情報処理学会論文誌 Vol.58 No.1 288–296 (Jan. 2017)

正則化処理を用いた特徴空間識別学習の高精度化と音響環境適応

福田 隆[1,a)] 市川 治[1,b)] 立花 隆輝[1,c)]

受付日 2016年4月20日, 採録日 2016年10月4日

概要：GMM/HMM システムにおいては，音響モデルを新たな音響環境に適応するため，MAP 法などのモデル空間適応がさかんに利用されている．一方で，近年の音声認識システムは特徴空間上の識別学習を行っているものの，識別的特徴変換行列が環境適応の対象になることは少ない．しかし，識別的特徴変換行列は大規模データから統計的に学習されるため，音響モデルと並んで識別的変換行列も対象ドメインに適応することが望ましい．本論文では，はじめに，特徴空間上の識別的適応において，小規模な適応データで良好に動作する方法を提案する．提案法では，適応学習時の間接的微分演算に正則化項を導入するとともに，音響モデルの更新に MAP 基準を用いる．そして，実証実験において提案法はモデル空間適応の MAP 法と比較して 4.4%の相対的改善を達成したことを示す．次に，本論文では通常の特徴空間識別学習の前後に事前・事後学習を導入することによって，少量の書き起こしデータのみが利用可能な状況において，音響モデルを高精度化する方法を提案する．提案法は大規模データを用いた検証実験において，平均して相対的に 1.5%以上の改善が得られたことを示す．

キーワード：大語彙連続音声認識，識別学習，正則化，適応，書き起こしデータ

Improving Feature-space Discriminative Training and Adaptation Using Regularization Process

Takashi Fukuda[1,a)] Osamu Ichikawa[1,b)] Ryuki Tachibana[1,c)]

Received: April 20, 2016, Accepted: October 4, 2016

Abstract: In GMM/HMM systems, model-space adaptation techniques such as MAP are often used for porting old acoustic models into new domains. Although modern ASR systems leverage feature-space discriminative training, adapting feature space transforms has not been much investigated. However, because the feature space transforms are statistically estimated with a large corpus, the transforms should also be adapted. This paper improves the feature-space discriminative adaptation by introducing a regularization term for an indirect differential computation of the fMPE objective function, and also by updating the acoustic models with MAP instead of ML criterion during the fMPE adaptation. The proposed method performed favorably for the adaptation conditions with small amounts of adaptation data, and yielded 4.4% relative improvement in comparison with MAP-adapted system without using the fMPE adaptation. Next, we introduce the regularization process to a standard feature-space discriminative training for situations when only limited amount of training data is available. The proposed method consists of pre- and post-training steps, and yielded more than 1.5% relative for various system configurations.

Keywords: LVCSR, discriminative training, regularization, adaptation, manual transcription

1. はじめに

音声認識におけるディープラーニングの利用は近年非常に活発になってきており [1]，音響モデルとしての利用に限定した場合，大きく (1) DNN/HMM ハイブリッドシ

[1] 日本アイ・ビー・エム株式会社
IBM Research, Chuo, Tokyo 103–8510, Japan
[a)] fukuda1@jp.ibm.com
[b)] ichikaw@jp.ibm.com
[c)] ryuki@jp.ibm.com

ステム，(2) DNN ボトルネック特徴システムの 2 つに分類することができる．(1) は DNN の出力層を音素環境決定木に結び付け，ソフトマックス関数を経由して DNN から直接入力フレームの事後確率を推定しようとするものである．一方，(2) の DNN ボトルネック特徴システムは，ネットワークの（中間層の）出力が音響的要因にともなう変形の少ない良質な特徴空間を構成していると仮定し，それをそのまま，もしくはさらに特徴空間変換を行った後，GMM/HMM システムへの入力として使う枠組みである．本論文は (2) のシステムに関するものである．

GMM/HMM システムの枠組みでは，識別学習が高精度な音響モデルの構築に利用され，音声認識の実用化に大きく貢献してきた．識別学習は GMM に基づく音声認識で中心的に用いられ，MFCC や PLP に加え，DNN から生成されるボトルネック特徴量などを入力として受付ける [2], [3]．近年では，モデル空間の識別学習だけでなく，識別的な基準で推定した特徴変換行列によって，特徴量をより正準化された特徴空間に写像する処理を含むケースが増えており [4], [5]，特徴空間，モデル空間にかかわらず千時間以上の大規模なデータから統計的に学習することが増えてきた．

しかしながら，音響モデル学習の全工程に大規模データを用いるとなると，モデルの構築時間に加えて，音声データの収集や，そのデータに対する書き起こしの整備などの事前作業が必要になり，全体としてモデルの完成までに多大なコストを要する．理想的にはアプリケーションごとに最尤推定から全工程を行うことが望ましいが，音声認識の利用場面が増えてきた現在では，アプリケーションごとにすべてを再学習するのは費用対効果の観点から現実的に困難である．そのため，新たな音響ドメインのモデル構築にあたって，新規に全学習を行うのではなく，既存の音響モデルについて適応処理を行うことによって，新ドメインに対応した音響モデルを用意することも少なくない．適応処理には MLLR や MAP 適応などのモデル適応法がよく用いられている [6], [7]．同様に，MPE や MMI といった識別的な基準で学習された音響モデルに対する適応処理（MPE-MAP，MMI-MAP）も広く検討されている [8]．これらの方法は話者適応のために用られることが多いが，音響モデルのパラメータを新しい音響ドメインに適合させるような場合にも広く利用される [9]．

