Vol.2012-MUS-94 No.21
Vol.2012-SLP-90 No.21
2012/2/4

特徴空間における長時間スペクトル変動成分の識別学習

福田 隆[†1] 市川 治[†1] 西村雅史[†1]

近年，特徴空間上の識別学習（fMMI）が注目され，多くの認識システムで効果を挙げている．通常，識別的特徴変換器には，MFCC+動的特徴やセグメント特徴量+LDAなどのスペクトル変動情報を含む特徴パラメータが入力され，それが正準化空間に写像される．特徴空間上の識別学習は近代音声認識において不可欠な要素であるが，低SNR 環境ではまだ改善の余地がある．本報告では，識別的特徴変換の枠組みに，雑音環境で頑健な性質を示す長時間スペクトル変動情報を組み込むことを提案する．提案手法は低 SNR 環境下で MFCC と動的特徴からなる標準的な特徴ベクトルセットと比較して 6.3%の性能改善を達成した．

Feature-space Discriminative Training for Long-term spectro-temporal Features

Takashi Fukuda,[†1] Osamu Ichikawa[†1]
and Masafumi Nishimura [†1]

Discriminative training of feature space using a maximum mutual information (fMMI) objective function has been shown to yield remarkable accuracy improvements. MFCC and dynamic features or LDA features are usually used for discriminative feature transform to map the features into canonicalized feature space. Discriminatively trained feature space transforms are essential for modern speech recognition but still need further improvement for low SNR conditions. In this paper, we show how noise-robust long-term temporal features can be combined with fMMI to build better discriminative models for noisy speech. The fMMI combined with long-term temporal features achieved 6.3% error reduction on average in low SNR environments when compared to the short-term temporal features alone.

†1 日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research - Tokyo, IBM Japan, Ltd.

1. はじめに

時系列データの分類に適した確率的分類器（HMM：Hidden Markov Model）が提案されて以来，音声認識では長らく HMM が音響モデルとして用いられてきた．対角共分散 HMM は無相関型のパラメータとの相性がよく，特徴量として MFCC や LPC メルケプストラム，PLP などがよく用いられている．音声認識ではスペクトルの時間変動も重要な情報であり，静的特徴の時間変化を捉えた動的特徴も特徴ベクトルの一部として広く利用されている．近年では，静的・動的特徴から成る特徴ベクトルセットをそのまま用いずに，LDA 変換や多層パーセプトロンを介してより頑健な特徴空間に写像した後，HMM で利用する例が増えてきた．中でも，特徴空間上の識別学習は特に注目されており，様々な認識タスク，雑音環境で大きな効果をあげている．特徴空間上の識別学習としては，相互情報量最大化基準に基づく fMMI や音素誤り最小化基準に基づく fMPE が有名である[1),2)]．特徴空間上の識別学習は多くの場面で認識性能を飛躍的に向上させるため，近代音声認識では欠かせない要素技術の一つとなった．識別的特徴変換[⋆1]は演算量がデコード処理と比較してとても小さいので，Embedded，Server-side を問わず様々な音声認識アプリケーションで用いられている．

識別的特徴変換過程では，まず，入力特徴ベクトルを事後確率から成る高次元のベクトル空間に変換する．その後，高次元の事後確率ベクトルを前後数フレームの同ベクトルと連結した後，識別的基準で学習された変換行列で差分ベクトルを算出する．最後に，差分ベクトルを入力特徴ベクトルに足しこむことによって特徴変換を実現する．このように現状の識別的特徴変換の枠組みは，事後確率ベクトル空間上で前後の音素環境を考慮しているものの，入力特徴ベクトルを連結するといったような，スペクトル変動情報の直接的な利用はなされていない．そのため，演算量を増やさずに識別的特徴変換処理の性能を向上させるためには，入力特徴ベクトルそのものに有益な時間変動情報を内在させることが望ましい．本報告では，識別的特徴変換器への入力特徴ベクトルに，雑音に頑健な性質を持つ長時間のスペクトル変動情報を組み込むことによって認識性能を改善させることを目指す．ここでは，組み込み型音声認識のような低リソースの音声認識環境でも好適に動作可能な方法を前提とし，5 dB や 0 dB といった極めて雑音が大きい環境でも良好に動作することを目標とする．また，識別的特徴変換と並んでよく利用されている特徴空間上の話者適応処理（fMLLR）に

