Vol.2015-SLP-105 No.6
2015/2/27

残響除去手法とシステム統合手法の種々の残響環境に対する有効性: REVERBチャレンジ

太刀岡 勇気[1,a)]　成田 知宏[1]　渡部 晋治[2]

概要：昨年公開された REVERB チャレンジには、残響音声の認識タスクが含まれる。本報では、ガウス混合モデル、部分空間ガウス混合モデルや深層回路網といった音響モデルの識別学習や、種々の特徴量変換手法といった最新の音声認識手法に焦点をあてた。その前段として、提案の単一チャンネルによる残響時間推定に基づく残響除去手法や、8 チャンネルのビームフォーミングにより直接音を間接音に比べて強調する手法に関して検討した。加えて、REVERB チャレンジでは種々の環境を扱う必要があり、環境ごとに最良のシステムが異なるため、異なる特徴量や異なる種類のシステムを統合する手法に関しても検討を加えた。さらに、補助システムを意図的に構築することで、システム統合の性能をより向上させる提案の識別学習法の有効性も検証した。実験によりこれらの手法の有効性が示され、REVERB チャレンジのシミュレーション・実測それぞれのデータに対して平均 6.76%、18.60%の単語誤り率を達成した。これはベースラインに比して、相対値で 68.8%、61.5%の向上に相当する。

キーワード：残響、残響除去、識別学習、特徴量変換、システム統合、REVERB チャレンジ

Effectiveness of dereverberation techniques and system combination approach for various reverberant environments: REVERB challenge

Tachioka Yuuki[1,a)]　Narita Tomohiro[1]　Watanabe Shinji[2]

Abstract: The recently released REVERB challenge includes a reverberant speech recognition task. This paper focuses on state-of-the-art ASR techniques such as discriminative training of acoustic models including Gaussian mixture model, sub-space Gaussian mixture model, and deep neural networks, and various feature transformations after the proposed single channel dereverberation method with reverberation time estimation and multi-channel beamforming that enhances direct sound compared with the reflected sound. In addition, because it is necessary to handle these various environments in the challenge and the best performing system is different from environment to environment, we perform a system combination approach using different feature and different types of systems. Moreover, we use our discriminative training technique for system combination that improves system combination by making systems complementary. Experiments show the effectiveness of these approaches, reaching 6.76% and 18.60% word error rate on the REVERB simulated and real test sets, which are 68.8% and 61.5% relative improvements over the baseline.

Keywords: Reverberation, Dereverberation, Discriminative training, Feature transformation, System combination, REVERB challenge

1. 序論

REVERB チャレンジは、残響信号処理の評価のため昨年公開された [1]。本報では、このチャレンジに含まれる中程度語彙の音声認識タスクを扱う。これは、残響環境下で

[1] 三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Kanagawa, 247-8501, Japan
[2] Mitsubishi Electric Research Laboratories
[a)] Tachioka.Yuki@eb.MitsubishiElectric.co.jp

の音声認識性能の評価を目的とする。残響環境下では、音声認識の前処理である音声強調が重要であり、すでに、単一チャンネルの残響除去法を提案している [2]。これは、初めに残響の程度を特徴づける最も重要なパラメータである残響時間を推定し、それに基づき残響の引き去り量を加減する手法である。これに加えて、REVERB チャレンジで提供されている 8 チャンネルのデータを有効に活用するために、到来方向推定を行った [3], [4] 後に、ビームフォーミング [5] により音声を強調する手法も合わせて検討する。

近年の音声認識性能の向上は、識別学習 [20] や種々の特徴量変換手法 [6], [7], [8], [9], [10] によるところが大きい。我々はすでに、騒音環境下での識別的手法の有効性を示している [11]。しかしながら、残響により音声認識性能が低下するため、残響への対処は騒音への対処と同じく重要である。REVERB チャレンジでは、8 種の異なる残響環境が提供されている。本報では、最先端の音声認識の手法が多様な残響騒音環境下において有効に働くことを示す。

識別的手法に加えて、いくつかの特徴量変換手法を扱う。特徴量変換は元の特徴量を、線形変換に基づき新しい特徴量に変換する手法 [*1] と識別的非線形特徴量変換 [9] がある。LDA は長いコンテキストを [*2] 扱うので、特徴量の動的特徴をモデル化でき、残響の影響を低減できると考えられる。MLLT は状態に紐づけられた特徴量の相関を低減するような特徴量変換を実現する。未知の条件に適応させることで音声認識の性能を向上させるためには、モデル適応が有効である。本報では、話者適応学習 (SAT) [8] と基底特徴量空間最尤線形回帰 (basis fMLLR) [12] を用いた。

