音声言語情報処理　34－2
（2000．12．21）

音声認識のための周辺特徴の検討

福田　隆　　瀧川　正史　　新田　恒雄

豊橋技術科学大学大学院工学研究科　知識情報工学専攻
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1　E-mail : nitta@tutkie.tut.ac.jp

あらまし　本報告では周波数領域およびケプストラム領域において，周辺特徴の抽出機構を特徴抽出器へ組み込むことを試みる．最初に，周波数領域における周辺特徴抽出法を説明し，周辺特徴を組み込んだ特徴抽出器が従来の特徴抽出器に比べて，高い認識性能を示すことを **HMM** に基づく単語認識実験から述べる．次に，これらの実験結果からケプストラム領域の周辺特徴抽出法を提案し，"**MFCC** と周辺特徴"のパラメータセットを"**MFCC** と動的特徴"のパラメータセットと性能比較した結果から，周辺特徴としてΔ_tにΔ_qパラメータを併用することの有効性を指摘する．

キーワード　音声認識，特徴抽出，ケプストラム分析，直交基底，写像演算子，周辺特徴

Peripheral Features for Speech Recognition

Takashi FUKUDA, Masashi TAKIGAWA, and Tsuneo NITTA

Department of Knowledge-based Information Engineering, Graduate School of Engineering,
Toyohashi University of Technology, 1-1 Hibariga-oka, Tempaku, Toyohashi, 441-8580 JAPAN
E-mail : nitta@tutkie.tut.ac.jp

Abstract　This paper describes an attempt to incorporate the mechanism of extracting the peripheral features into a feature extractor in frequency and quefrency domain. Firstly, the method of extracting peripheral features in frequency domain is shown, and then the feature extractor that combines the peripheral features with **MFCC** improves the performance in comparison with the standard feature extractor in word recognition experiments with an **HMM**-based ASR system. Secondary, the method of extracting the peripheral feature in quefrency domain is also provided. The proposed parameter set of **MFCC** with peripheral features shows significant improvements in comparison with the standard parameter set of **MFCC** with dynamic features in experiments.

key words　Speech Recognition, Feature Extraction, Cepstrum Analysis, Orthogonal Basis, Mapping Operator, Peripheral Feature

1. はじめに

時間-スペクトル(TS)パターン $\mathbf{x}(t,f)$は長らく，自動音声認識における音響特徴として使用されてきた．近年，Δケプストラム，Δパワーなどの動的特徴が ASR に導入され[1, 2]，**MFCC** と動的特徴のセットが広く用いられている．動的特徴は時間軸に沿った TS パターン上の点 $\mathbf{x}(t_i,f_j)$の周辺特徴を表現している．しかし，TS パターン上の n×n 近傍の点からはより多くの周辺特徴を得ることができると考えられる．

先行研究では[3]，複合音響特徴平面(**MAFP**)に基づいた特徴抽出を音素分類に適用し，性能を大きく改善することを示した．この方法では，$\mathbf{x}(t,f)$を要素とする集合 $\mathbf{X}$ を局所的な空間写像演算子 $\mathbf{G}_m$ ($\mathbf{G}_m \in \mathbf{G}$)により，さまざまな AFP(音響特徴平面 : Acoustic Feature Planes) $\mathbf{Y}_m=\mathbf{y}_m(t, f)$, m=1, 2, ..., M に写像した．

$$\mathbf{G}_m : \mathbf{X} \rightarrow \mathbf{Y}_m \qquad (1)$$

$\mathbf{G}_m$としては，まず TS パターン上の 3×3 近傍の直交基底$\{\mathbf{\Phi}_1, \mathbf{\Phi}_2, ..., \mathbf{\Phi}_9\}$を音声データから直接抽出する．次に，直交基底を対称化して，モデル化したものを使用している．今回は，この手法を周波数領域およびケプストラム領域における周辺特徴抽出に適用することを試みる[4]．周波数領域においては，最初に直交基底を抽出し，それを写像演算子として捉え，TS パターンを周辺特徴空間に写像する．一方，ケプストラム領域においても同様の直交基底を抽出できる．抽出した直交基底，すなわち写像演算子は単純化かつ対称化した後，**HMM** に基づく ASR システムの特徴抽出器に組み込む．最後に，今回提案する特徴パラメータセットを不特定話者単語認識実験を通して，**MFCC** と動的特徴の標準的なパラメータセットと比較する．