⋆1 本報告では，MMI や MPE 基準などの識別的基準で学習された特徴空間変換を，識別的特徴変換（Discriminative feature transformation）と呼ぶことにする．

Vol.2012-MUS-94 No.21
Vol.2012-SLP-90 No.21
2012/2/4

ついても，長時間スペクトル変動情報が高い性能改善を与えることを示す．本報告は，筆者らの過去の研究報告について比較実験を追加し，従来手法との関係，ならびに関連する考察を拡充したものである[3]．

本報告は以下のように構成される．2. で識別的特徴変換処理（fMMI）の概要を紹介した後，3. で識別的特徴変換器への入力となる長時間スペクトル変動情報について説明する．その後，4. と 5. で 評価結果を述べ，6. で結論をまとめる．

2. 特徴空間上の識別学習

2.1 識別的特徴変換の概要

$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ を d 次元の特徴ベクトル時系列とすると，特徴ベクトル $\boldsymbol{x}_t$ は次のように変換される[1]．

$$f(\boldsymbol{x}_t) = \boldsymbol{x}_t + \boldsymbol{M}\boldsymbol{h}_t \tag{1}$$

ここで $\boldsymbol{M}$ は識別的基準で事前に推定された変換行列，$\boldsymbol{h}_t$ は $\boldsymbol{x}_t$ と同じ特徴空間で学習された GMM による N 次元の事後確率ベクトルである．事後確率ベクトル $\boldsymbol{h}_t$ は，前後のフレームを結合する形で拡張してもよい．変換行列 $\boldsymbol{M}$ の推定は MMI および MPE 基準によるものが代表的である．各ガウス分布 g に対する事後確率を $p(g|\boldsymbol{x}_t)$，$g \in \mathcal{G}$ とすると，式 (1) は次のように書き換えることができる．

$$f(\boldsymbol{x}_t) = \sum_{g \in \mathcal{G}} p(g|\boldsymbol{x}_t)(\boldsymbol{x}_t + \boldsymbol{M}_g) \tag{2}$$

ここで $\boldsymbol{M}_g$ は変換行列 $\boldsymbol{M}$ における g 列目のベクトルあり，g 番目のガウス分布に対応する．各時刻 t において $\Sigma_{g \in \mathcal{G}} p(g|\boldsymbol{x}_t) = 1$ である．式 (2) から識別的特徴変換は，識別学習により事前に推定された定数ベクトル $\boldsymbol{M}_g$ を用いて，特徴ベクトル $\boldsymbol{x}_t$ を修正する処理と見なすことができる．識別的特徴変換は，変換成分が事後確率の高い定数ベクトル $\boldsymbol{M}_g$ に依存するため，GMM によって空間分割された区分的線形変換処理とも考えることができる．識別的特徴変換の基本的な枠組みは，SPLICE（Stereo Piecewise LInear Compensation for Environments）と深い関係があり，大雑把に言えば変換行列の求め方のみが異なる[4]．SPLICE はステレオデータで変換行列を推定する一方，識別的特徴抽出ではモノラルデータから誤り基準によって変換行列を推定する．なお，後に提案された MMI-SPLICE は fMMI と数学的にほぼ等価である[5]．

2.2 Compressed fMMI

識別的特徴変換の性能は事後確率ベクトルの次元数，変換行列 $\boldsymbol{M}$ のパラメータ数などに依存する．Povey らの論文では，ML 基準で学習された音響モデル (ガウス分布数 100000) から事後確率ベクトルを生成していたが，ガウス分布数の削減は認識性能に大きく影響し，パラメータ数の観点から組み込み用途の音声認識システムでは実現が難しかった[1]．そのため，Marcheret らによって fMMI（fMPE）変換にかかるパラメータ数を効果的に削減する方法が提案された[2]．本報告では，Marcheret らの方法を Compressed fMMI と呼び，我々の音声認識システムで利用する．