本報では、近年注目されている深層回路網 (DNN) [10] の有効性も検証する。DNN は、特徴量変換と音響モデリングを同時に最適化できる。騒音環境で有望な結果を示した [11] が、残響環境においてもその有効性を検証する。

上述の検討は、単一の音声認識システムが対象だが、複数のシステムの認識結果を統合することが、音声認識性能を向上させるためには有効である [13]。環境ごとに最適な音声認識システムが異なる場合、各々の結果を統合することで、より性能を向上させられる。本報では、ラティスに基づく識別学習の枠組みに依拠した、意図的に補助的なシステムを構築する手法の有効性も検証する [14]。システムの構築には、Kaldi ツールキット [15] を用いた。

2. 提案システムの概観

図 1 に、提案システムの概要図を示す。提案システムは 3 つの要素から構成されている。1 つ目の要素は、3 節に示す音声強調である。1) 到来方向推定後に、多チャンネル遅延和ビームフォーマーにより、反射音に比べて直接音を強調する手法、2) 残響時間を予測し、単一チャンネルの残響除去手法により後期残響成分を除去した手法、3) 正規化最小 2 乗誤差法 (NLMS) により短時間の歪を除去した手法より成る。2 つ目の要素は、4.2 節に示す特徴量変換である。これはいくつかの特徴量レベルでの変換 (LDA、MLLT と基底 fMLLR) と識別的特徴量変換より成る。これらの手法では、メル周波数ケプストラム係数 (MFCC) と知覚的線形予測 (PLP) の 2 種の特徴量を用いる。2 つの特徴量を使うことで、システム統合に使う補助システムが異なる傾向の仮説を出力することを期待できる。3 つ目の要素は、音声認識である。音響モデルに対して、マージン付きの識別学習 (ブーステッド相互情報量最大化法 (boosted MMI), cf. 4.1 節および 4.3 節) を適用し、3 種のシステム (GMM、SGMM と DNN) を構築した。さらに、4.4 節に示す識別的に学習された補助システムを使ったシステム統合手法を提案する。システムの出力結果は、ROVER により統合した。

[*1] 線形判別分析 (LDA) [6]、最尤線形変換 (MLLT) [7]
[*2] 例えば連続 9 フレームをコンテキスト拡張する。

3. 音声強調部

3.1 CSP 法を用いた到来音方向推定に基づく遅延和ビームフォーマー

音源からの直接音を強調するために、周波数領域での遅延和ビームフォーマー [5] を適用した。強調されたスペクトル $\tilde{\mathbf{y}}_t$ は、m 番目のマイクにより観測された短時間フーリエ変換 (STFT) によるスペクトル $\mathbf{x}_t(m)$ の和として得られる。

$$\tilde{\mathbf{y}}_t = \sum_m \mathbf{x}_t(m) \odot \exp(-\jmath\boldsymbol{\omega}\tau_{1,m}) \qquad (1)$$

t は現在フレームの番号、$\odot$ は要素ごとの積、$\boldsymbol{\omega}$ は角周波数の組である。1 番目のマイク基準の m 番目のマイクの到達時間遅れ $\tau_{1,m}$ は到来方向に関連している。時間遅れは、2 マイク間の相互パワースペクトルを用いる相互スペクトル位相 (CSP) 分析 [3] により推定される。

$$\tau_{1,m} = \arg\max \mathcal{S}^{-1}\left[\frac{\mathbf{x}_t(1) \odot \mathbf{x}_t(m)^*}{|\mathbf{x}_t(1)||\mathbf{x}_t(m)|}\right] \qquad (2)$$

$\mathcal{S}$ は、STFT による演算であり、*は複素共役を表す。CSP 法の性能を向上させるために、ピークホールド処理 [16] とノイズ成分の引き去り [4] を行った。3 つ以上のマイクを使いペアごとに CSP 係数を同期加算した [17]。

