周辺特徴抽出と CMN 制御を用いた認識タスクに依存しない音声認識性能の改善法

福田 隆† 新田恒雄†

本論文では，発話単位に実行される CMN 処理を組み込んだ音声認識システムの弱点を改善し，認識対象の発話内容にかかわらず，高い認識性能を与える特徴抽出方式を提案する．具体的には，TS パターンから局所特徴と周辺特徴を抽出し，これと MFCC を組み合わせた特徴パラメータセットに対して検討を行う．同時に発話単位の CMN 処理が，入力発話中の音韻の偏りが大きい場合，性能劣化を引き起こす事実に着目した修正 CMN（MCMN）処理方式を提案する．MCMN はケプストラム時系列から得た正規化分散値から，音韻の偏りの少なさに対する信頼度重みを計算して CMN 処理を制御する．提案する方式を組み合わせた特徴抽出器は，大語彙連続音声認識および孤立単語音声認識の双方で顕著に性能を改善することを，従来の標準方式と比較した実験結果から示す．

Improvement in Both Tasks of LVCSR and ISWR by Using Peripheral Feature Extraction and CMN Control

Takashi Fukuda† and Tsuneo Nitta†

In this paper, we propose a feature extractor that improves the performance of isolated-word and continuous speech recognition with CMN every utterance. Firstly, local features (LF) and peripheral features (PF) extracted from time-spectrum (TS) patterns and their roles in speech recognition are described. Then, the proposed feature extractors are implemented into a standard HMM-based speech recognition system with modified CMN (MCMN) in which CMN is controlled by a normalized variance of an utterance. Experiments were investigated both in an isolated spoken-word recognition (ISWR) system and a large vocabulary continuous speech recognition (LVCSR) system. Experimental results show that the feature set of MFCC with MCMN and novel PF outperforms the baseline feature set in an LVCSR task, and achieves significant improvement in an ISWR task.

1. はじめに

音声認識システムは，一般に分析器，特徴抽出器とそれに続く分類器から構成される．このうち分析器としては，帯域通過フィルタ（BPF: Band Pass Filter）群を用いて分析した結果を対数変換し，さらに DCT（Discrete Cosine Transform）で直交変換したパラメータ，すなわち MFCC（Mel Frequency Cepstrum Coefficient）が多用されている．一般的な音声認識システムでは，特徴抽出器を利用することが少なかったが，近年，MFCC の時間方向の変化（Δ パラメータ）が，動的特徴として抽出され用いられるようになった[1),2)]．

一方，音声認識では話者や集音環境などの音響的変動要因により性能が大きく低下する．このため，特徴抽出過程で CMN（Cepstrum Mean Normalization）処理が適用されることが多い．CMN はスペクトラムの乗法性歪みを少ない演算量で補正できることで知られている[3)]．我々がクリーン音声を対象に行った大語彙連続音声認識（以後 LVCSR（Large Vocabulary Continuous Speech Recognition）と呼ぶ）の実験では，**表 1** に示すように，MFCC の標準的特徴パラメータセットを用いた場合，発話単位で行う CMN 処理は単語誤り率を 3%低減した（実験条件は 3.1 節に同じ）．このように CMN は重要な役割を果たす．

他方，コマンド音声入力や 1 桁数字音声入力といった孤立単語音声認識（以後 ISWR（Isolated Spoken-Word Recognition）と呼ぶ）では，ケプストラム平均の計算区間が十分に確保できなかったり，発話内容に音韻的な偏りがある場合，発話単位の CMN 処理は逆に性能を劣化させてしまう（表 1 参照）．前の 1 つ

† 豊橋技術科学大学大学院工学研究科
Graduate School of Engineering, Toyohashi University of Technology

表 1 CMN 処理時の性能比較と周辺特徴の性能（混合数 8）
Table 1 Comparison between without CMN and with CMN, and the performance of additional peripheral features (mixture=8):
MFCC+Dyn.: MFCC+Δ_t+$\Delta_t\Delta_t$+ΔP+$\Delta\Delta$P
MFCC+Per.: MFCC+Δ_t+Δ_q+ΔP+$\Delta\Delta$P
Dimension = 38.

System	Word error rate [%]			
	MFCC+Dyn.		MFCC+Per.	
	Without CMN	With CMN	Without CMN	With CMN
LVCSR	19.3	16.3	31.2	23.1
ISWR	1.5	3.8	1.4	1.3

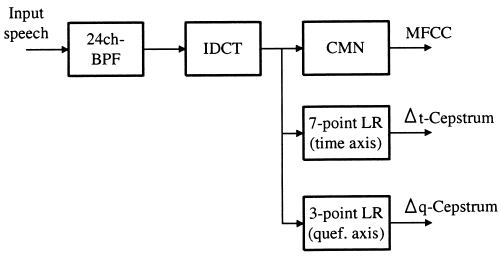

図 1 ケフレンシー領域における周辺特徴抽出
Fig. 1 Peripheral features in quefrency domain.

