音声言語情報処理 44－1
(2002. 12. 19)

音素弁別特徴ベクトルを用いた頑健な音声認識の検討

福田 隆　　山本 航　　新田 恒雄

豊橋技術科学大学 大学院工学研究科
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail: {fukuda, yamamoto}@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

あらまし 本報告では，入力音声から音素弁別特徴(DPF)を抽出し，頑健な音声認識システムを実現する方法を検討する．音素弁別特徴抽出過程では，まず入力音声を局所特徴(LF)に変換した後，LF と ΔP から成る音響特徴系列を多層ニューラルネットワーク(MLN)に通すことで，音素弁別特徴へ写像する．MLN の出力は，前後のコンテキストを含む 33 次元(11 次元×3)の音素弁別特徴を使用する．評価実験では，MLN の出力ユニットの構成に関する比較を行った後，提案の DPF パラメータと標準的な MFCC パラメータセットを比較する．実験の結果，clean speech ではほぼ同等の性能を達成することを不特定話者孤立単語認識実験から示す．また，DPF パラメータの耐雑音性能を 4 種類の加法性雑音を重畳して評価し，1 種類を除き標準パラメータセットと比較して良好な結果が得られることを示す．提案方法と MFCC との組み合わせについても評価を行う．

キーワード 音声認識，特徴抽出，弁別特徴，多層ニューラルネットワーク，局所特徴

A Study on Robust Speech Recognition by Using Distinctive Phonetic Feature Vectors

Takashi FUKUDA　Wataru YAMAMOTO　and　Tsuneo NITTA

Graduate School of Engineering, Toyohashi University of Technology
1-1, Hibariga-oka, Tempaku, Toyohashi, Aichi, 441-8580 Japan
E-mail: {fukuda, yamamoto}@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

Abstract This paper describes an attempt to extract distinctive phonetic features (DPFs) that represent articulatory gestures in linguistic theory by using a multi-layer neural network (MLN) and to apply the DPFs to noise-robust speech recognition. In the DPF extraction stage, after converting a speech signal to acoustic features composed of local features (LFs), an MLN with 33 output units corresponding to context-dependent DPFs of 11 DPFs, 11 preceding context DPFs, and 11 following context DPFs maps the LFs to DPFs. In experiments, firstly, the configuration of MLN output units is compared. The proposed DPF parameters without MFCC were secondly evaluated in comparison with a standard parameter set of MFCC and dynamic features on a word recognition task using clean speech and the result showed the same performance as that of the standard set. Noise robustness of these parameters was then tested with four types of additive noise and the proposed DPF parameters outperformed the standard set except one additive noise type. The combinatorial usage of DPFs and MFCC is also tested.

Keyword robust speech recognition, feature extraction, distinctive phonetic feature, multi-layer neural network, local feature

1. はじめに

近年，一般的な音声認識システムでは，特徴パラメータとして短時間パワースペクトラム情報に基づく MFCC(Mel Frequency Cepstrum Coefficient)と動的特徴を組み合わせたセットが多用されている[1]. MFCC は音声信号の対数スペクトラム包絡を効率よく表現する優れたパラメータであるが，伝達特性の差異や雑音などによるスペクトラム包絡の変形が，直接，認識性能に影響する.

一方，音韻論の分野では，調音様式（母音性，子音性，連続性，…）や調音位置（高舌性，前方性，舌端性，…）を表す音素弁別特徴(以後 DPF(Distinctive Phonetic-Feature)と呼ぶ)による音素分類が古くから提案されている[2]. また，音声認識においても古くから DPF の利用が研究され[3,4]，近年，再びこの積極的利用が検討され始めている[5,6,7,8,9]. 文献[5]では，複数の多層ニューラルネットワーク(以後 MLN(Multi Layer Neural network)と呼ぶ)を用いた DPF 抽出方法を提案している. この方法は，各 MLN を一つの弁別素性に対応させて学習を行った後，認識段階で各 MLN の出力，すなわち DPF を結合し，HMM 分類器の入力として用いている. 一方，文献[6]では，DPF 抽出に GMM (Gaussian Mixture Model) が利用されている. この方式では，各弁別素性に対して“弁別素性存在モデル”と“弁別素性不在モデル”の二つの尤度を比較することで DPF を抽出している. なお先行研究は，DPF 単独では高い性能が得られないため，従来の MFCC と組み合わせて利用している.

