

直交化音素弁別特徴ベクトルを用いた雑音に頑健な音声認識

福田 隆 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

Email: fukuda@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

あらまし 近年、実環境での利用を目指し、耐雑音性能の高い頑健な方式に焦点を当てた音声認識技術に関する研究が盛んになっている。我々は音素弁別特徴(DPF)ベクトルを用いた音声認識システムを提案し、これまでに孤立単語音声による評価実験を通して、DPF パラメータの雑音に対する頑健性を示した。本報告では、HMM の入力として、従来の DPF ベクトルを直交化して利用することを検討する。直交化 DPF は、38 音素の DPF に対する KL 変換から直交基底ベクトルを計算し、これと DPF との内積から求める。評価実験では、従来の DPF と直交化 DPF を比較した後、耐雑音性能を 4 種類の加法性雑音を付加した孤立単語音声を用いて評価する。実験の結果、直交化した DPF パラメータは、雑音に対して顕著な性能改善を達成することを示す。また、直交化 DPF に MFCC を併用した場合、さらに高い性能が得られる。

キーワード 音声認識, 特徴抽出, 雑音環境, 直交化音素弁別特徴

Noise-robust Automatic Speech Recognition Using Orthogonalized Distinctive Phonetic Feature Vectors

Takashi FUKUDA and Tsuneo NITTA

Graduate School of Engineering, Toyohashi University of Technology

1-1 Hibariga-oka, Tempaku-cho, Toyohashi, 441-8580, JAPAN

Email: fukuda@vox.tutkie.tut.ac.jp nitta@tutkie.tut.ac.jp

Abstract With the aim of using an automatic speech recognition (ASR) system in practical environments, various approaches focused on noise-robustness such as noise adaptation and reduction techniques have been investigated. We have previously proposed a distinctive phonetic feature (DPF) parameter set for a noise-robust ASR system, which reduced the effect of high-level additive noise. This paper describes an attempt to apply an orthogonalized DPF parameter set as an input of HMMs. In our proposed method, orthogonal bases are calculated using conventional DPF vectors that represent 38 Japanese phonemes, then the Karhunen-Loeve transform (KLT) is used to orthogonalize the DPFs, output from a multi-layer neural network (MLN), by using the orthogonal bases. In experiments, orthogonalized DPF parameters were firstly compared with original DPF parameters on an isolated spoken-word recognition task with clean speech. Noise robustness was then tested with four types of additive noise. The proposed orthogonalized DPFs can reduce the error rate in an isolated spoken-word recognition task both with clean speech and with speech contaminated by additive noise. Furthermore, we achieved significant improvements over a baseline system with MFCC and dynamic feature-set when combining the orthogonalized DPFs with conventional static MFCCs and ΔP .

Keywords ASR, feature extraction, noisy environment, orthogonalized distinctive phonetic 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)と動的特徴のセットが多用されている[1]．MFCC パラメータセットを利用した音声認識システムは，ディクテーション用ソフトウェアとして実用化され，静かな環境で明瞭に発声される読み上げ音声に対して，すでに高い性能を達成している．しかし実際の利用環境では，種々の雑音によりスペクトラム包絡が変形する結果，認識性能が低下する．

この問題への対処法として，従来から雑音適応法[2]や雑音除去法[3]など様々な方式が提案されている．また，近年では雑音対策の一つとして，音素弁別特徴(以後 DPF(Distinctive Phonetic Feature)と呼ぶ)の利用が再び検討され始めた[4, 5, 6]^{*1}．文献[4]では，28 種類の弁別素性を 5 つのグループに分割し，それぞれのグループについて多層ニューラルネットワーク(以後 MLN(Multi Layer Neural network)と呼ぶ)を用いることにより各弁別素性を抽出する．各 MLN の出力，すなわち DPF セットは高次の MLN に入力され，音素の音響尤度に変換される．この文献では，雑音に対する頑健性と共に，自由発話音声に対する改善効果も合わせて示されている．また文献[5]においても DPF 抽出に MLN が利用されている．この方式では，BPF(Band Pass Filter)の各チャンネルに対応する MLN により DPF が抽出され，各 MLN の出力は文献[4]と同様，高次の MLN に入力される．

