社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

複数残響特性下の音声を単一モデル学習に用いた未知残響環境に頑健な音声認識の検討

西亀　健太†　　渡部　晋治††　　西本　卓也†　　小野　順貴†　　嵯峨山茂樹†

† 東京大学情報理工学系研究科システム情報学専攻，〒 113-8656 東京都文京区本郷 7-3-1
†† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所，〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{nishiki,nishi,onono,sagayama}@hil.t.u-tokyo.ac.jp, ††watanabe@cslab.kecl.ntt.co.jp

あらまし　残響環境下では音声認識性能は著しく著しく劣化する．これに対し，人工残響インパルス応答をクリーン音声に畳み込んで学習することで，認識率が向上することが知られている．しかし，どのような残響インパルス応答を畳み込んで学習すべきか，という点に関しては十分に議論がなされていない．本研究では，人工残響インパルス応答の残響時間パラメータと認識率の関係を残響音声認識評価基盤 (CENSREC-4) を用いて詳細に調べ，認識率が大きく変わる残響時間パラメータの範囲がテスト環境より短い残響時間に分布していることを述べる．また，その範囲に基づいて人工残響インパルス応答を選択しマルチコンディション学習を行うことが，未知残響環境に対して頑健な音声認識であることを示す．

キーワード　残響環境，音声認識，モデル学習，人工残響，インパルス応答，残響時間

A Study on Robust Speech Recognition against Unknown Reverberation Using Single Speech Model Trained under Multiple Reverberant Environments

Kenta NISHIKI†, Shinji WATANABE††, Takuya NISHIMOTO†, Nobutaka ONO†, and Shigeki SAGAYAMA†

† Department of Information Physics and Computing, University of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
†† NTT Communication Science Laboratories 2-4 Hikaridai Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan.
E-mail: †{nishiki,nishi,onono,sagayama}@hil.t.u-tokyo.ac.jp, ††watanabe@cslab.kecl.ntt.co.jp

Abstract　In reverberant environment, speech recognition accuracy is seriously degraded. An effective method is to train acoustic models using reverberant speech which are generated from clean speech data and reverberant impulse response. It is not clear, however, what kind of impulse response is suitable for environmental variations. In this research, we show the range of reverberant time that changes speech recognition accuracy drastically through evaluation with CENSREC-4. Then we point out that the range of reverberant time which is effective for model training is shorter than that of test conditions. We also show the multi-condition training is effective for speech recognition under unknown environments using speech data of the range of reverberant time we showed.

Key words　reverberant environment, speech recognition, model construction, artificial reverberance, impulse response, reverberation time

1. はじめに

近年，会議録の自動作成や音声インターフェース等で，実環境における遠隔発話音声認識の必要性が高まっている．実環境遠隔発話音声認識においては，背景雑音や，残響の影響により，音響モデル学習時の音声と認識時の入力音声のミスマッチ

が生じ，音声認識の性能は著しく低下する [1]．実環境では大きく分けて，加法性雑音と乗法性歪みが音声認識性能を劣化させる要因となる．背景雑音などの加法性雑音に対してはスペクトル減算法などの雑音抑制手法やモデル合成法 [2] が用いられる．また，回線特性やマイクロホン特性，残響時間が分析フレーム長より短い場合の残響による乗法性歪みに対しては，従来 CMS(Cepstral Mean Subtraction) 法 [3] が幅広く用いられており，少ない計算量で高い効果が得られる．しかし，残響時間が分析フレーム長より長い場合，特徴量の歪みは単純に定常性の雑音や各フレームで同一の乗法性歪みとして扱うことはできず，CMS で完全に取り除くことはできない．従来，耐残響音声認識には主に以下のようなアプローチが存在する．

- 観測信号から残響を除去する方法
- 音響モデル適応
- 残響コンディション学習

観測信号から残響を除去する手法は，近年音響信号処理の分野で広く研究が行われており [6] [7]，残響音声のモデル化に基づいて聴覚上の明瞭性や S/N 比が大きく改善されている．しかし，残響の消し残りや過剰な抑圧により生じた音声の歪みによって，クリーン音声との間に新たなミスマッチが生じるため，残響除去法による改善が音声認識率向上につながらないケースも報告されており [8]，音声認識との高い親和性を確保するためには，後述の音響モデル適応法や残響コンディション学習法が必要となる(注1)．