提案方式は，単一の MLN から DPF を抽出すると共に，DPF パラメータ単独で，雑音に頑健な音声認識を達成することを目指す. 提案方式と先行研究との主な違いは以下のとおりである.

(1) MLN への入力音響特徴は MFCC ではなく，局所特徴(以後 LF (Local Feature)と呼ぶ)を用いる.

(2) 単一の MLN により抽出する DPF は前後のコンテキストを含む. すなわち，MLN の出力は先行および後続するコンテキストを含む 33 次元 (11 次元×3) ベクトルから成る.

本報告では，はじめに，MLN 出力ユニットの構造の違いを比較する. 次に不特定話者孤立単語音（clean speech）を用いて，提案手法と標準手法（MFCC+動的特徴）との比較評価実験を行う.

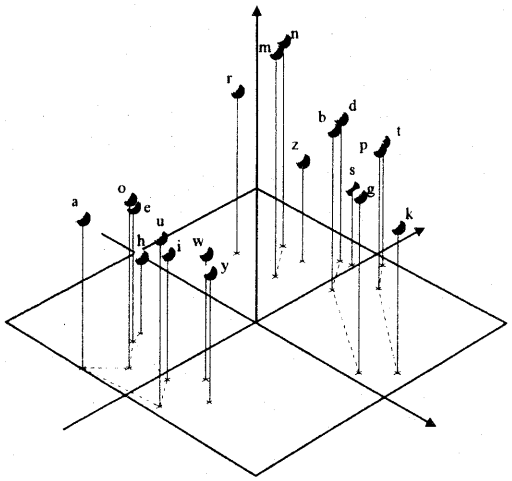


図1. MDS による三次元音素弁別特徴空間

実験では，MLN の入力特徴パラメータとして，MFCC と LF の 2 種類を比較する. 最後に，種々の雑音を重畳して，提案手法の耐雑音性能を評価する.

本報告は以下のように構成される. **2.**で DPF 抽出器の概要を説明した後，**3.**で評価実験結果と考察を述べる. 最後に **4.**で結論をまとめる.

2. 音素弁別特徴の抽出過程

2.1. 音素弁別特徴

図 1 は日本語音素の音素弁別特徴[10](母音性，子音性，高舌性，後方性，低舌性，前方性，舌端性，遮音性，有声性，連続性，鼻音性)を多次元尺度構成法(MDS)により，元の 11 次元空間から 3 次元空間に圧縮して，音素の布置を見たものである. 図から，母音グループと子音グループを初めとして，調音的に近い音素が近接して配置されていることがわかる. DPF の特徴としては，以下に示すものが挙げられる.

(A) 調音の類似した音素を距離の近いベクトルとして扱うことができる.

(B) パワースペクトルなどの音響特徴量ではなく，調音方式や調音位置を陽に表現した特徴量であるため，利用環境に影響され難いと推測される.

(C) 連続量である音響特徴ベクトルと離散量である音素との中間表現として，位置付けることができる.

