騒音環境下音声認識に対する識別的アプローチの有効性 第2回CHiMEチャレンジ*

◎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

第2回CHiME チャレンジ (以下 CHiME) は、高騒音下音声認識タスクである [1]。近年、音声認識のモデル学習法は、最大尤度法 (ML) から識別学習へと移り、加えて種々の特徴量変換が提案され、その有効性が示されてきた。これら最新の音声認識の手法が、クリーン音声に有効なことはよく知られているが、残響・騒音環境下ではさらなる検討が必要である。

そこで、特徴量変換と識別学習の残響・騒音環境下音声認識での有効性を、CHiME の Track2(中規模語彙の連続音声認識) を使って評価する。また、Kaldi ツールキットを使った公開ツールを CHiME のベンチマークとして構築する。加えて、提案法である音源方向の事前分布に基づくバイナリマスク、拡張識別的特徴量変換、識別的言語モデリングとベイズリスク最小化デコーディングの連結に関しても検討する [2, 3]。

2 システムの概観

全体システムの概要図を Fig. 1 に示す。提案法は3つの要素からなっている。1つめは、騒音抑圧部で、事前分布に基づくバイナリマスクにより、方向性の妨害音を抑圧する (3節)。2つめは、特徴量変換部で、一般の特徴量変換と、識別的特徴量変換 (4.2節) である。3つめは、デコード部で、音声認識には相互情報量最大化法 (MMI) と boosted MMI による識別的音響モデルを使う (4.1節)。N 位までの音声認識結果を、識別的言語モデリング (DLM) により並び替え (4.3節)、DLM の1位とラティス上の仮説を用いてベイズリスク最小化デコーディング (MBR) を行う (4.4節)。

3 事前分布に基づくバイナリマスク

CHiME では 2ch のデータが提供され、目的話者はマイクの正面とされる。到来時間差に基づくバイナリマスク [4] は、マイクの個数が少ない環境ではビームフォーミングよりも効果的である。正面方向に対しては理想的には目的話者からの信号の位相差は零になるので、マイク間の位相差がゼロから離れた時間周波数ビンでは目的話者以外の妨害音のエネルギーが大きいと考えられる。しかし、残響とダミーヘッドの回折により、位相差が零にならないこともある。例として、残響音声 (騒音なし) に対する 250 Hz と 1 kHz の場合の位相差のヒストグラムを Fig. 2 に示す。250 Hz の場合はヒストグラムはおおむね零対称で分散も小さいが、1 kHz の場合には平均は 0 からずれ、分散も大きい。騒音や残響の影響は各周波数ビンにより大きく異なるので、物理情報に基づく単純なバイナリマスクでは音声認識性能はベースラインよりも低下した。このような位相差のずれや分散を考慮するには、大量の学習 (残響音声) データから各周波数ビンでのそれらのパターンを推定する統計的モデルが必要である。本報では、上記により推定したパターンを事前分布として用いるバイナリマスク推定法を提案する。周波数ビン ω、時間フレーム t の時間周波数ビンの位相差 $\theta_{\omega,t}$ は $X^L_{\omega,t}/X^R_{\omega,t} = A_{\omega,t}\mathrm{e}^{\jmath\theta_{\omega,t}}$ のように表される。ここで、$\jmath$ は虚数単位、$A_{\omega,t}$ は正の実数、X^L と X^R はそれぞれ左チャネルと右チャネルの短時間フーリエ変換である。通常のバイナリマスクでは、マスク W は以下のように閾値を用いて設定される。

$$W_{\omega,t} = \begin{cases} \epsilon & \text{if } |\theta_{\omega,t}| > \theta_c, \\ 1 & \text{if } |\theta_{\omega,t}| \le \theta_c, \end{cases}$$