一方音響モデル適応法は，クリーン音声の特徴量により学習された音響モデルパラメータと残響音声の特徴量との間の変換をパラメトリックに表現し，適応データを用いてその変換パラメータを推定することにより，ミスマッチを緩和する手法である．適応学習の効果により，音声認識との高い親和性を保持することができる．例えば最尤線形回帰法 (MLLR [9]) は，HMM の出力確率分布の平均ベクトルに対して線形変換を，滝口らの方法 [10] は，ケプストラム領域での乗法性雑音の定式化に基づくモデル変換を仮定して，それらの変換パラメータを推定する手法である．また，これらの手法は原理的には分析フレームより短い時間の残響への対応をしたものであるが，文献 [11]- [13] では，分析フレームより長い時間の残響をモデル化することにより，残響下での高い音声認識率を実現している．このように，音響モデル適応法は高い認識性能を実現できるが，常に適応データの収集及び適応学習の工程を必要とするため，音声認識を使用するユーザーに負担を強いる．

一方残響コンディション学習 [14]- [19] は，あらかじめ残響インパルス応答を畳み込んだ音声をモデル学習に用いる方法であり，適応工程を必要とすることなく高精度の残響下音声認識を実現できる．ただし一般に，インパルス応答の性質は，部屋の環境やマイクと音源との位置関係などにより異なるため，これらの残響環境の変化に頑健な残響コンディション学習法が必要となる．文献 [15] では，音源とマイク間の距離が異なる複数の環境で収録された音声から，それぞれの音響モデルを作成し，認識時に最大尤度基準や特徴量総和基準に基づいてモデルを選択することにより，音源とマイク間の距離の変化に対して頑健な音声認識を実現している．また，文献 [14] [18] [19] では残響時間の異なる複数のモデルを作成することにより，残響時間の変化への対処を行っている．このような，異なる残響環境から複数音響モデルを作成し，認識時に適切なモデルを選択するアプローチと共に，複数の残響環境下の音声を用いて単一のモデルを作成し認識を行うマルチコンディション学習アプローチも広く研究が行われている [14] [16]．しかし，マルチコンディション学習アプローチにおいて，どのような残響インパルス応答を重畳し学習すべきか，またどのような環境を選択してマルチコンディションモデルを構築すべきか，という点については十分に考察されておらず，現状では試行錯誤により決定されている．

本研究では，残響環境の変化に頑健な汎用的音声認識を目指す上で，どのような残響インパルス応答を重畳しマルチコンディションモデルを構築すべきかに関して検討を行う．特に残響時間の違いは認識率に大きな影響を与えるため，どのような残響時間の残響インパルス応答を学習に用いるべきかについて検討する．そこで，我々は残響時間を制御しやすい人工的残響インパルス応答に着目し，認識率が大きく変化する残響時間パラメータの範囲を詳細に調べる．この残響時間パラメータの範囲内にある残響時間インパルス応答を用いてマルチコンディション学習をすることで，未知環境に対し頑健な残響音声認識を目指す．

2. 人工残響インパルス応答

残響コンディション学習による音声認識を行う上で，学習データに重畳する残響インパルス応答と利用環境における音声の残響時間・残響特性がどの程度異なるかという点は，音声認識率に大きく関係する．特に残響時間の違いは音声認識率に大きな影響を与えるパラメータであり，なおかつ必ずしもテスト環境と学習データの残響時間が同じときに認識率が高くなるとは限らない．

多くの未知残響環境に対し頑健な音声認識を行うためには，学習データにどのような残響時間を持った残響インパルス応答を畳み込むかという点に対する検討が必要である．しかし，検討のために任意の残響時間の実環境インパルス応答を用意するのは難しい．そこで，我々は人工残響インパルス応答を学習データに畳み込む手法に着目する．人工残響インパルス応答は残響時間を制御して設計することが可能であり，より多くの未知環境に対し頑健な残響コンディションモデルを構築するための残響インパルス応答となりうる．本研究では人工残響インパルス応答を以下の 2 つの方法で生成する．

