システム統合のための音響モデルの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

近年、音声認識の性能が大きく改善した要因として、音響モデルの学習法が最尤学習 (ML) から識別的学習 [1] に移行したことが挙げられる。これらの手法は単一システムでの性能向上を目的とする。これに対して、複数のシステムの仮説を統合するシステム統合に基づく方法が知られている。例えば Recognizer Output Voting Error Reduction (ROVER) [2] は、元のシステムと補助システムの仮説の多数決により、優れた仮説を得るもので、たとえ補助システムの性能が元のシステムの性能よりも低くとも、元のシステムを超える性能が得られることも少なくない。

システム統合の際には、傾向を異にする仮説の統合が効果的である。例えばランダムフォレスト法 [3] は、音響モデルの状態共有構造を無作為に変化させることで、異なる仮説傾向を持つ補助システムを構築する。しかしながら、補助システムの仮説が元のシステムと同傾向であったり、過度な誤りを含有する場合には、システム統合によって性能が向上しないこともありうる。通常は複数のシステムを試製し、最良の組み合わせを開発セットの性能に基づいて決定する。このような試行錯誤に頼る手法ではなく、理論的枠組みを備えた手法が望まれている。

機械学習の分野では、「ブースティング」がシステム統合の理論的裏付けとして広く研究されている。AdaBoost [4] は、教師ありの繰り返し学習中で、前ステップの識別器が誤った学習データに大きな重みを与えることで、違った仮説傾向を持つ識別器を構築することを可能にする。2 値分類のように単純な問題には有効であるが、複雑な系列識別問題である音声認識の問題に単純に適用することはできない。

音声認識の問題にブースティングを適用しようとする試みもいくつかみられる。例えば、ブースティング Baum-Welch [5] は、Baun-Welch アルゴリズムにおけるフレーム毎のブースティングであり、尤度の低い学習データの統計量を重点的に学習できる。ただしこの手法の目的は Baum-Welch アルゴリズムの繰り返しにおいて、前の識別器の出力傾向を考慮して当該のモデルを改善するところにあり、補助システムを構築するところにはない。

本報では、通常のラティスを使った識別学習の枠組みを拡張し、システム統合のための補助システムを構築する手法を提案する。文献 [6] において音素誤り最小学習を用いた枠組みが提案されているが、ここではブースティングとの関連が一層明確な相互情報量最大化 (MMI) 学習に焦点をあてる。提案法は、MMI の評価関数を一般化し、正解ラベルとの相互情報量を最大化すると同時に、元のシステムの仮説との相互情報量を最小化する。識別学習のためのラティスを使ってモデルを更新するため、識別学習法との関係性が明確である。従来の識別学習法と提案法を 2 節、3 節で述べ、4 節において提案法の有効性を示す。

2 相互情報量最大化 (MMI) 識別学習

MMI 学習では、初期モデル (例えば ML モデル) で生成されたラティス上の仮説 ω を参照して、以下の正解ラベル列 ω_r に対する評価関数を最大化する。

$$\mathcal{F}^{\mathrm{M}}_{\lambda}(\omega_r) = \ln \frac{P_\lambda(\omega_r, \mathbf{X})}{\sum_\omega P_\lambda(\omega, \mathbf{X})}, \tag{1}$$

$$= \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_\lambda\left(s_r, \mathbf{X}\right)^\kappa p_L(\omega_r)}{\sum_\omega \sum_{s \in \mathcal{S}_\omega} p_\lambda\left(s, \mathbf{X}\right)^\kappa p_L(\omega)}, \tag{2}$$

λ は最適化すべき HMM 変数の組、$\mathbf{X} = \{\mathbf{x}_t | t = 1, \cdots, T\}$ は T フレームからなる発話の特徴量ベクトルである。発話ごとの総和は可読性のために省略した。HMM 状態系列 s に対する音響モデルスコア p_λ(ただし音響スケール κ) と言語モデルスコア p_L の積を $P_\lambda(\omega, \mathbf{X})$ で表す。式 (2) において、音響スコアは正解 s_r、仮説 s とそれに対応する HMM 状態系列の組 $\mathcal{S}_{\omega_r}$、$\mathcal{S}_\omega$ が、正解ラベル ω_r と仮説 ω を各様に算出するスコアを合計して得られる。式 (2) は、GMM の平均 $\boldsymbol{\mu}_{jm}$ と共分散 $\boldsymbol{\Sigma}_{jm}$(HMM 状態 j、ガウス分布 m) に関する以下の更新式で具体化される。