ϵ は十分小さい定数 (スペクトルの非連続性を避けるため 0 でない定数)、θ_c は事前に定めておく閾値であり、事前分布に基づくバイナリマスクでは、マスク W' は周波数依存のヒストグラムを正規化した事前分布 q_ω を用いて以下のように定める。

$$W'_{\omega,t} = \begin{cases} \epsilon & \text{if } q_\omega(\theta_{\omega,t})/\bar{q}_\omega < q_c, \\ (q_\omega(\theta_{\omega,t})/\bar{q}_\omega)^\alpha & \text{if } q_\omega(\theta_{\omega,t})/\bar{q}_\omega \ge q_c, \end{cases}$$

$\bar{q}_\omega$ は $\max_\theta q_\omega(\theta)$ で、q_c は閾値、α は歪み係数。

4 識別的手法に基づく後段の処理

4.1 識別学習

識別学習は、正解ラベルと認識結果の情報からベイズリスク (以下は音声データが与えられた際の単語系列の事後確率で設計される) を最小化する。本報では音素正解率の重みである増幅係数 b を導入した boosted MMI (bMMI) [5] を用いる。目的関数は

$$\mathcal{F}_{\mathrm{bMMI}}(\lambda) = \log \frac{p_\lambda(\mathbf{x}_r|\mathcal{H}_{s_r})^\kappa p_L(s_r)}{\sum_s p_\lambda(\mathbf{x}_r|\mathcal{H}_s)^\kappa p_L(s)\mathrm{e}^{-bA(s,s_r)}},$$

*Effectiveness of discriminative approaches for speech recognition under noisy environments on the 2nd CHiME Challenge, by TACHIOKA, Yuuki (Mitsubishi Electric Corp.), WATANABE, Shinji, LE ROUX, Jonathan, HERSHEY, John R.(MERL)

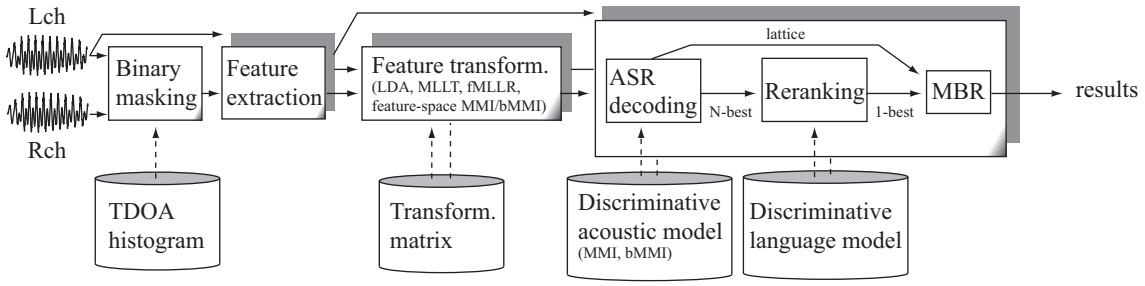

Fig. 1　Schematic diagram of the proposed system.

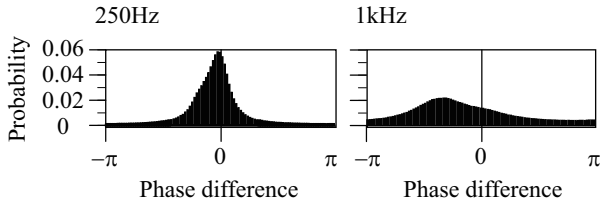

Fig. 2　Histogram of phase differences for two frequency bins.

で与えられる。$\mathbf{x}_r$ は r 番目の発話の特徴量系列であり、r に関する和は省略している。音響モデルパラメータ λ は拡張 Baum-Welch で最適化される。$\mathcal{H}_{s_r}$ と $\mathcal{H}_s$ は、それぞれ正解ラベル s_r と認識結果 s に対応する HMM 系列である。p_λ は音響モデル尤度、κ は音響スケール、p_L は言語モデル尤度、$A(s, s_r)$ は、正解 s_r に対する s の音素正解率である。実験では MMI($b = 0$ の場合) と bMMI の性能を、ML と比較する。