一方，特徴空間上の識別学習を考慮すると，識別的特徴変換行列は MPE や MMI 基準によって大量のデータから統計的に学習されるため，変換精度は学習データの音響特性に依存すると考えられる．したがって，本来であれば音響モデルと並んで識別的変換行列も対象ドメインに適応することが望ましい[*1]．この問題に対して，特徴空間上における識別学習法の拡張として，変換行列を対象ドメインに適合させる識別的特徴空間適応が提案されている [10]．この方法は別環境で推定された識別的変換行列を初期値として，適応データを用いた追加学習によって対象環境に合わせこもうとするものである．文献では，Broadcast News タスクから会議タスクへの適応実験において性能改善があることが明らかにされたが，十分な改善効果を得るためには数百時間以上の大規模な適応データを必要としていた．

本論文では，まずはじめに，特徴空間上の識別的適応において，より小規模な適応データで良好に動作する方法を提案する [11]．提案法は，適応学習時の間接的微分演算に正則化項を導入するとともに，音響モデルの更新に MAP 基準を用いることで実現する．そして，実証実験において提案法はモデル空間適応の MAP 法と比較して 4.4%の改善を達成したことを示す．

次に本論文では，正則化処理を通常の特徴空間識別学習に導入し，モデル構築を高精度化する方法を提案する．一般に，音響モデルの構築には大量の音声データとともに，発話内容を表す正確な書き起こしデータが必要であるが，各発話に対応する人手による書き起こしの付与は多大なコストを要する．したがって，現在では学習データの一部についてのみ人手によって正確な書き起こしを付与し，残りについては既存の音声認識システムによる擬似的な書き起こしを与える準教師付き学習が利用されている [12]．しかし，擬似書き起こしデータには音声認識上の誤りが必ず含まれるので，擬似書き起こしデータを音響モデルの構築に利用すると，特に識別学習の際に悪影響が出てしまう [13]．

そこで本論文では，擬似書き起こし精度の影響を軽減するため，書き起こしデータとして小規模なセットのみが利用可能な状況を想定のもと，特徴空間上の識別学習を高精度化する方法を提案する．提案法は，標準的な識別学習の前後に正則化処理を組み込んだ事前・事後学習を導入することで実現する．提案法は特徴空間識別学習の段階で性能改善があり，また後続で行われるモデル空間の識別学習後にも改善効果が維持されることを示す．

本論文は次のように構成される．2 章では，従来手法である特徴空間上の識別学習についてまとめる．続いて 3 章では，識別的特徴変換行列の適応処理に関する提案法を述べ，4 章でその効果を検証する．次に 5 章で，少量の書き起こし学習データを活用する識別学習について述べ，最後に 6 章で結論を述べる．

2. 特徴空間上の識別学習

本章では，従来の識別学習過程のうち，提案法に関連する部分にのみ焦点を当てて説明する．識別的基準としては MPE を採用する．

$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ を d 次元の特徴ベクトル時系列とすると，時刻 t における特徴ベクトル $\boldsymbol{x}_t$ は識別的

[*1] 識別的特徴空間変換は識別的基準に応じて fMPE 変換や fMMI 変換と称されることもある．

変換行列 M を用いて次のように写像される [4].

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{M}\boldsymbol{h}_t \tag{1}$$

ここで $\boldsymbol{h}_t$ 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 M は目的関数 $\mathcal{F}_{MPE}$ を最大化するように推定する．

$$M^* = \arg\max_{M} \mathcal{F}_{MPE}(y, \lambda) \tag{2}$$

ここで λ は新しいモデルのパラメータ集合である．行列 M の各要素は一次の最急降下法により更新される．

$$M_{ij} := M_{ij} + \nu_{ij}\frac{\partial \mathcal{F}}{\partial M_{ij}} \tag{3}$$

ここで ν_{ij} は行列 M の各要素に対応した学習係数である．そして i と j はそれぞれ入出力の特徴ベクトルと事後確率ベクトルの次元を表すインデックスである．上式の微分項 $\frac{\partial \mathcal{F}}{\partial M_{ij}}$ は，次式のように直接的微分と間接的微分の 2 つの項に分解することができる．

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial M_{ij}} &= \sum_{t=1}^{T}\frac{\partial \mathcal{F}}{\partial y_{ti}}h_{tj} \\ &= \sum_{t=1}^{T}\left[\frac{\partial \mathcal{F}}{\partial y_{ti}}^{direct} + \frac{\partial \mathcal{F}}{\partial y_{ti}}^{indirect}\right]h_{tj}\end{aligned} \tag{4}$$

このうち直接的微分は y_{ti} の変化にともなう目的関数の変化に影響しており，一方，間接的微分は y_{ti} の変化に間接的に影響する音響モデル（ガウス分布の平均，分散）の変動に関係する．式 (4) において，間接的微分項 $\frac{\partial \mathcal{F}}{\partial y_{ti}}^{indirect}$ はさらに次のように詳細化される．

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial y_{ti}}^{indirect} &= \sum_{s=1}^{S}\sum_{m=1}^{M_s}\xi(s,m) \\ \xi(s,m) &= \frac{\gamma_{sm}(t)}{\gamma_{sm}}\left\{\frac{\partial \mathcal{F}}{\partial \mu_{smi}} + 2\frac{\partial \mathcal{F}}{\partial \sigma^2_{smi}}(y_{ti} - \mu_{smi})\right\}\end{aligned} \tag{5}$$

ここで，S，M_s は最尤推定で学習された音響モデルの状態数および各状態に対応するガウス分布数であり，μ_{smi}，σ^2_{smi} は音響モデルの平均と分散を表す．そして $\gamma_{sm}(t)$ は標準的な前向き・後ろ向きアルゴリズムの際に用いられる占有確率である．γ_{sm} は全学習データから求まる占有確率を示す．特徴空間の識別学習時には，変換行列 M だけでなく，音響モデルも識別的特徴を用いて ML 基準で随時更新される．詳細は Povey らの論文を参照されたい [4].