$$\boldsymbol{\mu}'_{jm} = \frac{\sum_t \Delta_{jm,t}\mathbf{x}_t + D_{jm}\boldsymbol{\mu}_{jm}}{\sum_t \Delta_{jm,t} + D_{jm}}, \tag{3}$$

$$\boldsymbol{\Sigma}'_{jm} = \frac{\sum_t \Delta_{jm,t}\mathbf{x}_t\mathbf{x}_t^\top + D_{jm}(\boldsymbol{\Sigma}_{jm} + \mathbf{U}_{jm})}{\sum_t \Delta_{jm,t} + D_{jm}} - \mathbf{U}'_{jm},$$

$\Delta_{jm,t}$ は、$\gamma^{num}_{jm,t}$ と $\gamma^{den}_{jm,t}$ の差分、$\gamma^{num}_{jm,t}$ と $\gamma^{den}_{jm,t}$ は式 (2) と (4) の分子および分母の事後確率、$\top$ は転置を表す。$\mathbf{U}_{jm}$ と $\mathbf{U}'_{jm}$ は、各々 $\boldsymbol{\mu}_{jm}\boldsymbol{\mu}_{jm}^\top$ と $\boldsymbol{\mu}'_{jm}\boldsymbol{\mu}'^\top_{jm}$ で

*Maximum mutual information discriminative training of acoustic models for system combination, by TACHIOKA, Yuuki (Mitsubishi Electric Corp.), WATANABE, Shinji, LE ROUX, Jonathan, HERSHEY, John R.(MERL)

ある。ガウス分布に固有の学習率 D_{jm} は、Σ'_{jm} を正定値にするためのものである。GMM の混合重み π_{jm} も同時に最適化する [7]。

boosted MMI (bMMI) [7] においては、通常の MMI の評価関数は、誤りの多い仮説の効果を増幅する係数 b を含むように修正される。

$$\mathcal{F}^{\mathrm{b}}_{\lambda}(\omega_r) = \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_\lambda\left(s_r, \mathbf{X}\right)^{\kappa} p_L(\omega_r)}{\sum_{\omega} \sum_{s \in \mathcal{S}_{\omega}} p_\lambda\left(s, \mathbf{X}\right)^{\kappa} p_L(\omega) e^{-bA(s, s_r)}}, \tag{4}$$

$A(s, s_r)$ は、HMM の状態系列 s の状態/音素/単語正解率 (正解 s_r) であり、通常フレーム毎に算出される。

Algorithm 1 は、MMI もしくは bMMI の具体的なアルゴリズムを示している。拡張 Baum-Welch (EBW) は i_{eb} 回繰り返した。

Algorithm 1 Construct MMI or bMMI model

Require: Initial model λ, numerator (ω_r aligned) lattice $\mathcal{A}$, and denominator lattice $\mathcal{L}$ of (1) or (4)

for $i = 1$ **to** i_{eb} **do**

Rescore $\mathcal{A}$ and $\mathcal{L}$ with λ

$\gamma^{num}_{jm,t}$ and $\gamma^{den}_{jm,t}$ $\Leftarrow$posteriors of $\mathcal{A}$ and $\mathcal{L}$, respectively

$\Delta_{jm,t} \Leftarrow \gamma^{num}_{jm,t} - \gamma^{den}_{jm,t}$

$\lambda \Leftarrow$ Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by EBW (3).

end for

Ensure: MMI or bMMI model (λ)