以上の発話からケプストラム平均を計算する方法は，音韻の偏りの影響を受けにくいが，正規化対象音声の話者がケプストラム平均計算時の話者と異なる場合，認識性能の低下を引き起こすことがある．そこで本論文では，発話単位で CMN を行うことを前提に，認識対象に依存せず，性能を改善する新しい特徴抽出方式を提案する☆．

我々は，ケフレンシー領域における周辺特徴を先に提案し，表 1 に示すように CMN 処理下の ISWR の性能を大幅に改善させた[4]．**図 1** にこの方式を示す．図中の LR は線形回帰（Linear Regression）演算を示す．CMN 処理により失われる音韻情報は，図から理解されるように，CMN 処理を行う前に抽出した Δ_q パラメータに保存される．このため，認識性能を維持・向上できたと考えられる．しかし，この方式は LVCSR の場合，CMN 処理により音韻情報が失われることが少ないため，Δ_q パラメータを入れることのメリットがなくなり，MFCC と動的特徴だけの標準的なパラメータセットと比較して性能が低下してしまう（表 1 参照）．

本論文では，CMN 処理下で LVCSR および ISWR の性能を維持・向上できる特徴抽出方式を得ることを目指す．本論文では最初に，周波数領域で局所特徴と周辺特徴を求め，特徴抽出器に組み込む方法を提案する．局所特徴には音韻情報が保存されるため，CMN 処理にかかわらず，ISWR の性能が維持できると期待される．これまでに，時間–スペクトラムパターン（以後 TS パターンと呼ぶ）が持つ特徴的な幾何学構造（構造的特徴）を，独立した複数の音響特徴平面，すなわち局所特徴に分離して認識に利用する方式が提案されている[5),6)]．局所特徴は構造的特徴を分離表現することで，性質の良い acoustic cue を構成する．本論文では，主要な 2 種の局所特徴を抽出するとともに，さらに局所特徴から周辺特徴（より広い領域にわたる有用な構造的特徴）を抽出する方法を検討する．抽出した局所特徴および周辺特徴は MFCC，差分パワーと組み合わせて利用する．なお，2 種の局所特徴は，そのままでは次元数も倍になるため次元圧縮したものを使用する．本論文では，IDCT（Inverse DCT）と IDST（Inverse Discrete Sine Transform）を用いて次元圧縮する方法を説明する．

局所特徴と周辺特徴を用いたパラメータセットは，次元数が 50 と標準的なセットの 38 に比べて増加する．そこで，局所特徴を除き，周辺特徴のみを加えたセット（次元数は 38）についても評価を行う．

次に，本論文では CMN 処理自身を，入力音声信号の性質を用いて制御する方法を提案し，その評価結果を報告する．CMN 処理の目的は，平均化対数スペクトラムを引き去る操作によって，音響周波数特性上の差異（音響モデル設計時の音声データセットと評価用音声データ間の差異．主に利用環境の伝達特性および話者自身の音声スペクトル特性による）を補正することにある．このとき，評価用入力音声の特性を正確に推定するには，発話内容に音韻的な偏りが少ないことが要請される．すなわち，白色信号である必要はないが，入力発話から観測した信号の対数スペクトラム，もしくはケプストラムの平均値が，実際の発話集合全体のそれを代表していることが望ましい．以上の点を考慮して，本論文では評価用入力音声のケプストラム時系列の分散値により，CMN 処理を制御する方法を検討する．すなわち，観測信号の分散値が小さいうちは，CMN 処理をおさえ，分散値が大きくなると（発話集合を近似していると判断し），本来の CMN 処理に近づける．CMN 処理を制御することで，LVCSR に対しては性能を維持することを，同時に ISWR については音韻の偏りがある場合，CMN 処理による性能劣化を防ぐことを期待している．

本論文は以下のように構成される．2 章で局所特徴

☆ 以下，本論文中で CMN 処理と明記した場合，発話単位でケプストラム平均を計算し，元のケプストラムから引き去る CMN 処理を指す．

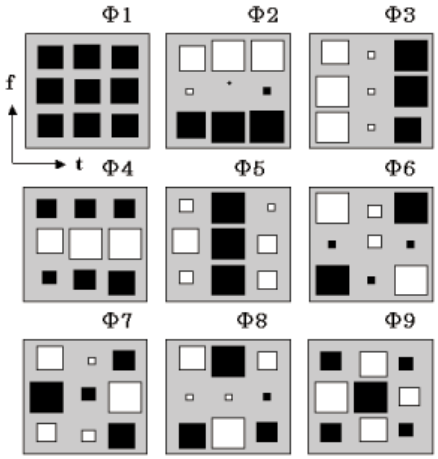

図 2 音声データから抽出される 3 × 3 直交基底
Fig. 2 3 × 3 orthogonal basis extracted from speech data.