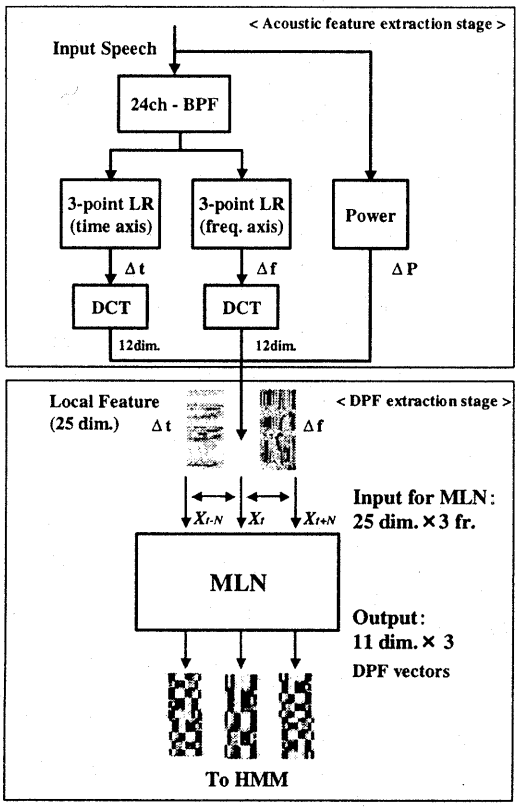


図2.　音素弁別特徴抽出器

DPFは，悪環境下の高精度音声認識を可能にする方式と期待される．なお以下では，予備実験の結果をもとに，“母音性／非母音性”と“子音性／非子音性”いう二つの弁別素性の代わりに“半母音性(/ j, w, r /)／非半母音性”と“摩擦性(/ s, z, h /)／非摩擦性”を使用している．

2.2. DPF 抽出器の設計

利用環境の違い，話者毎の調音器官の違いは，各音素の音響的イベントに影響するため，DPFを音声信号から直接抽出することには困難を伴う．本節では，MLN を用いて音響特徴を DPF に写像する方法を検討する．

図2に提案する DPF 抽出方式を示す．まず，入力音声をフレーム単位で局所特徴[11]に変換する．次に，局所特徴系列中の注目フレーム X_t と前後 N 点離れたフレーム（X_{t-N}, X_{t+N}）を結合して MLN に入力する．MLN は音素弁別特徴に対応する 11 個の出力ユニットを三つ(前後のコンテキストを含む)，すなわち計 33 個の出力ユニットを持つ．学習では，入力音素とその隣接音素の弁別特徴に対応する出力ユニットの値が 1 になるように重み係数を更新する．最後に，MLN の出力を DPF ベクトル時系列として，HMM 分類器に与える．

3. 評価実験

3.1. 音声試料

以下に示す三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(16kHz，16bit)のうち男性話者 30 名，合計 4503 文．

D2. 評価データセット:

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 名を使用． サンプリング周波数は 24kHz から 16kHz へ変換．

D3. 雑音データセット:

RWCP 実環境音声・音響データベースのうち以下に示す 3 種類の雑音

(A) Mobile Phone：携帯電話の着信音
(B) Particles：多数の粒子を金属箱に注ぐ音
(C) Whistle：笛を吹いた音

に加えて，白色雑音を使用する．

3.2. 雑音のスペクトラム構造

図 3 に，評価実験で使用する 3 種類の雑音のスペクトラムパターンを示す．図に示すように，“Mobile Phone”と“Whistle”は一定の周波数帯域で持続する雑音である．他方，“Particles”は白色雑音のように全周波数帯域に分布している．

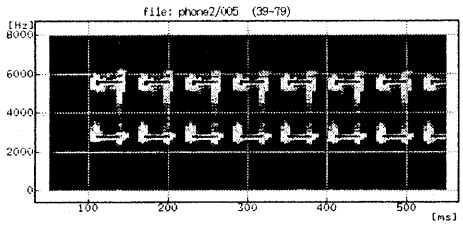
3.3. 実験の概要

3.3.1. 音響特徴パラメータ

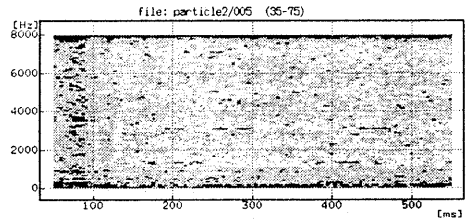
MLN に入力する音響特徴パラメータとして，以下の二つを比較検討する．

(I) MFCC

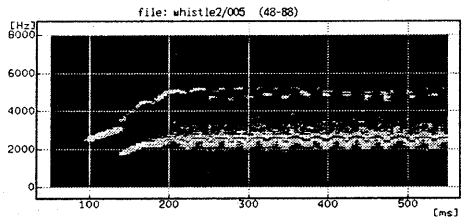
入力音声をメルスケール間隔の中心周波数を持つ 24 チャンネル BPF 群により分析した後(25ms ハミング窓，フレーム周期 10ms)，BPF 群の出力を DCT により MFCC に変換する．その後，1 次の動的特徴と差分パワーを組み合わせて 25 次元の特徴パラメータを構成する．MFCC については発話単位で CMN 処理を行っている．



