# 音声認識のための線形判別分析の系列相互情報量最大化識別学習 *

○太刀岡勇気, (三菱電機・情報総研), 渡部晋治, ルルージョナトン, ハーシージョン (MERL)

## 1 はじめに

音声認識を行う際には、複数のフレームにわたる特徴量の変動を直接捉えるために動的特徴量が使われる。他方で、線形判別分析 (linear discriminant analysis; LDA) [1] により、連結特徴量を次元縮減した特徴量を用いることもある。LDA は、クラス間分散のクラス内分散に対する比を最大化する。クラスは通常、コンテキスト依存音素モデルから導かれ、特徴量変換を推定するのに単純で効率的な閉解が存在する。一方で LDA の限界の一つは各クラスの共分散が等しいという仮定にある。この制約を緩和するために、異分散判別分析や異分散線形判別分析 (heteroscedastic LDA; HLDA) が提案されている [2, 3, 4]。

LDA のもう一つの限界は、音声認識器の出力を陽に考慮することができないところにある。特徴量変換の最終的な目的は、音声認識に適した特徴量得ることである。LDA は特徴量間の識別性を向上させることに寄与するが、認識器にとって識別しやすいクラスと識別しにくい (すなわち混同しやすい) クラスを同等に扱ってしまう。識別的手法の最近の進展によって、認識器の誤り傾向を考慮した系列の識別学習が、音響モデル、特徴量空間識別学習といったさまざまな従来法に有効であることが広く知られてきている。例えば、相互情報量最大化 (maximum mutual information; MMI) 基準 [5] もしくは音素誤り最小化 (minimum phoneme error; MPE) 基準 [6] が学習の基準としてよく使われる。これらは、音声認識のレベルでの誤りのパターンを考慮することで、認識器にとって最も重要な状態を識別することに焦点を当てることができる。パラメータ推定のための十分統計量は、認識単語系列の事後確率に基づいている。

線形特徴量変換は、一般的に射影行列とオフセット項により実現される。LDA は単一領域の線形写像でオフセット項はない。対照的に、領域依存線形変換 [7] では、初めに特徴量空間をいくつかの領域に分割し、各領域に異なる変換を適用する。HLDA の MPE 基準に基づく拡張である MPE-HLDA [8] や特徴量空間 MPE(feature space MPE; f-MPE) [9]、MMI-SPLICE [10] といった手法が提案されている。このような方法は、たいてい繰り返し最適化を必要とする。

提案法は特徴量の十分統計量を計算する際に、認識器の事後確率を考慮するために、LDA を MMI 評価関数に基づき拡張する。提案法の利点は、閉形式の解が得られることと、十分統計量の計算部分を変更するだけで済むことから、実装が簡便なことである。

本報では初めに、2 節で、従来の LDA [1] について、主に文献 [2, 4] に書かれている見地から述べる。次に MMI による拡張を、3 節に述べる。4 節において、2 つの異なるタスクで、提案法の有効性を示す。

## 2 最尤 LDA

$t$ 番目のフレームの $n$ 次元の入力特徴量を、$\boldsymbol{x}_t \in \mathbb{R}^n$ とする。LDA では通例、音響特徴量の連続数フレームを結合した特徴量を入力とする。LDA[1, 4] により、$\boldsymbol{x}_t$ はより低い次元の特徴量 $\boldsymbol{y}_t \in \mathbb{R}^p$ に変換される。

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t \tag{1}$$

ここで $\boldsymbol{A}$ は LDA による変換行列であり、その次元は $p \times n (p < n)$ である。LDA の評価関数は

$$\arg\max_{\boldsymbol{A}} \frac{|\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top|}{|\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^\top|} \tag{2}$$

のように与えられる。ここで $\top$ は転置、$\boldsymbol{B} \in \mathbb{R}^{n\times n}$ と $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ はそれぞれ、クラス間分散行列とクラス内分散行列であり、式 (3) により定義される。

$$\begin{aligned}\boldsymbol{B} &= \frac{1}{\sum_j N_j}\sum_j N_j \boldsymbol{\mu}_j^{\mathrm{x}}(\boldsymbol{\mu}_j^{\mathrm{x}})^\top - \bar{\boldsymbol{\mu}}^{\mathrm{x}}(\bar{\boldsymbol{\mu}}^{\mathrm{x}})^\top \\ \boldsymbol{W} &= \frac{1}{\sum_j N_j}\sum_j N_j \boldsymbol{\Sigma}_j^{\mathrm{x}}\end{aligned} \tag{3}$$