- ART1: 白色雑音に指数減衰する包絡を乗じる
- ART2: ART1 に時間遅れを与え単位インパルスを加える

ART1 は広林ら [5] による残響インパルス応答の近似で，以下の式で表わされ，後期残響のよい近似となっている．

$$h_1(t) = ae^{-\frac{6.9t}{T_R}} n_1(t) \tag{1}$$

(注1)：例えば，文献 [8] では，音響モデルの分散パラメータ補正と適応学習を組み合わせてこれに対処している．

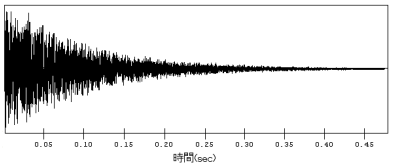

図 1　人工残響インパルス応答 (ART1) の時間波形

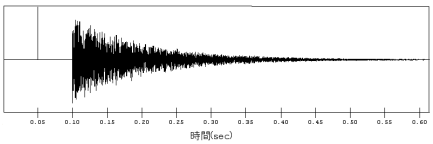

図 2　人工残響インパルス応答 (ART2) の時間波形

a は振幅，$n_1(t)$ は白色雑音である．T_R はこの残響インパルス応答が 60dB 減衰する時間である．ここでは便宜的に T_R を残響時間パラメータ (sec) と呼ぶ．後期残響は音声認識性能を劣化させる主要因となることが知られている [6]．ART1 を畳み込んだ音声で学習を行うことで，利用環境における音声が後期残響成分を含む場合に生じるモデルとのミスマッチを減らすことができる．$a = 1$，$T_R = 0.7$ のときの波形を図 1 に示す．また，ART2 の残響のインパルス応答は以下の式で書ける．

$$h_2(t) = b\delta(t) + ae^{-\frac{6.9(t-T_L)}{T_R}} n_2(t - T_L) \qquad (2)$$

a, b はそれぞれ初期反射および後期残響の振幅，$n_2(t)$ は白色雑音，T_R は残響時間パラメータ (sec) である．ART2 における単位インパルス関数 $\delta(t)$ は初期反射の近似である．T_L は初期反射から後期残響が開始するまでの時間 (sec) で，$n_2(t) = 0(t < 0)$ である．$a = 1$，$b = 0.1$，$T_R = 0.7$，$T_L = 0.05$ のときの波形を図 2 に示す．

3. 認 識 実 験

表 1　CENSREC-4 データの残響時間

名称	収録環境	残響時間 (秒)
incar	In-car	0.05
elevator_hall	Elevator hall	0.75
jp_style_bath	Japanese style bath	0.60
jp_style_room	Japanese style room	0.40
lounge	Lounge	0.50
living_room	Living room	0.65
meeting_room	Meeting room	0.65
office	Office	0.25

3.1 単一の人工残響インパルス応答を学習に用いた残響コンディションモデル

本実験で，単一の人工残響インパルス応答を学習に用いた際，その残響時間パラメータの違いにより，音声認識率がどのような傾向を示すかを調べる．学習および認識は HTK(Ver. 3.4) を用い CENSREC-1 [22][(注2)] 準拠の数字 HMM を用いて連続数字認識実験を行う．分析条件は CENSREC-4 のベースライン評価と同様である．ART1 の人工残響インパルス応答は式 (1) に基づいて以下のような手順で生成した．

（1） $n_1(t)$ をランダムに生成

（2） T_R の異なる 15 種類の人工残響インパルス応答を生成

以上のプロセスを 2 度繰り返すことで，15×2 セットの残響インパルス応答を得た．ただし $a = 1$ とした．残響時間パラメータ T_R は 0.05 秒から 0.6 秒までを 0.05 秒刻みで，0.6 秒から 0.9 秒まで 0.1 秒刻みで用意した．ART2 については式 (2) に基づいて，

（1） $n_2(t)$ をランダムに生成

（2） T_R の異なる 15 種類の人工残響インパルス応答を生成

（3） 時間遅れ T_L を与える

（4） 単位インパルス $\delta(t)$ を加える

というプロセスを 2 度繰り返すことによって 15×2 セットの残響インパルス応答を得た．ただし $a = 1$，$b = 0.1$，$T_L = 0.05$ とした．T_R の刻みは ART1 と同様である．それらを全て CENSREC-4 データベースに含まれるクリーン音声 110 名，8,440 発話 (男女 55 名，4,220 発話ずつ) 分に畳み込み，学習データを作成する．テストデータは CENSREC-4 のクリーン音声 104 名，4,004 発話 (男女 52 名，2,002 発話ずつ) 分を 2 分割した 2,002 発話に，同データベースに含まれる実環境インパルス応答 8 種類 (表 1) を畳み込んだものを用いた．また，より多様な残響環境での性能を評価するために，建築音響分野で収集・利用されている SMILE2004 [20] 室内音響データベースに含まれる残響インパルス応答を上と同じクリーン音声に重畳しテストデータを得た．SMILE2004「室内残響」58 種類のうち，表 2 にある 10 種類を使用した (ただし，44.1kHz から 16kHz にダウンサンプリングをした)．表 2 の残響時間は二乗積分法により計算した．ART1，ART2 ともに同じ残響時間パラメータのランダムに生成された 2 セットずつの人工残響インパルス応答が存在しており，それら 2 つの平均により単語正解率 (%) を求めた．