3. 識別的特徴空間適応の正則化

3.1 過学習

識別的基準で推定された特徴変換行列 M を新しい音響環境に適応する最も単純な方法は，適応データを用いて元の変換行列に対して追加学習を行うことである．**表 1** に自動車用音声認識を対象とした音響環境の追加学習の効果を示す．ここで元環境の音響モデルは最尤基準と識別的基準で学習された汎用大語彙連続音声認識（以下，汎用 LVCSR）のモデルである．元環境の音響モデルは前後 2 音素（quinphone）の音素環境を考慮した総数 20K の状態共有 HMM であり，ガウス分布数は 400K である．汎用 LVCSR の学習データは千時間以上である．一方，適応先である自動車内で収録された適応データは 60 時間である．この実験における汎用 LVCSR と自動車用音声認識の語彙セットおよび言語モデルは同じものを利用しており，音響環境のみが異なる状況を想定している．

表 1 fMPE 追加学習の効果
Table 1 Performance on additionally-trained fMPE.

Transforms	%WER
Original fMPE	17.66
Additionally-trained fMPE	23.20

表 1 に示すように，識別的変換行列に対する単純な追加学習は逆に性能を劣化させていることが分かる．この実験での環境適応データは 60 時間と少量である一方，元環境の音響モデルは千時間以上のデータから学習された比較的リッチな音響モデルであるため，環境適応を目的とした追加識別学習時に過学習が起きてしまっている．式 (5) を見ると，識別学習の目的関数は音響モデルの全状態・ガウス分布からの統計量に基づいていることが分かる．すなわち，リッチな音響モデルに対して少量の適応データではすべてのガウス分布を網羅するのに不十分であり，ある特定の状態から得られる統計量の信頼性が低いため過学習が発生しているといえる．

3.2 正則化項の導入

提案法では，まず小規模データに対する過学習を避けるため，式 (5) の間接的微分演算に正則化処理を導入する．適応データによって得られる各 HMM 状態からの統計量の信頼性にはばらつきがあるので，微分演算に状態あたりのデータサイズに応じたペナルティを付与することを目的としている．すなわち，次式のように信頼性が高いと思われる統計量のみを利用するようにペナルティを与える．

$$\frac{\partial \mathcal{F}}{\partial y_t}^{indirect} = \sum_{s=1}^{S}\sum_{m=1}^{M_s}\rho(\gamma_{sm}^{num}, \gamma_{sm}^{den})\xi(s,m) \tag{6}$$

ここで $\rho(\gamma_{sm}^{num}, \gamma_{sm}^{den})$ は正則化項であり，γ_{sm}^{num} と γ_{sm}^{den} は識別学習に用いられるラティスのカウント値である．この γ_{sm}^{num} と γ_{sm}^{den} は本来式 (5) の $\frac{\partial \mathcal{F}}{\partial \mu_{smi}}$ と $\frac{\partial \mathcal{F}}{\partial \sigma^2_{smi}}$ の計算に用いられるものである．詳細は文献 [4] を参照されたい．提案法において，式 (6) の正則化項 ρ には 0〜1 の間の値をとる任意の関数が利用できる．ただし，信頼性が低いと思われる統計量にペナルティを付与するという目的においては，γ_{sm}^{num}，γ_{sm}^{den} が小さいときに 0 に近い値を返すような関数であることが前提である．本論文ではシグモイド関数に基

づく正則化項を利用する．

$$\rho(\gamma_{sm}^{num},\gamma_{sm}^{den}) = \frac{1}{1+exp(-\alpha\beta_{sm}+\epsilon)} \quad (7)$$

$$\beta_{sm} = min(\gamma_{sm}^{num},\gamma_{sm}^{den}) \quad (8)$$

ここで α はゲイン調整係数，ϵ はシフト係数である．Povey らによれば，間接的微分が認識誤り削減において大きな役割を果たし，直接的微分のみを利用すると改善効果の大部分が失われたと報告している [4]．そこで提案法では，間接的微分についてのみ正則化処理を導入し，直接的微分は Povey らの定式をそのまま利用することとした．正則化処理はすべての HMM 状態に対して正則化関数 ρ が 1 の値をとる場合，通常の識別的適応を実施していることに相当する．この正則化関数は，ラティスのサイズに応じた統計量の信頼度（評価値）を付与するものであるともいえる．提案法の式 (7) において，ϵ は閾値処理の基準点を決定する重要なパラメータであるので，グリッドサーチやランダムサーチなどを活用して適切なオペレーティングポイントを探す必要がある．一方，α はシグモイド曲線の傾斜に関するパラメータであるので性能に対する感度が低く，ϵ と比較するとチューニングは容易である．

3.3 音響モデルの更新

本論文では次に音響モデルの更新について述べる．特徴空間上の識別学習では，まず特徴変換行列を推定し，その後，変換行列で写像した識別的特徴を用いて音響モデルを更新する．変換行列と音響モデルの更新は目的関数の値が収束するまで順次繰り返される．通常，音響モデルの更新には ML 基準が用いられる．しかし，正則化処理の有無に関係なく，適応後の識別的特徴は元環境の識別的特徴と比較して大きく変形するため，提案法では音響モデルの更新に MAP 基準を導入し，緩やかに更新するように調整する．更新式は次のとおりである．

$$\hat{\mu}_{sm} = \frac{\theta_{sm}^{num}(\mathcal{Y})+\tau\mu_{sm}}{\gamma_{sm}^{num}+\tau} \quad (9)$$

$$\hat{\sigma}_{sm}^2 = \frac{\theta_{sm}^{num}(\mathcal{Y}^2)+\tau(\mu_{sm}^2+\sigma_{sm}^2)}{\gamma_{sm}^{num}+\tau} - \hat{\mu}_{sm}^2 \quad (10)$$