Compressed fMMI は二段階（Level-1，Level-2）の処理に分解することができる[2]．まず第一段階（Level-1）では，d 次元の特徴ベクトル $\boldsymbol{x}_t$ を事後確率ベクトルに変換する．

$$o(t,g,i) = \begin{cases} \gamma_g \frac{(x_t(i) - \mu_g(i))}{\sigma_g(i)} & if \quad i \leq d \\ 5\gamma_g & if \quad i = d+1 \end{cases} \tag{3}$$

ここで t は時刻，i は特徴ベクトルの各要素を参照するインデックスである．また $\gamma_g = p(g|\boldsymbol{x}_t)$ はガウス分布 $g \in \mathcal{G}$（サイズ G）に対応する事後確率である．事後確率 γ_g を生成するガウス分布は，特徴ベクトル $\boldsymbol{x}_t$ と同じ特徴空間で学習された音響モデルについてガウス分布をクラスタリングし，分布数を削減したものである．$o(t,g,i)$ は各時刻 t で $(d+1)G$ の要素を持つ．計算の効率化のため，事後確率がしきい値 γ_{cut} を下回っている場合は $\gamma_g = 0$ とし，スパースな $o(t,g,i)$ を生成する．$o(t,g,i)$ は Level-1 変換行列 $M_1(g,i,j,k)$ によって中間ベクトルに変換される．

$$\begin{aligned} b(t,j,k) &= \sum_{g,i} M_1(g,i,j,k) o(t,g,i) \\ &= \sum_{g:\gamma_g > \gamma_{cut}} \sum_{i} M_1(g,i,j,k) o(t,g,i) \end{aligned} \tag{4}$$

ここで $j \in \{1, ..., 2J+1\}$ は Outer コンテキストと呼ばれている．

Compressed fMMI の第二段階（Level-2）では，中間ベクトル $b(t,j,k)$ を前後の時刻の同ベクトルと結合して $b(t+\tau,j,k)$，$\tau \in \{-\Lambda, \cdots, \Lambda\}$ の形で拡張する． 拡張された中間ベクトルは次式によって最終的な差分ベクトルに変換される．

$$\delta(t,k) = \sum_{j} \sum_{\tau} M_2(j,k,\tau+\Lambda+1) b(t+\tau,j,k) \tag{5}$$

ここで M_2 は Level-2 変換行列，τ は Inner コンテキストである．Level-2 の出力 $\delta(t,k)$ は，入力特徴ベクトル $x_t(k)$ を補正するベクトルとして利用する．これら Level-1，Level-2 変換行列は学習データから事前に推定しておく．学習の詳細は文献[2] を参照されたい．

Vol.2012-MUS-94 No.21
Vol.2012-SLP-90 No.21
2012/2/4

3. 長時間スペクトル変動抽出

3.1 先行研究

長時間のスペクトル変動に着目した例としては TRAP が代表的である[6]．これは，各臨界帯域に対応するスペクトルの時系列パターンを，多層パーセプトロンへの入力とすることにより，1 秒間という長い区間の変動をうまく扱っている．この方法を基礎とした研究は多数存在する[7]．一方 Poeppel らは，人間が音声言語を知覚する際に利用しているとされていた 20ms から 40ms の短時間変動に加え，150ms から 250ms 程度の比較的長い区間のスペクトル変動成分も合わせて利用している可能性を示唆した[8]．これらの先行研究は，長時間スペクトル変動成分が音声認識性能の改善に寄与する可能性を示唆しているが，長時間変動の利用を試みた従来手法は，特徴量の次元を大きくするなど構成が複雑なものが多く，必ずしも組み込み型の音声認識に適したものでなかった．そのため，これまでに我々は窓長を広げた線形回帰演算から得られる長時間の変動成分の利用を提案し，短・長時間の動的特徴の組み合わせが，音声認識性能を改善できることを示した[3),9)]．

