構成された特徴ベクトルを２段窓セグメント単位入力(DWWS)と呼ぶこととする。

DWWS を用いた特徴ベクトルの構成を図１に示す。

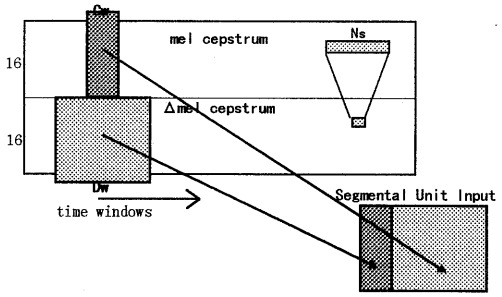



**図１．DWWS を用いた特徴ベクトルの構成法**

## ３．ＤＷＷＳのＨＭＭへの適用

セグメントを用いる方法は、主に NN や LVQ 等の時間的変化を表現できない手法に用いられてきた。そこで我々は、ベクトル空間内での DWWS の特性を直接的に観測するためにLVQ音声認識システム[13]を用いて種々の検討を行い、３フレーム程度のケプストラムセグメントと 15 フレーム程度のΔケプストラムセグメントの併用でかなり高い認識率が得られることを示してきた[5]。

DWWS の HMM への適用において、特徴ベクトル空間を連続的な確率密度で与える連続分布型 HMM では、DWWS のベクトル次元数が極めて大きいため、共分散行列の推定精度が低下するといった問題点がある。従って今回は、離散分布型 HMM を使用する。

離散分布型 HMM では、入力ベクトルを予めベクトル量子化してシンボル系列に変換し、モデルの推定および認識を行う。つまり、入力ベクトルとして DWWS を使用する場合、DWWS を用いて求めたコードブックによりベクトル量子化することになる。

コードブックの作成方法としては、全学習データを一括して LBG 等でクラスタリングし、K個の代表ベクトルを求めるカテゴリー非依存方式と、各音素毎に Ks 個の代表ベクトルを求め、最後にそれらをまとめて１つのコードブック(サイズ＝Ks×音素数)とするカテゴリー依存方式が考えられる。後者は、音声のラベリングデータが必要であるが、各音素毎に求まった代表ベクトルを LVQ で再度学習を行うことも可能であり、より有効的なコードブックが得られることが知られている(LVQ-HMM)。

本報告では、上記２つのコードブック各々についての検討を行う。

## ４．認識実験

DWWSQ-HMM の評価実験として、連続音素認識実験および孤立単語認識実験を行う。

ここで、ケプストラムに対する時間窓幅を Cw、Δケプストラムに対する時間窓幅を Dw とし、Cw=3, Dw=15 といったパターンを Nw=3-15 と表記する。さらに、Cw<DW を２段窓, Cw=Dw を１段窓, Cw>Dw を逆２段窓と呼ぶこととする。

また、DWWS 特徴ベクトルの時間的特性の効果を見るために、HMM モデルおよび認識には、継続時間等のの時間制御はいっさい使用しないものとする。

HMM は Left-right モデルを使用し、状態数等は各実験毎に示している。

認識率には次式(1)で定義される音素 Accuracy を用いる。

$$音素Accuracy(\%)=\frac{総音素数-(挿入+置換+脱落)}{総音素数}\times 100 \quad (1)$$

### ４．１　音声データ

音声データとして、東北大・松下単語音声データベースより男性話者の 212 単語データを使用する。音声の分析条件を表１に示す。ただし、各特徴パラメータは各次元毎に最大値と最小値により－10～10の範囲に正規化して使用している。この正規化はケプストラムとΔケプストラムに対する重み付き距離に対応しており、各特徴量を合わせた単一のコードブックで計算するためのものである。

**表１　音声の分析条件**

| | |
|---|---|
| サンプリング周波数 | 12kHz |
| プリエンファシス | $1-0.97z^{-1}$ |
| ハミング窓幅 | 25ms |
| フレーム周期 | 10ms |
| 特徴パラメータ | メルケプストラム　16次<br>Δメルケプストラム 16次<br>(Δ分析窓幅±30ms) |