3 補助システムのための識別学習

ここで、補助システムは初期モデル (ML) と元のシステムのモデル (MMI/bMMI) から構築されると仮定する。本節では、補助システムのための識別学習法を、(b)MMI のアルゴリズムを拡張し導出する。この手法は、大規模系列データのための「ブースティング」手法と関係がある。元のシステムの仮説を考慮するために、式 (1) の MMI の評価関数を、以下の評価関数 (5) のように一般化する。この関数は、右辺第 2 項において元のシステムの仮説との相互情報量を最小化すると同時に、右辺第 1 項において正解ラベルに対する相互情報量を最大化する。

$$\mathcal{F}^{\mathrm{c}}_{\lambda_c}(\omega_r, \omega_1) = \underbrace{\ln \left(\frac{P_{\lambda_c}(\omega_r, \mathbf{X})}{\sum_{\omega} P_{\lambda_c}(\omega, \mathbf{X}))} \right)^{1+\alpha}}_{\text{MI to the correct labels}} - \sum_{q=1}^{Q} \underbrace{\ln \left(\frac{P_{\lambda_c}(\omega_{q,1}, \mathbf{X})}{\sum_{\omega} P_{\lambda_c}(\omega, \mathbf{X}))} \right)^{\frac{\alpha}{Q}}}_{\text{MI to the 1-best (base system)}}, \tag{5}$$

λ_c は補助システムの HMM パラメータで最適化の対象である。$\omega_{q,1}$ は $q(1 \le q \le Q)$ 番目の元のシステムの 1 位の仮説である。α は補助システムのためのスケーリング係数で、α が 0 のときは、この評価関数は MMI のそれに一致する。簡単のために、以下は元のシステムの数 Q は 1 とし、指数 q は省略する。評価関数は解析的に以下のように導かれる。

$$\mathcal{F}^{\mathrm{c}}_{\lambda_c}(\omega_r, \omega_1) = \mathcal{F}^{\mathrm{M}}_{\lambda_c}(\omega_r) + \alpha \ln \frac{P_{\lambda_c}(\omega_r, \mathbf{X})}{P_{\lambda_c}(\omega_1, \mathbf{X})}. \tag{6}$$

新しい補助システムの評価関数は、MMI 識別学習の枠組みに入っており、付加的な対数尤度比の項を持つ。

式 (6) の単純な拡張として $\mathcal{F}^{\mathrm{M}}$ を $\mathcal{F}^{\mathrm{b}}$ で置換し、式 (4) の類比で対数尤度比の項に (逆符号の) 増幅係数を付すことで、以下の boosted 版の評価関数を得る。

$$\begin{aligned} \mathcal{F}^{\mathrm{c}}_{\lambda_c}(\omega_r, \omega_1) &= \mathcal{F}^{\mathrm{b}}_{\lambda_c}(\omega_r) \\ &+ \alpha \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p^{\kappa}_\lambda\left(s_r, \mathbf{X}\right) p_L(\omega_r)}{\sum_{s_1 \in \mathcal{S}_{\omega_1}} p^{\kappa}_\lambda\left(s_1, \mathbf{X}\right) p_L(\omega_1) e^{b_1 A(s_1, s_r)}}, \end{aligned} \tag{7}$$

s_1 は元のシステムの 1 位の仮説 ω_1 に対応する HMM 状態系列である。逆符号の増幅係数 b_1 に関しては後に詳述する。

ここで、提案の評価関数 (7) を使った場合の補助システムの更新式について述べる。GMM の平均と共分散の更新式は、元の (b)MMI の更新式 (3) に以下の 2 つの変数に修正を加えるだけでよい ($\gamma^{num}_{jm,t}$ は不変)。

$$\begin{aligned} \gamma^{den\,\prime}_{jm,t} &= \frac{\gamma^{den}_{jm,t} + \alpha \gamma^{1}_{jm,t}}{1+\alpha}, \\ \Delta'_{jm,t} &= (1+\alpha)\left(\gamma^{num}_{jm,t} - \gamma^{den\,\prime}_{jm,t}\right), \end{aligned} \tag{8}$$

$\gamma^{1}_{jm,t}$ は元のシステムの 1 位の仮説に対する当該システムの事後確率である。逆符号の増幅係数 b_1 項の効果を明確化するため、提案法の単一フレームでの分類問題を扱う。これは発話が 1 フレームである場合に対応している。単一フレームの場合は、HMM の状態遷移を考える必要がないため、HMM の初期状態の重みを省略すると、事後確率は音響と言語スコアの積に増幅係数が掛ったものに比例する。指数 t を省略すると、事後確率 γ^{1}_{jm} は以下の如く表される。

$$\gamma^{1}_{jm} = \begin{cases} C^{1}_{jm} e^{b_1} & (j,\ s.t.,\ s_1 = j \in \mathcal{S}_{\omega_r},\ \text{correct}), \\ C^{1}_{jm} & (j,\ s.t.,\ s_1 \neq j \in \mathcal{S}_{\omega_r},\ \text{incorrect}), \end{cases}$$