我々は，先に単一の MLN から DPF ベクトルを抽出し，HMM(Hidden Markov Model)で利用する方式を提案した[10]．提案方式は，入力音声を局所特徴(以後 LF(Local Feature)と呼ぶ)に変換した後，LF と ΔP から成る音響特徴系列を MLN に入力することで，音素弁別特徴を抽出する．前後のコンテキストを含む DPF

パラメータを MLN の出力とした場合，各種雑音を付加した孤立単語音声の認識性能を大きく改善した．

しかし，MLN により出力される DPF は多くが対数正規分布に近い分布を示し[11]，また各次元間で相関を持つ．対角共分散 HMM では，次元間の相関は考慮されないため，特徴パラメータの各次元間は無相関であることが望ましい．本報告では，対角共分散 HMM での DPF の利用を踏まえ，前後のコンテキストを含む DPF を直交化する方法を提案する．この処理により，DPF ベクトルは無相関正規分布を持つ特徴パラメータに変換される．評価実験では，clean な孤立単語音声を用いて，直交化 DPF と次元間に相関を持つ従来の DPF の性能を比較した後，標準的な MFCC パラメータとの比較検討結果を示す．その後，直交化 DPF の耐雑音性能を評価し，MFCC を併用した場合の結果を報告する．

本報告は以下のように構成される．2.でシステムの概要を示す．そして 3.で直交化 DPF を利用した HMM による評価実験結果を述べ，考察を示す．最後に 4.で結論をまとめる．

2. DPF ベクトルの直交化

図 1 に直交化 DPF の抽出過程を示す[12]．まず，入力音声をフレーム単位で LF に変換する．次に，LF と ΔP の系列の注目フレーム x_t と前後 3 点離れたフレーム(x_{t+3} ， x_{t-3})を結合して MLN に入力する．MLN の入力として MFCC を用いた場合，LF と比較して性能が大きく劣る[10]．MLN は 4 層構成(ユニット数は入力層から順に 75, 256, 64, 33)で，DPF 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

^{*1} 音韻論の分野では，DPF による音素分類[7]が古くから提案されており，音声認識においても古くから DPF の利用が研究されていた[8, 9]．

表1 音素弁別特徴セット

弁別特徴	a	i	u	e	o	N	w	y	j	my	ky	dy	by	gy	ny	hy	ry	py	p	t	k	ts	ch	b	d	g	z	m	n	s	sh	h	f	r	
高舌性	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	-	-	+	-	-	-	-	+	-	+		
低舌性	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	
後舌性	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	
前方性	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+	-	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	-	+	+	
舌端性	-	-	-	-	-	-	-	-	+	-	-	+	-	-	+	-	+	-	-	+	-	+	+	-	+	-	+	-	+	+	+	+	-	+	
破裂性	-	-	-	-	-	-	-	-	+	-	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	+
摩擦性	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-	-	+	-	-	+	+	+	+	+	-
有声性	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	+	+	-	-	-	-	+
連続性	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	+	+	+	+	-
鼻音性	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-
半母音性	-	-	-	-	-	-	+	+	-	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+