### 4.2 不特定話者における連続音素認識実験

Cw, Dw の様々な組み合わせに対し、男性10人の212単語データを使用して、以下の2つのコードブックを作成する（※単一コードブック）。

<カテゴリー非依存コードブック>

全サンプルデータを一括でLBGクラスタリング

<カテゴリー依存コードブック>

各音素毎にLBGクラスタリングしたものを、LVQを用いて修正

上記のコードブックおよびサンプルデータを用いてHMM初期モデルを学習し、さらに男性話者15人の212単語を追加した25人のデータで連結学習を行う。

認識評価用データは、学習用データに含まれない男性話者5人の212単語を使用する。

#### 4.2.1 カテゴリー非依存コードブック

HMMの状態数を4状態、コードブックサイズをK=64, 128, 256として、各窓幅と認識率の関係を見る。この結果を表2に示す。表は上段 K=64, 中段 K=128, 下段 K=256 であり、各々に対する最高認識率を太字で記している。また、認識率 70%以上のセルに網掛けを行っている。

コードブックサイズの増加に伴い認識率は向上している。しかし、全体的に振動的な結果となり、ケプストラムに対する時間窓幅 Cw=5, 7 に対して、Δケプストラムに対する時間窓幅 Dw=9-13 程度にピークが見られるものの、DWWS の効果ははっきりとは表れなかった。

ケプストラムのみでのセグメント単位入力は Cw=9 程度までは良好に働いおり、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ケプストラムのみでのセグメント単位入力では、時間窓幅の増加に伴いかなり認識率が低下している。これは、挿入誤りの増加によるものであり、K=256, Nw=0-17 において挿入誤り率 50.4%となった。しかし、1フレームのケプストラムを加えることで、認識率は大きく改善されている。

#### 4.2.2 カテゴリー依存コードブック

コードブックサイズを各音素毎に16（16×25＝40）、HMMの状態数をS=4, 5, 6と変化させ、各窓幅と認識率の関係を見る。この結果を表3に示す。表は上段 S=4, 中段 S=5, 下段 S=6 であり、各々に対する最高認識率を太字で記している。また、認識率 85%以上のセルに網掛けを行っている。

認識率は前述のカテゴリー非依存方式に比べかなり高い値をとり、状態数S=6, Nw=3-17において88.2%となった。傾向を見ると、ケプストラムに対する時間窓幅を小さく、Δケプストラムに対する時間窓幅を大きく取ったDWWSで高い認識率が得られている。特にDw=7, 9以上の使用で、状態数が小さいモデルでも良好な結果となっている。また、状態数の増加に伴い、入力ベクトルの次元数の高い領域での認識率が平均的に高くなり、また、時間情報の少ない領域(Cw, Dwが小)の認識率も向上した。

特に注目すべき点は、Δケプストラムのみでのセグメント単位入力が、非常に高い認識率となっており、時間窓幅の増加に伴いさらに向上している点である。これは、先ほどのカテゴリー非依存コードブックの結果と全く逆のものであり、挿入, 置換誤りの減少によるもので、Nw=0-1 において置換誤り率 20.5%, 挿入誤り率 12.0%が、Nw=0-17 において各々 8.5%, 6.3%と減少した。

ケプストラムセグメントはCw=7, 9程度でほぼ収束している。

#### 4.2.3 単一特徴量におけるセグメント単位入力

上の2つの実験において大きな違いとなった、単一特徴量におけるセグメント単位入力について、もう少し詳しく検討する。

ここでは、コードブックとして前述の2つに加え、カテゴリー依存で LVQ による修正を行う前のものも使用する。また、量子化誤差等も情報の一つとして検討するために、HMMモデルはコードブックを作成した男性10人で学習したもの（前述の初期モデル）とする。モデル状態数は4とした。

図2にケプストラムおよびΔケプストラムの単一特徴ベクトルに対する時間窓幅と認識率の関係を示す。ここで、C1, C2, C3 はケプストラムを特徴量としたカテゴリー非依存 LBG(K=256), カテゴリー依存 LBG(K=400), カテゴリー依存 LBG+LVQ(K=400) の各コードブックに対する結果であり、同様にΔ1, Δ2, Δ3はΔケプストラムを特徴量とした各コードブックに対するものである。また、図3, 4に時間窓幅と誤り率の変化を示す。