$$C^{1}_{jm} = \frac{p^{\kappa}_\lambda\left(j, m, \mathbf{x}\right) p_L(\omega_1)}{\sum_{m' j'_1 \in \mathcal{S}_{\omega_1}} p^{\kappa}_\lambda\left(j'_1, m', \mathbf{x}\right) p_L(\omega_1) e^{b_1 A(j'_1, j_r)}},$$

$$p_\lambda\left(j, m, \mathbf{x}\right) = \pi_{jm} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}),$$

$\mathcal{N}$ は単一のガウス分布の確率密度、j_1 と j_r は元のシステムの 1 位の仮説と正解ラベル各々に対応する HMM 状態である。元のシステムが不正解の場合に

Algorithm 2 Construct complementary system model

Require: Initial model λ (e.g., ML), base system models λ_q, numerator (ω_r aligned) lattice $\mathcal{A}$, and denominator lattice $\mathcal{L}$ of (1) or (4)

for $i = 1$ to i_{eb} **do**

Rescore $\mathcal{A}$ and $\mathcal{L}$ with λ

$\gamma_{jm,t}^{num}$ and $\gamma_{jm,t}^{den}$ $\Leftarrow$posteriors are gathered on $\mathcal{A}$ and $\mathcal{L}$, respectively

$\gamma_{jm,t} \Leftarrow -\gamma_{jm,t}^{den} + (1+\alpha)\gamma_{jm,t}^{num}$

for $q = 1$ to Q **do**

Rescore $\mathcal{L}$ with λ_q

$\mathcal{L}_1 \Leftarrow$best path of $\mathcal{L}$

Rescore $\mathcal{L}_1$ with λ

$\gamma_{jm,t}^{1} \Leftarrow$ posteriors are gathered on $\mathcal{L}_1$

$\gamma_{jm,t} \Leftarrow -\frac{\alpha}{Q}\gamma_{jm,t}^{1} + \gamma_{jm,t}$

end for

$\gamma_{jm,t}^{num}, \gamma_{jm,t}^{den}$ $\Leftarrow$positive and negative parts of $\gamma_{jm,t}$

$\lambda \Leftarrow$ Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by EBW ((3))

end for

Ensure: Complementary system model ($\lambda_c \leftarrow \lambda$)

は、正解の場合に比べて係数 $e^{b_1}(>1)$ がない分、γ_{jm}^{1} が小さくなる。γ_{jm}^{1} は、式 (8) において全体の差分統計量 Δ_{jm} から引き去られるので、γ_{jm}^{1} が小さくなると、これらの仮説に対する差分統計量が大きくなる。これは AdaBoost [4] といったブースティングのアルゴリズムと、元のシステムが誤った仮説を与えるデータに対して、より大きな重みを課し、対応するパラメータを大きく更新するという点で類似性がある。

発話が複数のフレームに亘る場合には、前向き・後ろ向きアルゴリズムによって事後確率を算出する必要があり、現フレームの事後確率は前後フレームのそれの影響を受けるため、事後確率と増幅係数の間の直接的な関係性を示すことは難しい。しかしながら、如上の議論を敷衍すれば、事後確率 $\gamma_{jm,t}^{1}$ は元のシステムの文平均正解率に関して増加関数であることが期待でき、複数フレームの場合でも、提案法はブースティングと関連があるといえる。Algorithm 2 には、提案の補助システムの音響モデルを更新するためのアルゴリズムを示す。

4 認識実験による評価

4.1 実験条件

提案法を第 2 回 CHiME チャレンジ (トラック 2) を用いて評価した。残響・非定常高騒音環境における中程度語彙の認識タスク (Wall Street Journal (WSJ0)) である [8]。語彙サイズは 5k(basic) とし、Kaldi[9] を使った。学習セットは 83 話者の 7,138 発話 (si84)、評価セット (si_et_05) は 12 話者の 330 発話 (Nov'92)、開発セット (si_dt_05) は 10 話者の 409 発話からなる。音響モデルは学習セットを用いて学習し、音響スケール κ は開発セットで調整した。騒音は他の話者の妨害、家庭内の騒音、音楽といった非定常性のもので、孤立発話に対して SNR $= \{-6, -3, 0, 3, 6, 9\}$dB で重畳されている。事前分布に基づくバイナリマスクにより騒音抑圧された音声を使って評価した [10]。