+ はポジティブな特徴 - はネガティブな特徴を示す

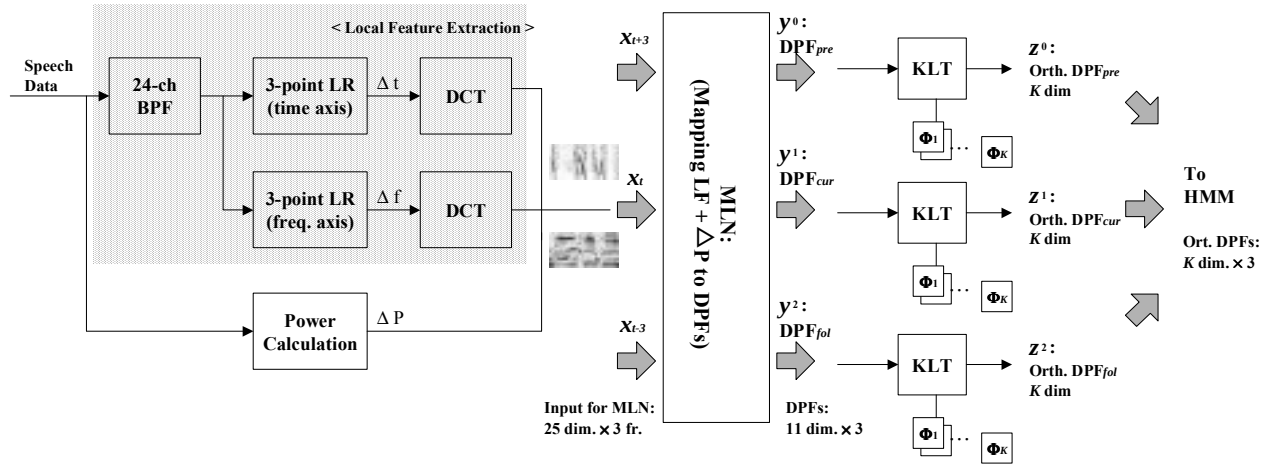


図1 直交化DPF抽出過程

なわち計 33 個の出力ユニットを持つ . 弁別特徴は表 1 に示す 11 種類を用いた[13] . 予備実験の結果を元に , “ 母音性 / 非母音性 ” と “ 子音性 / 非子音性 ” という二つの弁別素性の代わりに “ 半母音性 (/ y, w, r, ky, ny, hy, my, ry, gy, dy, by, py /) / 非半母音性 ” と “ 摩擦性 (/ s, z, sh, ts, ch, h, f, j, hy /) / 非摩擦性 ” を使用している . MLN の学習には , 誤差逆伝播法を用い , 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 . 学習データとしては , 3.1 に示す D1 データセット中に 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 により出力されるコンテキストを含む DPF は KLT を用いて無相関化する .

$$z_k^m = y^m \cdot \Phi_k \quad (m=0, 1, 2) \quad (1)$$

ここで , y^m は直交化前の DPF ベクトル , z_k^m ($k=1, 2, \dots, K$) は直交化後の k 次元目の DPF であり , Φ_k は第 k 番目の直交基底ベクトルを表す . また , $m=0$ は先行する DPF (DPF_{pre}) , $m=1$ は現在 (中心音素) の DPF (DPF_{cur}) , そして $m=2$ は後続の DPF (DPF_{fol}) を表す . 直交基底 Φ_k は , 38 音素の DPF から , 共分散行列 A を計算した後 , 固有値問題を解くことで求めた (すなわち , 表 1 において , “ + ” を 1 , “ - ” を 0 として計算する) .

$$(A - \lambda I)\Phi = 0 \quad (2)$$

ここで λ は直交基底ベクトル Φ に対応する固有値であり , I は単位行列である . 共分散行列 A は次式により計算した .

$$A = \frac{1}{38} \sum_{i=1}^{38} (x^i - \bar{x})(x^i - \bar{x})^T \quad (3)$$

ここで x^i は i 番目の音素の DPF ベクトルであり , $\bar{x}$ は 38 個の DPF の平均ベクトルである . また , 共分散行列の計算時には , DPF ベクトル x^i と平均ベクトル $\bar{x}$ の差分値を , そのノルムで正規化している . 直交化した各 DPF ($Orth.DPF_{pre}$, $Orth.DPF_{cur}$, $Orth.DPF_{fol}$) は結合し , DPF ベクトル時系列 (以後 直交化 DPF と呼ぶ) 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・松下单語音声データベース . 先頭の 100 語男性話者 10 名を使用 . サンプル周波数は 24kHz から 16kHz へ変換 .

