社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＳＬＰ－51　（４）
2004／5／27

音声認識のための特徴パラメータ正準化法の検討

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
E-mail: fukuda@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

あらまし　確率的分類器（HMM）に基づく認識システムは，性別，話速，音響環境等に起因する HMM の隠れ変数のバイアスにより性能が低下する．隠れ変数の影響を特徴抽出の段階で吸収することができれば，より頑健な認識システムを構成できると考えられる．その第一段階として，本報告では，男女の声質を対象とした特徴パラメータ正準化方式について述べる．正準化は音響特徴空間と音響モデルの間に，中間表現である音素弁別特徴（DPF）空間を導入することにより実現する．正準化器は，三つの DPF 抽出器と一つの DPF 選択器から成る．まず DPF 抽出部では，多層ニューラルネットワーク（MLN）に基づく DPF 抽出器から，話者の声質に対応する三つの DPF ベクトルを抽出する．次に，DPF 選択部では三種類の DPF ベクトルのうち，音響モデルに最も適合する DPF ベクトルを正準化 DPF ベクトルとして抽出する．評価実験では，単一の HMM 分類器に正準化 DPF ベクトルを入力する方式と，隠れ変数に対応させた複数の HMM 分類器に従来の音響特徴パラメータ（MFCC）を入力する方式とを比較する．提案方式は，少ない演算量にもかかわらず良好な性能が得られることを示す．

キーワード　音声認識，特徴抽出，正準化，隠れ変数，音素弁別特徴

Canonicalization of the Feature Parameters for Automatic Speech Recognition

Takashi FUKUDA　and　Tsuneo NITTA

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

Abstract　Acoustic models (AMs) of an HMM-based classifier include various types of hidden variables such as gender type, speaking rate, and acoustic environment. If there exists a canonicalization process that reduces the influence of the hidden variables from the AMs, a robust automatic speech recognition (ASR) system can be realized. In this paper, we describe the configuration of a canonicalization process targeting gender type as a hidden variable. The proposed canonicalization process is composed of multiple distinctive phonetic feature (DPF) extractors corresponding to the hidden variable and a DPF selector in which the distance between input DPF and AMs is compared. In a DPF extraction stage, an input sequence of acoustic feature vectors is mapped onto three DPF spaces corresponding to male, female, and neutral voice by using three multilayer neural networks (MLNs). Experiments are carried out by comparing (A) the combination of the canonicalized DPF and a single HMM classifier, and (B) the combination of a single acoustic feature (MFCC) and multiple HMM classifiers. The result shows that the proposed canonicalization method outperforms both of the conventional ASR with MFCC and a single HMM and the ASR with multiple HMMs in spite of less memories and computation time.

Keyword　Automatic Speech Recognition, Feature Extraction, Canonicalization, Hidden Variables, Distinctive Phonetic Features

1. はじめに

音声認識技術の進展により，近年，雑音の少ない環境で明瞭に読み上げた音声については，9 割以上の認識精度を達成できるようになった．これは，時系列データの分類に適した確率的分類器(HMM)が提案され，大規模音声コーパスの整備が組織的になされたことによる．しかしながら，性別，話速，音響環境等に起因する HMM の隠れ変数のバイアスにより，話者，発話スタイル，あるいは利用環境によっては，依然として低い認識精度にとどまっている．

この問題に対して，数百時間を超える膨大な音声データから，音響モデルを学習する方法が検討されている[1, 2, 3]．この手法は，広範囲に亘る変動をカバーできる反面，モデルの分布が広がることにより，音素分類能力そのものが低下する．

他方，変動要因に対応させた複数のパスを同一音響モデル内で定義し(マルチパス音響モデル)，デコーディングの際，最尤の経路を選択する方法が提案されている[4, 5, 6]（図 1 参照）．また，マルチパス音響モデルの代わりに，複数の HMM 分類器を隠れ変数に対応させ，尤度最大基準で仮説を選択する方法が検討され始めた[7, 8]（図 2 参照）．これらの方法は個々の音響モデルの音素分類能力を低下させることなく，頑健な音声認識システムを構築できる可能性がある．しかし，演算コスト，メモリ量の観点から実用的な方式とは言えない．HMM の隠れ変数の内，性能低下につながる大きな要因となるものを特徴抽出の段階で吸収することができれば，より頑健な音声認識を低コストで実現できると期待される．