表 2　SMILE2004「室内残響」

名称	収録環境	残響時間 (秒)
t20201	フラッタエコーが聞こえる講義室	1.13
t20317	クラシック音楽専用ホール 4：吸音カーテンなし	2.62
t20401	会議室	0.67
t20402	講義室	0.95
t20403	大講義室	1.16
t20404	講演・スピーチ主体のホール	1.66
t20405	教会 1	0.80
t20416	講演・スピーチ主体のホール 2, $d = 10$m(一階前方)	1.62
t20417	講演・スピーチ主体のホール 2, $d = 15$m(一階中央)	1.69
t20418	講演・スピーチ主体のホール 2, $d = 15$m(一階バルコニー下)	1.74

表 3〜6 に実験結果を示す．横軸が残響時間パラメータ，縦

(注2)：なお CENSREC-4 は音素 HMM を採用している．

軸がテスト環境である．また最下行に全テスト環境に関する単語正解率の平均を求め記載した．最も認識率の高いものは太字で示してある．表 3〜6 より，incar を除く多くのテスト環境 (表 1，2) に対して，テスト環境の残響時間より短い残響時間パラメータを持つ人工残響インパルスを学習に用いた残響コンディションモデルが有効であると言える．以上の傾向は ART1，ART2 双方にあてはまっている．

3.2 複数の人工残響インパルス応答を学習に用いたマルチコンディションモデル

前章の結果より，実際のテスト環境の残響時間より短い残響時間パラメータの人工残響インパルス応答を学習データに畳み込んだ残響コンディション学習がより多くの残響環境に対して有効であることがわかった．もし，ある単一の残響時間パラメータの人工残響インパルス応答を学習に用いた残響コンディションモデルが多くのテスト環境に対して有効であれば，それが汎用的な耐残響音響モデルとなる．しかし，実験 1 の結果 (表 3〜6) より，ある特定の残響時間の残響コンディションモデルによって多くのテスト環境において十分な認識性能を実現できていない．そこで実験 1 で求めた，学習データの残響時間パラメータに対する性能の変化が大きい 0.05 秒から 0.5 秒の範囲から，人工残響インパルス応答を複数選択し，マルチコンディションモデルを構築する．これにより，単一のモデルで多くの残響環境に対応することを目指す．

この実験で，以下のマルチコンディションモデルを構築し，評価する．

- MIX(ART1): ART1 の人工残響インパルス応答を用いたマルチコンディションモデル
- MIX(ART2): ART2 の人工残響インパルス応答を用いたマルチコンディションモデル

MIX(ART1)，MIX(ART2) ともに実験 1 で作成した 15 × 2 セット (乱数) の残響時間パラメータの各セットから 0.15，0.2，0.25，0.3 秒の 4 種類を選択し，それぞれを 2,002 発話ずつのクリーン音声に畳み込むことで学習データを得た．それらの学習データを用いてマルチコンディション学習を行う．つまりマルチコンディションモデルが MIX(ART1) に関して 1 × 2 セット (乱数)，MIX(ART2) に関して 1 × 2 セット (乱数) のモデルが生成される．MIX(ART1)，MIX(ART2) それぞれに関して 2 セットで認識実験を行い，それら 2 つの平均により単語正解率 (%) を求めた．参照のため，クリーンモデル，MLLR，CENSREC-4 に含まれるマルチコンディションモデル (multi)，マッチドコンディション学習との比較を行う．multi の学習データは 4 種類の実環境インパルス応答 (Office，Elevator hall，In-car，Living room) をそれぞれクリーン音声 2,002 発話ずつに畳み込むことで得た．MLLR の適応データはテスト環境と同じ残響インパルス応答を CENSREC-4 に含まれるクリーン音声 2,002 に畳み込むことで得た．テストデータは multi の学習に使用されていない 4 種類 (Lounge，Japanese style room，Meeting room，Japanese style bath) を用いる．SMILE2004 に関しては実験 1 と全く同じである．