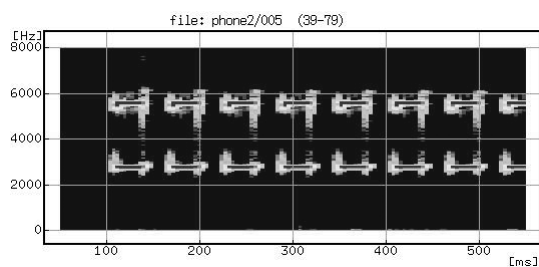
D3. 雑音データセット:

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

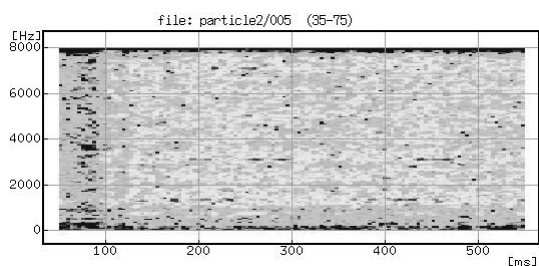
- (A) Mobile Phone : 携帯電話の着信音
 - (B) Particles : 多数の粒子を金属箱に注ぐ音
 - (C) Whistle : 笛を吹いた音
- に加えて , 白色雑音を使用する .

3.2 雑音のスペクトラム構造

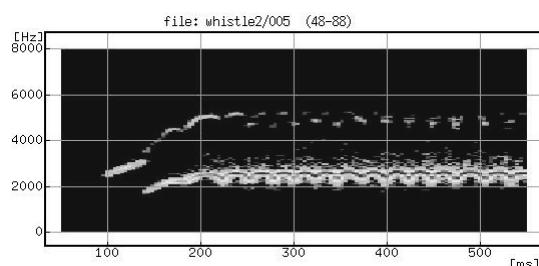
図 2 に評価で使用した 3 種類の雑音のスペクトラムパターンを示す . “ Mobile Phone ” と “ Whistle ” は一定の周波数帯域で持続する雑音である . “ Particles ” は白色雑音のように全周波数帯域に分布する .



(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2. 雑音のスペクトラムパターン

3.3 実験の概要

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

3.4.1 直交化の効果

(A) 直交化 DPF と非直交化 DPF の性能比較

前節に述べた実験条件下で不特定話者孤立単語認識 (clean 音声) を行った. 図 3 に実験結果を示す (図中の括弧内は次元数を表す). 直交化した DPF パラメータは, 次元間に相関を持つ DPF, すなわち MLN の出力をそのまま HMM に入力した場合と比較して, 性能を大き

く改善した. 基底ベクトルの数は $K = 6$ のときに最も誤り数が少なくなるが, $K = 4$ 以下の場合, 認識に必要な情報が欠落する結果, 低い性能にとどまった.

(B) MFCC パラメータとの比較

図 4 に標準的な MFCC パラメータとの比較結果を示す. 図中の Baseline は MFCC と 1 次と 2 次の動的特徴 (Δ_1, Δ_2), および差分パワー ($\Delta P, \Delta \Delta P$) を結合した 38 次元の特徴パラメータを直接 HMM に入力した場合の結果である. 混合数 4 以上の時, 直交化 DPF は Baseline よりも高い性能を示した.

(C) 直交化 DPF と MFCC の組み合わせ

図 5 に直交化 DPF ($K = 6$) と MFCC (12 次元), および差分パワー ($\Delta P, 1$ 次元) を併用した場合 (合計 31 次元) の認識結果を示す. 図に示すように, 直交化 DPF と MFCC を併用することで, 全ての混合数において Baseline を越える性能を得た. これは, 文献 [4, 14] でも報告されているように, MFCC と DPF では誤りを起こす語彙が大きく異なるため, 互いの弱点を補うことができたことによると考えられる.

3.4.2 耐雑音性能

(A) 直交化 DPF と非直交化 DPF の性能比較

D2 評価データセットに D3 雑音データセットを SN 比 10dB で重畳したときの実験結果を図 6 に示す. 結果が示すように, 雑音を付加した認識タスクにおいても, 直交化 DPF は性能を改善させた. KLT により次元数を変化させた場合の性能差は少なく, 直交化 DPF は $K = 11$ のとき平均して高い性能を示した.

(B) MFCC パラメータとの比較

D2 評価データセットに D3 雑音データセットを SN 比 5dB および 10dB で重畳したときの直交化 DPF の性能を図 7, 8 に示す. Baseline は 3.4.1 と同様である. 直交化 DPF は雑音に対して認識誤りを削減し, 特に, "White Noise" と "Mobile Phone" に対して高い効果を示した. しかし, "Particles" に関しては依然として, Baseline と比較して低い性能にとどまった [10]. "Particles" 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ため, 提案方式の性能が劣化したと推測される.