本報告では，第一段階として，男女の声質を対象に特徴パラメータを正準化する方式を提案する．提案方式では，まず，声質に対応した三つの音素弁別特徴(以後 DPF（Distinctive Phonetic Feature）と呼ぶ）に対応する抽出器（男性用，女性用，中性用）から，それぞれ DPF を抽出する（図 3 参照）．次に，男女の DPF ベクトルと音響モデルとの距離を比較し，距離の近い DPF 系列を正準化 DPF として選択する．両者の距離が近い場合には，中性の DPF を正準化 DPF として利用する．上述の正準化は，音響特徴空間と HMM 分類器の音響モデルとの間に DPF 空間を導入することによってはじめて達成される．

本報告は以下のように構成される．2.で DPF 抽出器の概要を示した後，3.で特徴パラメータ正準化法について説明する．4.で評価実験結果と考察を述べ，最後に 5.で結論をまとめる．

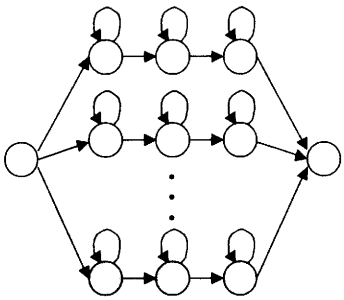

図 1 マルチパス音響モデルの例

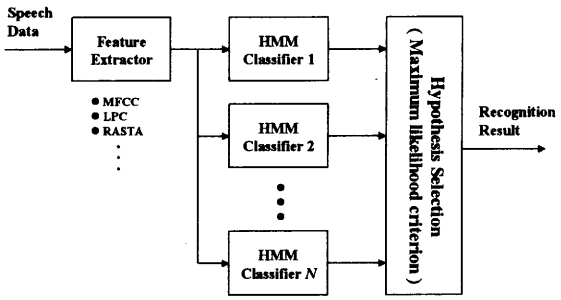

図 2 並列デコーディング方式の例

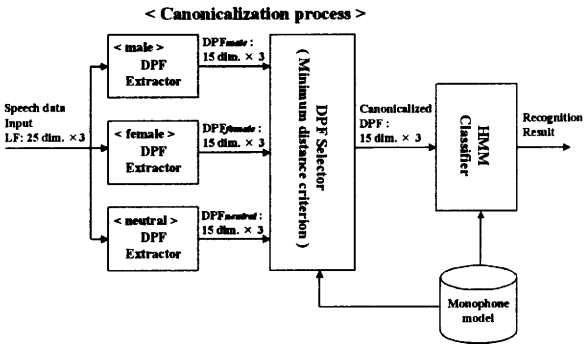

図 3 特徴パラメータ正準化器

2. 音素弁別特徴(DPF)抽出器

本節では，特徴パラメータ正準化の一役を担う DPF 抽出器の概要を説明する．図 4 に DPF 抽出過程を示す[9]．まず，入力音声をフレーム単位で局所特徴（以後

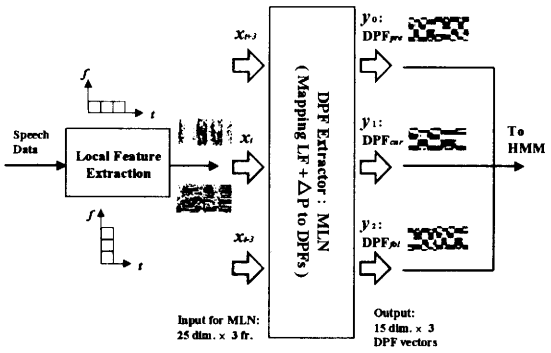

図 4 DPF 抽出器

LF (Local Feature) と呼ぶ) に変換する[10, 11]. 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})を結合して多層ニューラルネットワーク (以後 MLN (Multilayer Neural Network) と呼ぶ) に入力する. MLN は 4 層構成(ユニット数は入力層から順に 75, 256, 96, 45)で, DPF に対応する 15 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 45 個の出力ユニットを持つ. 弁別特徴は音素間の距離がなるべく均等になるように設計した DPF セット[12, 13]を用いる (表 1 参照). 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, 4.1 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).