ここで $\boldsymbol{\mu}^{\mathrm{x}}$ と $\boldsymbol{\Sigma}^{\mathrm{x}}$ は $\boldsymbol{x}$ の平均ベクトルと共分散行列である。$N_j$ は $j$ 番目のクラスに属する要素数、$\bar{\boldsymbol{\mu}}^{\mathrm{x}}$ は $\boldsymbol{\mu}_j^{\mathrm{x}}$ 全ての平均である。クラス $j$ に対して、$\boldsymbol{\mu}_j^{\mathrm{x}}$ と $\boldsymbol{\Sigma}_j^{\mathrm{x}}$ は以下のように計算される [4]。

$$\begin{aligned}N_j &= \sum_t \psi_t(j) \\ \boldsymbol{\mu}^{\mathrm{x}} &= \frac{1}{N_j}\sum_t \psi_t(j)\boldsymbol{x}_t \\ \boldsymbol{\Sigma}_j^{\mathrm{x}} &= \frac{1}{N_j}\sum_t \psi_t(j)\boldsymbol{x}_t\boldsymbol{x}_t^\top - \boldsymbol{\mu}_j^{\mathrm{x}}(\boldsymbol{\mu}_j^{\mathrm{x}})^\top\end{aligned} \tag{4}$$

---

*Sequential maximum mutual information discriminative training of linear discriminant analysis for speech recognition, by TACHIOKA, Yuuki (Mitsubishi Electric Corporation); WATANABE, Shinji; LE ROUX, Jonathan; HERSHEY, John (MERL).

ここで、$\psi_t(j)$ はクラス $j$ の重みであり、$\boldsymbol{x}_t$ をクラス $j$ に関連付ける。古典的な LDA では、クラスの割り当ては $j = l(t)$ で与えられることから、$\psi_t(j)$ は

$$\psi_t(j) = \delta(l(t), j) \tag{5}$$

のように定義される。ここで、$\delta$ はクロネッカーのデルタである。最も一般的な $j$ が HMM の状態番号と紐づけられている場合には、HMM モデルによるアライメントは、クラスラベルと対応する。

LDA の解は、一般化固有値問題 6 を解き [11]、$p$ 位までの固有値 $\lambda_{1:p}$ に対応する固有ベクトル $v_{1:p}^T$ で $\boldsymbol{A}$ の行を埋めることで得られる。

$$\boldsymbol{B}v = \lambda \boldsymbol{W}v \tag{6}$$

Kumar らによって、標準的な LDA は、最尤基準での最適化と同じ解を持つことが示されている [2]。この問題では、モデルは $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t$ において状態依存の分散を持つ。一方で、直交部分空間 $\boldsymbol{y}'_t = \boldsymbol{A}'\boldsymbol{x}_t$ での平均と分散は状態非依存である。ここで $\boldsymbol{A}' \in \mathbb{R}^{(n-p)\times n}$ の各行は $\boldsymbol{A}$ の各行と直交関係にある。

Kumar らの方法で、最尤の評価関数

$$\mathcal{F}^{\mathrm{MLK}} = \ln P(\boldsymbol{Y}, \omega_r) \tag{7}$$

を考えることで、この結果を HMM モデルの場合に一般化することができる。ここで $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots\}$ は変換特徴量ベクトルの系列、$\omega_r$ は正解単語ラベルである。モデルパラメータ $\theta_j$ に関する微分は以下のようになる。

$$\frac{\partial \mathcal{F}^{\mathrm{MLK}}}{\partial \theta_j} = \sum_t \sum_j \frac{\partial \mathcal{F}^{\mathrm{MLK}}}{\partial \ln p(\boldsymbol{y}_t|j)} \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j} \tag{8}$$

$$= \sum_t \sum_j \gamma_t(j) \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j} \tag{9}$$

ここで $p(\boldsymbol{y}_t|j)$ は HMM 状態 $j$ で条件づけられた確率である。この微分値を 0 として $\theta_j$ について解くことで、状態 $j$ に対して、 式 (4) で計算される状態依存の平均と分散が求まる。その際に式 (5) は

$$\psi_t(j) = \gamma_t(j) \tag{10}$$

のように変更される。ここで $\gamma_t(j)$ は正解単語ラベル $\omega_r$ に対する事後確率である。これは通常の LDA 同様、一般化固有値問題 (6) により求解できる。式 (5) と比較すると、状態事後確率によるソフトクラスタリングである点が異なる。Baum-Welch アルゴリズムにより推定されるモデルに対して、上記の LDA 統計量はより密接にそれらをモデル推定に使った場合に対応する。