と周辺特徴の概要を述べた後，局所特徴の次元圧縮方法を説明する．次に，3 章で局所および周辺特徴を組み込んだ特徴抽出器の構成を示し評価実験結果を述べる．最後に，4 章で CMN 処理を制御する方法を説明し実験結果と考察を述べる．

2. 局所特徴と周辺特徴

2.1 構造的特徴の表現

実験音声学の成果から，TS パターン上には音韻の種類に対応して，様々な幾何学構造が現れることが示されている[7]．現在のところ，TS 濃度パターンが持つ幾何学構造をうまく取り出す実用的な構造的特徴抽出系は得られていないが，これまで，KL 変換を用いて設計した 3 × 3 ブロックの写像演算子（**図 2** 参照）を用いて，TS パターンを複数の音響平面，すなわち局所特徴（LF）に写像する方法が提案され，音声セグメントの識別実験で高い性能が得られることが報告されている[5),6)]．図 2 において黒と白の正方形はそれぞれ正と負の値を表し，正方形の大きさは振幅を表す．

複数存在する局所特徴のうち，主要な 2 つは TS パターン上の各要素に対して，時間軸上および周波数軸上に各々3 点の線形回帰演算を行うことで得られる．**図 3** の (b)，(c) はこうして得た局所特徴の例を示している．図に示すように，(b) の局所特徴は子音区間の急激な変化特徴を表現し，(c) は定常音および比較的ゆっくりと変化するホルマント遷移特徴をとらえている．これら 2 つの局所特徴は TS パターン上の構造的特徴を明瞭に表現しているが，これをそのまま利用すると，次元数が 2 倍になるため，2.2 節で効率良く圧縮する方法を示す．

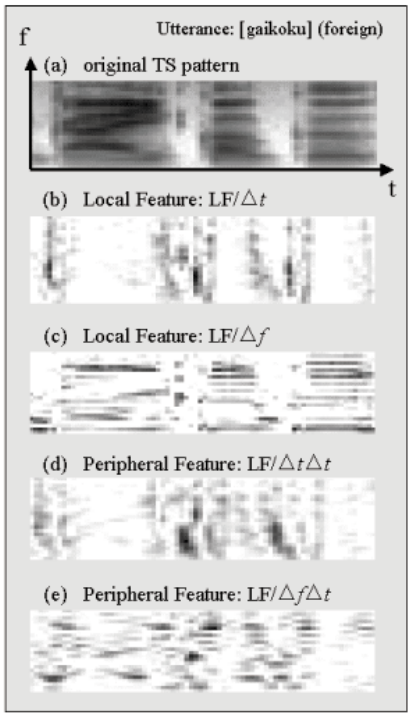

図 3 局所特徴の例
Fig. 3 An example of local features.

一方，TS パターンの n × n 近傍からはより多くの情報（周辺特徴）を得ることができる[4),8)]．本論文では，TS パターンの 3 × 7 近傍から周辺特徴を抽出して用いる．具体的には，まず 3 × 3 近傍から 2 つの局所特徴を抽出した後，時間軸上 7 点の線形回帰演算により周辺特徴を計算する．図 3 (d)，(e) に例を示す．このようにして得た周辺特徴は，TS パターンを過渡的な音声を代表する平面と，定常的な音声を代表する平面の 2 つに分離した後，動的特徴を抽出したものと見なすことができる．

2.2 局所特徴とケプスト ラム領域表現

局所特徴 $y_1(n,k)$ と $y_2(n,k)$ は直交していると仮定し，複素量 $y(n,k) = y_1(n,k) + jy_2(n,k)$ を定義する（n はフレーム番号，k はチャネル番号を示す）．次に，**図 4** に示すように $y_1(n,k)$ および $y_2(n,k)$ を，各々周期 $2K$ を持つ偶関数および奇関数と再定義☆すると（K はチャネル数を示す），$y(n,k)$ は次式で表現される．

☆ DCT では，変換対象の信号に対称性（偶関数）があることを仮定している．すなわち，DCT は信号を偶関数と再定義した後，DFT を行うのと等価である．同様に，DST では奇関数が仮定されている（文献 9) 参照）．

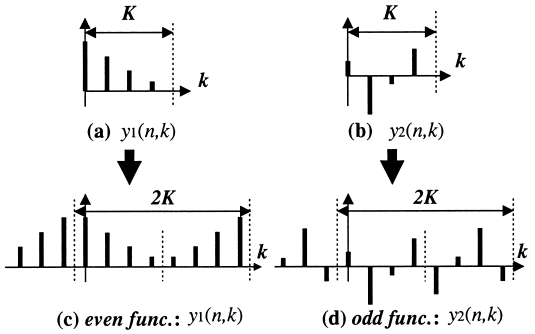

図 4 周波数領域局所特徴の複素量化
Fig. 4 Defining complex quantity of local features in frequency domain.