これまでに, 我々は HMM 中の DPF 分布を対数正規分布で近似する方法を提案した[14]. 以下, DPF を HMM 分類器への入力とする場合, 出力確率分布を対数正規分布で表現するものとする.

3. 特徴パラメータの正準化

本節では, 特徴パラメータ正準化方式について述べる. 図 3 に示したように正準化器では, まず, 三つの DPF 抽出器により声質に対応した三つの DPF を抽出する. 以後, 声質に対応させた DPF 抽出器をそれぞれ男性 DPF 抽出器, 女性 DPF 抽出器, および中性 DPF 抽出器と呼ぶ (各 DPF 抽出器から抽出される DPF を DPF_{male}, DPF_{female}, $DPF_{neutral}$ と呼ぶ). 男性 DPF 抽出器と女性 DPF 抽出器は, 4.1 に示す D1 データセットから独立に学習し, 中性 DPF 抽出器は, D1 データセット中の男女両方の音声を用いて学習する. 上記のように, 声質に応じて設計した各 DPF 抽出器は, 単一の DPF 抽出器 (男女全ての音声を単一の MLN で表現する場合) よりも歪みの少ない DPF を出力すると推測される. 話者の声質に適合する DPF を HMM の入力として利用することで, HMM の隠れ変数の影響を軽減させることができると考えられる. 次に, DPF 選択器に

表1 バランスを考慮した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
モーラ性	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
高舌性	−	+	+	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	−	+	−	−	+	−	−	−	−	+	−	+	−
低舌性	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−
nil	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	+	−	+	+	−	+	+	+	+	−	−	−	+
前方性	−	−	−	−	−	−	−	−	+	+	−	+	+	−	+	−	+	+	+	+	−	+	+	+	+	−	+	+	+	+	+	−	+	+
後方性	+	−	+	−	+	−	+	−	−	−	−	−	−	+	−	+	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−
nil	−	+	−	+	−	+	−	+	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−
舌端性	−	−	−	−	−	−	−	−	+	−	−	+	−	−	+	−	+	−	−	+	−	+	−	+	−	+	−	+	−	+	+	+	−	−	+
破裂性	−	−	−	−	−	−	−	−	−	−	+	+	+	+	−	−	−	+	+	+	+	−	−	−	+	+	+	−	−	−	−	−	−	−	−
破擦性	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	+	−	−	−	−	−	−	−	−
連続性	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	+	+	+	+	+	+	+	+	−
有声性	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	−	+	−	−	−	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
非有声性	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	+	+	+	+	+	+	+	−	−	−	−	−	−	+	+	+	+	−
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

ついて説明する．

DPF 選択器の構成としては，歪み尺度や情報量基準に基づく方法など様々なものが考えられるが，今回は音響モデルとの距離尺度に基づく方式を検討した．まず，次の手順で声質判別計算用の初期音響モデルを学習する．

(1) 4.1節のD1データセットから男女の声質に対応させた男性 DPF 抽出器と女性 DPF 抽出器を独立に学習する．
(2) D1 データセットのラベル情報に基づき，男性の音声については男性 DPF 抽出器から，女性の音声については女性 DPF 抽出器から DPF を抽出する．
(3) 上記の手順で得られる DPF を用いて初期音響モデルを学習する．

DPF 選択器は，以下に示す式(1)，(2)を用いて，初期音響モデルと男女 DPF（DPF_{male} と DPF_{female}）との距離を計算し，距離が小さい DPF を正準化 DPF として抽出する（正準化は単語単位で行う）．ただし，DPF_{male} と DPF_{female} の距離が近い場合，具体的には DPF_{male} と音響モデル間の距離が，DPF_{female} と音響モデル間の距離の±25%以内であるとき，$\mathrm{DPF}_{neutral}$ を正準化 DPF として利用する．

$$d = \frac{1}{N}\sum_{i=1}^{N} \min_{j} D_M\left(\mathbf{p}_i, \mathbf{q}_j\right) \quad (j = 1, 2, \ldots, M) \qquad (1)$$

$$D_M\left(\mathbf{p}_i, \mathbf{q}_j\right) = \left(\log(\mathbf{p}_i) - \boldsymbol{\mu}_j\right)' \Sigma_j^{-1} \left(\log(\mathbf{p}_i) - \boldsymbol{\mu}_j\right) \qquad (2)$$