更新式は対角共分散 HMM の利用を前提としている．ここで $\theta_{sm}^{num}(\mathcal{Y})$ は，識別的特徴変換後の特徴ベクトル時系列の集合 $\boldsymbol{Y}=[\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_n,\cdots,\boldsymbol{Y}_N]$ に対する，正解仮説の事後確率による重み付け総和を返す関数である．N は学習データに含まれる発話の総和を表す．そして，$\boldsymbol{Y}_n=[\boldsymbol{y}_{n1},\boldsymbol{y}_{n2},\cdots,\boldsymbol{y}_{nt},\cdots,\boldsymbol{y}_{nT}]$ であり，特徴変換後の d 次元特徴ベクトル時系列からなる．式 (10) において，τ は元のモデルからの更新度合いを調整するパラメータであり，適応前と適応後の音響環境が異なれば異なるほど適切なチューニングが必要である．これについてもグリッドサーチやランダムサーチによって値を決定することが望ましい．

3.4 平滑化パラメータ

文献 [10] では，次式のように間接的微分演算にも平滑化パラメータ τ を導入している．本提案手法にもこのパラメータ τ を導入し，正則化処理と併用する．

$$\xi(s,m)=\frac{\gamma_{sm}(t)}{\gamma_{sm}+\tau}\left\{\frac{\partial\mathcal{F}}{\partial\mu_{smi}}+2\frac{\partial\mathcal{F}}{\partial\sigma_{smi}^2}(y_{ti}-\mu_{smi})\right\} \quad (11)$$

4. 音響環境適応実験

識別学習を導入した不特定話者汎用 LVCSR システムを用いて実験を行った．データセットは独自に収集した LVCSR 用英語音声と自動車内発話データである．汎用 LVCSR データは収録環境が多岐にわたり，環境としてはオフィス，家，レストラン，駅などが含まれ，そのほとんどは自動車とは関係のない場所で発声された自由発話音声となっている．ただし，各発話に収録場所を示すタグは付与されていない．話者あたりのデータサイズは数秒から数時間と幅広い構成となっており，合計千時間以上の発話データから音響モデルを構築する．本章では，ネットワーク非依存の音響特徴（MFCC）を用いて提案手法の効果を検証する．

4.1 音響特徴量

学習・評価データはともにサンプリング周波数 16 kHz で収録されており，音声データから 256 点 FFT および 24 チャネルのバンドパスフィルタ処理を経由して，13 次元の MFCC 特徴量を抽出する．MFCC は発話区間のみで求めたケプストラム平均から発話単位の CMN を行っている．その後，MFCC 特徴量について前後 4 フレームを結合して合計 9 フレームの音声セグメントを構成し，LDA（Linear Discriminant Analysis）変換を経由して 24 次元の特徴量（LDA 特徴量）に圧縮する．LDA 特徴量は STC（Semi-Tied Covariance）によって近似的に対角化している [16]．STC 処理後の特徴量はさらに 2 章で述べた識別的変換行列を用いて，より正準化された空間に写像する．識別的特徴空間変換（fMPE 処理）のための事後確率ベクトル $\boldsymbol{h}_t$ は 512 次元であり，Inner と Outer コンテキストはそれぞれ 8 と 4 とした[*2]．

4.2 音響モデル

音響モデルは話者独立であり，実験には HMM/GMM システムを用いている．音響モデルは ML 基準でモデルを学習した後，MPE 基準を用いた特徴空間およびモデル空間上の識別学習によってモデルの識別性能を高めている．音素コンテキストは前後 2 フレームを考慮した quinphone

[*2] fMPE 変換はより高精度な変換を実現するため，中間的な特徴空間を介し，2 段階の変換処理で実現されることがある．本実験ではその実装を採用しており，1 段目と 2 段目の特徴空間変換の際に用いられるのが Inner と Outer コンテキストである．詳しくは，たとえば文献 [18] を参照されたい．

表 2 適応実験の比較
Table 2 Comparisons of adaptation methods.

System	Model Update in f-DT stage	%WER	%Relative (Baseline)	%Relative (Model-MAP)
Baseline	–	17.66	–	–
Model-MAP	–	16.04	9.17	–
fMPE-Adaptation	ML	23.20	−31.37	−44.64
fMPE-Adaptation (fMPE-MAP [10])	MAP	18.96	−7.36	−18.20
Regularized fMPE-Adaptation	ML	17.31	1.98	−7.92
Regularized fMPE-Adaptation	MAP	15.43	12.63	3.80
Regularized fMPE-Adaptation + Model-space DT	MAP	15.33	13.19	4.43 (t-test)

であり，状態スキップなしの 3 状態 Left-to-right HMM を用いた．HMM の各状態は決定木によって構造化され，各ノードに対する質問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 20K で，ガウス分布の総数は 400K である．デコーダには時間同期 Viterbi ビーム探索をベースとする有限状態トランスデューサを用いた．