以下，2. で TS パターン上の 7×7 近傍の幾何学的な特徴(構造特徴)について述べ，TS パターンから得られる周辺特徴を用いた認識実験結果を示す．　3. では TS パターンに代えて，TQ パターン上の 7×3 近傍の直交基底を観測し，TQ パターンから抽出される周辺特徴を用いた認識実験結果を述べる．

2. 周波数領域における周辺特徴

2.1 TS パターン上の周辺特徴

TS パターン上にはさまざまな幾何学的構造が観測される．図 1 に TS パターン $\mathbf{x}(t_i,f_j)$, j=1, 2, ..., 24 の 7×7 ブロックにおける上位 9 個，$\mathbf{\Phi}_1$, $\mathbf{\Phi}_2$, ..., $\mathbf{\Phi}_9$ の直交基底を示す．図では，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．7×7 直交基底は **2.3.1** で示す音声データから KL 変換(**KLT** : Karhunen-Loeve transform)を用いて抽出した．

これらの直交基底は空間写像演算子として観察すると，まず$\mathbf{\Phi}_1$ は平滑化演算子と見なせる．この演算子は **MFCC** の濃度情報を別に使用するなら，特徴抽出効果はないと考えられる．$\mathbf{\Phi}_2$, $\mathbf{\Phi}_3$ はそれぞれ時間軸(Δ_t)と周波数軸(Δ_f)に関しての 1 次微分演算子，$\mathbf{\Phi}_4$, $\mathbf{\Phi}_9$ は時間軸($\Delta_t\Delta_t$)と周波数軸($\Delta_f\Delta_f$)についての 2 次微分演算子，また$\mathbf{\Phi}_5$, $\mathbf{\Phi}_6$, $\mathbf{\Phi}_7$, $\mathbf{\Phi}_8$ は TS パターン上の尾根や谷を表す部分空間である．

時間-周波数空間演算子$\{\mathbf{\Phi}_m\}$($\mathbf{\Phi}_m \in \mathbf{\Phi}$)を用いると，TS パターン $\mathbf{x}(t, f)$ からさまざまな周辺特徴 $\mathbf{Y}_m= \mathbf{y}_m(t, f)$, m=1, 2, ..., M を抽出することができる．周辺特徴の構成要素 $\mathbf{y}_m$ (t,f)は 7×7 近傍の $\mathbf{x}(t,f)$と $\mathbf{\Phi}_m= \boldsymbol{\phi}_m(t_i,f_j)$により次式から計算される．