3.2 フィルタリング処理としての動的特徴抽出

本節では，Compressed fMMI の入力となる長時間スペクトル変動抽出を変調スペクトル変換の観点から考察する．スペクトルの時間変動成分を周波数次元で表したものは変調スペクトルと呼ばれているが，見方を変えると，動的特徴抽出はスペクトルの時間軌跡に対するフィルタリング処理と見なすことができる．

$$H(z) = \sum_{k=1}^{K} \left\{ k \cdot (z^{k} - z^{-k}) \right\} \Big/ 2 \sum_{k=1}^{K} k^{2} \tag{6}$$

ここで，K は動的特徴抽出（線形回帰演算）の窓長である．一方，雑音に頑健な特徴量としてよく利用される RASTA も人間の聴覚特性を反映させた変調スペクトル強調方式と捉えることができる．これらの方法では，スペクトルを効果的に変形させることにより，音声認識にとって重要な成分を強調している．

図 1 は短時間窓（K=3，計 7 フレーム，10ms シフト）及び長時間窓（K=8，計 17 フレーム）を用いたときの線形回帰演算による変調周波数応答を示している．図から，従来の音声認識でよく用いられている短時間窓の線形回帰演算は 10Hz 付近の変調スペクトルを強調する一方，これまであまり利用されてこなかった長時間窓による線形回帰演算は 2Hz 付近の変調スペクトルに焦点を当てていることがわかる．文献[10] にもあるとおり，音声認識

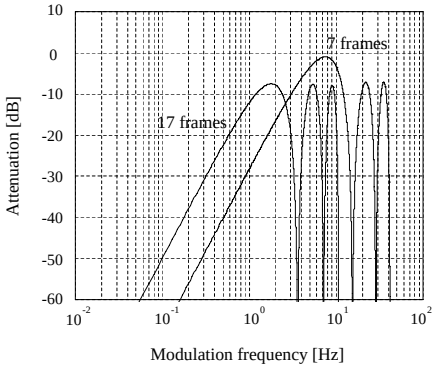


図 1 線形回帰演算の変調周波数応答

とっては 2Hz から 10Hz にわたる幅広い範囲の変調スペクトル成分が重要であるにもかかわらず，従来の短時間窓の線形回帰演算では，10Hz 付近の変調スペクトル成分のみしか利用されてこなかった．しかし，長時間窓の線形回帰演算を併用することによって，2Hz 付近の変調スペクトル成分も特徴パラメータに組み込むことができ，変調スペクトルの観点から，重要な成分をほぼ網羅することが可能になる．他方，文献[8] において，人間が音声言語を知覚する際にスペクトルの短時間変動，及び長時間変動の双方を利用していることが導かれたことも，長時間変動の重要性を示唆している．我々の提案している短・長時間窓の動的特徴はこれらの知見に基づいており，演算量を増やすことなく性能を改善できるので，低リソースの音声認識に適している．なお，長時間スペクトル変動成分は発話区間検出（VAD）の情報としても有効であり，平均音素長以上の窓幅でデルタケプストラムを抽出することによって，VAD の性能を飛躍的に向上できることがわかっている[11)]．

3.3 識別的特徴変換と長時間スペクトル変動量

Compressed fMMI の性能を左右する要素には様々なものが存在するが，事後確率を生成する GMM や，Level-1，Level-2 変換行列のコンテキストサイズは特に大きな影響を与える．また，入力特徴ベクトル $\boldsymbol{x}_t$ の性質自身も重要な要素である．

ここでは，識別的特徴変換で時間変動成分がどのように扱われるかを考察する．2 章で述べたように，Compressed fMMI は Level-1，Level-2 変換行列における Inner コンテキストと Outer コンテキストを調整することにより時間変動を考慮する．しかし，式 (3)-(5) に示すように，特徴ベクトル $\boldsymbol{x}_t$ はまず事後確率からなる高次元空間に写像され，その空間で