(A) Mobile Phone：携帯電話の着信音



(B) Particles：多数の粒子を金属箱に注ぐ音



(C) Whistle：笛を吹いた音

図3. 雑音のスペクトラムパターン

(Ⅱ) 局所特徴(LF)

BPF の出力を各々時間軸および周波数軸に沿って 3 点の線形回帰演算を行うことにより 2 種類の LF を得る[11]. 続いて, DCT を用いてそれぞれの LF を 12 次元に圧縮した後, 差分パワー(ΔP)と連結して 25 次元の特徴パラメータを構成する.

3.3.2. MLN の構造

前項の(Ⅰ),(Ⅱ)で示した音響特徴パラメータを用いて MLN を学習する. (Ⅰ) については, 連続した 3 フレームを結合したものを, (Ⅱ) については注目フレームと前後 3 点離れたフレームを結合したものを MLN の入力とした(すなわち, 入力音響特徴を構成する元のフレーム数は等しい). MLN の学習データとしては, D1 データセット中に連続した 3 フレームで実際に

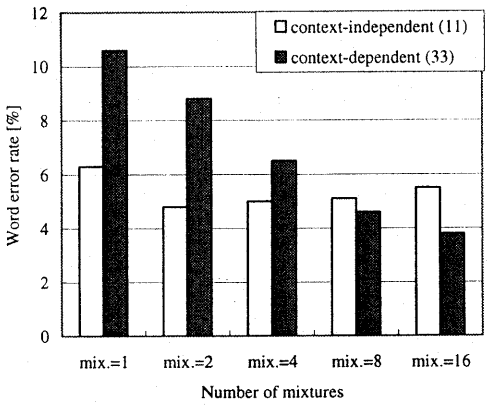


図4. MLN の構造の比較結果

出現した 3 つ組音素(LF については, 3 フレーム間隔で出現する 3 つ組音素)の内, 重心からの距離が最も近い上位 30 個を利用した(30 個に満たない 3 つ組音素はそのまま利用した). MLN の学習には誤差逆伝播法を使用している. また, MLN は 4 層構成で, 各層のノード数は入力層から順に 75, 256, 64, 33 とした.

3.3.3. 音響モデル

音響モデルは 5-state, 3-loop, 日本語 43 音素 monophone-HMM を使用し, 学習には D1 データセットの内, ASJ 男性話者 30 名(4503 文)を用いた. HMM は出力確率をガウス混合分布で表現すると共に, 共分散行列を対角化している.

3.4. 実験結果と考察

前節に示した実験条件下で, 不特定話者孤立単語認識を行った.

3.4.1. MLN 構造の比較

MLN の出力ユニット構成の違いを比較した結果を図 4 に示す. ここでは, MLN 入力特徴として LF を用いた. 出力に前後のコンテキストを持つ MLN は, 前後のコンテキストを持たない MLN と比較して, 混合数が大きい場合(混合数 8 以上) 高い性能を示した.

前後のコンテキストを持たせたことによる性能改善は, 以下の理由によると推測される. コンテキストを持たない MLN は, LF を 11 次元の DPF に写像する際, 音素境界付近で写像時の誤りを生じる. 一方, コンテキストを持つ MLN においても, 音素境界付近では同様に写像誤りを生じるが, DPF を前後のコンテキストに対応させたことで, 写像時の誤り (歪み) を軽減させることができた.