$$y(n,k)=\begin{cases} y_1(n,k)+jy_2(n,k) \\ \qquad k=0,1,\cdots,K-1 \\ y_1(n,2K-k-1) \\ \qquad -jy_2(n,2K-k-1) \\ \quad k=K,K+1,\cdots,2K-1 \end{cases} \tag{1}$$

すなわち，$y(n,2K-k-1)=y^*(n,k), k=0,1,\cdots,K-1$ となる（$*$ は共役複素数を示す）．以下，$y_1(n,k)$ と $y_2(n,k)$ が $k=-0.5$ に関して対称であることを利用して IDFT を行う☆．

$$\begin{aligned} & c(n,m) \\ =& \sum_{k=0}^{2K-1} y(n,k+0.5)\exp\left(j\frac{2\pi km}{2K}\right) \\ =& \sum_{k=0}^{K-1}\Big\{y_1(n,k+0.5)+jy_2(n,k+0.5)\Big\} \\ & \qquad \times\exp\left(j\frac{2\pi km}{2K}\right) \\ & +\sum_{k=K}^{2K-1}\Big\{y_1(n,2K-k-0.5) \\ & \qquad -jy_2(n,2K-k-0.5)\Big\}\exp\left(j\frac{2\pi km}{2K}\right) \\ =& \sum_{k=0}^{K-1} y_1(n,k)\Big[\exp\left\{j\frac{2\pi m(k+0.5)}{2K}\right\} \\ & \qquad +\exp\left\{-j\frac{2\pi m(k+0.5)}{2K}\right\}\Big] \\ & +j\sum_{k=0}^{K-1} y_2(n,k)\Big[\exp\left\{j\frac{2\pi m(k+0.5)}{2K}\right\} \\ & \qquad -\exp\left\{-j\frac{2\pi m(k+0.5)}{2K}\right\}\Big] \end{aligned}$$

☆ 信号を偶関数（奇関数）に拡張する方法はいくつかあるが，ここでは $k=-0.5$ で対称となる方法を採用した（文献 9) 参照）．

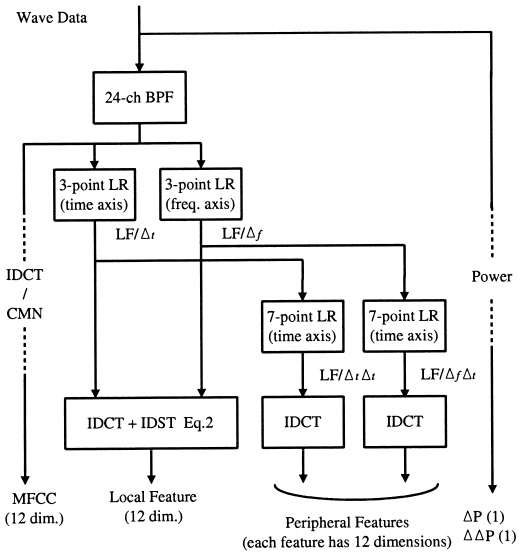

図 5 局所特徴と周辺特徴（MFCC+LF+PF）
Fig. 5 MFCC with local and peripheral features.

$$\begin{aligned} =& \sum_{k=0}^{K-1} 2y_1(n,k)\cos\left\{\frac{\pi m(k+0.5)}{K}\right\} \\ & -\sum_{k=0}^{K-1} 2y_2(n,k)\sin\left\{\frac{\pi m(k+0.5)}{K}\right\} \\ & \qquad m=1,2,\cdots,K \end{aligned} \tag{2}$$

上式第 1 項は MFCC の計算にも使用される IDCT[9]，第 2 項は IDST である．式 (2) は局所特徴がケフレンシー領域において，標準的な MFCC と同じ次元数を持つパラメータとして表現できることを示している．

3. 特徴抽出方式間の性能比較

3.1 局所特徴と周辺特徴を組み合わせた場合

3.1.1 特徴抽出方式

図 5 に局所特徴と周辺特徴の 2 つを組み込んだ特徴抽出過程を示す．まず TS パターンに対して，時間軸と周波数軸に沿った線形回帰演算を行い，2 つの局所特徴（LF: Local Features）を求める．続いて局所特徴を式 (2) により IDCT – IDST 変換して，ケプストラム（12 次元）を得る．次に，周辺特徴（PF: Peripheral Features）を 2 つの局所特徴から，各々時間軸上 7 点の線形回帰演算により求め，IDCT によりケプストラムに変換する（2×12 次元）．最後に，TS パターンを IDCT して求めた MFCC および差分パワーと結合して，50 次元の特徴パラメータを構成する（以

