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.

音響モデル変換による残響環境中の音声認識

槻　　武也　　西本　卓也　　嵯峨山茂樹

東京大学大学院情報理工学系研究科
〒 113–8656 東京都文京区本郷 7–3–1
E-mail: {saikachi, nishi, sagayama}@hil.t.u-tokyo.ac.jp

あらまし　本稿では、残響に頑健な音声認識を実現するため、クリーン音声のモデルと残響特性を与えて、残響環境に適応させた音響モデルを作成する手法について議論する。実環境では、音源からの直接音に加えて壁からの反射などによる残響成分が重畳した音声信号が観測される。残響時間がフレーム長に対して長い場合、観測信号には観測フレーム以前の信号が伝達歪みを受けて残響成分として重畳される。このため、残響による歪みは観測フレーム以前のフレームの音声に依存している。そこで、本手法では変換する音素に対してその直前にある音素列の可能性を場合分けし、それぞれの場合で残響モデルを求める。そして残響モデルを音素列の出現確率によって重ね合わせて、変換結果とする。残響モデルの求め方としては、各フレームの残響成分を独立した分布とみなしてモデル合成をする方法と、HMM から MFCC の出力系列を構成し、直接計算した残響を残響分布の平均とする 2 通りの方法を提案する。実験評価のため、残響環境下の音声の特定話者孤立単語音声認識実験を行い、認識率の向上を確認した。
キーワード　音声認識, 耐残響, モデル適応

Reverberant Speech Recognition Based on Acoustic Model Conversion

Takeya SAIKACHI, Takuya NISHIMOTO, and Shigeki SAGAYAMA

Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: {saikachi, nishi, sagayama}@hil.t.u-tokyo.ac.jp

Abstract　This paper discusses acoustic model conversion techniques for robust speech recognition under a reverberant environment given an acoustic model for clean speech recognition and the characteristics of reberveration. In real environments, acoustic signals are often distorted by overlapping reflective sounds. If the reverberation time is significantly longer than the frame length, the observed signal is affected by reverberant signals derived not only from the present frame but also from preceding frames. Therefore, reverberant distortion depends on signals in preceding frames. In our approach, we calculate reverberant distributions of all possibilities of phoneme that may arise at the past frames before the present frame. Then, we build reverberant HMM by adding these distributions. For building reverberant HMM models, we propose two methods: one is to add the effect of reverberation of preceding frames by model composition; the other is to compose output pattern of MFCC from HMM and then convolve reverberant characteristics. We evaluated the proposed method with speaker dependent isolated word speech recognition. A slight improvement was found in word accuracy rate.
Key words　Robust reverberant speech recognition, model adaptation

1. はじめに

本稿では、残響に頑健な音声認識を実現するため、クリーン音声のモデルと残響特性を与えて、残響環境に適応させた音響モデルを作成する手法について議論する。

近年、実環境における音声認識が重要な研究テーマとなっている。実環境中では、周囲雑音や残響などの要因によってモデル学習時と認識時の環境に不整合が生じ認識性能が劣化する [1]。実環境におけるノイズには大きく分けて乗法性雑音と加法性雑音の二種類があり、周囲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 [9] が用いられる。また乗法性歪みの中でも、回線特性やマイクロ

ホン特性による歪みに対しては、ケプストラム平均減算法などによって対処できる。

しかし、残響による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ない。なぜならば、残響環境中では音源からの直接音のほかに壁や床からの反射音が観測されるが、これらは現在見ている分析フレームよりも過去のフレームの音声信号に依存して変動するからである。

従来、残響に対しては大きく分けて 3 通りの方法が提案されてきた。それぞれ、頑健な特徴量を認識に用いる方法、観測信号から残響を除去する方法、音響モデルを残響環境に適応させる方法である。

残響に頑健な特徴量として RASTA [2] が提案されている。RASTA では、フロントエンド処理によって残響に影響されにくい特徴量の抽出を行う。

観測信号から残響を除去する方法が提案されている。多くの方法は文献 [3] のように聴覚的な明瞭性の改善を目的としているので、音声認識に用いる特徴量の次元ではクリーン音声に近いとは限らない。また、文献 [4] では音声認識のために特徴量の補正を行っているが、認識率の向上が十分でないことが報告されている。