表２．カテゴリー非依存コードブックを用いたHMMにおける時間窓幅と認識率の関係

| Cw \ Dw | 0 | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|
| | — | 35.3 | 34.8 | 31.8 | 42.1 | 26.2 | 15.5 | -16.5 | 1.5 | -5.4 |
| 0 | — | 48.0 | 39.4 | 44.2 | 42.9 | 33.5 | 11.4 | 14.5 | 5.8 | 1.4 |
| | — | 55.0 | 46.7 | 49.3 | 44.1 | 42.3 | 18.9 | 21.3 | 12.1 | 0.4 |
| | 28.0 | 53.7 | 51.4 | 58.8 | 55.9 | 58.9 | 54.3 | 54.6 | 53.2 | 54.5 |
| 1 | 39.1 | 59.5 | 63.2 | 66.8 | 65.4 | 67.8 | 65.9 | 65.5 | 62.1 | 61.4 |
| | 40.6 | 69.7 | 66.9 | 71.2 | 68.7 | 74.1 | 73.4 | 71.6 | 72.3 | 71.4 |
| | 32.2 | 40.3 | 56.1 | 60.2 | 58.6 | 62.3 | 62.4 | 56.8 | 54.5 | 55.5 |
| 3 | 39.6 | 39.4 | 57.7 | 58.5 | 64.9 | 66.4 | 63.4 | 61.8 | 64.8 | 66.7 |
| | 44.6 | 59.8 | 65.4 | 66.5 | 70.4 | 70.2 | 68.4 | 70.6 | 72.3 | 69.7 |
| | 41.9 | 42.5 | 50.7 | 55.3 | 62.4 | 55.9 | 61.3 | 60.1 | 61.3 | 61.9 |
| 5 | 52.5 | 54.8 | 59.4 | 65.9 | 70.5 | **70.8** | 65.7 | 68.9 | 67.5 | 66.9 |
| | 58.3 | 59.1 | 68.3 | 70.5 | 74.5 | 75.0 | 75.2 | 74.6 | 74.5 | 73.3 |
| | 44.5 | 48.9 | 55.2 | 57.0 | 56.3 | 58.9 | 59.4 | **64.8** | 62.4 | 59.3 |
| 7 | 51.5 | 62.1 | 62.9 | 65.5 | 69.1 | 66.5 | 68.8 | 65.5 | 67.5 | 67.8 |
| | 62.9 | 65.8 | 70.1 | 72.9 | 72.7 | 73.4 | 74.8 | 75.4 | 75.2 | 72.3 |
| | 55.4 | 56.4 | 59.0 | 59.0 | 60.7 | 61.7 | 56.1 | 63.0 | 57.8 | 61.1 |
| 9 | 61.2 | 63.7 | 64.7 | 64.4 | 68.1 | 68.9 | 67.3 | 68.4 | 62.3 | 67.1 |
| | 69.9 | 69.5 | 71.4 | 70.3 | 73.3 | 74.9 | 68.4 | 73.8 | 70.9 | **75.6** |
| | 54.6 | 51.1 | 56.6 | 56.4 | 56.1 | 57.5 | 59.6 | 62.4 | 64.3 | 61.0 |
| 11 | 61.4 | 64.3 | 64.2 | 64.7 | 64.7 | 61.5 | 66.7 | 67.6 | 68.8 | 70.0 |
| | 70.2 | 59.7 | 70.6 | 72.7 | 72.0 | 66.3 | 73.5 | 73.2 | 73.9 | 72.3 |
| | 49.7 | 53.0 | 55.8 | 51.4 | 56.0 | 57.0 | 62.7 | 63.8 | 64.7 | 61.9 |
| 13 | 62.6 | 62.3 | 65.0 | 61.9 | 67.3 | 66.6 | 65.6 | 68.1 | 66.4 | 68.8 |
| | 69.8 | 58.8 | 69.3 | 71.6 | 71.9 | 68.9 | 73.1 | 70.6 | 72.2 | 74.0 |
| | 48.6 | 48.4 | 55.3 | 60.2 | 59.6 | 61.1 | 59.4 | 61.1 | 61.8 | 60.8 |
| 15 | 63.6 | 63.5 | 60.5 | 61.8 | 68.1 | 67.0 | 66.1 | 67.4 | 62.7 | 69.4 |
| | 68.9 | 67.4 | 69.6 | 62.9 | 71.3 | 71.2 | 71.5 | 71.1 | 72.9 | 72.7 |