$$y_m(t,f) = \sum_{i=-3}^{3}\sum_{j=-3}^{3} x(t+i, f+j)\phi_m(i,j) \qquad (2)$$

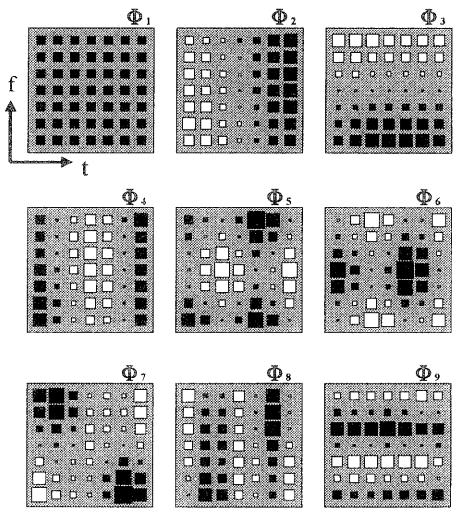

図 1　TS パターン上の 7×7 直交基底

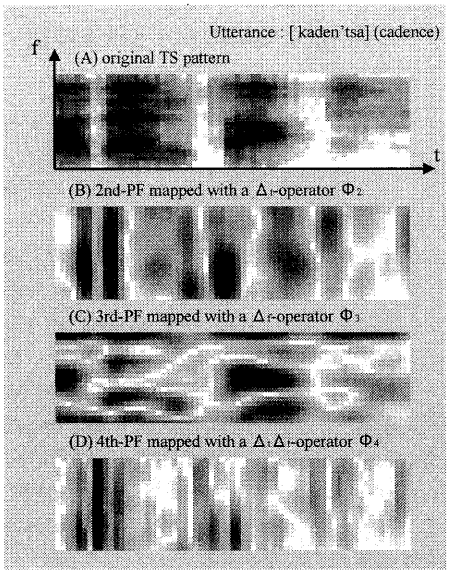

図 2　周波数領域の時間 – スペクトルパターンと周辺特徴

図 2 は発話[kaden'tsa](カデンツァ)の上位 3 つの周辺特徴の例である．図において，(A)は **TS** パターン，(B)，(C)および(D)はそれぞれΔ_t演算子$\boldsymbol{\Phi}_2$を用いて写像した 2 番目の **PF**(PF : Peripheral Feature)，Δ_f演算子$\boldsymbol{\Phi}_3$による 3 番目の **PF**，そして$\Delta_t\Delta_t$演算子$\boldsymbol{\Phi}_4$ による 4 番目の **PF** である．$\mathbf{y}_m(t, f)$の正の値は正の傾斜を表しており，負の値は負の傾斜を示す．たとえば，破裂音の明瞭な時間変化は 2 番目の **PF** 上で正負の値の対で表現され，同様に，定常部の明瞭なスペクトルピークは 3 番目の **PF** 上で正負の値の対で表現される．なお，図では絶対値の大きさが示されている．

2.2　特徴抽出器への組込み

この節では周辺特徴の抽出方法，および周辺特徴を **MFCC** と結合する方法を述べる．図 3-A は連続 **HMM** に基づく ASR システムで標準的に用いられている特徴パラメータの抽出方法を示している．特徴抽出器では，入力音声を 16KHz でサンプリングし，窓掛けした音声(25ms ハミング窓)を 10ms 毎に 512 点 **FFT** した後，パワースペクトルをメルスケール化した 24 チャンネルの **BPF** を通して得る．その後，**BPF** 出力を 12 次元のケプストラム係数(**MFCC**)に **DCT** 変換した後，対数パワー(ΔP，$\Delta\Delta$P)，および 24 次元の動的特徴(Δ_t，$\Delta_t\Delta_t$)を含む 38 次元の特徴パラメータを抽出する．

図 3-B は **TS** パターン上のさまざまな周辺特徴を抽出し，それらを統合，圧縮する手順を示す．まず，24 チャンネルの **BPF** 出力 $\mathbf{x}(t,f)$から式(2)により周辺特徴 $\mathbf{y}_m(t,f)$, m=1,2,...,8; f=1,2,...,20 を抽出する．続いて，それぞれの周辺特徴を **DCT** によりケプストラム $\mathbf{c}(m,q)$, m=1,2,...8; q=1,2,...10 に変換する．最後に 80 次元の $\mathbf{c}(m,q)$を次式を用いてKL変換し，周辺特徴ベクトル $\mathbf{z}(k)$, k=1,2,..., 24 に圧縮する．

$$z(k)=\sum_{m=1}^{8}\sum_{q=1}^{10}c(m,q)\phi_k(m,q) \qquad k=1,2,\cdots,24 \qquad (3)$$

ここで，$\boldsymbol{\varphi}_k(m,q)$は **KLT** のための k 番目の固有ベクトルである．24 次元の周辺特徴 $\mathbf{z}(k)$を，静的特徴を主として表す **MFCC**(12MFCC+Δp,$\Delta\Delta$p)と結合する(計 38 次元).

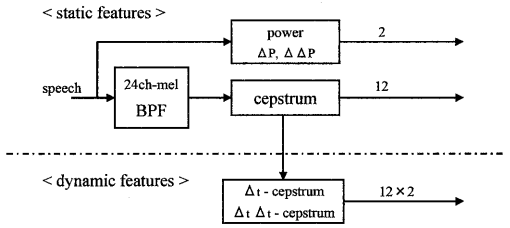

図 3-A　MFCC と動的特徴 : Baseline

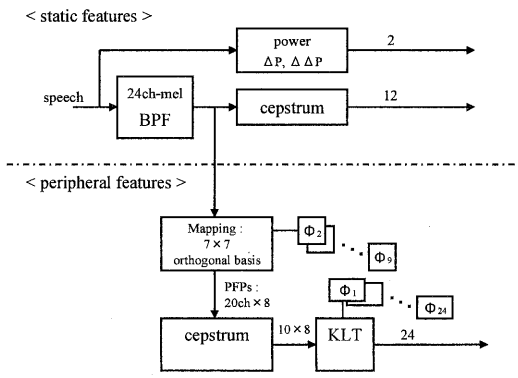

図 3-B　MFCC と周辺特徴 (TS) : PF-KL(7×7 直交基底)

2.3 評価実験

2.3.1 音声試料

以下に示す4種類のデータセットを使用した.