## 3 系列相互情報量最大化 LDA

2 節のような最尤基準では認識タスクに対して最適なモデルとはならず、音響モデルパラメータの学習と同様、認識器の誤り傾向に基づいたデータの重みづけと選択が必要である。LDA の学習の際にも、最尤基準ではなく、音響モデルの MMI 識別学習に類似して、評価関数

$$\mathcal{F}^{\mathrm{MMI}} = \ln \frac{P(\boldsymbol{Y}, \omega_r)}{\sum_\omega P(\boldsymbol{Y}, \omega)} \tag{11}$$

の分母のラティスの事後確率 $\gamma_t^{\mathrm{den}}$ を考慮するべきであり、これを実現する提案法を、系列の MMI LDA (sequential LDA; sLDA) と呼ぶことにする。ここで、$\omega$ は認識器より出力される仮説である。式 (11) の $\theta_j$ に関する微分は、MMI-SPLICE [10] の場合と同様、

$$\begin{aligned}\frac{\partial \mathcal{F}^{\mathrm{MMI}}}{\partial \theta_j} &= \sum_t \sum_j \frac{\partial \mathcal{F}^{\mathrm{MMI}}}{\partial \ln p(\boldsymbol{y}_t|j)} \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j} \\ &= \sum_t \sum_j \Delta_t(j) \frac{\partial \ln p(\boldsymbol{y}_t|j)}{\partial \theta_j}\end{aligned} \tag{12}$$

のようになる。平均と分散の推定は、式 (4) と同様だが、$\psi_t(j) = \Delta_t(l(t))$ となる点が異なっている。$\Delta_t$ は、式 (11) の分子に関する事後確率 $\gamma^{\mathrm{num}}$ と分母に関する事後確率 $\gamma^{\mathrm{den}}$ の差分 $\Delta_t = \gamma^{\mathrm{num}} - \gamma^{\mathrm{den}}$ である。$\Delta_t(j)$ は負になりうるので、通常の拡張 Baum-Welch アルゴリズムでは、$\Delta_t(j)$ が正値になるように工夫している。ここでは簡便に、パラメータ $\alpha$ $(0 \le \alpha \le 1)$ を導入し、分母項 $\gamma_t^{\mathrm{den}}(j)$ の影響度合いを弱める。

$$\psi_t(j) = \gamma_t^{\mathrm{num}}(j) - \alpha \gamma_t^{\mathrm{den}}(j) \tag{13}$$

式 (10) は $\psi_t(j) = \gamma_t^{\mathrm{num}}(j)$ に対応するので、$\alpha = 0$ のときは、これは式 (10) と一致する。

提案法は LDA にソフトな特徴量選択 [12] を加えたものと解釈できる。認識器が正解の場合に対応する $\gamma_t^{\mathrm{den}}(j)$ が 1 に近いデータに対しては、小さい重みが課せられる。これにより、認識器による誤差に応じて学習データの重みが調整されることになる [1]。

### 3.1 I-smoothing の解釈

式 (13) は $\psi_t(j) = (1-\alpha)\gamma_t^{\mathrm{num}}(j) + \alpha\Delta_t(j)$ のように書き直すこともできる。この式は、差分統計量 $\Delta_t(j)$ とクラスラベルの事後確率 $\gamma_t^{\mathrm{num}}(j)$ を内挿比 $\alpha$ で平滑化したと解釈できる。よって、パラメータ $\alpha < 1$ とすれば、$\alpha$ により過学習を防ぐことができる。これは音響モデルの識別学習で広く使われている I-smoothing の手法 [6] と関連する。

[1] クラス間分散 $\boldsymbol{B}$ は MMI に基づく重みに多少影響を受けるのみであって、依然として大域的なものである。

## 4 認識実験

### 4.1 実験条件

提案法の有効性を、2 つのタスクを使って確かめた。日本語話し言葉コーパス (CSJ) と第 2 回 CHiME チャレンジ トラック 2[13] である。前者は広く使われている大語彙 (語彙サイズは 70k) 連続音声認識タスクである。3 種のテストセットが提供されており、各 10 話者の講演調の発話が含まれる。テストセット 1、2、3 は、それぞれ 22,682、23,226、14,896 語から成る。音響モデルは最尤推定により学習した。0∼12 次の MFCC とその 1 次・2 次の動的特徴量を使った。コンテキスト依存 HMM の状態数は 3,500 とし、ガウス分布の総数は 96,000 である。