4.2　識別的特徴量変換とその拡張

識別学習の指標に基づく特徴量変換は、特徴量空間 MMI(f-MMI[6]) と呼ばれ、豊富な情報を持つ高次元特徴量 $\mathbf{h}$(J 次元) を低次元特徴量 (I 次元) に写像する行列 $\mathbf{M}$ を学習し、変換特徴量 $\mathbf{y}$ を得る。

$$\mathbf{y} = \mathbf{x} + \mathbf{M}\mathbf{h},$$

$\mathbf{h}$ は通常 $\mathbf{x}$ を入力とする GMM の事後確率値からなり、$\mathbf{x}$ に依存している。本報では、f-MMI と f-boosted MMI(f-bMMI) の有効性を検討する。

騒音下では、タンデム手法のように異なる種類の特徴量の併用が有効なので、$\mathbf{h}$ に特徴量 $\mathbf{h}'$ を加え、新しい特徴量 $\mathbf{y}'$ を得る方法 (拡張 f-MMI) を提案する。

$$\mathbf{y}' = \mathbf{x} + [\mathbf{M}\ \mathbf{M}'] \begin{bmatrix} \mathbf{h} \\ \mathbf{h}' \end{bmatrix}.$$

$\mathbf{M}$,$\mathbf{M}'$ の結合行列は、$\mathcal{F}$ を最大にするよう最適化する。

$$\mathcal{F}_{\text{af-MMI}}([\mathbf{M}\ \mathbf{M}']) = \log \frac{p_\lambda(\{\mathbf{y}'\}_r|\mathcal{H}_{s_r})^\kappa p_L(s_r)}{\sum_s p_\lambda(\{\mathbf{y}'\}_r|\mathcal{H}_s)^\kappa p_L(s)}.$$

拡張 f-MMI では、$\mathbf{h}'$ の選択により幅広い特徴量変換を扱える。(例えば回線歪みに頑健な PLP 特徴量、バイナリマスク値、音声区間検出の際の情報)

4.3　識別的言語モデリング (DLM)

DLM はデコーダーの出力仮説に現れる誤りパターンを学習し、誤りを削減するように仮説のスコアを補正する [7]。スコアは、次式のようにデコーダーの仮説 $\mathcal{H}_s$ から得られる素性ベクトル $\phi(\mathcal{H}_s)$(例えば N-gram の出現個数) と重みベクトル $\mathbf{w}$ の内積で補正し、N-best リストを並び替え、新しい仮説 s' を得る。

$$s' = \arg\max_s \left[w_0 \ln p_\lambda(\{\mathbf{y}\}_r|\mathcal{H}_s)^\kappa p_L(s) + \mathbf{w}^\top \phi(\mathcal{H}_s) \right],$$

$\top$ は転置を表す。$\mathbf{w}$ は、発話ごとオンライン学習する。パーセプトロン法では、学習則は $\mathbf{w} = \mathbf{w} + (\phi(\mathcal{H}_{s_o}) - \phi(\mathcal{H}_s))$ のようになる。平均化パーセプトロン法では、全発話で集積した $\mathbf{w}$ を平均する。$\mathcal{H}_{s_o}$ は、N-best リスト中の単語誤り率最低の仮説 (オラクル) である。

4.4　DLM とベイズリスク最小化 (MBR) デコーディングの接続

MBR デコーディングでは、1 位の仮説とラティス上の仮説との間の編集距離 (ベイズリスクに関連) を最小にするように単語系列を選ぶ [8]。

$$\mathcal{H}_{\hat{s}} = \arg\min_{s'} \sum_{s \in \mathcal{L}} p_\lambda(\{\mathbf{y}\}_r|\mathcal{H}_s)^\kappa p_L(s) L(\mathcal{H}_s, \mathcal{H}_{s'}),$$

$L(\mathcal{H}_s, \mathcal{H}_{s'})$ はラティス中の仮説 $\mathcal{H}_s$ と対象の仮説である $\mathcal{H}_{s'}$ の間の編集距離である。編集距離は、(ϵ を含む) シンボルが単語列 $\mathcal{H}_{\hat{s}}$ 中におかれた時の確率から計算され、目的関数は反復的に更新される。