モデル適応法では、クリーン音声で学習を行った音響モデルを何らかの環境情報によって残響音声モデルに変換することで、学習環境と認識の環境の不整合を解消する。MLLR 法 [5] では少量の残響音声を再学習データとして与えることで、デコーダに与える音響モデルを残響音声に適応させる。モデル適応は音声認識とは独立しており、特徴量の線形写像を求めることによって行う。滝口らの方法 [6] では、環境のインパルス応答を与える。ケプストラム領域における乗法性雑音の定式化に基づいてモデルを変換する。ともに、原理的には分析フレームより短い時間の残響への対処が可能である。山本らの研究 [7] では、環境の残響特性 (インパルス応答) を与えることで残響成分を予測してフレームごとにモデル適応を行う。モデル適応は、認識時にデコーダが行う。分析フレームより長い残響を扱うことが可能であるが、フレームごとに適応を行うため認識は低速である。本研究では、インパルス応答を入力とした音響モデル変換を行う。音響モデルの変換は音声認識とは独立して行い、デコーダには変更を加えない。また、分析フレームよりも長い残響を扱うことが可能である。

以下、2 章で本手法の枠組と残響に関する基本的な表現について述べ、3 章で提案手法であるモデル変換法について述べる。そして 4 章で音声認識実験による評価結果を報告し、考察を加える。

2. 残響音声の定式化

2.1 提案手法の枠組

音声認識において耐残響などの環境適応手法を用いる場合、その手法の認識性能のほかに、扱いやすさも重要な要素となる。たとえばその手法が、非常に手間がかかったり、音声認識システム全体を大きく変更しなければならないとすれば、多少認識性能が高くてもあまり良い手法であるとは言えない。音声認識システム自体に与える影響を最小限に、ある程度頑健な音声認識を行なうことができれば、それは非常に扱いやすい手法であると言える。

前章でも述べたように、残響に対しては大きく分けて 3 通りの手法が提案されてきたが、扱いやすさという目的に対してはモデル適応法がもっとも適している。文献 [5] [6] のように音声認識とは独立してモデル適応を行なう方法ならば、システムには変更を加えずに、これに与える音響モデルの変換のみで残響への対応が可能となるからである。しかしながら、既存のモデル適応法では残響のインパルス応答が分析フレームよりも短いと仮定しており、フレーム外からの長い残響を考慮していない。そのため、実際の残響環境で認識性能が劣る要因になっている。したがって、フレーム外からの残響を考慮することで性能の高い耐残響モデル適応が可能になると考えられる。ところが、このような残響は観測フレームよりも過去のフレームに発声された音素に依存しており、通常はこれを知ることはできない。そこで、音素列を確率的に扱うなどして、解決する必要がある。

以上を踏まえ、本研究ではフレーム外の長い残響を考慮したモデル適応を行なう。本研究で対象とする音響モデルは、状態間の遷移確率と、各状態ごとの MFCC と ΔMFCC の出力確率分布によって構成される一般的な音素 HMM である。出力確率分布は平均と分散で与えられる正規分布である。

また、変換の際環境の特性を表す情報として、環境のインパルス応答を与える。インパルス応答は事前に測定することが可能である。ただし、インパルス応答には音源や観測点の位置、気温や湿度など室内の些細な状況変化によって敏感に変動するため、波形領域では環境変動に対するロバスト性が低いという問題がある。これに対しては、インパルス応答を直接使うのではなく、よりロバストな、ここでは短時間スペクトルの形で使用する。

2.2 スペクトル領域における残響表現

まず準備として、短時間スペクトル領域における残響の表現について述べる。残響環境の影響は、波形領域では環境のインパルス応答を用いて記述することができる。すなわち、残響環境における観測信号を $o(t)$ とすると、これは以下のように環境のインパルス応答 $h(t)$ と原音声 $s(t)$ の畳み込みで表すことができる。

$$o(t) = h(t) * s(t) \tag{1}$$

短時間スペクトル領域において、インパルス応答 $h(t)$ が分析フレームよりも短い場合には観測信号 $o(t)$ の短時間スペクトル $O(n, w)$ は次の式で表される。

$$O(\omega, n) = H(\omega)S(\omega, n) \tag{2}$$

ここで、H と S はそれぞれインパルス応答と原音声の短時間スペクトル、n と ω はフレーム番号と周波数（フィルタバンク分析においては、フィルタバンク番号）である。多くの従来法では、残響を含め乗法性の歪みを式 (2) によってモデル化している。