3.2 残響時間推定に基づく単一チャンネル残響除去法

単一チャンネルの残響除去のために、文献 [2] のアルゴリズムを用いた。残響時間 T_r がフレーム長より十分長い場合には、観測されたパワースペクトル $|\mathbf{x}|^2$ は、音源のパワースペクトル $|\hat{\mathbf{y}}|^2$ の重み付き和でモデル化できる。音源のパワースペクトルはノイズのパワースペクトル $|\mathbf{n}|^2$ が定常ならば、以下のように求められる。

$$|\mathbf{x}_t|^2 = \sum_{\mu=0}^{t} w_\mu |\hat{\mathbf{y}}_{t-\mu}|^2 + |\mathbf{n}|^2 \qquad (3)$$

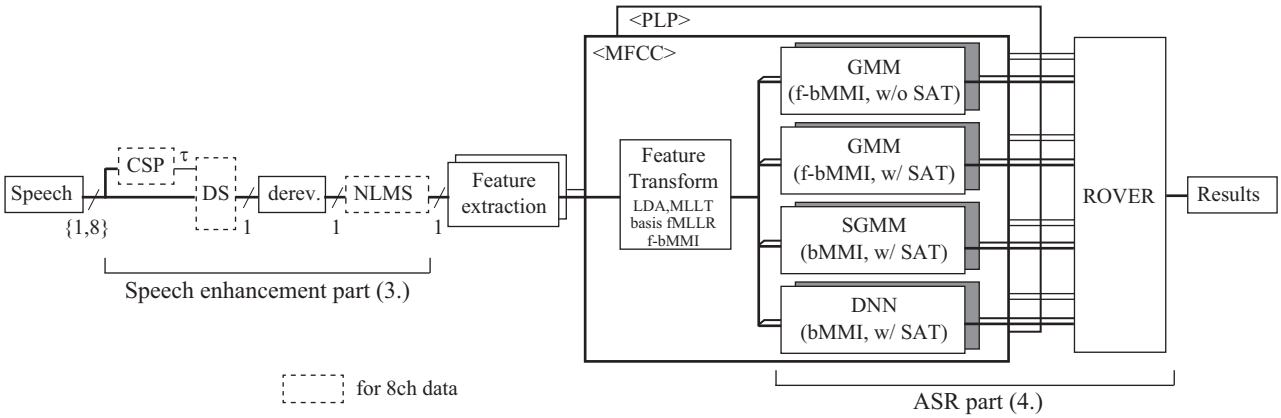

図 1 Schematic diagram of the proposed system. (CSP: cross spectrum phase analysis, DS: delay-and-sum beamformer, derev.: proposed dereverberation method, NLMS: normalized least-mean-squares algorithm, gray blocks are complementary systems for each system type)

μ と w_μ は、それぞれ遅れフレームと重み係数である。音源のパワースペクトル $|\hat{\mathbf{y}}|^2$ は、以下の如く $|\mathbf{x}|^2$ と関連する。

$$|\hat{\mathbf{y}}_{t-\mu}|^2 = \eta(T_r)|\mathbf{x}_{t-\mu}|^2 - |\mathbf{n}|^2 \tag{4}$$

ここで η は直接音と反射音の和の内の直接音が占める割合であり、T_r が長くなるにつれ、反射音成分のエネルギー割合が増えるため、T_r に関する減少関数となっている。w_0 は 1 と仮定すると、上述の関係より Eq. (5) が導かれる。

$$|\hat{\mathbf{y}}_t|^2 = |\mathbf{x}_t|^2 - \sum_{\mu=1}^{t} w_\mu \left[\eta(T_r)|\mathbf{x}_{t-\mu}|^2 - |\mathbf{n}|^2\right] - |\mathbf{n}|^2 \tag{5}$$

残響は、初期残響と後期残響の 2 つの段階に分けて考えられる。直接音到来後、それらの残響を分ける閾値が D(フレーム) である。音声認識の性能に悪影響を与えるのは主に後期残響なので、複雑な初期残響は無視できる。後期残響では、拡散音場が仮定できる場合、音響エネルギー密度が指数的に減衰することが知られており、Polack の統計モデル [18] によりモデル化でき、w_μ は $D < \mu$ の場合、以下のように決定される。$\mu < D$ の場合 w_μ は 0。

$$w_\mu = \alpha_s / \eta(T_r) e^{-2\Delta\varphi\mu} \tag{6}$$

φ はフレームシフト、α_s は引き去り係数である。上段・下段の条件はそれぞれ、初期・後期残響に対応している。η が定数と仮定すると、Eq. (5) は、スペクトルサブトラクション (SS)[19] と類似の働きをする。SS では、引き去られたパワースペクトル $|\hat{\mathbf{y}}|^2$ が $\beta|\mathbf{x}|^2$ より小さい場合には、$\beta|\mathbf{x}|^2$ で置き換える (フロアリング)。β はフロアリング係数である。フロアリング率を、時間周波数ビンの総数中、フロアリングしたビン数の比率と定義する。いくつか仮定した残響時間に対して、フロアリング率を計算した後、最小 2 乗法で傾きを算出し、あらかじめ開発セットで定めた傾きと残響時間の関係から、実際の残響時間 T_r を求めることができる [2]。