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)の男性話者30名(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. 男性話者10名(981単語). サンプリング周波数は24KHzから16KHzへ変換.

D3. 7×7直交基底抽出用データセット: 日本音響学会, 新聞記事読み上げ音声コーパス(ASJ-JNUS, 16KHz, 16bit)の男性話者53名(2662文)

D4. KL変換用固有ベクトル学習データセット: 直交基底抽出に用いたデータASJ-JNUSの内, 前出とは異なる男性話者53名(5569文)

2.3.2 実験仕様

5状態3ループ, 43音素の日本語monophone-**HMM**を**D1**データセットを使用して設計した. **HMM**は出力確率をガウス混合分布で表現し, 共分散行列を対角化している(混合数 = 1, 2, 4). 学習の後, **D2**のデータセットを用いて不特定話者単語認識実験を行った.

2.3.3 実験仕様

表1は**Baseline**の特徴抽出器と周辺特徴を組み込んだ特徴抽出器(**PF-KL**)の単語認識率である. 周辺特徴を用いた**PF-KL**の認識率は, 動的特徴のみを組み込んでいる**Baseline**の特徴抽出器よりも高い. 特に, **PF-KL**の混合分布1のモデルの認識率が, 同じ次元数の**Baseline**モデルの認識率よりきわめて高いことが注目される. この結果は, 周辺特徴を付加した**MFCC**がロバストな特徴パラメータセットを構成していることを推測させる.

表1 特徴抽出器の比較結果

method	word correct rate [%]		
	mix. = 1	mix. = 2	mix. = 4
Baseline	93.58	93.99	95.82
PF-KL	97.25	97.15	97.15

3. ケプストラム領域における周辺特徴

3.1 TQパターン上の周辺特徴

2節では, さまざまな幾何学的構造が**TS**パターン上で観測されることを示した. ここでは, **TS**パターンの代わりに**TQ**パターン上の周辺特徴を観測する. 図4は**TQ**パターン $\mathbf{c}(t_i,q_j)$, j=1, 2, ..., 12の7×3ブロックにおける上位9個, Φ_1, Φ_2, ..., Φ_9の直交基底を示している. 直交基底の表現方法は図1と同様である. 空間演算子の観点から, Φ_1は平滑化演算子と見なせる. Φ_2, Φ_3はそれぞれケフレンシー軸に沿った1次と2次の微分演算子(Δ_q演算子, $\Delta_q\Delta_q$演算子), Φ_4, Φ_7, Φ_9は時間軸に沿っての1次, 2次および3次の微分演算子(Δ_t演算子, $\Delta_t\Delta_t$演算子および$\Delta_t\Delta_t\Delta_t$演算子), そしてΦ_5, Φ_6およびΦ_8は**TQ**パターン上の谷や尾根を表す部分空間である.

TQパターン上の直交基底は**TS**パターン上の直交基底と異なり, ケフレンシー軸に沿った変化を示している基底(Φ_2, Φ_3)が時間軸に沿った変化を示している基底(Φ_4, Φ_7)よりも高い寄与を示した.

TQ空間演算子, すなわち写像演算子Φ_mは**TQ**パターン $\mathbf{c}(t,q)$を周辺特徴 $Y_m=y_m(t,q)$,m = 1, 2, ...,Mに写像する. 周辺特徴の構成要素 $y_m(t_i,q_j)$は7×3近傍の$\mathbf{c}(t,q)$と$\Phi_m=\phi_m(t_i,q_j)$により式(4)で計算される.