4.3 適応実験概要

図 1 は実験の概略図である．実験で用いたベースラインの音響モデルは汎用 LVCSR のために構築されたものであり，識別的特徴変換行列も音響モデルと同じ LVCSR データで推定されている．実験ではこの汎用 LVCSR モデルについて，提案法を含むいくつかの適応手法で自動車環境に適応し，認識性能を比較する．ここで，適応データサイズは合計 60 時間である．テストデータは，男女各 22 名の話者からなる 8,976 文の自動車内自由発話であり，停車時，市街地走行，高速走行時に far field マイクロフォンで音声を収録している．提案法に関連するパラメータ（式 (7) の α と ϵ，および式 (10) の τ）を決定するための開発セットとして，評価データとは異なる 500 文を用意した．評価実験では，開発セットで最良の性能を示した $\alpha = 0.1$，$\epsilon = 100$，$\tau = 500$ を用いている．なお，汎用 LVCSR および自動車用音声認識には，ともに同じ言語モデルを用いた．ここで使用した言語モデルは標準的な 3-gram モデルであり，ゼロ頻度問題に対して Knerser-Ney スムージングを採用している．

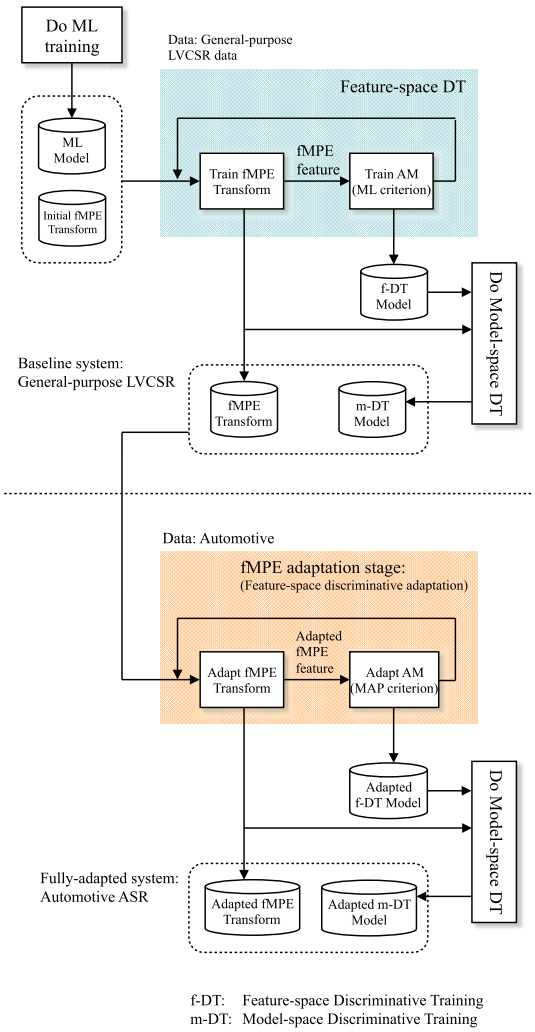

図 1 特徴空間適応実験の概要図
Fig. 1 Schematic diagram of adaptation experiments.

4.4 実験結果

実験結果を**表 2** に示す．表において “Baseline” は LVCSR データで学習されたベースライン，すなわち適応処理を施していないソースドメインの音響モデルによる結果である．次に “Model-MAP” は MAP 法による音響モデルのみの適応結果であり，特徴空間の適応処理は行っていない．続いて “fMPE-Adaptation” は正則化なしの特徴空間適応法の結果を示している．ここでは ML 基準と MAP 基準による音響モデルの更新を比較した．また，正則化なしの特徴空間適応法には，文献 [10] で提案されたスムージングパラメータ τ を間接的微分項に適用した．最後に提案法の結果は “Regularized fMPE-Adaptation” に示している．

結果を見ると，Model-MAP は種々の先行研究で証明されているとおり大きな改善を示しており，ベースラインと比較して相対的に 9.17%の改善を示した．続いて正則化なしの特徴空間適応法を見ると，ML 基準，MPE 基準の

どちらの場合も性能が劣化しているが，これは 3.1 節で述べたとおり音響モデルのサイズに対して適応データが少量であるため，過学習が起きた結果と見なすことができる．一方，正則化項を導入した提案法はどちらも WER の削減に貢献しているが，ML 基準による性能改善はわずかであり，音響モデルのみを対象とした Model-MAP 単独に劣る結果となった．その一方，提案法である “Regularized fMPE-Adaptation w/ MAP” は Model-MAP 単独を大きく上回る改善を得ており，ここからさらにモデル空間の識別学習を積み重ねることで，相対的に 13.19%にまで改善を向上させることができた．また，これは Model-MAP 法との比較においても 4.43%の相対的改善であり，t 検定を行った結果，危険率 5%で提案法は Model-MAP 法と比較して有意な差であった．これらの結果は音響モデルだけでなく，識別的特徴空間についても音響環境適応を行うことの重要性を示している．

5. 特徴空間識別学習の高精度化

5.1 小規模書き起こしデータの利用

3，4 章では少量データによる識別学習の適応処理について述べた．正則化処理の基本的な考え方は，ベースの音響モデル構築時に利用した学習データサイズに対して，新学習データのサイズが小さな場合のすべてについて適用できる．本章では少量の書き起こしデータを用いた（適応処理ではない）通常の識別学習の高精度化について議論する．