4. 音声認識部

4.1 音響モデルの MMI 識別学習

MMI 識別学習では、正解ラベルと認識仮説の間の相互情報量を最大化する。以下、改良版のブーステッド MMI (bMMI) 学習 [20] について述べる。この手法は、ブースティング係数 $b(\geq 0)$ により、音素正解率によって学習データの重みを変化させることを可能にしている。評価関数は

$$\mathcal{F}_{\text{bMMI}}(\lambda) = \sum_r \log \frac{p_\lambda(\boldsymbol{x}^r|\mathcal{H}_{s_r})^\kappa p_L(s_r)}{\sum_s p_\lambda(\boldsymbol{x}^r|\mathcal{H}_s)^\kappa p_L(s) \mathrm{e}^{-bA(s,s_r)}} \tag{7}$$

のように表される。$\boldsymbol{x}^r$ は、r 番目の発話の特徴量系列である。音響モデルパラメータ λ は、拡張バウム・ウェルチ法で最適化される。$\mathcal{H}_{s_r}$ と $\mathcal{H}_s$ は、各々、正解ラベル s_r と仮説 s に対する HMM の系列である。p_λ は音響モデル尤度、κ は音響スケール、p_L は言語モデル尤度であり、$A(s, s_r)$ は s の s_r に対する音素正解率である。

4.2 識別的特徴量変換

特徴量空間ブーステッド MMI (f-bMMI) は、高次元特徴量 $\boldsymbol{h}_t$ を低次元特徴量に写像する行列 $I \times J$ の行列 $\boldsymbol{M}$ を、識別的基準により推定する [9]。

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{M}\boldsymbol{h}_t \tag{8}$$

$\boldsymbol{x}_t$ は、t フレームにおける元の I 次元特徴量、$\boldsymbol{y}_t$ は変換された同じく I 次元の特徴量、$\boldsymbol{h}_t$ は $J(\gg I)$ 次元の補助的な特徴量である。行列 $\boldsymbol{M}$ は、評価関数 $\mathcal{F}_{\text{f-bMMI}}(\boldsymbol{M})$ を最大化するように最適化される。評価関数は、Eq. (7) の $\boldsymbol{x}^r$ を、r 番目の発話の変換特徴量系列 $\boldsymbol{y}^r$ で置き換えて、

$$\mathcal{F}_{\text{f-bMMI}}(\boldsymbol{M}) = \sum_r \log \frac{p_\lambda(\boldsymbol{y}^r|\mathcal{H}_{s_r})^\kappa p_L(s_r)}{\sum_s p_\lambda(\boldsymbol{y}^r|\mathcal{H}_s)^\kappa p_L(s) \mathrm{e}^{-bA(s,s_r)}} \tag{9}$$

のように得られる。

4.3 DNN の識別学習

DNN-HMM システムにおいて、通常のクロスエントロピー (CE) 学習に加えて、bMMI 基準 (7) に基づく系列の識別学習法 [21] が提案されている。DNN は HMM 状態 j に対する事後確率を出力し、音響モデル尤度 p_θ は、事後確率から求められる疑似的な尤度

$$p_\theta\left(\boldsymbol{x}^r|j\right) = \frac{p_\theta\left(j|\boldsymbol{x}^r\right)}{p_0\left(j\right)} \tag{10}$$

で置き換えられる。$p_0\left(j\right)$ は状態 j に対する事前確率であり、強制整列した学習データから求められる。評価関数 $\mathcal{F}_{\mathrm{bMMI}}\left(\theta\right)$ は、Eq. (7) の λ を θ に置き換えて得られる。

4.4 システム統合のための補助システム

効果的なシステム統合のためには、補助システムを意図的に構築することが有効である [14]。提案の補助システムの識別学習法は、識別学習の原理を拡張したものである。Q 個の元となるシステムが既に構築されているときに、識別学習の評価関数 $\mathcal{F}$ は、

$$\mathcal{F}^c(\varphi) = (1+\alpha_c)\mathcal{F}(\varphi) - \frac{\alpha_c}{Q}\sum_{q=1}^{Q}\mathcal{F}(\varphi) \tag{11}$$