後，これを MFCC+LF+PF と呼ぶ）．

3.1.2 音声試料

以下に示す 2 つのデータセットを使用する．

D1. 音響モデル学習データセット：

日本音響学会（ASJ）研究用連続音声データベース（16 kHz，16 bit）のうち男性話者 30 名（4,503 文），および新聞記事読み上げコーパス（ASJ-JNAS，16 kHz，16 bit）のうち男性話者 103 名（15,911 文）．合計 20414 文．

D2. 評価データセット

(a) 孤立単語認識用（ISWR セット）： 東北大・松下単語音声データベース．先頭の 100 語男性話者 10 名．サンプリング周波数は 24 kHz から 16 kHz へ変換．

(b) 連続音声認識用（LVCSR セット）： ASJ-JNAS，D1 で使用していない男性話者 23 名からなる 100 文．

3.1.3 実験の概要

入力音声は 16 kHz でサンプリング後，512 点の FFT 分析処理を行った（25 ms ハミング窓，フレーム周期 10 ms）．続いてパワースペクトラムを，メルスケールの中心周波数を持つ 24 チャネル BPF 群により求めた．この後，3.1.1 項に説明した方式を用いて特徴パラメータを構成する．Baseline として，MFCC，Δ_t，$\Delta_t\Delta_t$ ケプストラム，および差分パワー（ΔP，ΔΔP）を結合した 38 次元の特徴パラメータを用意した．

音響モデルは 5 状態 3 ループ，日本語 43 音素 monophone-HMM を使用し，学習には発話単位（文頭・文中・文末の無音部分を含む）で CMN 処理を行った D1 データセットを用いた．HMM は出力確率をガウス混合分布で表現するとともに，共分散行列を対角化している（混合数は 4〜16）．評価には，発話単位の CMN 処理後の D2 データセットを使用し，不特定話者に対する ISWR および LVCSR を対象とした実験を行った．LVCSR セットについては無音部分を含め，発話全体を CMN 処理に使用した．一方，ISWR セットについてはヘッダ情報に基づき，語頭・語尾の無音部分をあらかじめ除いている．

連続音声認識にはデコーダとして Julius を使用した[10]．デコーダは 2 パス構成で，1 パス目に bi-gram を，また 2 パス目には tri-gram を用いている．言語モデルは，毎日新聞の記事データ 75ヵ月分（1991 年 1 月〜1994 年 9 月，1995 年 1 月〜1997 年 6 月，約 118 M 単語）を使用して構築したものを用いた．語彙数は 20k である．

表 2 実験結果（LVCSR）：MFCC+LF+PF
Table 2 The result in LVCSR: MFCC+LF+PF.

Model	Word error rate [%]		
	mix.=4	mix.=8	mix.=16
Baseline	21.3	16.3	14.7
MFCC+LF+PF	21.2	17.4	15.7

表 3 実験結果（ISWR）：MFCC+LF+PF
Table 3 The result in ISWR: MFCC+LF+PF.

Model	Word error rate [%]		
	mix.=4	mix.=8	mix.=16
Baseline	4.8	3.8	3.7
MFCC+LF+PF	2.0	2.0	1.6

3.1.4 実験結果と考察

表 2，**表 3** に実験結果を示す．MFCC+LF+PF は，ISWR セットに対する認識性能を顕著に改善する一方（誤り削減率約 50%），LVCSR 評価セットについては Baseline と比較して若干性能が落ちる結果を得た．局所特徴を直接利用することで，CMN 処理による単語認識性能の劣化は補償できるが，ケプストラム平均による音響特性の推定を，十分長い発話区間から推定できる LVCSR では，局所特徴を加えた効果がなく，次元数が増えた分かえって性能が下回ったと考えられる．なお，周辺特徴を除いた特徴パラメータ，すなわち MFCC と局所特徴（LF/Δ_t，LF/Δ_f）および差分パワー（ΔP，ΔΔP）を組み合わせた場合（38 次元）についても実験を行ったが，同様の結果を得た．

3.2 周辺特徴のみを組み合わせた場合

3.2.1 特徴抽出方式

図 6 に MFCC+LF+PF 方式から局所特徴を除いた方式の構成を示す（以後，これを MFCC+PF と呼ぶ）．特徴パラメータは，2 つの周辺特徴（PF; 2×12 次元），MFCC（12 次元），および差分パワー（2 次元）の計 38 次元から構成される．