(C) 直交 DPF と MFCC の組み合わせ

直交化 DPF ($K = 11$) と MFCC (12 次元) および差分パワー ($\Delta P, 1$ 次元) の計 46 次元を連結した場合の性能を比較評価した. 図 9, 10 に実験結果を示す. 直交化 DPF と MFCC を連結したパラメータを適用することで, 顕著な性能改善が得られた. また, "Particles" に関しても Baseline と同等の性能を示した. 特に, "White Noise" と "Mobile Phone" に関して, 極めて高い効果を示しており, "White Noise" について, SN 比 10dB で 9.8% から

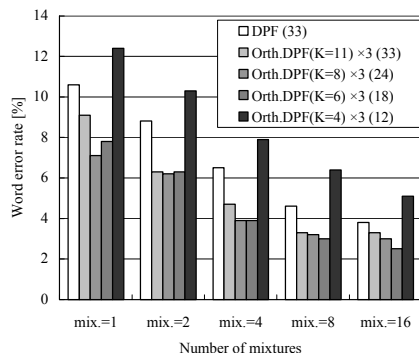


図3 直交化 DPF vs. 非直交化 DPF

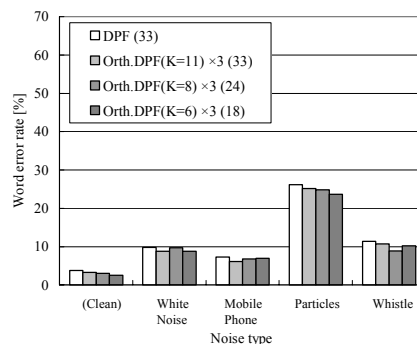


図6 耐雑音性能: SNR=10 dB
直交化 DPF vs. 非直交化 DPF

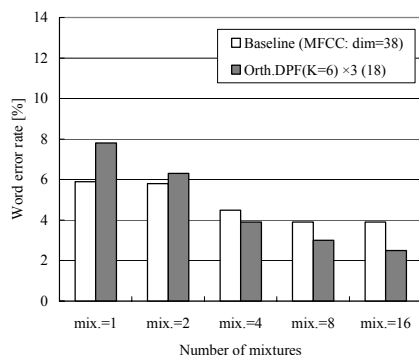


図4 直交化 DPF vs. MFCC

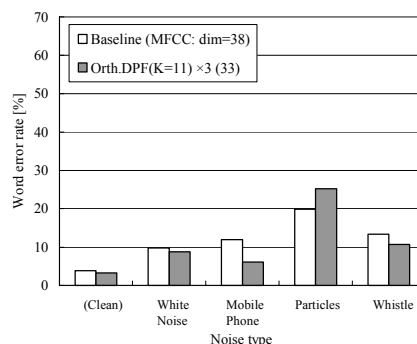


図7 耐雑音性能: SNR=10 dB
直交化 DPF vs. MFCC

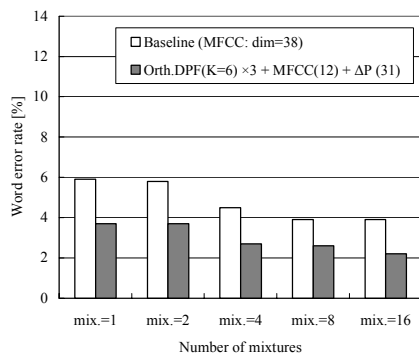


図5 MFCCを併用した場合の性能改善

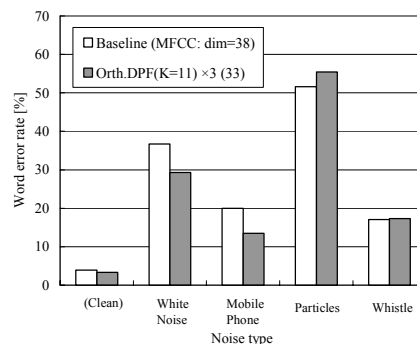


図8 耐雑音性能: SNR=5 dB
直交化 DPF vs. MFCC

5.5%, SN 比 5dB で 36.7%から 26.8%, また, ”Mobile Phone” について, SN 比 10dB で 12.0%から 4.4%, SN 比 5dB で 20.0%から 12.0%と誤りを大きく改善した.