のように、一般化される。これは、正解ラベル s_r に関連する元の評価関数から、q 番目の元のシステムの 1 位の仮説 (ラティス)$s_{q,1}$ に関連する評価関数を引き去ったものである。φ は最適化されるべき補助システムのモデルパラメータの組 (例えば、λ、$\boldsymbol{M}$ や θ)、α_c はスケーリング係数である。識別的基準 $\mathcal{F}$ は、bMMI や f-bMMI が選択できる。もし α_c が零の時には、この評価関数は元々の $\mathcal{F}$ に一致する。Eq. (11) の第 1 項は、識別学習の基準に従って性能を向上させ、第 2 項は今構築しているシステムが、元のシステムと異なる傾向の仮説を出力するようにさせる効果がある。

5. 実験条件

提案法の有効性を REVERB チャレンジ [1] で提供されている残響音声を用いて検証した。1、2、8 チャンネルのデータのうちの 1 と 8 チャンネルを検討した。タスクは中程度語彙 (5k) の連続音声認識である (WSJCAM0)。“SimData” は、残響時間が 0.25、0.5、0.75 秒の 3 種類のオフィス (Room 1〜3) において、音源とマイクの距離が 0.5 m (near) もしくは 2 m (far) の 6 種の室内伝達関数をクリーン音声に畳み込み、SNR が 20dB になるように騒音を重畳してある。一方、“RealData” は、実測により、騒音が存在する状況で、1 つの部屋 (Room 1) において、同じく音源とマイクの距離が 2 条件で収録されたデータである。8 本のマイクは半径 0.1 m の円状に配置されている。学習セット (**tr**) は、92 話者の 7,861 発話、評価セット (**eva**) は、SimData が 28 話者の 2,176 発話、RealData が 10 話者の 372 発話、開発セット (**dev**) が SimData が 20 話者の 1,484 発話、RealData が 5 話者の 179 発話よりなる。音響モデルは **tr** により学習し、パラメータは、**dev** の単語誤り率 (WER) に基づき調整した。tri-gram 言語モデルを使った。実験のすべては「発話単位の一括処理」である。

1 チャンネルの場合には、残響時間推定を行う提案の残響除去法のみを用いた [*3]。8 チャンネルの場合には、残響除去の前に、推定した到来方向情報に基づく遅延和ビームフォーミングを行った。到来方向推定とビームフォーミングには、全ペアのマイクを用いた。残響除去後に、200 タップの NLMS フィルタにより、短時間の歪を除去した。

元となる音響特徴量は、0 から 12 次の MFCC、PLP とその 1 次 2 次の動的特徴量である。9 連続フレームの静的 MFCC を結合した 117 次元の特徴量を、LDA を用いて 40 次元に圧縮した。LDA のクラスはトライフォンの HMM の状態とした。これに加えて、MLLT を用いた [11]。

音響モデル適応には、短い発話で性能を発揮しやすい基底 fMLLR [12] を用いた。さらに、話者間の多様性の影響を低減するために、SAT [8] を行った。

bMMI と f-bMMI におけるブースティング係数は 0.1 とし、補助システムを構築するためのパラメータは、Eq. (11) の第 2 項の増幅係数が 0.3、α_c は 0.75 とした。

始めに、クリーン音響モデルを学習する [*4]。次に、クリーンモデルによる強制整列結果や tri-phone のツリー構造を使って、残響音響モデルを最尤 (ML) で学習する。最後に、ML モデルから学習を進めて、識別学習や識別的特徴量変換を行う。DNN は、Kaldi [15] の Povey の実装を使った。隠れ層 2 層でパラメータ数は 2M である。初期の学習率は 0.02 であり、学習の最後には 0.004 まで低減した。

GMM、SGMM [22] と DNN の 3 種類の異なる音響モデリング手法を試した。それぞれに、識別学習を加えた。GMM には、f-bMMI を、SGMM と DNN[21] に関しては bMMI を用いた。各々のシステムに対して、組となる補助システムを構築した。これらは、MFCC と PLP 特徴量で構築されているので、システム総数は 16 となる。