3.2.2 実験結果と考察

音声試料および実験の概要は 3.1 節と同様である．**表 4**，**表 5** に実験結果を示す．LVCSR セットに対して，MFCC+PF は Baseline よりも高い性能を達成している．性能が向上したのは以下の理由によると推測される．全共分散 HMM では，ケプストラム係数の間の相関が表現されているが，本論文で用いている対角共分散 HMM では，ケプストラム係数相互の相関が考慮されていない．周波数領域の局所特徴から周辺特徴（特に LF/$\Delta_f\Delta_t$）を抽出した場合，隣接するパワースペクトラム成分間の相関情報が，IDCT 変換後のパラメータに含まれる．そのため，LVCSR の性能向上に寄与した．

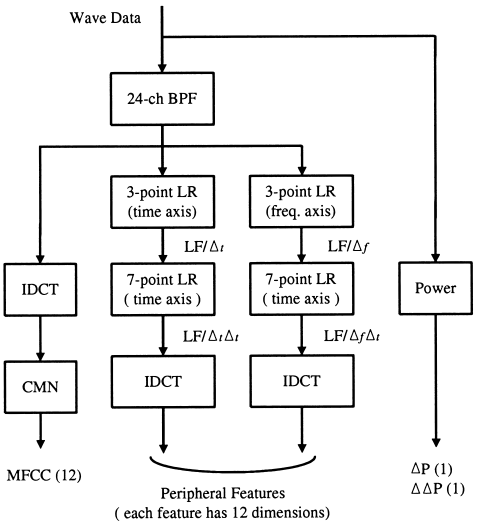

図 6 MFCC と周辺特徴（MFCC+PF）
Fig. 6 MFCC with peripheral features.

表 4 実験結果（LVCSR）：MFCC+PF
Table 4 The result in LVCSR: MFCC+PF.

Model	Word error rate [%]		
	mix.=4	mix.=8	mix.=16
Baseline	21.3	16.3	14.7
MFCC+PF	19.4	15.8	12.5

表 5 実験結果（ISWR）：MFCC+PF
Table 5 The result in ISWR: MFCC+PF.

Model	Word error rate [%]		
	mix.=4	mix.=8	mix.=16
Baseline	4.8	3.8	3.7
MFCC+PF	6.1	4.9	5.1

一方 ISWR セットに対しては，Baseline および MFCC+LF+PF と比較して劣った性能となった．これは，局所特徴を使用していないためで当然の結果といえる．4 章では，ここで得られた LVCSR における性能の優位を維持すると同時に，ISWR における性能を向上させる方策について検討する．

4. CMN 処理の制御

4.1 分散値を用いた制御方法

CMN は入力発話内容に音韻的偏りがあるとき，音響特性に対する推定精度が落ち，認識性能の低下を引き起こすと推測される．これを防ぐため，ケプストラム時系列の分散計算を通して，発話内容の音韻的偏りを推定し，CMN 処理の働きを制御する仕組み（以下，修正 CMN と呼び，MCMN と略す）を特徴抽出器に組み込む方法を検討する．

CMN は次式で表される．

$$c_{CMN,ij} = c_{ij} - \frac{1}{N}\sum_{j=1}^{N} c_{ij} \quad (i = 1, 2, \cdots, M) \tag{3}$$

ここで c_{ij} は CMN 処理前，$c_{CMN,ij}$ は CMN 処理後のケプストラム係数であり，i，M はそれぞれ（メル）ケプストラム係数の番号と次元数，また j，N はそれぞれフレーム番号と処理に用いたフレームの総数である．上式の右辺第 2 項は，音響環境や話者の違いから生じた音響伝達特性の偏りを代表することが期待されたケプストラムの平均値である．したがって，平均化の際に利用した音声データが，音響伝達特性の偏りを十分に代表していない場合，CMN 処理は性能劣化を招く．そこで以下では，この右辺第 2 項を計算する際に用いた c_{ij} から分散値を計算し，この値を推定に対する信頼度として，CMN を制御する方法を試みる．具体的には，次式に示すように信頼度重み w をかけて CMN を制御する．

$$c_{MCMN,ij} = c_{ij} - w\frac{1}{N}\sum_{j=1}^{N} c_{ij} \quad (i = 1, 2, \cdots, M) \tag{4}$$

信頼度重み w は以下の手順で求めた．まず，分散計算はケプストラムの振幅に大きく影響されるため，以下のようにフレームごとにノルムで正規化して用いる．

$$c'_{ij} = \frac{c_{ij}}{\sqrt{\dfrac{1}{M}\displaystyle\sum_{i=1}^{M} {c_{ij}}^2}} \quad (j = 1, 2, \cdots, N) \tag{5}$$

ここで c'_{ij} は正規化されたケプストラムである．

次に，正規化ケプストラム c'_{ij} から各次元の分散値を求めた後，ケプストラム係数方向に平均値を計算する．

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=1}^{N}\left(c'_{ij} - \frac{1}{N}\sum_{j=1}^{N} c'_{ij}\right)^2 \tag{6}$$

この値はフレーム方向に正規化したケプストラムから分散を計算しているため，以後，正規化分散値と呼ぶ．続いて，次式のシグモイド関数により信頼度重み w を決定し，式 (4) へ代入する．

$$w = \frac{1}{1 + \exp\left(-\alpha\sigma^2 + \beta\right)} \tag{7}$$

ここで α，β は正定数である．

MCMN 処理により，LVCSR ではケプストラム平均値の推定に信頼がおけるため，w の値が 1 に近づ

表 6 実験結果（LVCSR）：修正 CMN
Table 6 The result in LVCSR: Modified CMN.