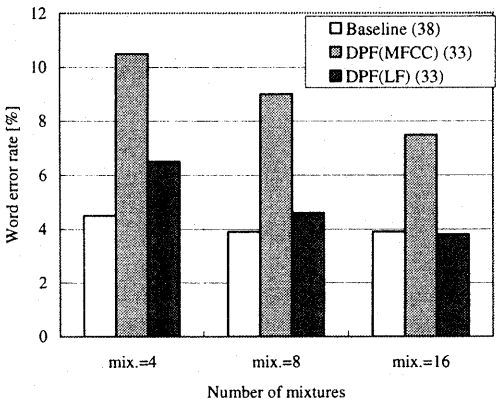


図5. MLN 入力特徴の比較結果

3.4.2. 入力音響特徴パラメータの比較

図 5 に実験結果を示す．図中の Baseline は，MFCC と 1 次，2 次の動的特徴(Δ_t, $\Delta_t\Delta_t$)，および差分パワー(ΔP, ΔΔP)を連結した 38 次元の音響特徴パラメータを，直接 HMM に入力したときの結果である．提案手法の内，DPF(MFCC)は音響モデルの混合数に関わらず性能が大幅に低下した．一方 DPF(LF)では，混合数が 16 の時 Baseline と比較してほぼ同程度の性能を達成した．実験結果から，音素特徴空間への写像は，局所特徴が入力パラメータとして適していることがわかる．

3.4.3. 耐雑音性能

(A) DPF のみを利用

D2 評価データセットに D3 雑音データセットを，それぞれ SN 比 5dB および 10dB で重畳したときの実験結果を図 6，7 に示す．Baseline は **3.4.2** と同様である．提案方式は，DPF パラメータ単独で，"Particles"を除く 3 種類の雑音に対して誤りを削減した．特に，"Mobile Phone"については，SN 比 10dB で 12.0%から 7.3%，SN 比 5dB で 20.0%から 15.2%と顕著な性能改善を達成した．

局所特徴は，対数パワースペクトラムの時間軸および周波数軸上の変化を表現している．このため，"Particle"のように全周波数帯域に亘って変動する雑音が音声に重畳した場合，局所特徴に雑音の変動が大きく現れることになり，DPF(LF)の性能がより劣化したと推測される．

(B) DPF と MFCC の組み合わせ

DPF(33 次元)と MFCC (12 次元)および差分パワー(ΔP，1 次元)の計 46 次元を連結した場合

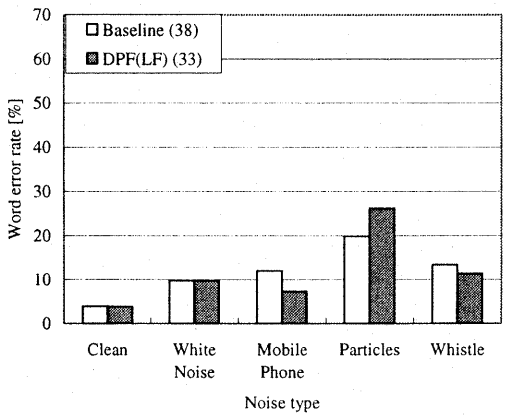


図6. 耐雑音性能評価結果
(without MFCC, SNR=10 dB, Mixture=16)

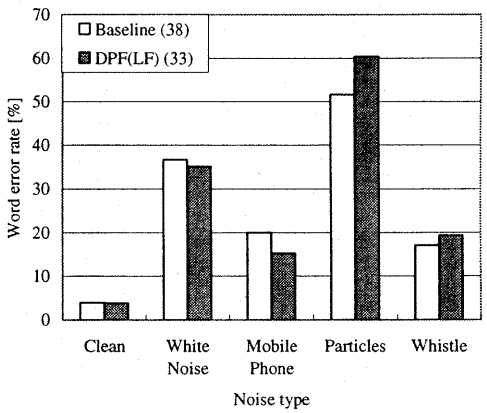


図7. 耐雑音性能評価結果
(without MFCC, SNR=5 dB, Mixture=16)

の性能を比較評価した．図 8，9 に実験結果を示す．DPF(LF)と MFCC を連結したパラメータを適用することで，DPF(LF)単独と比較したとき，性能上の改善がみられた．特に“Particle”では，SN 比 10dB で 26.2%から 23.7%，SN 比 5dB で 60.4%から 55.9%まで誤りを改善できた．