音響特徴量と特徴量変換の設定に関して述べる [11]。ベースの音響特徴量は MFCC と PLP (0-12 order MFCCs/PLPs $+ \Delta + \Delta\Delta$) である。加えて、特徴量変換手法 (線形判別分析 (LDA)、最尤線形変換 (MLLT)) と話者適応手法 (話者適応学習 (SAT)、特徴量空間最尤線形回帰 (fMLLR)) を使った。音響モデルの学習手順と特徴量変換の設定に関しては、文献 [10, 11] を参照されたい。HMM の状態数は 2,500 で、ガウス分布の総数は 15,000 である。ランダムフォレスト的な効果を期待して、MFCC と PLP 特徴量で木構造を 2 種構築した。変数 α と b_1 は開発セットで調整し、0.75 と 0.3 に設定した。EBW は 4 回繰り返した。

4.2 結果と考察

Table 1 には、MFCC と PLP 特徴量での開発セット (dt) と評価セット (et) の単語誤り率 (WER) を示す。上段、上中段、下中段、下段はそれぞれ単一システム (S_1-S_4)、いくつかのシステムに対する旧来の ROVER(R_1,R_2)、提案法による補助システム (P_1,P_2)、提案法を含む ROVER(RP_1-RP_3) の結果である。開発セットについては、PLP 特徴量での性能は MFCC 特徴量のそれより低いが、bMMI(MFCC、PLP) とシステム統合することで、WER が 2.2% ($S_2 \rightarrow R_2$) 向上した。これは、異なる特徴量を使ったシステムの統合が有効であることを示している。補助システム ($bMMI_c$) の性能は、MFCC の場合、ML より低く、PLP の場合、ML と bMMI の中間であるが、bMMI と $bMMI_c$ の統合により、bMMI に比べて 0.3%認識率が向上した (MFCC and PLP, $S_2 \rightarrow RP_1$)。さらに、ML、bMMI と $bMMI_c$ との統合により、ML と bMMI を統合した場合に比べて、認識率は 0.3%向上した ($R_1 \rightarrow RP_2$)。MFCC と PLP 特徴量を併用し、$bMMI_c$ を ML と bMMI と組み合わせることで、WER は 0.3%改善し、最も高い性能を得た ($R_2 \rightarrow RP_4$)。これにより提案法の有効性が示された。

Table 2 には、MFCC と PLP 特徴量に対して、特徴量変換 (LDA+MLLT) と話者適応 (SAT+fMLLR) を施した場合の WER を示す。傾向は上述のものと同

Table 1 Average WER[%]. (MFCC/PLP) (upper: conventional Single systems, upper middle: ROVER among conventional multiple systems, lower middle: single Proposed systems, and lower: ROVER including Proposed system)

ID	MFCC			PLP			WER	
	ML	bMMI	prop	ML	bMMI	prop	dt	et
S_1	✓						46.9	42.5
S_2		✓					45.6	40.7
S_3				✓			48.2	44.4
S_4					✓		46.7	42.1
R_1	✓	✓					45.1	40.0
R_2	✓	✓		✓	✓		**43.4**	**38.6**
P_1			✓				47.0	42.9
P_2						✓	47.5	43.7
RP_1		✓	✓				45.3	40.6
RP_2	✓	✓	✓				44.8	40.2
RP_3		✓	✓		✓	✓	43.3	38.7
RP_4	✓	✓	✓	✓	✓	✓	**43.1**	**38.3**

Table 2 Average WER[%] for isolated speech. (MFCC/PLP with LDA+MLLT & SAT+fMLLR)

ID	MFCC			PLP			WER	
	ML	bMMI	prop	ML	bMMI	prop	dt	et
S_1	✓						38.2	32.2
S_2		✓					35.9	29.5
S_3				✓			38.1	32.2
S_4					✓		36.4	30.0
R_1	✓	✓					36.1	29.2
R_2	✓	✓		✓	✓		**35.0**	**28.0**
P_1			✓				36.2	30.1
P_2						✓	36.7	30.5
RP_1		✓	✓				35.7	28.8
RP_2	✓	✓	✓				35.6	28.8
RP_3		✓	✓		✓	✓	**34.4**	**27.4**
RP_4	✓	✓	✓	✓	✓	✓	34.6	27.5