実験結果を表 7〜9 に示す．min(ART1) は ART1 の単一の残響コンディションモデルにおいて残響時間パラメータが 0.05 秒から 0.3 秒のモデルのうち最も認識率が低かったものであり，max(ART1) は最も認識率が高かったものである．min(ART2)，max(ART2) に関しても同様である．表 8 および図 3 より，MIX(ART2) の認識率は Japanese style bath 以外の 3 つの環境で max(ART2) より高く，MIX(ART1) の認識率は Japanese style bath，Japanese style room 以外の 2 つの環境で max（ART1) より高い．Japanese style room においては MIX(ART1) と max（ART1) は同等の認識率である．また，MIX(ART1)，MIX(ART2) ともに Japanese style bath でも min(ART1)，min（ART2) よりは認識率が高い．表 9 においても，MIX(ART1) は max(ART1) と，MIX(ART2) は max(ART2) と同等の性能であることがわかる．また，表 7 および図 4 より，MIX(ART1) と MIX(ART2) での認識率は，CENSREC-4 のマルチコンディション学習やマッチドコンディション学習に比べると認識率は低いものの，クリーンや MLLR に比べて認識率が高く，実収録のインパルス応答を必要としないため有効であることが確かめられた．最後に，MIX(ART1) と MIX(ART2) では，すべての環境において MIX(ART1) の方が認識率が高く，より汎用性の高いモデルとなっていると言える．

表 7　複数の人工残響インパルス応答を学習に用いたマルチコンディションモデルと他手法の比較 (CENSREC-4)

テスト環境	clean	MLLR	multi	MIX (ART1)	MIX (ART2)	matched
lounge	29.1	83.6	96.7	84.2	78.1	99.8
jp_style_bath	26.0	67.7	96.4	88.7	84.5	98.5
meeting_room	33.5	80.4	99.5	98.5	96.9	99.7
jp_style_room	33.9	82.0	98.1	97.1	95.5	99.5

表 8　人工残響インパルス応答を用いたマルチコンディション学習の効果 (CENSREC-4)

テスト環境	min (ART1)	max (ART1)	MIX (ART1)	min (ART2)	max (ART2)	MIX (ART2)
lounge	70.6	82.3	84.2	61.5	77.1	78.1
jp_style_bath	64.3	90.9	88.7	70.1	86.0	84.5
meeting_room	87.6	98.2	98.5	90.7	96.7	96.9
jp_style_room	82.7	97.2	97.1	79.5	95.2	95.5

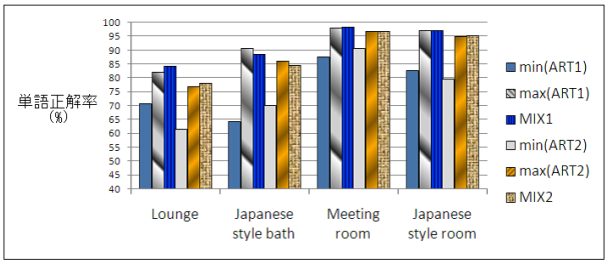

図 3　人工残響インパルス応答を用いたマルチコンディション学習の効果 (CENSREC-4)

表 3　学習データの残響時間と認識率の関係 (モデル: ART1，データベース: CENSREC-4)

テスト環境	残響時間パラメータ T_R(sec)															
	clean	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.70	0.80	0.90
clean	99.6	**56.8**	52.3	44.9	39.5	36.3	33.8	33.3	29.6	27.7	26.5	25.0	23.2	26.8	21.0	20.0
incar	49.3	97.5	**98.0**	95.9	93.3	88.7	84.4	79.9	74.4	72.5	71.0	68.0	64.8	73.5	56.0	51.7
elevator_hall	18.4	54.8	59.6	63.7	65.4	65.5	66.5	67.9	**68.1**	67.1	67.9	67.6	66.6	68.1	66.0	63.2
jp_style_bath	26.0	64.3	74.6	79.5	85.3	88.8	90.8	91.7	91.5	91.7	**91.8**	90.9	90.5	89.0	86.2	82.9
jp_style_room	33.9	82.7	91.0	93.5	95.9	96.8	97.2	**97.2**	96.6	96.3	95.8	94.8	93.8	94.0	87.9	84.2
lounge	29.1	70.6	77.7	79.0	82.0	82.0	**82.3**	81.4	78.0	75.9	74.2	71.1	69.1	72.5	66.3	52.0
living_room	25.4	75.7	83.6	87.9	92.1	94.5	95.7	**95.9**	95.7	95.6	95.3	94.5	93.7	93.7	90.0	87.6
meeting_room	33.5	87.6	94.0	96.2	98.0	**98.2**	97.9	97.2	95.8	94.8	93.7	91.7	90.0	91.9	81.2	76.5
office	42.9	97.2	99.0	**99.0**	98.9	98.4	97.0	95.9	93.6	91.5	89.1	85.9	83.1	87.2	70.0	64.7
平均	32.3	78.8	84.7	86.83	88.9	**89.1**	89.0	88.4	86.7	85.7	84.9	83.1	81.5	83.7	75.4	70.4