Model	Word error rate [%]		
	mix.=4	mix.=8	mix.=16
Baseline: CMN	21.3	16.3	14.7
Baseline: MCMN	21.7	16.7	14.9
MFCC+PF: MCMN	19.5	16.7	12.8

表 7 実験結果（ISWR）：修正 CMN
Table 7 The result in ISWR: Modified CMN.

Model	Word error rate [%]		
	mix.=4	mix.=8	mix.=16
Baseline: CMN	4.8	3.8	3.7
Baseline: MCMN	2.3	1.7	1.7
MFCC+PF: MCMN	2.6	2.5	2.3

き，また ISWR で音韻的な偏りのある場合は 0 に近づくことになるため，入力発話内容にかかわらず性能を維持できると期待される．

4.2 実験結果と考察

音声試料および実験の概要は 3.1 節と同様である．**表 6**，**表 7** に MCMN を適用したときの実験結果を示す．実験は，LVCSR セットで高い性能を示した MFCC+PF と Baseline について行った．なお音響モデル作成の際は，文発声データを用いたため，通常の CMN 処理を行っている．また予備実験結果から，式 (7) 中の定数は $\alpha = 18$，$\beta = 5$ とした．

MCMN を適用することで，ISWR セットは両方式とも誤り率を大きく改善した．一方，LVCSR セットについても双方の方式で性能が維持され，MFCC+PF 方式の優位も保持された．

これまでの議論は以下の 2 つを仮定して進めてきた．1 つは，入力発話中には音韻の偏りがあり，これが性能の低下をもたらすという仮定，もう 1 つは，c_{ij} の正規化分散値とこれから計算される信頼度重みが，音韻の偏りを推定するのに役立つという仮定である．そこで，以下でこの仮定を検証する．

発話内容に偏りがある ISWR に用いたセットについて，この中から音韻の偏りが比較的少ない単語と，偏りが大きい単語を各々数個選び，前者をグループ A，後者をグループ B とした後，CMN と MCMN 処理時の誤認識率，および正規化分散値と信頼度重みの関係を調べた．特徴抽出は標準方式（Baseline）を使用した．**表 8** は両グループの単語誤認識率を示している．これから，グループ B のように音韻の偏りが大きい場合には，CMN 処理による性能低下が大きく，一方，発話区間が短くても音韻的な偏りが小さい場合には，CMN 処理による性能低下が少ないことが分かる．また表から，MCMN 処理によって,「アオアオ」以外は誤りがほとんど解消されていることが見てとれる．

表 8 CMN と MCMN の誤認識率（Baseline model 使用）
Table 8 The error rate of CMN and MCMN: Baseline model.

グループ	発話内容	音素記号	誤り率 [%](mix.=1)	
			CMN	MCMN
A	イロガミ	irogami	0	0
	ウォッカ	woqka	0	0
	ノハラ	nohara	10	0
B	アワ	awa	100	0
	アオアオ	aoao	100	90
	チリ	chiri	50	10
	ロウドウ	ro:do:	50	0

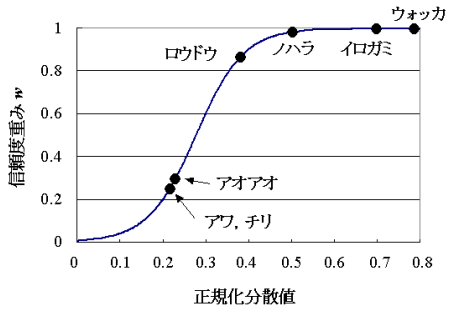

図 7 正規化分散値と信頼度重み w の関係
Fig. 7 The relationship between normalized variance and confidential weight.