4. おわりに

DPF を KLT により直交化し, 対角共分散 HMM の入力として利用する方式を提案した. 孤立発話音声を対象とした不特定話者音声認識実験において, 直交化

DPF は MFCC に基づく標準的なパラメータと比較して, clean 音声および雑音重畳音声の双方で, 良好な性能を示した. また, 直交化 DPF と MFCC を結合したパラメータを利用したとき, 提案方式は Baseline を大きく上回る性能を達成した.

今後は, 音素弁別特徴セットの再検討を含め, DPF 抽出方式を改良する共に, この方式を効果的に利用する音声認識法を検討したい.

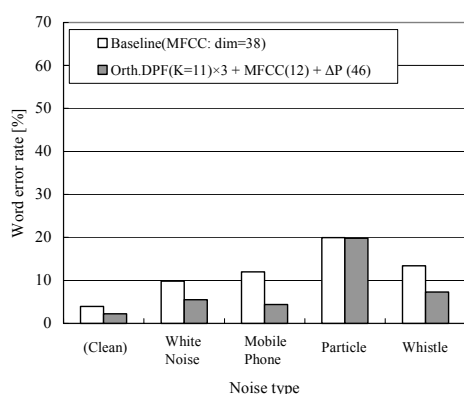


図9 耐雑音性能: SNR=10 dB
直交化 DPF + MFCC

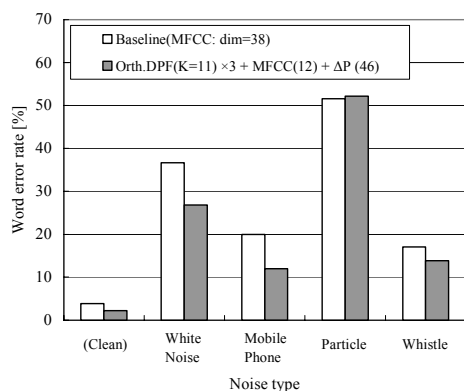


図10 耐雑音性能: SNR=5 dB
直交化 DPF + MFCC

謝辞

本研究は文部科学省 21 世紀 COE プログラム「インテリジェント ヒューマン センシング」の援助により行われた。また、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。

参考文献

- [1] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process, ASSP-34, pp.522-529, 1986.
- [2] M. J. F. Gales and S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination" IEEE Trans. Speech and Audio Processing, Vol.4, No.5, pp.352-359, 1996.
- [3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoustic

Speech Signal Processing, Vol.27, No.2, pp.113-120, 1979.

- [4] 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.
- [5] P. Jain, H. Hermansky and B. Kingsbury, "Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features," Proc. ICSLP'02, pp.473-476, 2002.
- [6] E. Eide, "Distinctive Features for Use in an Automatic Speech Recognition System," Proc. Eurospeech'01, pp.1613-1616, 2001.
- [7] N. Chomsky and M. Halle, "The Sound Pattern of English," New York, Harper and Row, 1968.
- [8] T. B. Martin, "Practical Application of Voice Input to Machine," Proc. IEEE, 64-4, 1976.
- [9] 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.
- [10] T. Fukuda, W. Yamamoto and T. Nitta, "Distinctive Phonetic Feature Extraction for Robust Speech Recognition," Proc. ICASSP'03, Vol. , pp.25-28, 2003.
- [11] 福田 隆, 新田恒雄, "音素弁別特徴ベクトルの対数正規分布近似を用いた雑音環境下音声認識," 信学技報, SP2003-23, pp.19-24 (2003) .
- [12] T. Fukuda and T. Nitta, "Noise-robust Automatic Speech Recognition Using Orthogonalized Distinctive Phonetic Feature Vectors," Proc. Eurospeech'03, 2003.
- [13] 比企静雄 編, "音声情報処理," 東京大学出版会, 1973 .
- [14] B. Launay, O. Siohan, A. Surendran and C. H. Lee, "Towards Knowledge-based Features for HMM Based Large Vocabulary Automatic Speech Recognition," Proc. ICASSP'02, pp.817-820, 2002.