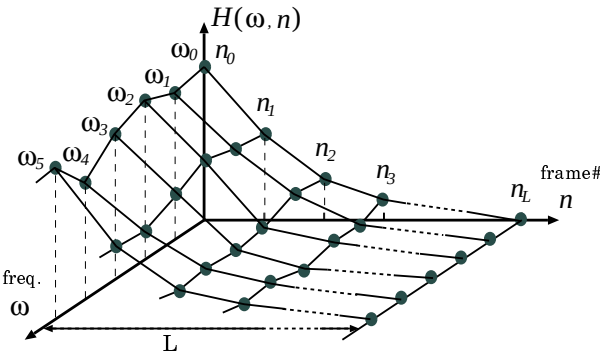


図 1　残響特性の短時間周波数応答による表現

Fig. 1　Characteristics of reverberation expressed by a response of short-time spectra.

しかし、インパルス応答が分析フレームよりも長い場合には式 (2) は成り立たない。この場合短時間スペクトル領域では、あるフレーム m に対して、そのフレームとそれ以前のフレームのスペクトル $S(\omega, m)$, $S(\omega, m-1)$,$\cdots$, がそれぞれ伝達特性 $H(\omega, 0)$, $H(\omega, 1)$,$\cdots$, による歪みを受けて重なっている。伝達特性 H は、図 1 に模式的に示すようなフレームおよび周波数帯域ごとのパワーの減衰を表している。ここで、各フレームのスペクトルの和をとった時に位相成分の和の期待値が 0 になると仮定すれば、この重なりは期待値の意味でスペクトルの和で表せる。これは、加法性雑音と同様の関係である。

以上から、残響は乗法性雑音と加法性雑音の両者を併せた形である、フレーム単位の畳み込みによって定式化される。これを式で表すと、次のような形になる [8]。

$$O(\omega, n) = H(\omega, n) * S(\omega, n) \tag{3}$$

$$= \sum_{l \geqq 0} H(\omega, l) S(\omega, n-l) \tag{4}$$

右辺で加算される各項の $H(\omega, l)S(\omega, n-l)$ は、フレーム n からみて l フレーム前に発声した音声に起因する残響成分であると言える。この式では通常残響と呼ぶ反射音のほかに、伝送路歪みを受けた直接音も内部に含んでいる。このため、以下では直接音も残響成分として等価に扱うことができる。

3. 残響音声のためのモデル変換

3.1 先行音素列と継続時間が既知の場合

本節では、任意のフレームに対して先行するフレームの音素とその継続長が既知であるような場合のモデル適応法について述べる。このような状況は一般的ではないが、例えば 2 パス探索の第 2 パスにおいて精度の高い尤度比較を行う場合に使うことができる。

式 (4) で示したように、あるフレームにかかる全残響はそれ以前のフレームに起因する残響成分をスペクトル領域で加法したものである。先行する各フレームからの残響成分が互いに独立であると仮定すれば、残響全体のモデルは各フレームからの残響成分のモデルを加法性雑音のためのモデル合成法によって合成することで得られる。モデル合成法としては PMC 法 [9] や JA 法 [10]、LPA 法 [11] などが利用でき、計算速度や精度の要請によって選択をすることができる。本研究では、PMC 法を用いる。なお、この方法では複数の残響成分の分布を加法するので合成モデルの分布が非常に大きくなる場合がある。大きすぎる分散は認識に悪影響を与えるので、モデル合成後に分散を減らすことが考えられる。本研究では、変換後に分散をクリーン音声モデルの分散まで戻す。

続いて先行する各フレームからの残響成分のモデルを与える。ここでは、n フレームからみて l フレーム前の音声による残響成分の分布、すなわち式 (4) の右辺における $H(\omega, l)S(\omega, n-l)$ について考えるとする。これは、ケプストラム領域で和に分解することができる。任意のスペクトル X の MFCC を C_X と表すことにすると、次のように書ける。

$$C_{H(\omega,l)S(\omega,n-l)} = C_{H(\omega,l)} + C_{S(\omega,n-l)} \tag{5}$$

問題設定より各フレームの音素が既知なので、$C_{S(\omega,n-l)}$ はクリーン HMM から $n-l$ フレームの音素の出力確率分布として得られる。また、$C_{H(\omega,l)}$ はインパルス応答の MFCC である。従ってこの残響成分は、$n-l$ フレームの音素の出力確率分布に平均を $C_{H(\omega,l)}$ 加算し、分散は変更しない分布で表されることになる。ここではインパルス応答をスペクトルの形で用いているので、2.1 節で述べたインパルス応答のロバスト性の低さが解決されているはずである。