6. 結果と考察

表 1 は、開発セット (**dev**) の WER である。4 種の部屋において、音源・マイク間の距離が 2 種ある。“Kaldi baseline” は、残響音声より学習された音響モデルで、音声強調手法なしで認識した場合の WER である。“derev.” は、提案の残響時間推定に基づく単一チャンネルの残響除去法である。平均的に、性能が向上している。8 チャンネルの場合は、到来方向推定は安定していたため、ビームフォーミングと “derev.” を併用することで認識性能の大

[*3] 残響除去法のパラメータは、$(D = 9, \alpha = 5, \beta = 0.05, a = 0.005, b = 0.6)$ のように設定した。

[*4] mono-phone は、無音のモデル (“sil”) を含み 44 である。tri-phone は、状態数は 2,500、ガウス分布総数は 15,000 である。

表 1 WER [%] by room and microphone distance on the REVERB Challenge (**dev**).

			SimData							RealData		
			Room 1		Room 2		Room 3		Avg	Room 1		Avg
	Feature	Type	near	far	near	far	near	far		near	far	
1ch												
Kaldi baseline	MFCC	ML	10.96	12.56	15.70	34.21	19.61	39.24	22.05	48.53	47.37	47.95
derev.			12.41	14.68	14.03	27.16	16.39	33.85	19.75	47.04	44.57	45.81
GMM	+LDA+MLLT	ML	9.46	11.01	11.51	22.04	13.08	28.09	15.87	39.99	40.67	40.33
	+basis fMLLR		7.77	10.00	9.76	19.28	11.05	24.90	13.79	33.00	35.54	34.27
		bMMI	7.13	9.61	9.12	16.19	10.46	21.98	12.42	30.69	35.20	32.95
		f-bMMI	6.27	8.73	8.28	14.89	9.37	19.54	11.18	28.32	31.31	29.82
		f-bMMI$_c$	7.06	9.05	8.58	14.96	10.16	20.43	11.71	29.01	31.72	30.37
	+SAT	ML	8.87	11.21	9.71	19.89	10.95	24.04	14.11	36.06	36.23	36.15
		bMMI	6.56	8.51	7.76	16.24	9.03	19.88	11.33	34.19	37.53	35.86
		f-bMMI	5.88	7.60	7.25	14.59	8.09	17.51	10.15	31.63	34.72	33.18
		f-bMMI$_c$	6.07	7.82	7.22	14.89	8.43	17.51	10.32	32.38	35.27	33.83
SGMM		ML	6.47	9.07	8.18	17.11	9.55	20.40	11.80	33.13	34.93	34.03
		bMMI	5.53	7.23	7.00	14.44	7.76	17.48	9.91	31.50	33.36	32.43
		bMMI$_c$	5.68	7.28	7.02	14.44	7.94	17.68	10.01	30.94	33.08	32.01
DNN		CE	6.71	8.85	8.70	15.58	9.15	19.07	11.34	30.88	35.82	33.35
		bMMI	5.29	7.06	6.95	13.09	7.57	15.53	9.25	28.45	32.67	30.56
		bMMI$_c$	5.14	6.74	6.51	12.37	7.27	15.50	8.92	28.32	33.49	30.91
ROVER			**4.67**	**5.88**	**6.31**	**11.93**	**6.63**	**14.89**	**8.39**	**26.58**	**28.91**	**27.75**
8ch												
CSP+BF+derev.	MFCC	ML	10.79	12.19	11.02	16.71	11.47	20.43	13.77	40.36	42.83	41.60
+NLMS			11.11	12.27	11.81	17.40	12.34	21.46	14.40	38.37	40.74	39.56
GMM	+LDA+MLLT	ML	8.38	10.30	9.91	14.94	10.19	17.28	11.83	34.06	37.18	35.62
	+basis fMLLR		7.74	9.22	8.80	13.33	9.05	15.28	10.57	27.39	30.14	28.77
		bMMI	6.64	8.21	7.25	11.39	7.10	11.50	8.68	24.89	27.96	26.43
		f-bMMI	6.19	7.40	7.39	10.13	6.58	10.24	7.99	22.58	**26.25**	24.42
		f-bMMI$_c$	6.39	7.33	7.44	9.86	6.70	10.44	8.03	22.71	27.41	25.06
	+SAT	ML	7.25	9.32	8.70	12.79	8.33	13.80	10.03	28.88	32.88	30.88
		bMMI	5.24	7.10	6.56	9.93	5.98	10.98	7.63	26.58	30.83	28.71
		f-bMMI	5.01	6.76	5.96	9.07	5.84	9.40	7.01	24.27	29.60	26.94
		f-bMMI$_c$	5.16	6.93	6.11	9.49	5.96	9.67	7.22	24.27	29.73	27.00
SGMM		ML	5.65	7.62	7.47	10.97	7.00	11.45	8.36	25.27	30.35	27.81
		bMMI	4.57	6.05	6.19	9.27	6.01	9.89	7.00	24.70	30.01	27.36
		bMMI$_c$	4.72	6.10	6.09	9.56	6.18	10.01	7.11	24.39	30.01	27.20
DNN		CE	6.49	7.45	7.84	11.44	7.25	11.97	8.74	25.27	29.32	27.30
		bMMI	5.56	6.27	6.24	9.29	5.71	10.44	7.25	23.27	28.84	26.06
		bMMI$_c$	5.26	6.05	6.21	9.10	5.61	10.06	7.05	22.65	28.50	25.58
ROVER			**4.18**	**5.11**	**5.50**	**7.74**	**4.85**	**8.23**	**5.94**	**21.90**	26.52	**24.21**