※上段：K=64，　中段：K=128，　下段：K=256

表３．カテゴリー依存コードブックを用いたHMMにおける時間窓幅と認識率の関係

| Cw \ Dw | 0 | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|
| | — | 53.3 | 72.4 | 76.6 | 80.5 | 80.3 | 83.2 | 83.5 | 84.1 | 84.5 |
| 0 | — | 64.7 | 75.1 | 78.0 | 81.8 | 82.3 | 83.9 | 84.6 | 84.1 | 84.8 |
| | — | 67.4 | 76.6 | 78.1 | 82.8 | 84.0 | 84.9 | 85.4 | 84.3 | 86.2 |
| | 29.8 | 74.4 | 77.9 | 83.1 | 85.4 | 85.6 | 85.6 | 86.8 | **87.5** | 87.2 |
| 1 | 32.8 | 74.7 | 81.1 | 83.3 | 85.6 | 86.3 | 86.1 | 86.4 | **87.8** | 87.6 |
| | 49.4 | 77.1 | 82.4 | 84.8 | 86.7 | 87.1 | 87.0 | 87.5 | 87.9 | 87.7 |
| | 54.4 | 67.7 | 75.2 | 78.9 | 80.7 | 85.6 | 85.2 | 85.9 | 86.8 | 87.1 |
| 3 | 56.9 | 70.3 | 79.1 | 82.6 | 85.0 | 86.0 | 85.5 | 85.6 | 86.7 | 87.4 |
| | 61.8 | 74.5 | 81.4 | 84.5 | 84.8 | 86.8 | 86.5 | 87.2 | 87.7 | **88.2** |
| | 70.4 | 68.4 | 78.7 | 77.5 | 83.0 | 83.8 | 84.8 | 84.5 | 86.7 | 86.2 |
| 5 | 72.9 | 72.1 | 79.3 | 79.1 | 83.7 | 84.6 | 84.7 | 85.4 | 86.6 | 86.4 |
| | 78.4 | 79.6 | 82.6 | 82.0 | 85.1 | 86.4 | 86.1 | 86.0 | 87.3 | 87.3 |
| | 75.5 | 74.7 | 77.8 | 80.3 | 82.5 | 83.4 | 84.7 | 85.5 | 84.4 | 86.1 |
| 7 | 76.5 | 76.3 | 78.9 | 81.4 | 83.4 | 85.0 | 84.6 | 85.9 | 84.2 | 86.5 |
| | 80.2 | 78.8 | 80.3 | 83.6 | 84.7 | 86.5 | 85.5 | 87.3 | 86.9 | 87.4 |
| | 78.4 | 79.8 | 79.4 | 79.7 | 82.5 | 84.1 | 84.4 | 84.9 | 85.2 | 84.9 |
| 9 | 77.3 | 79.9 | 77.8 | 81.8 | 83.1 | 83.9 | 85.1 | 84.7 | 85.1 | 84.8 |
| | 81.0 | 79.8 | 80.0 | 83.5 | 84.4 | 85.3 | 86.1 | 86.2 | 86.4 | 86.1 |
| | 79.8 | 79.8 | 81.2 | 82.0 | 82.8 | 83.5 | 83.7 | 84.5 | 85.0 | 85.1 |
| 11 | 80.4 | 80.6 | 81.8 | 82.3 | 82.1 | 83.8 | 85.1 | 84.3 | 85.5 | 86.2 |
| | 82.1 | 81.8 | 83.7 | 83.6 | 84.2 | 85.3 | 85.6 | 86.1 | 86.2 | 86.8 |
| | 77.9 | 81.1 | 81.4 | 81.2 | 83.1 | 83.4 | 84.3 | 85.1 | 84.9 | 84.8 |
| 13 | 79.3 | 81.6 | 81.7 | 81.5 | 83.4 | 83.7 | 84.4 | 84.5 | 84.9 | 84.8 |
| | 80.8 | 82.6 | 82.9 | 83.0 | 83.8 | 84.1 | 84.8 | 85.2 | 85.9 | 86.0 |
| | 79.7 | 81.0 | 79.8 | 82.7 | 81.7 | 82.8 | 83.2 | 83.9 | 84.5 | 84.2 |
| 15 | 80.5 | 80.5 | 81.9 | 82.4 | 82.0 | 83.2 | 83.4 | 83.4 | 84.7 | 84.9 |
| | 82.3 | 81.6 | 82.1 | 83.5 | 83.8 | 85.0 | 83.7 | 84.7 | 85.2 | 85.7 |