表 4　学習データの残響時間と認識率の関係 (モデル: ART2，データベース: CENSREC-4)

テスト環境	残響時間パラメータ T_R(sec)															
	clean	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.70	0.80	0.90
clean	99.6	29.7	**41.8**	34.8	31.2	26.5	25.0	23.2	20.5	18.0	16.6	16.1	15.5	15.9	16.4	16.0
incar	49.3	88.6	**94.7**	91.4	85.8	76.7	76.0	71.5	68.9	62.6	56.8	54.1	54.0	52.0	50.9	47.8
elevato_hall	18.4	43.6	60.8	61.4	62.6	62.1	63.9	64.6	**65.2**	64.5	64.3	64.4	63.9	63.4	61.9	60.5
jp_style_bath	26.0	70.1	74.6	78.1	81.8	84.7	86.0	86.5	**86.8**	87.6	87.3	86.7	86.3	83.9	79.4	77.4
jp_style_room	33.9	79.5	90.7	93.2	94.8	95.1	**95.2**	94.8	94.5	94.0	93.2	91.9	91.0	86.9	82.2	79.5
lounge	29.1	61.5	75.0	75.9	**77.1**	75.8	76.4	73.4	71.7	68.8	67.3	64.8	61.8	56.5	50.8	46.7
living_room	25.4	82.9	84.8	88.6	91.6	93.3	94.0	**94.2**	94.0	93.7	92.7	92.2	91.1	89.0	85.8	83.3
meeting_room	33.5	90.7	94.5	96.3	**96.8**	96.2	95.6	94.4	93.0	91.2	89.1	87.0	84.9	79.0	72.5	68.6
office	42.9	93.4	**98.2**	98.1	97.3	96.0	94.3	91.7	89.0	85.4	81.5	76.9	73.9	65.3	59.1	56.0
平均	39.8	71.1	79.5	79.8	**79.9**	78.5	78.5	77.1	76.0	74.0	72.1	70.4	69.2	65.8	62.1	59.5

表 5　学習データの残響時間と認識率の関係 (モデル: ART1，データベース: SMILE2004)

テスト環境	残響時間パラメータ T_R(sec)															
	clean	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.70	0.80	0.90
t20201	23.8	79.6	88.35	91.6	93.7	**94.1**	93.7	93.0	92.0	90.9	89.5	87.7	85.9	81.1	75.7	71.9
t20317	21.0	65.5	73.85	78.8	83.8	87.3	88.9	89.9	**90.3**	89.9	89.9	88.6	87.8	86.1	83.1	80.8
t20401	32.2	94.3	96.47	**96.3**	95.2	93.7	91.88	87.0	82.4	78.2	74.6	69.4	65.8	58.02	52.2	48.8
t20402	21.5	84.3	91.57	93.0	**93.7**	93.2	92.03	89.6	87.6	84.3	81.5	78.9	76.1	68.39	61.6	57.7
t20403	24.8	78.5	86.47	90.0	**93.6**	95.4	95.92	95.8	95.1	94.9	94.2	93.0	91.7	88.67	84.4	80.4
t20404	23.1	74.4	83.72	88.4	92.5	94.6	95.17	95.1	**95.2**	95.1	94.4	94.0	93.3	92.01	89.9	87.6
t20405	42.5	97.5	98.89	99.2	**99.3**	99.2	98.80	98.2	96.8	96.1	94.3	92.2	90.3	83.66	77.6	72.4
t20416	30.4	84.5	90.99	93.7	96.2	96.9	**97.39**	97.1	96.6	96.2	95.6	95.0	94.1	92.25	89.2	86.8
t20417	19.5	70.8	82.63	88.2	91.3	92.8	**92.94**	92.7	92.2	91.6	90.9	89.5	88.6	85.44	82.0	79.4
t20418	21.9	75.8	85.50	88.7	90.8	**91.4**	91.20	89.6	89.9	87.3	85.2	83.7	81.2	75.61	69.5	65.3
平均	26.1	80.5	87.8	90.79	93.0	**93.9**	93.79	92.8	91.7	90.4	89.0	87.2	85.5	81.11	76.5	73.1

4. ま と め

本稿では，人工残響インパルス応答を畳み込んだ音声を用いて残響コンディションモデルを構築する手法について，学習に用いた人工残響インパルス応答の残響時間パラメータと認識率の関係を CENSREC-4，SMILE2004 両データベースを用いて評価を行うことで詳細に調べた．その結果，認識率が大きく変わる残響時間パラメータの範囲がテスト環境より短い残響時間に分布していることを示した．また，その範囲内にある人工残響インパルス応答を用いてマルチコンディション学習を行うことが，多くの残響環境に対して有効であることを示した．