N-best リストの並び替えを行う DLM と、ラティスに基づく MBR を効率的に結合する方法を提案する。MBR はラティス上の 1 位の仮説を初期値として、ラティス中の系列のアライメントをとるが、これには初期値依存性があるので、この初期値の代わりに DLM により並び替えられたリストの 1 位で置き換える。

5　実験の設定

5.1　タスクの記述

第 2 回 CHiME[1] の Track2 を評価した。これは残響騒音下の音声認識タスクであり、Track2 は残響騒音下における中程度の語彙サイズで、発話は Wall Street

Journal のデータベース (WSJ0) から取られている。学習セット (si_tr_s) は、83 話者 (7138 発話)(si84)、評価セットは (si_et_05) は、12 話者 (330 発話)(Nov'92)、開発セットは (si_dt_05) は、10 話者 (409 発話) である。音響モデルは si_tr_s で学習し、言語モデル重みといったパラメータは開発セットで調整した。言語モデルのサイズは 5 k である。データベースは、2 種類のデータからなる。"reverberated" は、クリーン発話に居間におけるマイクの 2 m 前方話者に対応するバイノーラルの室内のインパルス応答を畳みこんだものである。"isolated" は、家庭内の騒音を、信号対雑音比 (SNR) が −6、−3、0、3、6、9 dB になるように、正規化せずに重畳したものである。騒音は他の話者の発話や音楽といった非定常性のものである。

5.2 特徴量抽出および特徴量変換

音響的な特徴量と特徴量変換について述べる。基本特徴量は 1-13 次の MFCC/PLP とその Δ,ΔΔ である。これを線形判別分析 (LDA) と最尤線形変換 (MLLT) により変換する。LDA はあるクラスの特徴量が他のクラスの特徴量に対して、識別性が高くなるよう変換行列を決定する。連続 9 フレームの 13 次元の静的特徴量を結合した 117 次元の特徴量を、LDA により 40 次元に圧縮する。LDA のクラスはトライフォンの状態 (2500 状態) とした。特徴量の次元間の相関は、対角共分散モデルにとって問題となる。次元間の相関を低減するためには MLLT が広く使われる。

さらに、話者間の特徴量のばらつきが大きいと音響モデルの性能が低下する。この問題に対処するのに、話者適応化学習 (SAT) と特徴量空間最尤線形変換 (fMLLR) がよく用いられる。SAT では、学習データを fMLLR で標準話者空間に変換し、話者間の特徴量のばらつきを低減して学習を行う。本報では、LDA と MLLT、SAT と fMLLR の有効性を検討する。

5.3 識別的手法

識別的特徴量変換 (4.2 節) では、400 個のガウス分布を使い、特徴量は、その事後確率とオフセット特徴量 (39 次元の MFCC) の計 40 次元から計算される。特徴量は連続 9 フレーム、コンテクスト拡張されるので、特徴量 $\mathbf{h}_t$ の次元は、$400 \times 40 \times 9$ である。上位 2 つの事後確率を持つ特徴量だけを選択する。

本報では Deep Neural Networks(DNN) の予備的検討を行った。Kaldi の CPU バージョンを使い、3 つの隠れ層と 1M パラメータを学習した。学習率は初期は 0.01 とし、終盤には 0.001 になるように低減した。

DLM の重み $\mathbf{w}$ は、学習データの 100 位までの認識候補から学習した。学習時の元のスコアの重み w_0 は 20 とした。認識時は $\mathbf{w}$ を用いて、結果を並び替える。その際の w_0 は 13 とした。$\mathbf{w}$ は平均化パーセプトロン法 (3 回繰り返し) により求めた。素性は、1,2,3 グラムの出現個数とした。

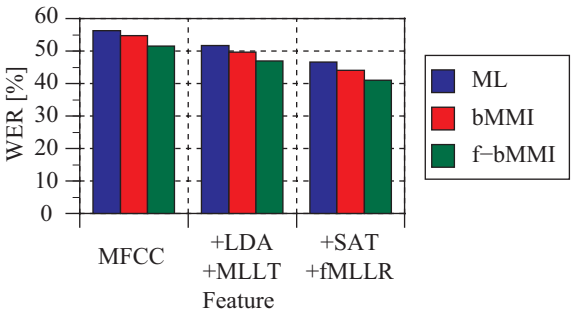

Fig. 3 Effectiveness of feature transformation (FT) and discriminative training (DT) without noise suppression (**si_dt_05**).