4. おわりに

DPF を用いた頑健な音声認識方式を検討した．DPF 抽出過程に用いた MLN では，音響特徴を

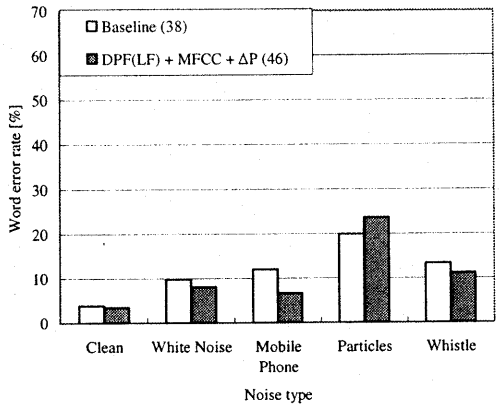


図8. 耐雑音性能評価結果
(with MFCC, SNR=10 dB, Mixture=16)

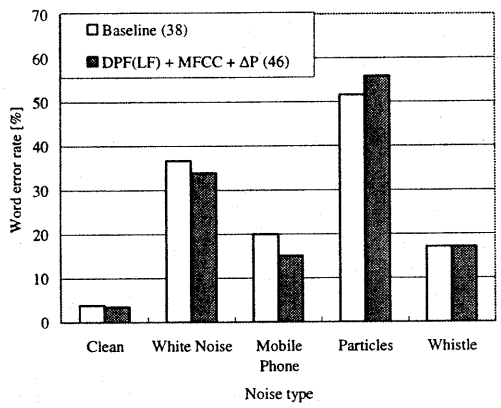


図9. 耐雑音性能評価結果
(with MFCC, SNR=5 dB, Mixture=16)

前後のコンテキストを含む DPF ベクトルに写像した．また MLN の入力音響特徴比較では，LF が MFCC と比較して高い性能を示した．提案方式は，clean speech を対象とした不特定話者孤立単語認識実験において，従来手法と比較して同等の性能を得ると共に，各種雑音を重畳した認識タスクに対しても性能を改善した．

今後は，DPF 抽出方式をさらに改良するとともに，この方式を効果的に利用する音声認識方法を検討したい．

謝辞

本研究の一部は，RWCP 実環境音声・音響データベースの非音声音の無響室測定データを利用した．

文 献

[1] S.Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process1, ASSP-34, pp.522-529, 1986.

[2] N. Chomsky and M. Halle, "The Sound Pattern of English," New York, Harper and Row, 1968.

[3] T. B. Martin, "Practical Application of Voice Input to Machine," Proc. IEEE, 64-4, 1976.

[4] S. Makino, S. Homma and K. Kido, "Speaker independent word recognition system based on phoneme recognition for a large size (212 words) vocabulary," J. Acoust. Soc. Jpn., (E) 6, 3, pp.171-180, 1985.

[5] B. Launay, O. Siohan, A. Surendran and C. H. Lee, "Towards Knowledge-based Features for HMM Based Large Vocabulary Automatic Speech Recognition," Proc. of IEEE ICASSP'02, pp.817-820, 2002.

[6] E. Eide, "Distinctive Features For Use in an Automatic Speech Recognition System," Proc. of Eurospeech'01, pp.1613-1616, 2001.

[7] K. Kirchhoff, G. A. Fink and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," Speech Communication, 37, pp.303-319, 2002.

[8] P. Jain, H. Hermansky and B. Kingsbury, "Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features," Proc. of ICSLP'02, pp.473-476, 2002.

[9] H. Tolba, S. A. Selouani and D. O'Shaughnessy, "Comparative Experiments to Evaluate the Use of Auditory-based Acoustic Distinctive Features and Formant Cues for Automatic Speech Recognition Using a Multi-stream Paradigm," Proc. of ICSLP'02, pp.2113-2116, 2002.

[10]比企静雄 編，"音声情報処理，" 東京大学出版会，1973.

[11]福田 隆，新田恒雄，"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，" 情処論，Vol.43，No.7, pp.2022-2029, 2002.