前後のコンテキストが考慮されている．Compressed fMMI で扱っているコンテキスト情報は標準的に数音素にわたる長いものであり，我々の実験環境では Inner コンテキスト=8 が最適であった．すなわち合計 17 フレーム分もの情報を扱っていることに相当する．しかし，これは必ずしもスペクトルの時間変動を考慮したものではなく，音声知覚処理における長時間変動の性質を反映した処理になっているとは言えない．よって，音声認識にかかる演算量を増やさないという前提で識別的特徴変換を利用するならば，長時間変動成分は入力特徴ベクトル $\boldsymbol{x}_t$ そのものに含まれることが望ましい．本報告では，短・長時間の動的特徴を fMMI への入力として用いる．すなわち $\boldsymbol{x}_t$ を次のように構成する．

$$\boldsymbol{x}_t = [\boldsymbol{S}_t, \boldsymbol{D}_{S_t}, \boldsymbol{D}_{L_t}]^T \tag{7}$$

ここで $\boldsymbol{S}_t$ は静的特徴（ケプストラム），$\boldsymbol{D}_{S_t}$ と $\boldsymbol{D}_{L_t}$ は短・長時間のデルタケプストラムを表す．

4. 長時間スペクトル変動量の評価

4.1 音声データ

IBM Research で独自に収集した車載音声認識用の英語音声を実験に用いる．ここでは，短・長時間デルタケプストラムの組み合わせそのものに効果があるをことを実証した後，先に特徴空間上の適応技術（fMLLR）との併用の効果を示す．その後，5 章で fMMI との組み合わせを検証する．認識タスクは，住所入力，連続数字，オーディオ制御コマンド，POI，フリーフォームコマンド（自然言語に近いコマンドフレーズ群），およびその他孤立単語である．走行速度はそれぞれ 0，30，60 mph であり，サンプリング周波数は 16kHz である．学習データは 796.1 時間，評価データは 133 名の話者による 38,905 発話，約 39.2 時間分のデータである．

4.2 特徴抽出

フレームサイズとシフト幅はそれぞれ 25ms および 15ms である．入力音声はフレーム毎に $1 - 0.97z^{-1}$ の高域強調処理，ハミング窓掛け処理を行った後，FFT を通じて線形領域のパワースペクトル系列を得る．そして，パワースペクトル時系列に対して，24 チャネルのメルフィルタバンク分析を行った後，c0 を含む 13 次元のケプストラムを抽出する．その後ケプストラム時系列に対して 3 種類の動的特徴を抽出し，比較実験を行う．

(A) LDA: 13 次元のケプストラムについて前後 4 フレーム（合計 9 フレーム）を結合し，117 次元のスーパーベクトルを構成した後，LDA 変換によって 40 次元のベクトルに圧縮する．LDA 変換行列は，音響モデルの学習と同じデータセットで推定する．

(B) Short delta & ddelta: 標準的な 1 次と 2 次のデルタケプストラム（ShortΔ，ShortΔΔ）を線形回帰演算によって抽出する．ShortΔ と ShortΔΔ の窓長はそれぞれ 5 フレーム (75 ms) と 9 フレーム (135 ms) である．特徴ベクトルの次元数は 13 次元の MFCC，ShortΔ，ShortΔΔ から成る 39 次元で，パワー項は使用しない．

(C) Short & Long delta: ShortΔ ケプストラムは前項のように抽出する．そして，(B) の特徴パラメータセットから ΔΔ ケプストラムを取り除き，代わりに長時間のデルタケプストラム（LongΔ）を特徴パラメータセットに加える．ただし，LongΔ は c4 から c12 についてのみ適用し，c0 から c3 については，ケプストラムの時間変動が揺るかやなことを考慮して，通常の ShortΔΔ ケプストラムを利用する．先行研究から LongΔ の窓幅としては 170ms 程度（10ms シフトで 17 フレーム）が最適であることがわかっているので，ここでは LongΔ は 11 フレーム（165ms）から抽出することとした．最終的な特徴ベクトルは，MFCC，ShortΔ，ShortΔΔ（c0-c3），LongΔ（c4-c12）の 39 次元である．