ここで，$\mathbf{p}_i$ は i フレーム目の DPF ベクトル，$\mathbf{q}_j$ は j 番目の音響モデルにおける中央の状態の分布集合，D_M は $\mathbf{p}_i$ と $\mathbf{q}_j$ のマハラノビス距離を表す．また，N は距離計算に用いるフレーム数，M は音素数を表す，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ はそれぞれ音響モデルの平均ベクトルおよび共分散行列である．上式の D_M は DPF の抽出性能が高い母音区間で計算し，混合数 1 の音響モデルに対して行う．今回は，モーラ性を示す DPF の値が 0.5 以上の区間を母音区間とした．なお，音響モデルの出力確率分布表現として対数正規分布を利用しているため，式(2)中には入力ベクトル $\mathbf{p}_i$ に対する対数演算が含まれる．

上述の手順で得られる正準化 DPF は，HMM 分類器に入力される．HMM 分類器は，正準化 DPF を用いて尤度が収束するまで学習を繰り返す．

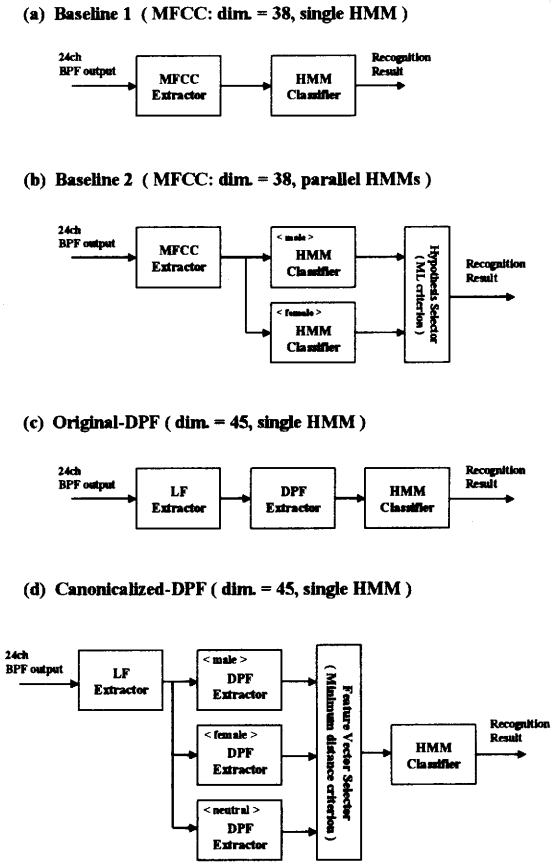

図 5 実験で用いる ASR システム

4. 評価実験

4.1. 音声試料

以下に示す二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[15] (16kHz, 16bit)のうち男女各 30 名，合計 9003 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[16]．先頭の 100 語男女各 10 名，合計 1962 文を使用．サンプリング周波数は 24kHz から 16kHz へ変換．