後者は、残響および騒音がある環境での音声認識の性能を評価することを目的としている。中程度語彙 (5k) の連続音声認識で、Wall Street Journal の読み上げである。騒音は非定常性であり、音声区間が切り出された (isolated) 音声に騒音が SN 比 $= \{-6, -3, 0, 3, 6, 9\}$dB で重畳されている。このタスクにより、提案の sLDA の騒音下音声認識タスクでの有効性を検証するとともに、音響モデル (ガウス混合モデル (Gaussian mixture model; GMM) と深層神経回路網 (deep neural networks; DNN)) の識別学習や、特徴量空間識別学習 (feature-space boosted MMI; f-bMMI [9]) と提案法を組み合わせた際の有効性を検証する。事前分布に基づくバイナリマスク [14] により騒音抑圧された、単一チャネルの音声を対象とした。Kaldi ツールキット [15] と我々が CHiME チャレンジのために提供したベースラインの評価ツール [14] を利用した。学習セット (83 話者、6,921 発話、125,095 単語) は全 SNR で共通であり、開発セット (si_dt_05)(10 話者、409 発話、6,779 単語) および評価セット (si_et_05)(12 話者、330 発話、5,353 単語) は SNR ごとに用意されている。HMM の状態数は 2,500 で、ガウス分布の総数は 15,000 である。DNN の学習には Kaldi の nnet2 の実装を使った。DNN の隠れ層数は 3 とし、全部で 1M パラメータある。初期の学習率は 0.01 とし、学習の終盤に向けて 0.001 まで低減させた。ベースラインの特徴量は、0∼12 次元の MFCC とその 1 次・2 次の動的特徴量である。LDA 後に最尤線形変換 (maximum likelihood linear transformation; MLLT) [16, 17] を行った。MLLT は、LDA とともによく用いられる特徴量変換手法である。CHiME タスクには、話者適応手法 (話者適応学習 (speaker adaptive training; SAT) と特徴量空間最尤線形回帰 (feature space maximum likelihood linear regression; fMLLR)) を用いた。

Table 1 WER of the conventional LDA ($\alpha = 0$) and the proposed sequential maximum mutual information LDA (sLDA) with different smoothing factors $\alpha$ on the CSJ database.

| | $\alpha$ | test1 | test2 | test3 | Avg. |
|---|---|---|---|---|---|
| LDA | 0 | 20.42 | 17.95 | 19.22 | 19.20 |
| | 0.1 | **20.39** | 17.81 | 19.49 | 19.23 |
| | 0.3 | 20.47 | 17.93 | 19.28 | 19.23 |
| | 0.5 | 20.44 | 17.81 | 19.14 | 19.13 |
| | 0.7 | 20.40 | 17.83 | 19.03 | 19.09 |
| | 1.0 | 20.51 | **17.68** | **18.77** | **18.99** |
| LDA | 0 | 19.09 | 16.31 | 17.21 | 17.54 |
| +MLLT | 0.1 | 19.13 | 15.96 | 17.23 | 17.44 |
| | 0.3 | 19.08 | **15.91** | 17.07 | **17.35** |
| | 0.5 | 19.04 | 16.12 | 17.25 | 17.47 |
| | 0.7 | 19.09 | 16.03 | 17.11 | 17.41 |
| | 1.0 | **18.90** | 16.24 | **16.94** | 17.36 |

Table 2 WER[%] for isolated speech (**si_dt_05**) of the CHiME challenge with different $\alpha$s using ML acoustic model for noisy speech recognition with noise suppression by prior-based binary masking (sLDA+MLLT).

| $\alpha$ | −6dB | −3dB | 0dB | 3dB | 6dB | 9dB | Avg. |
|---|---|---|---|---|---|---|---|
| 0 | 64.64 | 54.24 | 46.35 | 37.91 | 32.75 | 28.96 | 44.14 |
| 0.1 | 64.64 | 53.81 | 46.45 | 38.65 | 32.75 | 29.15 | 44.24 |
| 0.3 | 64.88 | **53.72** | 45.58 | **37.13** | **31.89** | **28.43** | **43.61** |
| 0.5 | 64.71 | 53.84 | 46.20 | 37.81 | 32.25 | 28.81 | 43.94 |
| 0.7 | 64.48 | 54.43 | 45.88 | 37.51 | 32.44 | 28.69 | 43.91 |
| 1.0 | **64.36** | 54.29 | **45.01** | 37.81 | 32.59 | 28.96 | 43.84 |