上述した 3 つの変換行列（LDA，Short delta & ddelta，Short & Long delta）は，semi-tied covariance (STC) 行列を用いて近似的に対角化する[12]．STC の推定は，音響モデルの学習時（ML 基準）に実行する．MFCC は動的特徴抽出に先立って発話単位で CMN 処理を行っている．

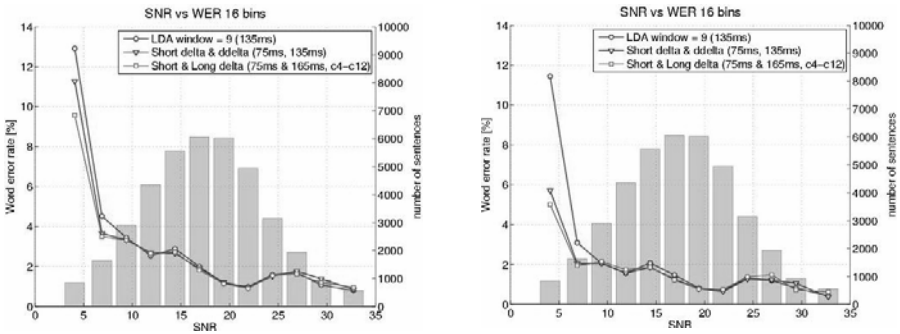
4.3 音響モデル

音響モデルには状態スキップなしの 3 状態 Left-to-right Quinphone HMM を用いた．HMM の各状態は決定木によって構造化され，各ノードに対する Question によって目的のリーフが選択される．決定木の各リーフは音響モデルの状態に相当する．状態数（リーフサイズ）は 830 で，ガウス分布の総数は 10000 である．デコーダには有限状態トランスデューサを用いる．

4.4 実験結果と考察

図 2-(a) に実験結果を示す．ここでは各種認識タスクの発話をマージした後，SNR でソートした場合の性能を図示している．折れ線グラフは SNR に対する認識性能の推移を表し，棒グラフは各 SNR ビンでのテスト発話数を示す．車載音声認識では低 SNR 環境での性能改善が重要な研究課題であるので，ここでは 10 dB 以下の高騒音環境下での認識性能の改善に注目する．図 2-(a) を見ると，LongΔ は高 SNR での性能を維持しつつ低 SNR での性能を改善し，LDA や Short delta & ddelta よりも高い認識性能を示していることがわかる．

Vol.2012-MUS-94 No.21
Vol.2012-SLP-90 No.21
2012/2/4



(a) w/o fMLLR　　(b) w/ fMLLR

図 2　長時間スペクトル変動の効果

次に fMLLR との組み合わせを考察する．今回は組み込み型音声認識を対象としているが，組み込み型音声認識でも fMLLR が使われることもあるので，長時間スペクトル特徴との組み合わせについて検討が必要である．ここでは，話者単位で fMLLR を適用した場合の性能を比較する（図 2-(b)）．図 2(a)，(b) を比較すると，fMLLR は入力特徴量にかかわらず，全ての SNR ビンで性能を改善している．その中でも，LongΔ を入力とした fMLLR が 10 dB 以下で高い改善を示している．fMLLR を適用すると，ML 空間では有意だった性能差が打ち消されてしまうことがよくある一方，LongΔ を入力とする fMLLR は有意差を維持し，着実な改善を果たしている．fMLLR はランタイムでは単純な線形変換になるので，事前に変換行列を推定するような利用法であれば演算量を低く抑えることができる．fMMI と同様に fMLLR も通常は特徴ベクトル間のコンテキストを考慮せず，注目フレームのみを処理の対象としている．逆に言えば，前後のフレームを連結して fMLLR に入力するといった直接的な利用法ができなかった．そのため，fMLLR で前後のコンテキストを考慮する Context filtering という手法が Huang らによって提案されたが，演算量が多いので低リソース音声認識向きとは言えない[13]．したがって，fMLLR においても入力特徴ベクトルに長時間変動成分を組み込んでおくことが重要であると言える．