次に，両グループ中の単語に対する正規化分散値と信頼度重みを調べた．**図 7** は式 (7) で表される信頼度曲線上で，各単語がどこに位置するかをプロットしたものである．グループ A に属するイロガミ，ウォッカ，ノハラは重みが 1 に近く，CMN の効果が最大限に引き出される．一方，グループ B に属する単語はロウドウを除き，信頼度重みが低くおさえられている．以上から，MCMN 処理は発話中の音韻の偏りを信頼度重みの形で抽出し，この値に基づき CMN 処理を制御することで性能を大きく改善していることが分かる．

5. ま と め

発話単位で行う CMN 処理を前提とした音声認識方式において，認識対象の発話内容にかかわらず，高い性能を得る特徴抽出方式を提案した．具体的には，TS パターンから局所特徴と周辺特徴を抽出し，これと MFCC を組み合わせた特徴パラメータセットに対して検討を行った．同時に，CMN 処理による性能劣化を軽減する MCMN 処理を提案した．この処理は，入力発話中のケプストラム時系列から正規化分散値を

観測し，これから音韻の偏りの少なさに対する信頼度重みを得て，CMN 処理を制御する．大語彙連続音声および孤立単語音声を対象に，標準方式（Baseline）と比較評価した実験の結果から以下の結論を得た．

(1) 局所特徴と周辺特徴の双方を組み込んだ MFCC+LF+PF 方式は，CMN 処理下の ISWR において顕著な性能改善を示した．しかし，LVCSR では標準方式を下回る．

(2) MFCC と周辺特徴のみを組み合わせた MFCC +PF 方式は，LVCSR セットで標準方式を上回る性能を示した．しかし，ISWR では低い性能にとどまる．

(3) ケプストラム時系列の正規化分散値を用いて CMN を制御する MCMN 処理方式は，LVCSR と ISWR の双方で高性能を実現する．

(4) MCMN 処理を組み込んだ MFCC+PF 方式は，標準方式に対して LVCSR での優位を維持した．また，ISWR でも高い性能改善を示した．

(5) MCMN 処理は，発話中の音韻の偏りを信頼度重みの形で抽出し，この値に基づき CMN 処理を制御することで性能を大きく改善していることが示された．

本論文では，発話全体を CMN 処理の対象とした．しかし実際の応用，特に LVCSR では，前の複数の発声からケプストラム平均を求めるスムージング手法が利用されることが多い．したがって，発話の途中に計算されるケプストラム平均を評価する必要がある．一方，スムージング手法を用いた場合，話者交代で問題を生ずることがしばしばある．今後は，こうした実用的な観点からも特徴抽出方式の改善を行っていきたい．

参 考 文 献

1) Elenius, K. and Blomberg, M.: Effect of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system, *IEEE Proc. ICASSP'82*, pp.535–538 (1982).
2) Furui, S.: Speaker-independent isolate word recognition using dynamic features of speech spectrum, *IEEE Trans. Acoust., Speech & Signal Processing, ASSP-34*, pp.522–529 (1986).
3) Liu, F.H. and Stern, R.M.: Efficient cepstral normalization for robust speech recognition, *ARPA Human Language Technology Workshop*, pp.69–74 (1993).
4) Fukuda, T., Takigawa, M. and Nitta, T.: Peripheral Features for HMM-based Speech Recognition, *IEEE Proc. ICASSP'01*, pp.129–132 (2001).
5) Nitta, T.: A novel feature-extraction for speech recognition based on multiple acoustic-feature planes, *IEEE Proc. ICASSP'98*, pp.29–32 (1998).
6) Nitta, T.: Feature extraction for speech recognition based on orthogonal acoustic feature planes and LDA, *IEEE Proc. ICASSP'99*, Phoenix, Vol.1, pp.421–424 (1999).
7) Ladefoged, P.: *A course in phonetics*, 2nded., New York, Harcourt Brace, Jovanovich (1982).
8) Nitta, T., Takigawa, M. and Fukuda, T.: A Novel Feature Extraction Using Multiple Acoustic Feature Planes for HMM-based Speech Recognition, *Proc. ICSLP'00*, Vol.1, pp.385–388 (2000).
9) Makhoul, J.: A Fast Cosine Transform in One and Two Dimensions, *IEEE Trans. Acoust., Speech & Signal Processing, ASSP-28*, No.1, pp.27–34 (1980).
10) Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *Eurospeech'01*, pp.1691–1694 (2001).

(平成 13 年 11 月 19 日受付)
(平成 14 年 4 月 16 日採録)

福田 隆

平成 12 年豊橋技術科学大学知識情報工学科卒業．平成 14 年豊橋技術科学大学大学院工学研究科修士課程修了．現在，同大学院電子情報工学専攻博士後期課程在学中．音声認識に関する研究に従事．日本音響学会会員

新田 恒雄（正会員）

昭和 44 年東北大学工学部電気工学科卒業.（株）東芝勤務を経て，平成 10 年より豊橋技術科学大学大学院工学研究科教授．工学博士．音声認識・合成・文字認識，およびマルチモーダル対話システムの研究に従事．著書に「マルチメディアとデジタル信号処理」（共著),「マルチメディア処理入門」（共著）等．電子情報通信学会論文賞受賞．電子情報通信学会，日本音響学会，人工知能学会，IEEE 各会員．