幅な改善が見られた。“NLMS” は、RealData においては WER を 2.04%改善したものの、 SimData に対しては 0.63%悪化する結果となった。しかしながら、悪化の方が少なかったので、以後は NLMS を使った場合の結果を示す。

LDA と MLLT により、WER は大幅に改善した。表 1 より、識別学習が残響環境に対しても有効であることがわかる。全ての場合で、f-bMMI 学習の性能が、bMMI 学習を上回った。提案の補助システムの性能は、元のシステムの性能よりも若干悪い程度で、システム統合に適している。

SGMM モデルは、SimData に対しては GMM を上回ったが、RealData に対しては GMM よりも性能が低かった。

DNN モデルは SimData に対して、最良の性能を得た。RealData に対する最良のシステムは SAT を使わない GMM で、DNN は 2 番目であった。SimData と RealData の平均では、DNN は最良の性能を得た。DNN に対しても系列の識別学習は有効であった。

システム統合した場合の結果を最下段に示している。全 16 システムの出力結果を統合により、最良の性能を得た [*5]。

[*5] PLP を使った場合には、MFCC を使った場合の結果よりも若干性能が悪化したものの、それらの誤り傾向は相当異なっていた。

表 2 WER [%] on the REVERB Challenge (**eva**). MFCC feature was used for single system and MFCC and PLP features were used for ROVER).

		SimData							RealData		
		Room 1		Room 2		Room 3		Avg	Room 1		Avg
		near	far	near	far	near	far		near	far	
1ch	Kaldi baseline	13.23	14.13	15.54	29.69	20.06	37.44	21.68	50.62	45.98	48.30
	derev.	12.50	13.43	14.61	24.71	17.09	32.62	19.16	44.75	43.32	44.04
	f-bMMI	7.27	8.17	8.82	14.11	10.54	18.76	11.28	28.65	29.54	29.10
	SAT+f-bMMI	6.44	7.22	7.57	13.97	9.52	18.44	10.53	28.87	29.78	29.33
	SGMM+bMMI	5.81	6.54	7.22	13.84	8.70	18.17	10.05	27.75	28.36	28.06
	DNN+bMMI	5.90	6.84	7.35	12.57	9.40	16.55	9.77	25.97	25.69	25.83
	ROVER	**5.30**	**5.61**	**6.30**	**11.16**	**7.76**	**14.95**	**8.51**	**23.79**	**23.60**	**23.70**
8ch	CSP+BF+derev.	10.94	11.69	10.98	16.33	12.79	21.39	14.02	34.33	36.93	35.63
	+NLMS	10.94	12.32	11.38	17.59	13.46	22.96	14.78	35.32	35.28	35.30
	f-bMMI	6.57	6.93	6.80	9.93	7.47	12.76	8.41	20.22	23.19	21.71
	SAT+f-bMMI	6.17	6.64	6.51	10.13	7.40	13.15	8.33	20.63	23.67	22.15
	SGMM+bMMI	5.86	6.44	6.29	9.23	6.96	12.83	7.94	20.66	23.50	22.08
	DNN+bMMI	5.64	6.18	6.16	9.29	7.08	12.40	7.79	19.35	22.28	20.82
	ROVER	**4.96**	**5.62**	**5.58**	**8.18**	**5.73**	**10.47**	**6.76**	**16.90**	**20.29**	**18.60**