※上段：S＝4，　中段：S＝5，　下段：S＝6

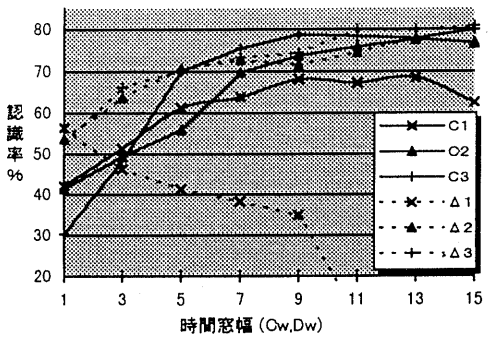



図２．単一特徴量における時間窓幅と認識率の関係

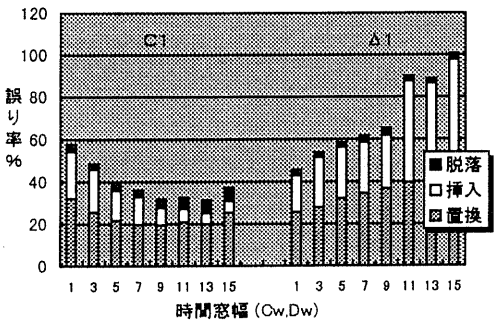



図３．単一特徴量における誤り率の変化１

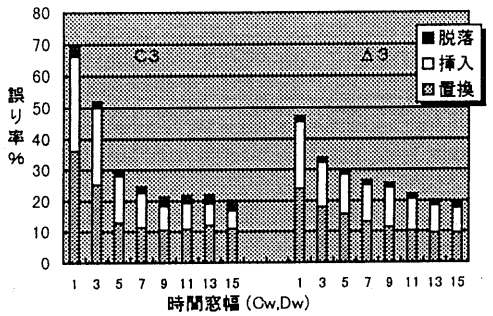



図４．単一特徴量における誤り率の変化２

ケプストラムセグメントでは、ど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,挿入誤りが大きく減少している。これは、ベクトル次元数の増加による量子化誤差の減少と、入力ベクトル内への時間特性の導入によるものであると考えられる。また、Cw が小さい範囲では、コードブックサイズが小さいにもかかわらず、カテゴリー非依存コードブックを用いた方が高い認識率となっている。

Δケプストラムセグメントでは、カテゴリー非依存コードブックにおいて、置換,挿入誤りの増加により認識率が大きく減少している。それに対し、カテゴリー依存コードブックでは、全く逆の現象が起こっている。これは、Δケプストラムセグメントのベクトル空間が単純な線形クラスタリングで分類不可能であり、この特性が時間窓幅の増加に伴い強くなっているためと考えられる。

また、カテゴリー依存コードブックでは、全体的に 80%前後の認識率に収束している。このうち、Δケプストラムセグメントの結果(Δ３)は、表３に示す前実験に比べ低い値となっており、Nw=0-17 で約 3.2%の差があった。逆に、ケプストラムセグメントでは大きな差は見られなかった。これは、連結学習による学習サンプルの増加によるものであり、量子化誤差を連結学習により吸収できるためであると考えられる。