様である。この場合には、ML の性能が顕著に bMMI よりも低かったため、ML と統合することで認識率は低下した。システム数は 2 つで同じであっても、ML と bMMI の統合 (R_1) は、bMMI と $bMMI_c$ の統合 (RP_1) に比べて認識率が低かった。加えて、提案の $bMMI_c$ の性能は中程度であり、システム統合を効果的なものにしており、提案法は WER を 0.6%改善した ($R_2 \rightarrow RP_3$)。これは提案法の性能が調整できることによる利点である。

表には同時に、評価セット (et) の WER を示している。提案法は評価セット (et) に対しても有効で、WER を 0.3% (MFCC/PLP, $R_2 \rightarrow RP_4$)、0.6% (変換特徴量, $R_2 \rightarrow RP_3$) 改善した。

Table 3 と 4 は WER を SNR の観点から R_1 と RP_2 のペアと R_2 と RP_3 のペアで比較したものである。ほとんどすべての場合で、提案法は WER を改善した。就中、この傾向は SNR が低い場合に顕著であり、最大で 1%認識率を改善した。よって、提案法は環境変化に対しても頑健であるといえる。

Table 3 WER[%] in terms of SNR[dB] (**si_et_05**). (MFCC/PLP)

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
R_1	56.2	49.2	42.6	34.7	30.9	26.6	40.0
R_2	55.8	47.9	40.8	33.4	29.0	**25.0**	38.6
RP_2	55.8	49.0	42.6	35.2	31.5	27.4	40.2
RP_3	54.6	47.9	40.9	33.7	29.2	26.0	38.7
RP_4	**54.8**	**47.4**	**40.3**	**33.2**	**29.0**	25.3	**38.3**

Table 4 WER[%] in terms of SNR[dB] (**si_et_05**). (MFCC/PLP with LDA+MLLT & SAT+fMLLR)

	−6dB	−3dB	0dB	3dB	6dB	9dB	Avg.
R_1	47.1	38.1	31.0	23.7	19.1	16.6	29.3
R_2	45.9	36.6	29.2	22.4	18.6	15.4	28.0
RP_2	46.4	38.0	30.3	23.2	18.7	16.2	28.8
RP_3	**44.6**	36.0	29.0	**22.2**	**17.7**	**14.6**	**27.4**
RP_4	44.8	**35.8**	**28.9**	22.3	18.1	15.1	27.5

5 まとめと今後の課題

システム統合のための音響モデルの識別学習法を提案した。提案法は識別学習法の枠組みの中で補助システムを構築することができ、残響・高騒音下の音声認識タスクで単語誤り率の低減が見られた。今後の課題としては、提案法を特徴量空間識別学習や識別的言語モデリングといった種々の識別学習法 [1] と統合することがあげられる。

参考文献

[1] M.J.F. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured discriminative models for speech recognition: An overview," *IEEE Signal Processing Mag.*, **29**, 70–81 (2012. 11).

[2] J.G. Fiscus, "A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER)," in *Proc. of ASRU*, pp.347–354 (1997).

[3] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *Proc. of ICASSP*, 197–200 (2005).

[4] Y. Freund and R. Schapire, "A dicision-theoretic generalisation of online learning and an application to boosting," *J. Computer and System Sci.*, **55**, 119–139 (1997).

[5] H. Tang, M. Hasegawa-Johnson, and T.S. Huang, "Toward robust learning of the Gaussian mixture state emission densities for hidden Markov models," in *Proc. of ICASSP*, 5242–5245 (2010).

[6] F. Diehl and P. Woodland, "Complementary phone error training," in *Proc. of INTERSPEECH* (2012).

[7] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. of ICASSP*, pp.4057–4060 (2008).

[8] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. of ICASSP*, 126–130 (2013).

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, 1–4 (2011).

[10] Y. Tachioka, S. Watanabe, J. Le Roux, and J.R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proc. of the 2nd CHiME Workshop*, pp.19–24 (2013).

[11] Y. Tachioka, S. Watanabe, and J.R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proc. of ICASSP*, pp.6935–6939 (2013).