本論文では，千時間以上の音声データが存在する一方，人手による書き起こしデータは数十時間から百時間程度しか存在しないという現実的に起こりうる状況，すなわち準教師付き学習を想定する．本論文の冒頭で述べたように，識別学習の精度は発話データの書き起こし精度に大きく依存する．その理由は，識別学習が正解音素列に対する認識器の誤識別を軽減するように学習が進められるためである．そのため，理想的にはすべての音響モデル学習過程が書き起こしデータのみで実施されることが望ましいが，大規模データに対する書き起こしデータをアプリケーションごとに用意するのは現実的に非常に困難である．その一方で，少量の書き起こしデータだけでは話者変動や音響環境変動を十分にモデル化できないという問題点があった．我々が行った予備実験において，百時間の書き起こしに千時間以上の擬似書き起こしデータを加えた学習セットで作成した音響モデルが，百時間の書き起こしデータのみで構築した音響モデルと比較して，相対的に約 3%の改善があることを確認している．すなわち，書き起こしデータがきわめて少量の場合には，大規模な擬似書き起こしデータの利用が重要であることを示している．

5.2 提案手法

本論文では，標準的な特徴空間識別学習の前後に，正則化処理を組み込んだ事前・事後学習を導入して，人手による書き起こしデータを効果的に活用する方法を提案する．具体的な処理ステップを次に示す．

(a) 書き起こしと擬似書き起こしデータの両方を用いて，最尤推定により ML モデルを学習する．
(b) ML 学習された音響モデルについて，少量の書き起こしデータのみを用いて特徴空間上の識別学習を行う（事前学習：Pre-training step）．ただし，ここで想定している ML モデルは数千時間以上のデータから作成されたリッチな音響モデルであるため，3.2 節で述べた正則化処理を導入する．この事前学習は，識別的特徴変換行列に対する書き起こしデータを用いた初期化処理という位置づけである．
(c) Step(b) で得られた音響モデルおよび特徴変換行列を初期値とし，書き起こしと擬似書き起こしデータの両方を用いて，通常の特徴空間上の識別学習（通常学習：Standard-training step）を行う．
(d) Step(c) で得られた音響モデルと特徴変換行列のセットに対して，最後にもう一度，書き起こしデータのみを用いた識別学習（事後学習：Post-training step）を実行する．ただし，この際にも正則化処理を導入する．事後学習は，Step(c) の擬似書き起こしに含まれる間違った正解音素列に対する誤学習を，人手の書き起こしデータで再洗練することを意図している[*3]．

特徴空間上の識別学習における各発話の事後確率の推定には，弱言語モデル（unigram LM）から生成されたラティスが用いられるが，ここでは上記 Step(b)〜(d) の手順に先立って，音響モデルの学習データに使われる発話全体から推定された弱言語モデルでラティスを生成し，Step(b)〜(d) で共通に利用する．もしくは，擬似書き起こしには平均して 10%–30%程度の誤りが含まれるので，人手による少量の書き起こしデータが擬似書き起こしも含めた学習データ全体の傾向をカバーしていることが事前に分かっているなら，書き起こしデータのみを用いた弱言語モデルによってさらに効果を高めることができる．これについては 5.3.3 項で比較結果を示す．

5.3 実験

4 章と同様，識別学習高精度化の検証実験にも不特定話者汎用 LVCSR システムを用いた．ただし，ここでのデータセットは独自に収集した LVCSR 用日本語音声である．収録環境は英語のデータセットと同様多岐にわたり，千時間以上の自由発話音声から構成される．その他の条件も英語データと同じである．人手による書き起こしデータは，LVCSR セットから構成された 100 時間のサブセットである．本実験で利用する LVCSR は DNN から生成される特

[*3] 人手の書き起こしデータにも誤りは含まれるが，相対的に見て，音声認識による擬似書き起こしよりもはるかに誤りは少ない．

徴量を入力とするHMM/GMMシステムである．

5.3.1 DNN特徴量

音声データから256点FFTおよび31チャネルのバンドパスフィルタ処理を経由して，DNNに入力するための31次元のMFCCを抽出する．動的特徴量は利用せず，MFCCは発話区間のみで求めたケプストラム平均から発話単位のCMNを行っている．実験では，2種類のDNN構造（DNN-A，DNN-B）を利用してそれぞれDNNに基づく特徴量を抽出し，識別学習の効果を比較する．

DNNの入力には31次元のMFCCについて前後5フレームを連結した合計11フレームからなる341次元の特徴量を用いる．DNN-Aは各層1,024個のユニットからなる5層の隠れ層を持つネットワークである．出力層のユニット数は512であり，各ユニットは前後2音素のコンテキストを有する音素決定木のリーフに対応している．512次元の出力は主成分分析（PCA: Principal Component Analysis）に基づく特徴変換技術を通じて40次元の特徴量に圧縮する．ただし，PCAの前にSoftmaxによる非線形変換は行わない．DNN-Aの実験では，出力層からの出力を特徴量として利用するので，出力層の数は512と小さめのサイズにしている．次にDNN-Bについて述べると，DNN-BもDNN-Aと同じ入力層を有するが，5層目の隠れ層と出力層の間に40次元のボトルネック層を挿入し，ボトルネック層からの出力をDNN特徴量として用いる構成とした．なお，DNN-Bの出力層のユニット数は512から3,000に増やしている．隠れ層の非線形変換には，両DNNともシグモイド関数を用いる．

DNNの学習データはフレーム単位でランダムに混ぜ合わせ，損失関数としてクロスエントロピー基準を用いた250フレーム単位のミニバッチ処理により，最急降下法で重みの更新を行っている．DNNの学習データサイズは1,500時間である．ネットワーク全体の学習に先立って，識別的な事前学習によってネットワークを1層ずつ拡張する．ネットワーク全体の学習（Fine Tuning）は20回繰り返した．