$$y_m(t,q)=\sum_{i=-3}^{3}\sum_{j=-1}^{1}c(t+i,q+j)\phi_m(i,j) \qquad (4)$$

図5に発話[kaden'tsa](カデンツァ)の上位3つ

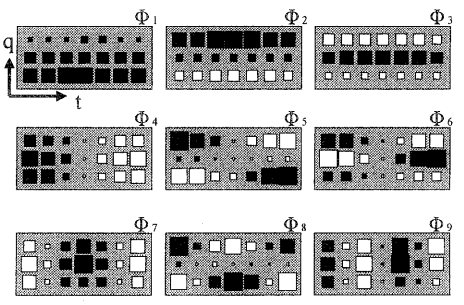

図4 TQパターン上の7×3直交基底

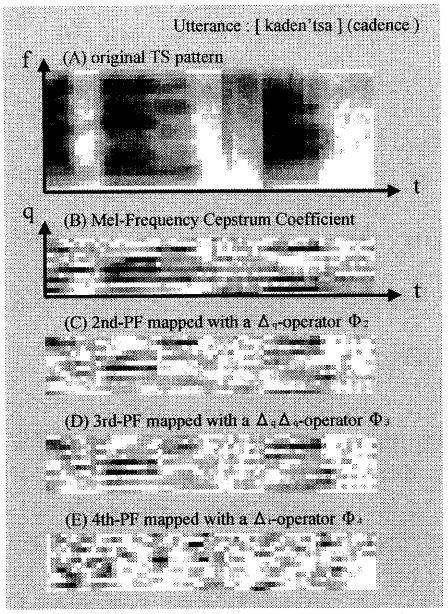

図 5 ケプストラム領域の周辺特徴

の周辺特徴の例を示す．写像演算子$\{\Phi_m\}$，m=2, 3, 4 は，単純化・対称化した後，**TQ** パターンに適用した．この場合，3 つの写像演算子はそれぞれ Δ_q 演算子，$\Delta_q\Delta_q$ 演算子および Δ_t 演算子に相当する．図において，(A)は**TS**パターン，(B)は**MFCC**，(C)，(D)および(E)はそれぞれ，Δ_q 演算子Φ_2，$\Delta_q\Delta_q$ 演算子Φ_3，Δ_t 演算子Φ_4 により抽出した周辺特徴である．周辺特徴と **MFCC** のパターンは絶対値で表示されている．図の(C)，(D)，(E)に示した周辺特徴には，写像演算子の性質を反映してケフレンシー方向，および時間方向の特徴的なパターンが現われている．

3.2 特徴抽出器への組込み

3. 1 節では，**TQ** パターン上の 7×3 直交基底を調べ，上位の主要な基底にはケフレンシー軸に沿う微分演算子(Φ_2，Φ_3)とみられる基底が存在することを見出した．一方，**MFCC** に基づく標準的な特徴抽出器はケフレンシー方向の変化を表す特徴(Δ_q パラメータ)を含んでいない．図 6 はケプストラム領域の周辺特徴抽出過程を示す．図は **MFCC** および動的特徴の抽出過程を含んでいる．

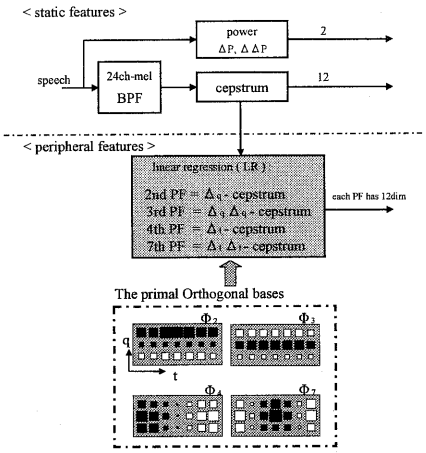

図 6 MFCC と周辺特徴 (TQ)

図では，4 つの周辺特徴(Δ_q-，$\Delta_q\Delta_q$-，Δ_t-，$\Delta_t\Delta_t$-ケプストラム)を与える空間演算子を 1×3 ブロックの演算子(Δ_q，$\Delta_q\Delta_q$)と 7×1 ブロックの演算子(Δ_t，$\Delta_t\Delta_t$)の形式に簡単化している．また，微分演算は線形回帰計算に置き換えている．周辺特徴は **MFCC** 特徴(12MFCC+ΔP+ΔΔP)と合わせて，38 次元の特徴パラメータセットとして使用される．