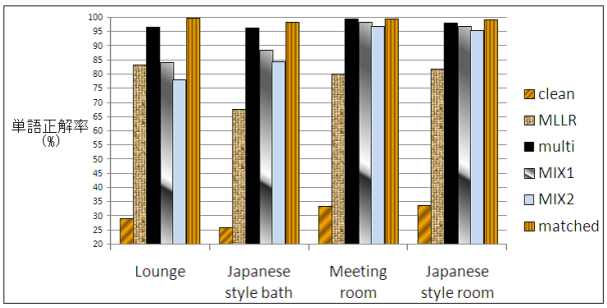

図 4　複数残響特性下の音声で学習した単一音響モデルの認識率 (CENSREC-4)

表 6　学習データの残響時間と認識率の関係 (モデル: ART2，データベース: SMILE2004)

	残響時間パラメータ T_R(sec)															
テスト環境	clean	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.70	0.80	0.90
t20201	23.8	82.5	85.7	90. 1	**90.7**	90.6	89.8	89.0	87.4	85.6	84.0	82.7	79.7	72.9	66.0	61.9
t20317	21.0	72.2	75.4	79.6	83.3	85.4	86.4	**87.2**	87.0	86.4	85.5	85.2	84.1	81.8	77.9	75.5
t20401	32.3	92.2	**93.3**	91.9	89.5	84.8	81.4	75.6	70.0	62.4	55.5	51.8	47.5	43.8	41.3	38.6
t20402	21.5	88.8	86.6	**88.9**	88.0	86.0	83.9	81.5	77.7	73.6	70.0	66.8	62.6	55.3	48.9	45.6
t20403	24.8	77.3	86.5	90.7	92.7	**93.5**	93.2	93.1	92.7	91.4	90.1	89.2	87.5	83.3	77.3	73.1
t20404	23.1	79.6	86.6	90.2	93.1	94.4	**94.5**	94.4	94.1	94.0	93.2	92.7	91.9	89.7	86.5	83.7
t20405	42.5	82.4	98.7	**98.9**	98.6	98.1	97.5	96.5	95.0	93.0	90.1	87.1	84.4	75.6	67.6	64.5
t20416	30.4	84.7	90.8	94.1	95.6	**95.8**	95.6	95.4	95.2	94.6	94.1	93.1	92.0	88.3	83.6	80.2
t20417	19.5	82.9	83.5	88.6	90.7	**91.6**	91.4	90.9	90.0	89.2	88.1	87.7	85.8	81.4	75.8	72.2
t20418	21.9	85.9	82.4	85.8	**86.7**	86.1	84.8	83.8	82.1	80.0	78.1	76.7	72.5	66.8	58.5	54.7
平均	26.1	82.8	86.9	89.9	**90.9**	90.6	89.9	88.7	87.1	85.0	82.9	81.3	78.8	73.9	68.3	65.0

表 9　人工残響インパルス応答を用いたマルチコンディション学習の効果 (SMILE2004)

テスト環境	min (ART1)	max (ART1)	MIX (ART1)	min (ART2)	max (ART2)	MIX (ART2)
t20201	79.6	94.1	94.1	82.5	90.7	91.1
t20317	65.5	88.9	87.3	72.2	86.4	86.5
t20401	91.9	96.5	95.3	81.4	93.3	88.8
t20402	84.3	93.7	93.8	83.9	88.8	88.1
t20403	78.5	95.9	95.3	77.3	93.4	94.1
t20404	74.4	95.2	94.3	79.6	94.5	94.4
t20405	97.4	99.3	99.2	82.4	98.9	98.6
t20416	84.5	97.4	96.9	84.7	95.8	96.1
t20417	70.9	92.9	92.3	82.9	91.6	92.0
t20418	75.8	91.4	91.6	82.4	86.7	87.0

謝辞　本研究の一部は東京大学と NTT の共同研究として行われた．情報提供や有益な議論等に協力していただいた NTT コミュニケーション科学基礎研究所の信号処理研究グループのメンバーに感謝する．本研究は CENSREC-4 の音声データ・残響インパルス応答データおよび評価スクリプト，SMILE2004 の残響インパルス応答データを利用した．

文　献

[1] 中村哲: “実音響環境に頑健な音声認識を目指して, ” 電子情報通信学会技術研究報告, SP 2002-12, pp. 31-36, 2002.