DNNから40次元の特徴量を生成した後は，特徴空間上で識別的に学習された変換行列（fMPE変換）を用いて，特徴量をより正準化された空間に写像する．fMPE変換のための事後確率ベクトルは512次元であり，InnerとOuterコンテキストはそれぞれ8と4とした．本章では，このfMPE変換をより効果的に行うために導入した事前・事後学習の比較結果を示す．

5.3.2 音響モデル

音響モデルは4章と同様に話者独立であり，千時間以上の音声データを用いてLVCSRの音響モデルを構築する．ただし，入力特徴量がDNNに由来する特徴量である点で異なる．状態数は3,000で，ガウス分布の総数は120Kである．音響モデルはML基準で学習した後，MPE基準を用いた特徴空間およびモデル空間上の識別学習を続けて行う．

表 3 特徴空間識別学習のための事前・事後学習の効果（DNN-A）
Table 3 Pre/Post-training for feature-space discriminative training (f-DT) using DNN-A bottleneck feature.

f-DT Process	KER	%Relative
Baseline f-DT	13.38	–
+ Pre-Training	13.26	0.90
+ Pre/Post-Training	13.20	1.35

表 4 特徴空間識別学習のための事前・事後学習の効果（DNN-B）
Table 4 Pre/Post-training for feature-space discriminative training (f-DT) using DNN-B bottleneck feature.

f-DT Process	KER	%Relative
Baseline f-DT	12.21	–
+ Pre-Training	12.07	1.15
+ Pre/Post-Training	12.01	1.64

表 5 モデル空間識別学習後の比較（DNN-A）
Table 5 Comparisons feature-space DT (f-DT), followd by model-space DT using DNN-A bottleneck feature.

f-DT Process	KER	%Relative
Baseline f-DT	11.96	–
+Pre/Post Training	11.79	1.42 (t-test)

5.3.3 実験結果

表 3，**表 4** にDNN-A，DNN-Bによる実験結果を示す．ここで識別学習の事前学習は10回，標準学習は20回，事後学習は1回の反復処理とした．公平な比較のため，標準学習のみと事前・事後学習込みの識別学習には，学習データ全体から構築した弱言語モデルを利用し，共通のラティスを用いた．音響的な性能を重視するため，表では読み単位の比較結果，すなわち漢字混じりの認識結果を片仮名に変換した後の誤り率をKERとして記している．すなわち，同音異義語としての認識間違いは置換誤りとしてカウントしないこととした．表に示すとおり，提案手法はDNN-A，DNN-Bの両方のケースでともに認識誤りを軽減させており，DNN-Aで相対的に1.35%，DNN-Bで1.64%の改善が得られていることが分かる．これは提案法がDNN由来の特徴量の抽出方法に依存した方法ではないことを意味している．また，事前学習・事後学習のどちらも性能改善に貢献していることが確認された．**表 5** は，DNN-Aのシステムについて，特徴空間の識別学習に続けてモデル空間の識別学習を行った結果である．表に示すとおり，提案手法はモデル空間の識別学習の後でも改善効果が得られており，事前・事後学習なしの識別学習と比較して，1.42%の改善があることが分かった．t 検定を行った結果，提案法（表5の改善率1.42%に対応）は，危険率5%で事前・事後学習を行わない従来法と比べて有意に効果があった．

最後に，**表 6** は人手による書き起こしデータのみで構築した弱言語モデルによる効果を示している．ここでの数字

表 6 人手による書き起こしデータのみから推定した弱言語モデルの効果（DNN-A）

Table 6 Effect of weak LM trained only with manually transcribed data using DNN-A bottleneck feature.

f-DT Process	KER	%Relative
Pre/Standard/Post f-DT	13.20	1.35
+ Weak LM w/ transcribed data	13.15	1.72

は DNN-A による特徴空間識別学習の結果である（表 3 と対応）．今回の実験で利用した 100 時間の書き起こしデータは，おおむね擬似書き起こしを含む学習データ全体の傾向を反映しているため，弱言語モデルを人手による正確な書き起こしデータから推定することによって，さらに性能を改善できることを確認した．

6. おわりに

本論文では，第 1 に，音響モデルに対する適応データのスパースネスに焦点を当てて，識別的空間適応法を改善する方法を示した．提案法では，正則化項を識別学習の間接的微分に適用し，音響モデルの更新には ML 基準に代えて MAP 基準を導入した．先行研究において，音響モデルを大規模コーパスから ML 推定で学習し，その後，特徴空間上の識別学習の段階で学習データを対象環境の音声データに変更する方法が検討されている．識別的特徴変換行列の初期値としては 0 やランダムな値を付与するのが一般的であると考えられるが [4]，我々の提案手法では，大規模コーパスから学習された既存の識別的変換行列を初期値として追加の適応学習を実施することに相当する．本論文では，汎用 LVCSR から自動車環境という音響環境のミスマッチが大きい条件において検証実験を行い，提案法は代表的な適応法である MAP 法と比較して相対的に 4.4%の改善が得られることを示した．近年では，識別的基準は DNN のシーケンストレーニングにも導入され大きな効果をあげている [17]．識別的適応法について，今後は，小規模データを用いたシーケンストレーニングに関する適応処理の正則化を検討するとともに，L1，L2 正則化と併用した正則化法について検討する予定である．

第 2 に，本論文では特徴空間識別学習の高精度化のため，限られたサイズの書き起こしデータを利用するという観点から，正則化処理を導入した事前・事後学習を提案し，大規模データによる LVCSR システムにおいて統計的に有意な改善が得られることを示した．大規模データを用いた音響モデルの構築は標準的な作業になりつつあるが，大規模データによる統計モデルが，これまでに提案された種々のモデル化技術の効果を打ち消すことも少なくない．そうしたなか，我々が提案した方法はコンスタントな性能改善を示しており，実用的な方法であると考えている．