### 4.2 CSJ

表 1 には、CSJ タスクの実験結果を示している。性能はパラメータ $\alpha$ に依存するものの、全体的に提案の sLDA は従来の LDA ($\alpha = 0$) よりも性能が向上しており、MLLT と組み合わせた場合にも有効である。太字で示した最良の場合では、sLDA 単独では 0.21%、MLLT と組み合わせた場合には 0.19%の絶対値での WER の低減が見られる。

### 4.3 第 2 回 CHiME チャレンジ トラック 2

表 2 には CHiME チャレンジ トラック 2(開発セット) での結果を示している。本節では、MLLT も併用している。平均的に、$\alpha \geq 0.3$ のときに、音声認識の

Table 3 WER[%] for isolated speech (**si_dt_05**) using ML and discriminatively trained acoustic model (bMMI) with feature-space discriminative training (f-bMMI). LDA+MLLT (upper), sLDA+MLLT (lower).

| | −6dB | −3dB | 0dB | 3dB | 6dB | 9dB | Avg. |
|---|---|---|---|---|---|---|---|
| ML | 64.64 | 54.24 | 46.35 | 37.91 | 32.75 | 28.96 | 44.14 |
| bMMI | 63.39 | 52.54 | 44.56 | 35.60 | 30.98 | 28.10 | 42.53 |
| f-bMMI | 60.92 | 50.41 | 41.76 | 33.59 | 29.56 | 25.90 | 40.36 |
| ML | 64.88 | 53.72 | 45.58 | 37.13 | 31.89 | 28.43 | 43.61 |
| bMMI | 62.75 | 51.78 | 44.24 | 35.92 | 30.80 | 27.32 | 42.14 |
| f-bMMI | **60.27** | **49.26** | **41.08** | **32.95** | **28.63** | **25.17** | **39.56** |

Table 4 WER[%] for isolated speech (**si_dt_05**) with speaker adaptive training, speaker adaptation (fMLLR), and minimum Bayes risk decoding (MBR). LDA+MLLT (upper), sLDA+MLLT (lower).

| | −6dB | −3dB | 0dB | 3dB | 6dB | 9dB | Avg. |
|---|---|---|---|---|---|---|---|
| ML | 59.94 | 47.93 | 39.83 | 33.01 | 28.00 | 23.47 | 38.70 |
| bMMI | 56.90 | 45.79 | 37.60 | 30.31 | 26.15 | 21.74 | 36.42 |
| f-bMMI | 52.93 | 42.62 | 34.59 | 27.63 | 24.27 | 20.24 | 33.71 |
| +MBR | 52.65 | 42.04 | 33.75 | 27.05 | 23.74 | 19.91 | 33.19 |
| DNN | 52.78 | 42.50 | 34.08 | 27.05 | 24.13 | 20.12 | 33.44 |
| bMMI | 47.34 | 36.33 | 28.96 | **23.40** | 20.03 | 17.05 | 28.85 |
| ML | 59.21 | 48.40 | 39.28 | 32.41 | 27.72 | 22.86 | 38.31 |
| bMMI | 56.14 | 45.51 | 36.69 | 29.55 | 26.08 | 21.33 | 35.88 |
| f-bMMI | 53.09 | 43.34 | 33.71 | 27.16 | 23.93 | 19.78 | 33.50 |
| +MBR | 52.60 | 42.51 | 33.03 | 26.38 | 23.34 | 19.18 | 32.84 |
| DNN | 52.91 | 41.81 | 32.56 | 27.73 | 24.31 | 19.68 | 33.17 |
| bMMI | **47.31** | **36.13** | **28.49** | 23.50 | **20.00** | **16.57** | **28.67** |

Table 5 WER[%] for isolated speech (**si_et_05**) with speaker adaptive training and speaker adaptation (fMLLR). LDA+MLLT (upper), sLDA+MLLT (lower). In this table, DNN is DNN with boosted MMI.