3.3 評価実験

3.3.1 実験仕様

実験に使用した音声試料は**2.3.1**節と同じである．また音響モデルも，**2.3.2** 節に説明したと同じものを用いて，不特定話者単語認識実験を行う．特徴パラメータはΔ_t，$\Delta_t\Delta_t$，Δ_q および$\Delta_q\Delta_q$，を **MFCC** とさまざまな組み合わせで結合して評価する．

3.3.2 実験結果

表 2 および図 7 は種々の周辺特徴を **MFCC** に付加したときの実験結果である．表 2 の特徴パラメータセットの項では，**MFCC** の表記を省略し，付加した周辺特徴のみを記した．実験結果は以下のことを示している．

- 周辺特徴Δ_q ケプストラムはΔ_t ケプストラムと同等の改善を与える．
- Δ_t ケプストラムとΔ_q ケプストラムの両方を付加したとき，最も高い認識率を得る．

なお，50 次元の特徴である "**MFCC**+Δ_t+Δ_q+$\Delta_q\Delta_q$+ΔP+$\Delta\Delta$P" パラメータセットについても実験を行ったが，"**MFCC**+Δ_t+Δ_q+ΔP+$\Delta\Delta$P" のセットよりも認識率が低かった．これは学習データセットが小さいためと考えられ，今後，再検討が必要である．

ケフレンシー軸に沿った動的特徴量を組み込んだ特徴パラメータセットは，動的特徴のみを付加した特徴パラメータと比較して顕著な改善を示した．この結果は，ケフレンシー軸に沿った動的特徴量を標準的な **MFCC** パラメータセットへ付加することの重要性を示している．

表 2 各特徴パラメータセットの認識率

特徴パラメータセット	word correct rate [%] 混合数		
	mix.=1	mix.=2	mix.=4
MFCC	81.96	81.86	81.04
Δt	91.54	92.25	92.86
Δq	93.68	94.29	93.58
Δt, ΔtΔt	93.37	93.48	94.09
Δq, ΔqΔq	93.17	94.19	94.19
Δt, ΔtΔt, ΔP, $\Delta\Delta$P	93.58	93.99	95.82
Δq, ΔqΔq, ΔP, $\Delta\Delta$P	95.21	95.21	95.31
Δt, Δq, ΔP, $\Delta\Delta$P	97.96	98.37	98.47

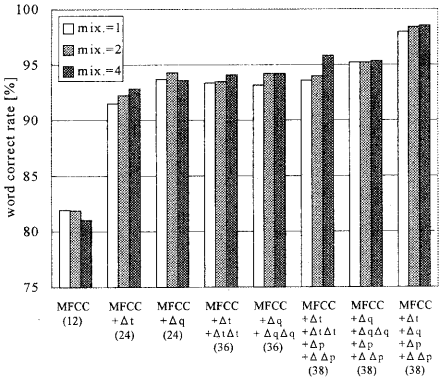

図 7 MFCC パラメータセットの比較

提案した特徴抽出器の性能が標準的な特徴抽出器よりも高い理由としては，以下の点が考えられる．Δ_t-パラメータは時間軸に沿った変動を表現するため子音の認識には有効に働くが，一方，定常的な母音等の認識には寄与しない．他方，Δ_q-パラメータはケフレンシー方向の変動を表現するため，主に母音認識に大きく寄与する．こうした仮説は今後，実験を通して実証する必要がある．

4. まとめ

さまざまな幾何学的構造を ASR システムの特徴抽出器に組み込む方法を提案した．周辺特徴に関わる特徴抽出器の設計法としては，音声の直交基底を観測し，主要な基底を単純化・対称化した形で特徴抽出器に組み込む方法を示した．周辺特徴と **MFCC** を結合した特徴パラメータセットは，**HMM** に基づく単語認識実験において標準的な特徴セット(**MFCC**+動的特徴)と比較して顕著な改善を示した．

参考文献

[1] K. Elenius and M. Blomberg, "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] S. Furui,"Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust. Speech Signal Process. ASSP-34, pp. 522-59 (1986).

[3] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. IEEE ICASSP'99, Phoenix, Vol.1, pp.421-424 (1999-3).

[4] T. Nitta, M. Takigawa, and T. Fukuda, "A Novel Feature Extraction Using Multiple Acoustic Feature Planes for HMM-based Speech Recognition " Proc. ICSLP'00, Vol. 1, pp. 385-388 (2000).