謝辞 本研究の遂行にあたり，IBM ワトソンリサーチセンターの Vaibhava Goel 氏と Steven J. Rennie 氏に有益な助言をいただいた．ここに感謝の意を表する．

参考文献

[1] Sainath, T.N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P. and Mohamed, A.R.: Making deep belief networks effective for large vocabulary continuous speech recognition, *IEEE ASRU*, pp.30–35 (2011).
[2] Sainath, T.N., Kingsbury, B. and Ramabhadran, B.: Auto-encoder Bottleneck Features Using Deep Belief Networks, *IEEE ICASSP*, pp.4153–4156 (2012).
[3] Yu, D. and Seltzer, M.L.: Improved Bottleneck Features Using Pretrained Deep Neural Networks, *Interspeech*, pp.237–240 (2011).
[4] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H. and Zweig, G.: fMPE: Discriminatively trained features for speech recognition, *Proc. ICASSP*, pp.961–964 (2005).
[5] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah, K.: Boosted MMI for Model and Feature Space Discriminative Training, *Proc. IEEE ICASSP*, pp.4057–4060 (2008).
[6] Leggetter, C.J. and Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech and Language*, Vol.9, pp.171–185 (1995).
[7] Gauvain, J.L. and Lee, C.H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains, *IEEE Trans. Speech and Audio Processing*, Vol.2, No.2, pp.291–298 (1994).
[8] Povey, D., Gales, M.J.F., Kim, D.Y. and Woodland, P.C.: MMI-MAP and MPE-MAP for acoustic model adaptation, *Interspeech/Eurospeech*, pp.1981–1984 (2003).
[9] Bippus, R., Fischer, A. and Stahl, V.: Domain adaptation for robust automatic speech recognition in car environments, *Interspeech/Eurospeech*, pp.1943–1946 (1999).
[10] Zheng, J. and Stolcke, A.: fMPE-MAP: Improved discriminative adaptation for modeling new domain, *Proc. Interspeech*, pp.1573–1576 (2007).
[11] Fukuda, T., Ichikawa, O., Nishimura, M., Rennie, S.J. and Goel, V.: Regularized Feature-space Discriminative Adaptation for Robust ASR, *Proc. Interspeech*, pp.2185–2188 (2014).
[12] Yu, D., Varadarajan, B., Deng, L. and Acero, A.: Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, *Computer Speech and Language*, Vol.24, Issue 3, pp.433–444 (2010).
[13] Yu, K., Gales, M.J.F., Wang, L. and Woodland, P.C.: Unsupervised Training and Directed Manual Transcription for LVCSR, *Speech Communication*, Vol.52, pp.652–663 (2010).
[14] Fukuda, T., Tachibana, R., Chaudhari, U., Ramabhadran, B. and Zhan, P.: Constructing ensembles of dissimilar acoustic models using hidden attributes of training data, *IEEE ICASSP*, pp.4141–4144 (2012).
[15] Beaufays, F., Vanhoucke, V. and Strope, B.: Unsupervised discovery and training of maximally dissimilar cluster models, *Proc. Interspeech*, pp.66–69 (2010).
[16] Gales, M.J.F.: Semi-tied covariance matrices for hidden Markov models, *IEEE Trans. Speech and Audio Processing*, Vol.7, No.3, pp.272–281 (1999).

[17] Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling, *IEEE ICASSP*, pp.3761–3764 (2009).
[18] Marcheret, E., Goel, V. and Olsen, P.: Optimal quantization and bit allocation for compressing large discriminative feature space transforms, *IEEE ASRU*, pp.64–69 (2009).

福田 隆 （正会員）

1998 年神戸市立工業高等専門学校電子工学科卒業．2000 年豊橋技術科学大学知識情報工学科卒業．2005 年同大学大学院工学研究科電子情報工学専攻博士後期課程修了．同年日本アイ・ビー・エム株式会社入社．以来，音声言語処理の研究に従事．現在，IBM 東京基礎研究所専任研究員．2010 年日本音響学会粟屋潔学術奨励賞，同年 IBM プロフェッショナル論文年間最優秀賞，2012 年電子情報通信学会音声研究会研究奨励賞，2013 年情報処理学会山下記念研究賞各賞受賞．本会シニア会員，日本音響学会，電子情報通信学会，IEEE，ISCA 各会員．

市川 治

1986 年東京大学工学部航空学科卒業．1988 年同大学大学院工学系研究科航空学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社大和研究所にて，オペレーティング・システム，音声認識ソフトウェア，マルチメディア製品，システム管理ソフトウェア，衛星通信ソフトウェア等の開発に従事．1999～2001 年文部省宇宙科学研究所受託研究員として数値シミュレーションの研究に従事．2001 年より日本アイ・ビー・エム（株）東京基礎研究所研究員としてロバスト音声認識の研究に従事，現在に至る．2008 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．2015 年法政大学理工学部創生科学科非常勤講師．電子情報通信学会シニア会員，日本音響学会，IEEE 各会員．

立花 隆輝

1996 東京大学工学部航空宇宙工学科卒業．1998 同大学大学院工学系研究科航空宇宙工学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて音楽電子透かしと，音声合成や音声認識等の音声言語情報処理の研究に従事．2007 年大阪大学大学院工学研究科電気電子情報専攻博士課程修了．2016 年豊橋技術科学大学客員教授．電子情報通信学会，日本音響学会各会員．