| | −6dB | −3dB | 0dB | 3dB | 6dB | 9dB | Avg. |
|---|---|---|---|---|---|---|---|
| ML | 50.91 | 41.64 | 33.89 | 26.30 | 21.61 | 18.85 | 32.20 |
| f-bMMI | 44.54 | 35.91 | 29.24 | 22.31 | 17.77 | 15.88 | 27.61 |
| DNN | **37.98** | 28.26 | **21.86** | 17.71 | **12.61** | 11.75 | **21.70** |
| ML | 50.46 | 42.05 | 32.80 | 26.42 | 21.22 | 18.61 | 31.93 |
| f-bMMI | 44.85 | 35.05 | 27.69 | 21.43 | 17.34 | 14.74 | 26.85 |
| DNN | 38.63 | **27.54** | 22.55 | **17.37** | 13.23 | **11.69** | 21.84 |

性能が向上しており、$\alpha = 0.3$ の場合に最も性能改善 (0.53%の絶対値での WER の低減) が見られた。表 1 および 2 の検討から、2 つの異なる音声認識タスクにおいて、提案の sLDA が従来の LDA に比べて優れていることが示された。

表 3 には、音響モデルの識別学習 (bMMI) および特徴量空間識別学習 (f-bMMI) の結果を示している。双方の場合でも、提案法は音声認識性能を改善している。とりわけ f-bMMI の場合に顕著であり、絶対値で 0.8%の WER の低減が見られた。提案法と f-bMMI の組み合わせにより、付加的な性能向上が見られたことになる。これは提案法による予備的な識別的分類が f-bMMI のよい初期値となっており、より正確な領域に基づくモデリングにより識別的特徴量変換が実現されたことを示唆している。

表 4 と 5 は開発セットと評価セットでの SAT、fMLLR と DNN モデルを検討している。DNN システムの評価セットに対する音声認識性能は低下したが、提案法により、5 つの SNR 条件で開発セットの認識性能を向上させるとともに、評価セットにおいても半数の SNR 条件 (−3、3 および 9dB) では認識性能の向上が見られた。提案法により、平均で 0.9%(絶対値) の WER の低減が見られた。

## 5 おわりに

本報では、系列の MMI 学習法に基づき、LDA を拡張する方法を提案した。提案法は、計算量が通常の LDA と同等で小さく、ツールの改変が十分統計量を求める部分の修正のみで済むという利点がある。大語彙連続音声認識と騒音下音声認識のタスクにおいて、提案法の有効性を確認した。提案法では、一般化固有値問題により閉形式の解が得られる点が、他の拡張 Baum-Welch 法や勾配法による最適化手法に基づく識別的特徴量変換と異なる。今後は、それらの関係性を理論的に考察する。

## 参考文献


[1] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," Proc. of ICASSP, pp.13–16 (1992).

[2] N. Kumar, Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition, PhD JHU (1997).

[3] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," Speech Com., **26**, 283–297 (1998).

[4] G. Saon et al., "Maximum likelihood discriminant feature spaces," Proc. of ICASSP, pp.1129–1132 (2000).

[5] L. Bahl et al., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," Proc. of ICASSP, pp.49–52 (1986).

[6] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," Proc. of ICASSP, pp.105–108 (2002).

[7] B. Zhang et al., "Recent progress on the discriminative region-dependent transform for speech feature extraction," Proc. of INTERSPEECH, pp.1573–1576 (2006).

[8] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," Proc. of ICASSP, pp.925–928 (2005).

[9] D. Povey et al., "Boosted MMI for model and feature-space discriminative training," Proc. of ICASSP, pp.4057–4060 (2008).

[10] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," Proc. of INTERSPEECH, pp.989–992 (2005).

[11] K. Fukunaga, Introduction to statistical pattern recognition, Academic Press (1990).

[12] B. Chen et al., "Training data selection for improving discriminative training of acoustic models," Pattern Recognition Letters, **30**, 1228–1235 (2009).

[13] E. Vincent et al., "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," Proc. of ICASSP, pp.126–130 (2013).

[14] Y. Tachioka et al., "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," Proc. of the 2nd CHiME Workshop, pp.19–24 (2013).

[15] D. Povey et al., "The Kaldi speech recognition toolkit," Proc. of ASRU, pp.1–4 (2011).

[16] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," Proc. of ICASSP, pp.661–664 (1998).

[17] M. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE Trans. on Speech and Audio Processing, **7**, 272–281 (1999).