4.5　LDA による長時間変動表現

ここでは LDA のコンテキストサイズを広げた場合の性能の推移を見てみる．実験結果を図 3 に示す．音響モデルは LDA の各窓長ごとに学習している．図 3 に示すように，N=13,

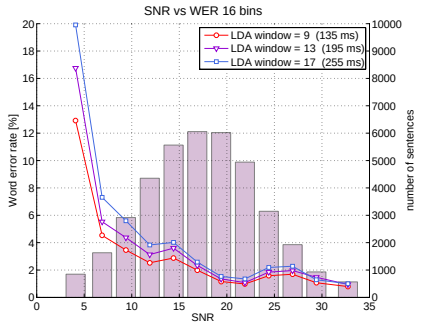


図 3　LDA のコンテキストサイズによる比較

N =17 の LDA は性能が大幅に低下していることがわかる．本実験では，LDA のクラス定義として HMM の状態単位を用いたが，13 フレーム（195ms）や 17 フレーム（255ms）というのは，平均音素長を大きく超えているので，長時間の変動を捉えるという観点からは適切ではないかもしれない．言い換えれば，大きな窓長の LDA はクラス単位を超えて複数の音素にまたがっているので，クラス内分散とクラス間分散の比が小さくなりすぎている可能性がある．一般に，LDA は長い区間にわたる情報を表現するのに適していないという指摘があり，いつくかの解決策が提案されている[14]．しかしながら，これらのアルゴリズムはシンプルな LDA や線形回帰演算による動的特徴抽出と比べて演算量が多いので，LDA 関連の技術は本稿の比較対象としないこととする．

5.　識別的特徴変換の評価

5.1　実験の概要

本章では，fMMI の入力特徴ベクトルとして，短・長時間の動的特徴を組み合わたパラメータセットを検証する．ここでは，前節と異なる環境で収集されてた英語音声コーパスを実験に用いた．学習データは 170 時間で，テストデータは 10,755 コマンドフレーズである．サンプリング周波数は 16kHz である．実験では，4.2 章で述べたデルタケプストラムに関する二つのパラメータセット（Short delta & ddelta と Short & Long delta）を比較した．フロントエンドの仕様および HMM のトポロジーは前節と同様である．本実験にかかる音響モデルのパラメータは，リーフサイズが 1000，ガウス分布数が 16,000 である．

Vol.2012-MUS-94 No.21
Vol.2012-SLP-90 No.21
2012/2/4

表 1 fMMI 変換と長時間スペクトル変動の組み合わせ効果

	Word error rate [%]							
	w/o fMMI transform				w/ fMMI transform			
SNR bin number	bin 0	bin 1	bin 2	bin 3	bin 0	bin 1	bin 2	bin 3
SNR center bin [dB]	1.0	9.5	18.0	26.5	1.0	9.5	18.0	26.5
Short delta & ddelta	10.6	5.1	2.5	1.7	7.9	3.2	1.5	1.0
Short delta & Long delta	9.8	4.8	2.2	1.3	7.4	3.0	1.4	1.0
Relative gains [%]	7.5	5.9	12.0	23.5	6.3	6.3	6.7	0.0

5.2 実験結果

表 1 に実験結果を示す．テストデータは SNR に応じて 4 つのビンに分割している．fMMI の Inner と Outer コンテキストは，それぞれ 8 と 4 とした．また，Level-1 変換行列に入力する事後確率ベクトルのサイズは $N = 512$ である．事後確率ベクトルを生成するガウス分布セットは，fMMI 変換前のケプストラムドメインで学習した音響モデルをクラスタリングして 512 個の分布に圧縮したものである．