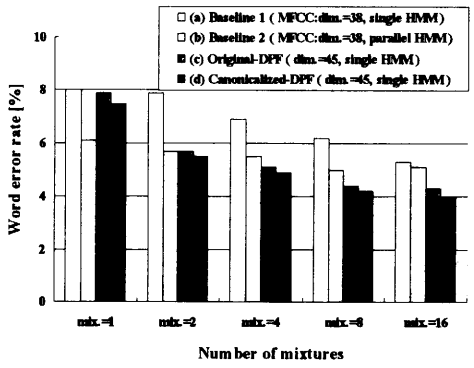

図 6 実験結果

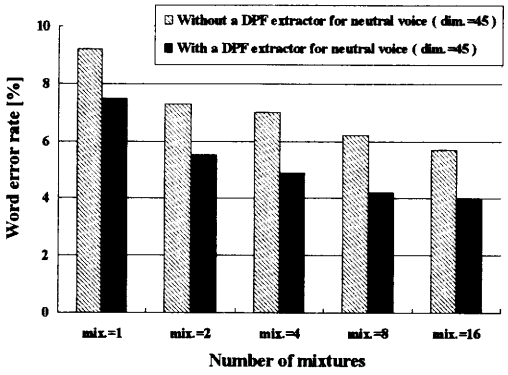

図 7 中性 DPF 抽出器の有無による性能差

4.2. 実験の概要

入力音声は 16kHz でサンプリング後，512 点の FFT 分析処理を行う（25ms ハミング窓，フレーム周期 10ms）．続いてパワースペクトルをメルスケールの中心周波数を持つ 24 チャンネル BPF 群により求めた．この後，時間軸及び周波数軸上 3 点の線形回帰演算により局所特徴を求め，DPF 抽出器の入力パラメータとする．

音響モデルは 5-state，3-loop，日本語 43 音素 monophone- HMM を使用し， D1 データセットから抽出した正準化 DPF を用いて学習する．出力確率分布表現として，MFCC には正規分布を，DPF には対数正規分布を用いる[14]．なお，今回は負の歪度を持つ分布だけを扱い，共分散行列は対角成分のみを利用する．

4.3. 実験結果

前節に示した実験条件で不特定話者孤立単語認識を行った．実験では，図 5 に示す四つの音声認識(ASR)システムについて比較した．図中の Baseline 1 は MFCC と 1 次と 2 次の動的特徴(Δ_t, $\Delta_t\Delta_t$)，および差分パワー(ΔP, $\Delta\Delta P$)を結合した 38 次元の特徴パラメータを直接 HMM に入力する方法を示す．また，図中 (b) の Baseline 2 は男女の声質に対応させた HMM の出力を尤度最大基準で選択する方式を示す．そして図中 (c) の Original-DPF は，D1 データセットにおける男女両方の音声から学習した単一の DPF 抽出器を利用する方法を指す．

図 6 に実験結果を示す．正準化 DPF パラメータは，混合数 1 の場合を除き，Baseline1，2 と比べて同等以上の性能を示した．また，単一の DPF 抽出器から得られる Original-DPF についても高い性能を達成していることがわかる．これは，DPF そのものが音素分類に適した特徴であるのに加え，正準化 DPF が男女の声質を吸収した特徴空間を構成したためと考えられる．

表 1 各 DPF の選択率と性別判別精度 [%]
（中性 DPF 抽出器を利用しない場合）．

	D1 data set (ASJ database)		D2 data set (Tohoku Univ. database)	
	選択率	正解精度	選択率	正解精度
DPF*male*	48.5	99.5	51.2	86.9
DPF*female*	51.5	96.6	48.8	88.3

表 2 各 DPF の選択率と性別判別精度 [%]
（中性 DPF 抽出器を利用する場合）．

	D1 data set (ASJ database)		D2 data set (Tohoku Univ. database)	
	選択率	正解精度	選択率	正解精度
DPF*male*	34.7	100.0	34.8	94.9
DPF*female*	31.5	99.9	26.1	97.8
DPF*neutral*	33.8		39.1	

4.4. 考 察

提案した特徴パラメータ正準化方式は三つの DPF

抽出器（男性用，女性用，中性用）を利用している．ここでは，中性 DPF 抽出器の効果について検討する．図 7 に，正準化器から中性 DPF 抽出器を取り除いた場合の認識性能を示す．また，表 1，2 に DPF 選択器が出力する各 DPF（DPF_{male}, DPF_{female}, $DPF_{neutral}$）の割合と，その正解精度を示す（$DPF_{neutral}$ は DPF_{male} と DPF_{female} が近い場合に選択されるため，表 2 で $DPF_{neutral}$ の正解精度は存在しない）．