これにより、先行する音素とその継続長が与えられている場合のモデル変換法が得られる。以下に変換のアルゴリズムを示す。ここでは、フレーム m の音素 (状態) を $p(m)$ とし、音素 p のモデルの平均を μ_p、分散を σ_p とおく。

（1） フレーム n の残響のうち、フレーム $n-l$ の音声に起因する成分の音響モデルを求める。これは、平均 $\hat{\mu}(n-l) = \mu_{p(n-l)} + C_{H(\omega,n-l)}$、分散 $\hat{\sigma}(n-l) = \sigma_{p(n-l)}$ の分布である。

（2） $l \geqq 0$ の各フレームに対して (1) のモデルを求める。

（3） (2) の各残響成分モデルをモデル合成法により合成し、これを音素 $p(n)$ の残響モデルとする。

3.2 先行音素列が既知で継続時間が未知の場合

前節の議論から一歩進めて、音素の継続長の情報を用いずにモデル適応を行う方法を述べる。先行音素のトランスクリプションは依然既知である。これは、木構造モデルや単語 HMM を残響に適応させる場合に利用可能である。

音素の継続長が与えられていないとすると、あるフレームの前に来る音素系列は音素の長さの伸縮によって様々なパターンをとりうる。ここでは多様なパターンを一つのモデルで扱うために、これらを確率的に統合する。すなわち、継続長の長さによって場合分けしてモデル変換を行い、得られたモデルに各々の場合の出現確率を掛けて重ね合わせることでミクスチャモデルを構成する。各音素継続長の出現確率をなんらかの方法で用意する必要があるが、これには全ての可能性を等確率であるとしたり、モデルの状態遷移確率を確率変数とする負の二項分布によって与えるなどの近似的な方法が考えられる。本研究では、より簡単に継続フレーム数が音素の遷移確率の逆数である場合

が確率 1 で出現すると考える。

これを選んだのは、負の二項分布の期待値が確率変数の逆数となるからである。

構成されたミクスチャモデルをそのまま認識に用いることも可能ではあるが、多すぎる場合には少数のミクスチャに統合する。これには EM アルゴリズムによる GMM 推定法などが利用できる。

3.3 先行音素列が未知の場合

前節の議論をさらに進め、先行音素の情報が内容と継続長ともに与えられていない場合のモデル変換法を述べる。本研究で扱うモデル変換がこの状況に該当する。

まず、継続長に関しては前節の方法で扱う。そして、同様に先行音素の内容についても確率的に扱う。すなわち可能なすべての場合のモデル変換を行い、出現確率をかけて重ね合わせる。ただし、先行音素列のパターンを網羅することは計算量や実現性の面から見て困難なので、ここでは削減を行う。

ここでは、残響の主要な成分が現在とその直前の音素の 2 音素がカバーする時間内のみに由来し、それ以前は無視できると仮定する。これは、残響のパワーには時間的な偏在性があり、残響成分の大部分は残響時間に比べ短い時間内に存在しているという考え方に基づく。音素出現率は音素 bi-gram としてコーパスから用意する。なお、この仮定によってある長さより長い残響への対応が難しくなる可能性があるが、その場合はさらに前の音素を含めることで解決できると考えられる。ただし、この場合はクラスタリング等によって削減する必要があると考えられる。

実際の変換のアルゴリズムを以下に示す。

（1） 変換対象とする音素に対して、それとその直前の音素よりなる 2 音素の音素列を構成する。例えば、変換する音素が/a/の場合は/sa/や/ka/のような音素列である。各音素の継続フレーム数は、3.2 節の方法を用いる。

（2） 各音素列に対して 3.1 節の方法で残響のかかった/a/の音響モデルを求める。

（3） 各モデルに音素列の出現率を掛けて和をとり、ミクスチャモデルを構成する。

（4） GMM 推定によって適切な数のミクスチャに統合し、最終的な/a/の残響モデルとする。

3.4 スペクトル領域の畳み込みによるモデル変換

本節では 3.1 節の条件に戻り、これとは別のモデル変換法を提案する。3.1 節では、先行する各フレームからの残響成分が独立であると仮定して残響全体の分布を求めた。しかし隣接するフレームの特徴量が連続的に変化するなどの理由で、残響成分は各フレーム間で完全に独立ではないと考えられる。このため 3.1 節の方法では実際にはありえない特徴量系列による残響をモデルの中に含むことになり、認識率の低下要因となる可能性がある。ここでは HMM パラメータを直接に出力確率分布として扱うのではなく、パラメータから出力系列を生成することで残響特性を畳み込んで直接残響パラメータを計算する。出力系列は無数に考えることができるが、ここでは最尤系列に対してのみ残響計算を行い、得られた残響パラメータをモデルの平均として使う。分散については計算できないため、クリーン音声モデルの分散をそのまま使用する。