5.4 実験手順

以下に上述の設定に基づく実験の手順を述べる。初めにクリーンの音響モデルを学習した。モノフォンは、無音のモデル ("sil") を含み 40 とした。トライフォンでは、状態数は 2500 とし、ガウス分布の全体数は 15000 とした。次に、クリーンモデルによるアライメントと木構造を使って、"reverberated" セットにより残響音響モデルを学習した。その後、"isolated" セットにより、騒音音響モデルをマルチコンディション学習で学習した。実験に用いた設定は、Kaldi に付属の WSJ のチュートリアルを参考に決定した。

6 結果と考察

6.1 騒音抑圧をしない場合 (開発セット)

MFCC における識別学習による ML からの単語誤り率 (WER) の向上を Fig. 3(左) に示す。以下特に断らない限り、SNR に関して平均した WER を示している。識別学習によって誤認識が効果的に修正できたと考えられる。(f-)boosted MMI は、(f-)MMI と比べて WER を 1%改善した。増幅係数はおおよそ 0.1〜0.2 が最適であり、本実験では 0.1 とした。f-bMMI は WER を 3%改善した。特徴量空間を目的話者に適応させることで、騒音の影響が低減され、性能改善が図られたと考えられる。なお本実験において、識別学習の分母のラティスは、ML で生成した。

次に、MFCC を LDA と MLLT で変換した。Fig. 3 (中) に WER を示す。LDA により、他の音素との混同性を低減させるような特徴量が得られたと考えられる。CHiME は他者の多数の妨害発話を含んでおり、この種の騒音には LDA が適しているといえる。音声が混合し複数音素が同一フレームにある場合にも、これらの音素を識別可能なモデルが学習できると考え

Table 1　Effectivenss of augmented FT with PLP (P) features without noise suppression (**si_dt_05**).

	ML(M)	ML(P)	f-MMI	+P
WER[%]	56.37	57.35	52.79	**52.62**

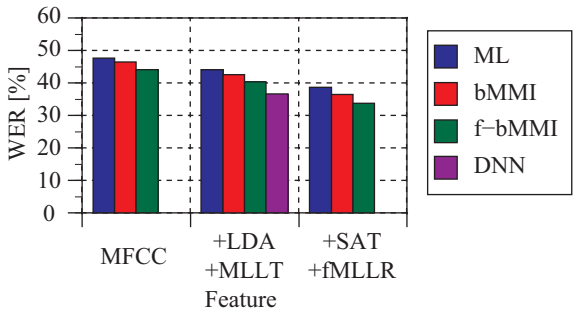

Fig. 4　Effectiveness of FT and DT with noise suppression (**si_dt_05**).

られるからである。長いコンテキストも非定常騒音と残響の影響を低減するのに有効である。騒音により特徴量の次元間の相関が高まるが、MLLT により相関が低減されたと考えられる。なお識別学習用の分母のラティスは同様に ML で再生成した。

さらに、SAT と fMLLR を上述の変換に加えた。Fig. 3 (右) に、WER を示す。学習データの量が限られているため、標準話者空間への変換は実質的な学習データの増加をもたらし、音響モデルの推定精度を向上させる。加えて、fMLLR の話者適応により、騒音の影響を低減できた。この場合特徴量変換と識別学習は相補的に働いており、ともに有効である。

Table 1 は、拡張 f-MMI において、ML と拡張 f-MMI において Eq. (4.2) の特徴量 $\mathbf{h}'_t$ として PLP (13 次元) を用いた結果を示している。ML では、絶対値で 1%、PLP の方が MFCC より性能が低いが、PLP を f-MMI に加えると、性能が向上した。特徴量 $\mathbf{h}_t$ とは違う情報を含む特徴量を併用することが有効である。