まず，fMMI の ON/OFF で比較すると，表に示すように両特徴ベクトルセットは，fMMI を適用しないシステムと比較して性能を改善した．入力特徴ベクトルについて，Short & Long delta は Short delta & ddelta と比較して fMMI の性能を改善し，特に低 SNR 環境での改善が大きかった．提案手法は，bin 0 で相対的に 6.3% の改善を得た．Inner および Outer コンテキストをそれぞれ 4，2 として別途実施した実験においても bin0 で WER=7.7%，bin1 で WER=3.1%を確認した．今回行った実験では，Level-1 と Level-2 変換行列のコンテキストサイズは大きい方が性能が良かったが，この傾向は他のコーパスでも同様に見られる．fMMI は事後確率空間でのコンテキストを考慮するものであるが，その空間においても複数音素にまたがる長いコンテキストを併用する方が良いという結果であった．

6. おわりに

本報告では，識別的特徴変換の入力として長時間スペクトル変動を利用することを提案した．識別的特徴変換は環境を問わず性能を大きく改善できるため，近代音声認識では不可欠な技術である．fMMI および fMLLR 処理そのものは基本的にスペクトルの変動成分，すなわち特徴ベクトル間のコンテキストを考慮しない．そのため，入力特徴量自身に長時間変動情報を組み込んでおくことが望ましい．英語音声を用いた評価実験において，長時間変動情報を組み込んだパラメータは fMMI，fMLLR の性能を向上させることができた．今後は識別的特徴変換と学習データの関係について検討を進めていきたい．

参考文献

1) D.Povey, B.Kingsbury, L.Mangu, G.Saon, H.Soltau, G.Zweig, “fMPE: Discriminatively trained features for speech recognition,” *Proc. ICASSP*, pp. 961-964, 2005.
2) E.Marcheret et al., “Optimal quantization and bit allocation for compressing large discriminative feature space transforms,” *Proc. ASRU*, pp. 64-69, 2009.
3) T. Fukuda, O. Ichikawa, and M. Nishimura, “Combining feature space discriminative training with long-term spectro-temporal features for noise-robust speech recognition,” *Proc. Interspeech*, pp. 229-232, 2011.
4) J.Droppo, L.Deng, and A.Acero, “Evaluation of SPLICE on the Aurora 2 and 3 tasks,” *Proc. ICSLP*, pp. 29-32, 2002.
5) J. Droppo and A. Acero, “Maximum mutual information SPLICE transform for seen and unseen conditions,” *Proc. Interspeech*, pp. 989-992, 2005
6) H.Hermansky and S.Sharma, “TRAPS - Classifiers of temporal patterns,” *Proc. ICASSP*, Vol. I, pp. 280-292, 1999.
7) F.Grezl, M.Karafiat, S.Kontar, and J.Cernocky, “Probabilistic and bottle-neck features for LVCSR of meetings,” *Proc. ICASSP*, pp. 757-760, 2007.
8) D. Poeppel, “The analysis of speech in different temporal integration windows: cerebral lateralization as asymmetric sampling in time,” *Speech Communication*, Vol. 41, pp. 245-255, 2003.
9) T.Fukuda, O.Ichikawa, and M.Nishimura, “Short- and long-term dynamic features for robust speech recognition,” *Proc. Interspeech*, pp. 2262-2265, 2008.
10) N.Kanedera, T.Arai, H.Hermansky, and M. Pavel, “On the relative importance of various components of the modulation spectrum for automatic speech recognition,” *Speech Communication*, Vol. 28, No. 1, pp. 43-55, 1999.
11) T.Fukuda, O.Ichikawa, and M.Nishimura, “Long-term spectro-temporal and static harmonic features for voice activity detection,” *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 5, pp. 834-844, 2010.
12) M.J.F.Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 3, pp. 272-281, 1999.
13) J.Huang, K.Visweswariah, P.Olsen, V.Goel, “Front-end feature transforms with context filtering for speaker adaptation,” *Proc. ICASSP*, pp. 4440-4443, 2011.
14) M.Sugiyama and S.Roweis, “Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis,” *Journal of Machine Learning Research*, Vol. 8, pp. 1027-1061, 2007.