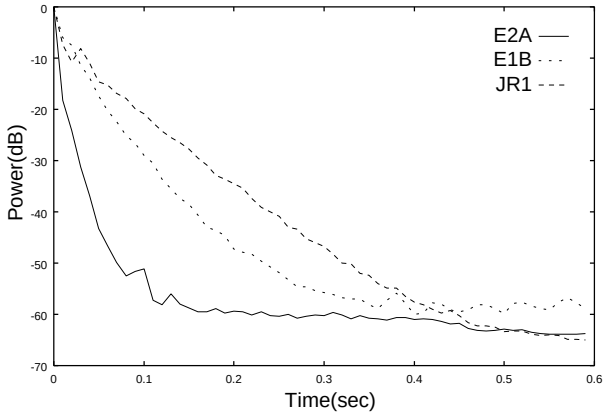


図 2 評価実験に使用した 3 種類のインパルス応答の特性
Fig. 2 Characteristics of impulse response used in evaluation.

問題設定により、先行フレームの音素は既知である。ΔMFCC などの動的特徴を考慮しないでよいと仮定すれば、最も尤度の高い出力ベクトル系列は各フレームに対応する音素 HMM の出力平均によって与えられる。この出力ベクトル系列は、これを逆フーリエ変換し、exp を施すことによってフィルタバンク領域まで戻すことができる。すると、式 (4) によって残響特性の畳み込みが可能になる。残響特性を畳み込んだのち再度 log をかけフーリエ変換することで、残響 MFCC 系列が得られることになる。

変換のアルゴリズムを示す。3.1 節同様先行する音素と長さは既知である。このため、本研究で対象とする条件のためには、3.2 節以下の方法をあわせて行う必要がある。ここではフレーム m の音素 (状態) を $p(m)$ とし、音素 p のクリーン HMM の平均を μ_p とおく。

（1） クリーン HMM から、音素列に対する出力値系列 $\hat{S}(n) = \mu_{p(n)}$ を生成する。

（2） $\hat{S}(n)$ をフィルタバンク領域に逆変換し、式 (4) によって残響を畳み込む。

（3） (2) を再度 MFCC 領域に変換して、フレーム n の残響 MFCC を音素 $p(n)$ のモデルの新しい平均とする。

4. 評　　価

上記提案手法によりフレーム外の残響に対処できる音響モデルを構築できたかを検証するため、残響を畳み込んだ音声による特定話者の孤立単語音声認識実験を行った。

4.1 実 験 条 件

評価データには ATR 音声データベース A セットの 655 単語を用い、計算機上で残響環境のインパルス応答を畳み込んで残響音声とした。本手法では比較的長い残響を対象としているが、3.3 節で述べたように残響時間がある程度長くなると性能向上が低下する可能性がある。そこで、残響時間の異なる 3 種類のインパルス応答を選び比較を行う。インパルス応答には、RWCP 実環境音声・音響データベースから E2A(残変室・パネル)、E1B(残変室・シリンダ)、JR1(畳部屋・大) を用いた。図