表 2 には、評価セット (**eva**) の結果を示す。識別学習を行った DNN は、単一のシステム中、最良の性能を得た。これは DNN の未知条件に対する頑健性を示すものといえる。さらに、システム統合 (ROVER 5) により WER が、それぞれ SimData と RealData に対して、1ch の場合、1.26%、2.13%、8ch の場合 1.03%、2.22%改善した。

7. 結論

残響除去や複数マイクによる音声強調が有効であった。特徴量変換と識別学習が、種々の残響環境下で有効であることが示された。望ましい補助システムを構築するための提案のシステム統合法により、さらに性能が向上した。

参考文献

[1] Kinoshita, K. *et al.*: The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech, *Proc. of WASPAA* (2013).
[2] Tachioka, Y. *et al.*: Dereverberation Method with Reverberation Time Estimation Using Floored Ratio of Spectral Subtraction, *Acoust. Sci. & Tech.*, Vol. 34, No. 3, pp. 212–215 (2013).
[3] Knapp, C. and Carter, G.: The Generalized Correlation Method for Estimation of Time Delay, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 24, pp. 320–327 (1976).
[4] Tachioka, Y. *et al.*: Direction of Arrival Estimation by Cross-power Spectrum Phase Analysis Using Prior Distributions and Voice Activity Detection Information, *Acous. Sci. & Tech.*, Vol. 33, pp. 68–71 (2012).
[5] Johnson, D. and Dudgeon, D.: *Array Signal Processing*, Prentice-Hall (1993).
[6] Haeb-Umbach, R. and Ney, H.: Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition, *ICASSP*, pp. 13–16 (1992).
[7] Gopinath, R.: Maximum Likelihood Modeling with Gaussian Distributions for Classification, *ICASSP*, pp. 661–664 (1998).
[8] Anastasakos, T. *et al.*: A Compact Model for Speaker-adaptive Training, *ICSLP*, pp. 1137–1140 (1996).
[9] Povey, D. *et al.*: fMPE: Discriminatively Trained Features for Speech Recognition, *ICASSP*, pp. 961–964 (2005).
[10] Hinton, G. *et al.*: Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Processing Mag.*, Vol. 28, pp. 82–97 (2012).
[11] Tachioka, Y. *et al.*: Discriminative Methods for Noise Robust Speech Recognition: A CHiME Challenge Benchmark, *the 2nd CHiME Workshop on Machine Listening in Multisource Environments*, pp. 19–24 (2013).
[12] Povey, D. and Yao, K.: A Basis Representation of Constrained MLLR Transforms for Robust Adaptation, *Computer Speech and Language*, Vol. 26, pp. 35–51 (2012).
[13] Fiscus, J.: A Post-processing System to Yield Reduced Error Word Rates: Recognizer Output Voting Error Reduction (ROVER), *Proc. of ASRU*, pp. 347–354 (1997).
[14] Tachioka, Y. *et al.*: A Generalized Framework of Discriminative Training for System Combination, *Proc. of ASRU*, pp. 43–48 (2013).
[15] Povey, D. *et al.*: The Kaldi Speech Recognition Toolkit, *ASRU*, pp. 1–4 (2011).
[16] Suzuki, T. and Kaneda, Y.: Sound Source Direction Estimation Based on Subband Peak-hold Processing, *The Journal of the Acoust. Soc. of Jpn*, Vol. 65, pp. 513–522 (2009).
[17] Nishiura, T. *et al.*: Localization of Multiple Sound Sources Based on a CSP Analysis with a Microphone Array, *ICASSP*, pp. 1053–1056 (2000).
[18] Habets, E.: Speech Dereverberation Using Statistical Reverberation Models, *Speech Dereverberation* (Naylor, P. and Gaubitch, N., eds.), Springer (2010).
[19] Boll, S.: Suppression of Acoustic Noise in Speech Using Spectral Subtraction, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113–120 (1979).
[20] Povey, D. *et al.*: Boosted MMI for Model and Feature-space Discriminative Training, *ICASSP*, pp. 4057–4060 (2008).
[21] Veselý, K. *et al.*: Sequence-discriminative Training of Deep Neural Networks, *INTERSPEECH* (2013).
[22] Povey, D. *et al.*: The Subspace Gaussian Mixture Model –A Structured Model for Speech Recognition, *Computer Speech and Language*, Vol. 25, No. 2, pp. 404–439 (2011).