[2] M. J. F. Gales, S. J. Young: “Robust continuous speech recognition using parallel model combination, ” IEEE Trans. on Speech and Audio Process, . 4, pp. 352-359, 1996.

[3] F. H. Liu, R. M. Stern, X. Huang, A. Acero: “Efficient cepstral normalization for robust speech recognition, ” Proc. ARPA Workshop on Human Language Technology, pp.69-74, Princeton, USA, March 1993.

[4] B. Kingsbury, N. Morgan: “Recognizing Reverberant Speech with RASTA - PLP, ” Proc. ICASSP1997, Vol. 2, p. 1259, 1997.

[5] 広林茂樹, 野村博昭, 小池恒彦, 東山三樹夫: “パワーエンベロープ伝達関数の逆フィルタ処理による残響音声の回復, ” 電子情報通信学会論文誌, J81-A, No. 10, pp. 1323-1330, 1998.

[6] 木下慶介, 中谷智広, 三好正人: “マルチステップ線形予測を用いた 1ch 残響除去法の検討, ” 日本音響学会 2006 年春季発表講演論文集 1-5-3, 2006.

[7] T. Yoshioka, T. Nakatani, T. Hikichi, M. Miyoshi: “ Maximum likelihood approach to speech enhancement for noisy reverberant signals, ”Proc. of ICASSP, pp.4585-4588, Apr., 2008.

[8] M. Delcroix, S. Watanabe, T. Nakatani: “Combined Static and Dynamic Variance Adaptation for Efficient Interconnection of Speech Enhancement Pre-Processor with Speech Recognizer, ” Proc. of ICASSP, pp.4075-4076, Apr., 2008.

[9] M. J. F. Gales, P. C. Woodland: “Mean and variance adaptation within the MLLR framework, ” Computer Speech and Language, vol. 10, pp. 249-264, 1996.

[10] 滝口哲也, 中村哲, 鹿野清宏 : “雑音と残響のある環境下でのHMM合成によるハンズフリー音声認識法, ” 電子情報通信学会論文誌, Vol.J79-D-2, No.12, pp. 2047-2053, 1996.

[11] 山本仁, 西本卓也, 嵯峨山茂樹: “モデル合成法を用いた複数フレームにまたがる残響下の音声認識, ” 日本音響学会 2003 年秋季発表講演論文集 1-6-7, 2003.

[12] 槻武也, 西本卓也, 嵯峨山茂樹: “ 音響モデル変換による残響環境中の音声認識, ”電子情報通信学会技術研究報告, SP2004-150, pp.31-36, Jan., 2005.

[13] C. K. Raut, T. Nishimoto, S. Sagayama: “Adaptation for long convolutional distortion by maximum likelihood based state filtering approach, ”Proc. of ICASSP, May., 2006.

[14] L. Couvreur, C. Couvreur, C. Ris:“A corpus-based approach for robust ASR in reverberant environments, ” Proc. of ICSLP, Beijing, China, 2000, vol. 1, pp. 397-400.

[15] 清水泰博, 梶田将司, 武田一哉, 板倉文忠: “空間音響特性を考慮したスペースダイバシチ型音声認識, ” 電子情報通信学会論文誌, J83-DII, No. 11, pp. 2448-2456, 2000.

[16] T. Haderlein, E. Noeth, W.. Herbordt, W. Kellermann: “Using artificially reverberated training data in distant-talking ASR, ” Proc. Text, Speech, and Dialogue, Carlsbad, Czech Republic, Sep. 2005.

[17] V. Stahl, A. Fischer, and R. Bippus: “Acoustic synthesis of training data for speech recognition in living room environments, ” Proc. of ICASSP, Salt Lake City, 2001.

[18] 馬場朗, 李晃伸, 猿渡洋, 鹿野清宏: “残響適応音響モデルを用いた音声認識, ” 日本音響学会 2002 年秋季研究発表会講演論文集, 1-9-14, pp. 27-28, 2002.

[19] L. Couvreur, C. Couvreur:“Blind Model Selection for Automatic Speech Recognition in Reverberant Environments: Special Issue on Real World Speech Processing, ” The Journal of VLSI Signal Processing, Volume 36, Numbers 2-3, pp. 189-203, Feb. 2004.

[20] DVD 版 建築と環境のサウンドライブラリ, 日本建築学会編 (SMILE2004)

[21] 残響下音声認識評価環境（CENSREC-4）

[22] 雑音重畳日本語連続数字 音声認識評価環境 (CENSREC-1)