図から，正準化器における中性 DPF 抽出器の有無で認識性能が大きく変化することがわかる．これは表 1 に示すように，中性 DPF 抽出器を利用しない場合は，DPF 選択器の性別判別精度が低下するためと考えられる（特に D2 データセットで性能低下が顕著である．）．一方，中性 DPF 抽出器を正準化器に導入する場合は，これらの選択誤りが $DPF_{neutral}$ として吸収されるため，性能の改善につながったと考えられる．

5. おわりに

音素弁別特徴の特徴パラメータの正準化方式を提案した．正準化は，音響特徴空間と HMM 分類器の音響モデルとの間に DPF 空間を導入することではじめて実現される．評価実験では，孤立単語音声を用いた不特定話者認識実験により，正準化パラメータの有効性を示した．

今後は，発話スタイル，音響環境に関しても正準化方式を検討したい．

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，堀情報科学振興財団研究助成の支援を受け実施された．

文　献

[1] K. W. Church, "Speech and Language Processing: Where have we been and where are we going?," Proc. Eurospeech'03, vol.1, pp.1-4, 2003.

[2] 南條浩輝，加藤一臣，李晃伸，河原達也，"大規模な日本語話し言葉データベースを用いた講演音声認識，" 信学論，Vol.J86-D-II, No.4, pp.450-459, 2003.

[3] 河原達也，住吉貴志，李晃伸，坂野秀樹，武田一哉，三村正人，伊藤克亘，伊藤彰則，鹿野清宏，"連続音声認識コンソーシアム 2002 年度版ソフトウエアの概要，" 2003-SLP-48，pp.1-6, 2003.

[4] A. Lee, Y. Mera, K. Shikano and H. Saruwatari, "Selective multi-path acoustic model based on database likelihoods," Proc. ICSLP'02, pp.2661-2664, 2002.

[5] 伊田政樹，中村哲，"雑音 GMM の適応化と SN 比別マルチパスモデルを用いた HMM 合成による高速な雑音環境適応化，" 信学論，Vol. J86-D-II, No. 2, pp. 195-203, 2003.

[6] 伊藤彰則，喜嶋朋令，鈴木基之，牧野正三，"雑音マルチパスモデルによる非定常雑音下音声認識の検討，" 信学技報，SP2003-20，pp.1-6, 2003.

[7] 松田繁樹，實廣貴敏，コンスタンティン マルコフ，中村哲，"雑音や発話スタイルの変動に頑健な日本語大語彙連続音声認識，" 2004-SLP-50，pp.37-44, 2004.

[8] 篠崎隆宏，古井貞熙，"超並列デコーダによる話し言葉音声認識，" 第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.67-72, 2004.

[9] T. Fukuda, W. Yamamoto and T. Nitta, "Distinctive Phonetic Feature Extraction for Robust Speech Recognition," Proc. ICASSP'03, vol.II, pp.25-28, 2003.

[10] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. of Proc. ICASSP'99, pp.421-424, 1999.

[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, 2002-7.

[12] 福田隆，新田恒雄，"頑健な音声認識のためのバランスを考慮した日本語音素弁別特徴セットの検討，" 音講論，1-6-5, pp.9-10, 2003-9.

[13] T. Fukuda and T. Nitta, "Orthogonalized Distinctive Phonetic Feature Extraction for Noise-Robust Automatic Speech Recognition," The Institute of Electronics, Information and Communication Engineerings (IEICE) Trans. Information and Systems, 2004.

[14] T. Fukuda and T. Nitta, "Noise-robust ASR by Using Distinctive Phonetic Features Approximated with Logarithmic Normal Distribution of HMM," Proc. Eurospeech'03, Vol.III, pp.2185-2188, 2003

[15] T. Kobayashi, S. Itahashi, S. Hayamizu and T. Takezawa, "ASJ continuous speech corpus for research," Acoustic Society of Japan (ASJ) Trans. vol.48. no.12. pp.888-893, 1992.

[16] S. Makino, K, Niyada, Y. Mafune and K. Kido, "Tohoku University and Matsushita isolated spoken word database," Acoustic Society of Japan (ASJ) Trans. vol.48, no.12, pp.899-905, 1992