LVQ によるコードブック修正は、各々に対して効果的に働いており、特に時間窓幅の大きいパターンに有効的であることが分かる。

### 4.2.4 考 察

以上の結果より、HMM 音声認識システムにおいてもカテゴリー依存型のコードブックを用いることで、DWWS が有効的に働くことが確かめられた。

ここで、各特徴量およびそのセグメントに関して以下のようなことが考察される。

Δケプストラムセグメントは、音声の時間情報を取り入れた有効的な特徴ベクトルである。しかし、時間窓幅の増加に伴い、当該時刻の入力に対して前後に(2)式で定義される区間の情報が含まれていることになり、現時刻の音素と同時に隣接する他の音素の情報まで取り入れた入力ベクトルとなっている。

$$\frac{(Dw-1)+(Ns-1)}{2} \qquad (2)$$

(Dw：Δケプストラム時間窓幅,　Ns：Δ分析窓幅)

よって、ベクトル空間内では隣接する音素境界が明確ではなく、線形分離型のカテゴリー非依存コードブックでは分類不可能であり、認識率の急激な低下を引き起こしたと考えられる。そこで、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上での現時刻の音素を特徴付けることで、音

素環境依存の入力ベクトルとして分類可能となる。このような、ベクトル空間の教師付き非線形分離は、LVQ や NN の得意な分野であり、これらの手法と組み合わせて使用することが効果的である。またこれは、ケプストラムに比べ音素性が小さいΔケプストラムで有効的に働くものと考えられる。

逆に、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がある程度固有の位置に存在しており、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。これは、非常に分布の小さいΔケプストラムの空間を引き延ばすのに効果的である。また、ケプストラムにおいても、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有効的であると言えるが、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逆効果となる。さらに、Δケプストラムを導入することで、ケプストラムに線形従属な区間が存在することとなり、重複した情報としてさほどの効果は得られないため、ケプストラムに対する時間窓幅は小さく取る方が好ましくなる。

また、ケプストラムに比べΔケプストラムは、フレーム間の相関が小さく、HMM の入力にも適していると考えられる。

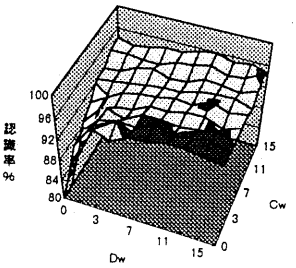
### 4．3　孤立単語認識実験

DWWS では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に、ある程度のコンテキスト依存性があると考えられる。よって、サブワード等の時間的に幅のある認識対象に強い入力であると言える。そこで、212 単語の単語 HMM を作成し、孤立単語認識実験を行う。

学習には、男性 25 人の 212 単語データを使用し、HMM の状態数を各単語の音素数×１，２で検討を行う。認識には、学習に使用していない男性５人の 212 単語データを使用する。ただし、コードブックは先ほど良好な結果の得られた、カテゴリー依存(LVQ で修正)を使用する。

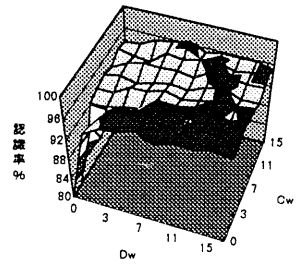
図５，６に各状態数での単語 HMM における時間窓幅と認識率の関係を示す。

この結果、単語 HMM においても DWWS が有効的に働いているのが分かる。最高認識率は、状態数が音素




（状態数＝音素数×１）

**図５．単語 HMM における時間窓幅と認識率の関係１**




（状態数＝音素数×２）

**図６．単語 HMM における時間窓幅と認識率の関係２**

数×２の Nw=1-13 で 97.8%となった。

一般に、音素環境依存型の HMM モデル(ここでは単語 HMM)を構成する場合、各音素毎に３状態程度の状態数を設定する。本実験では、各音素 1,2 状態での検討しか行わなかったが、２つの差がさほど見られなかった。これは、セグメント単位の入力ベクトルが音素環境を取り入れており、小さな状態数でも十分にモデルを表現できるためと考えられる。ただし、今回の実験では、単語の学習データが話者数分の 25 サンプルしかなかったため、モデルを十分に推定できなかったことも、１つの要因であると考えられる。