6.2　騒音抑圧した場合 (開発セット)

Fig. 4 は、バイナリマスキングを行った後の WER である。すべての SNR で、絶対値で 7%から 9%向上した ($\alpha = 0.25$)。方向性雑音はある程度除けるが、音楽のような拡散性雑音は依然として残っている。特徴量変換と識別学習は有効であった。DNN は、bMMI と f-bMMI を最大で 2.8%上回り、話者適応を行った場合と同程度の性能であった。残響騒音下の音声認識における DNN の潜在的な有効性を示している。

Table 2 に f-bMMI の結果に対して、DLM と MBR を適用した場合 (**dt**) を示す。DLM により WER は平均で 0.23% (9dB の場合には 0.77%) 向上した。MBR により、ML (MFCC) から 0.77%改善し、f-bMMI (LDA+MLLT と SAT+fMLLR) に対しても 0.52%改善した。(SNR によらず効果有。) さらに DLM と MBR の結合 (4.4 節) により性能が向上した。

Table 2　Effectiveness of the DLM, MBR, and their combination (**si_{dt,et}_05**).

	f-bMMI	+DLM	+MBR	+both
WER[%] (**dt**)	33.71	33.48	33.19	**33.11**
WER[%] (**et**)	27.61	27.14	27.10	**26.86**

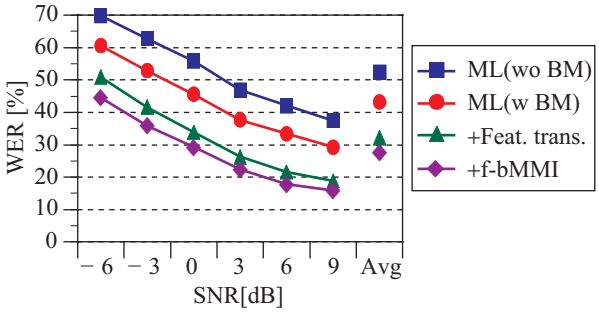

Fig. 5　Effectiveness of FT and DT (**si_et_05**). Comparison of ML (MFCC) without and with binary masking(BM), ML with FT (LDA+MLLT and SAT+fMLLR), and f-bMMI with FT.

6.3　評価セット

Fig. 5 は、評価セットの WER である。ML は騒音抑圧あり・なしで比較した。騒音抑圧により 10%程度 WER が低減した。変換特徴量での ML と f-bMMI の結果も示している。f-bMMI は騒音抑圧後の ML(43.23%) に比べて、すべての SNR で 15%程度の WER の低減がみられ、9dB では誤りが半減した。特徴量変換と識別学習が残響騒音環境下において有効であり、最良では WER 26.86%を達成した。Table 2(**et**) に示す通り、f-bMMI に加えて DLM、MBR とそれらの結合により WER がさらに低減した。

7　まとめ

残響・騒音環境で、最新の音声認識手法 (特徴量変換と識別学習) の有効性を確認した。いくつかの提案法を含む本報のシステムは、第 2 回 CHiME チャレンジ Track 2 において首位を獲得した。

参考文献

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

[2] Y. Tachioka *et al.*, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," *ICASSP*, 6935–6939 (2013).

[3] Y. Tachioka *et al.*, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," *The 2nd International Workshop on Machine Listening in Multisource Environments*, 19–24 (2013).

[4] H. Sawada *et al.*, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. on ASLP*, **19**, 516–527 (2011).

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

[6] D. Povey *et al.*, "fMPE: Discriminatively trained features for speech recognition," *ICASSP*, 961–964 (2005).

[7] B. Roark *et al.*, "Discriminative language modeling with conditional random fields and the perceptron algorithm," *ACL*, 47–54 (2004).

[8] H. Xu *et al.*, "An improved consensus-like method for minimum Bayes risk decoding and lattice combination," *ICASSP*, 4938–4941 (2010).