## 5．ま と め

音声の時間情報を効果的に表現した２段窓セグメント単位入力(DWWS)を離散分布型 HMM に適用し、各種認識実験によりその有効性を確かめ、セグメント単位入力における各特徴空間の特性を考察した。

この結果、離散分布型 HMM に適用する場合、カテゴリー依存コードブックを用いることで、Δケプス

トラムのセグメント特徴空間を非線形分離することが可能となり、音声の時間情報を取り入れる強力な入力ベクトルとなることが分かった。また、ケプストラムは、全体のベクトル空間の引き延ばしと、現時刻の音素性を明確に特徴付けるために、小さな時間窓幅で使用することが効果的であり、これらが、DWWS の本質となっていることが分かった。

以上のことから DWWS は、biphone や triphone HMM のような性能をもつ音素環境依存型の入力ベクトルであると考えられる。一方、タスク open のコンテキスト独立なデータに対しても有効的に働くことが確かめられており[5]、類似した隣接音素クラスを上手くクラスタリングして代表ベクトルで表現しているものと考えられる。よって、DWWS は非常に簡単に構成できる、音声認識全般での有効的な入力ベクトルであると言える。また、セグメント単位入力の対雑音性, Δケプストラムの不特定話者に強い性質等、様々な有効性が考えられる。

ただし、このように大きなセグメントを構成した場合、発話様式に大きく依存してくる恐れがある。また、各音素毎・フレーム毎に最適な時間窓幅が存在すると考えられる。これらのことより、セグメント長を入力に合わせ自由に変化することのできる可変セグメント単位入力への拡張が望まれる。その他、DWWS の連続分布型 HMM への適用等を今後の課題とする。

## 文　献


[1] Furui S.,"Speaker-independent isolated word recognition useing dynamic features of speech spectrum", IEEE Trans. Acoust., Speech & Signal Process., Vol. ASSP-34, No. 1, pp. 52-59, (1986-2)

[2] 相川清明, 河原英記, 東倉洋一,"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", 信学論 (A), vol. J76-A, no. 11, pp. 1514-1421, (1993-11)

[3] 井出和之, 牧野正三, 城戸健一: "時間周波数パタンを用いた子音の認識", 音響学会音声研資, S82-23 (1982-6)

[4] ｴﾘｯｸ・ﾏｯｸﾀﾞｰﾓｯﾄ, 片桐滋: "コホネンのネットワークに基づいたシフトインバリアントな音韻認識", 音響学会秋期全大, 2-P-8, pp. 217-218 (1988-10)

[5] 二宮和則, 大槻恭士, 大友照彦 :"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, (1996-6)

[6] 中川聖一, 平田好充, 橘本康秀,"連続分布型 HMM による日本語音韻認識", 音響誌, vol. 46, no. 6, pp. 486-496, (1990-6)

[7] M. J. Ressell and A. E. Cook:"Experimental evaluation of duration modelling techiques for automatic speech recognition", ibid., pp. pp. 2376-2379

[8] F. J. Smith, J. Ming, P. O'Boyle, and A. D. Irvine,"A hiden Markov model with optimized inter-frame depedence", ICASSP-95, Detroit, USA., vol. I, pp. 209-212, (1995-5)

[9] 山本一公, 中川聖一 : "セグメント単位入力HMMとその評価", 信学技報, SP95-104 (1995-12)

[10] 大倉計美, 杉山雅英:"セグメント統計量を用いた HMM 音声認識", 信学技報, SP91-55, (1991-5)

[11] 二宮和則, 大槻恭士, 大友照彦,"離散型 HMM における DWWS の効果", 音響学会秋期全大, 2-1-10, pp. 67-68 (1997-9)

[12] 大槻恭士, 大友照彦,"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, (1997-11)

[13] Makino S., Endo M., Kido K.,"Recognition of phonemes in continuous speech using a modified LVQ2 method", J. Acoust. Soc. Jpn. (E) 13. 6, pp. 351-360 (1992-11)