表 1　提案手法による残響音声の単語認識率（%）

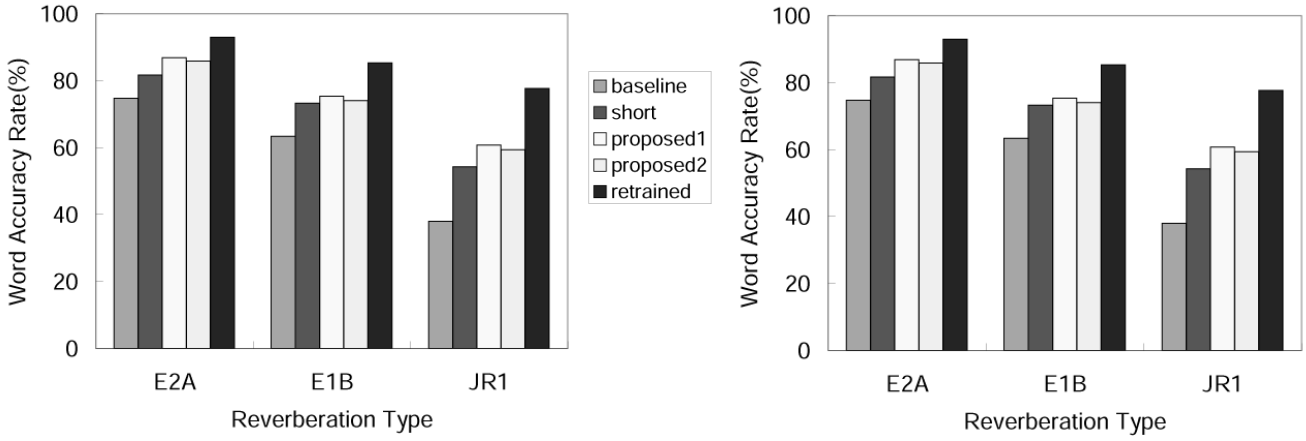
Table 1　Word recognition rate（%） for reverberant speech.

(a) Speaker: MAU(male)

	baseline	short	proposed1	proposed2	retrained
E2A	74.81	81.68	86.87	85.95	92.98
E1B	63.36	73.28	75.42	74.05	85.34
JR1	38.02	54.35	60.76	59.39	77.71

(b) Speaker: FFS(female)

baseline	short	proposed1	proposed2	retrained
72.52	74.50	81.37	81.83	94.35
57.56	75.27	75.11	75.42	85.50
37.25	58.32	60.15	62.90	80.31



(a) Speaker: MAU(male)　　(b) Speaker: FFS(female)

図 3　提案手法による残響音声の単語認識率（%）

Fig. 3　Word recognition rate (%) for reverberant speech.

2 に各インパルス応答の特性を示す。なお、マイクロフォンと音源の距離は約 2m である。

音声を 16kHz、量子化ビット数 16bit でディジタル化し、フレーム長 25ms、フレーム周期 10ms で分析した。高域強調の定数 $\alpha = 0.97$ とし、窓関数にはハミング窓を用いた。音響特徴量は 0 次を含む 13 次元の MFCC と ΔMFCC の計 26 次元とし、24 個のフィルタバンクによって計算した。音響モデルは 41 音素で 3 状態単混合のモノフォン HMM とし、評価データと同一話者でかつ評価データと重複しない 2620 単語によって学習した。学習したクリーン音響モデルに、評価音声に畳み込んだものと同一のインパルス応答を用いて提案手法のモデル変換を施した。ただし、変換時に用いる各音素列の出現確率は認識対象の単語から事前に求めて与えた。デコーダには Julian [13] を用いて、以上の条件で孤立単語音声認識を行った。

比較手法として、フレーム外の残響に対応できているかを検証するため、インパルス応答を最初の 1 フレームだけ与えて、モデル変換を行った。他のフレームの残響は存在しないと考える。これは直接音の伝送路歪みのみに対応したモデル変換であると考えられる。また、2.1 節で述べたインパルス応答の変化に対するロバスト性を検証するため、評価音声に畳み込んだものとは若干異なるインパルス応答によるモデル変換も行った。これには、評価音声と同一環境だが約 30cm 離れたマイクロフォンで観測したインパルス応答を用いた。

4.2　実 験 結 果

結果を表 1 及び表 2 に示す。表 1(図 3) は 2 名の話者 MAU(男性)・FFS(女性) について、各インパルス応答を畳み込んだ音声の単語認識率 (%) である。用いた手法を以下に示す。

- クリーン音声モデル (baseline)
- 最初の 1 フレームの残響のみで変換したモデル (short)
- 3.1 節の手法を施したモデル (proposed1)
- 3.4 節の手法を施したモデル (proposed2)
- 評価音声と同一の残響特性下の音声で学習したモデル (retrained)

表 2(図 4) は 1 名の話者 MAU(男性) について、残響音声に畳み込んだものと若干異なるインパルス応答を与えて 2 つの提案手法を施した結果である。それぞれの意味は以下である。

- クリーン音声モデル (baseline)
- 評価音声用と同一のインパルス応答を与えて変換したモデル (matched)
- 評価音声用と異なるインパルス応答を与えて変換したモデル (mismatched)

4.3　考察と今後の展望

表 1 から、proposed1、proposed2 ともに提案手法による認識率の向上が確認されたが、retrained に比べ低く、改善の余地がある。残響成分を入れたことによる改善効果は環境によって異なり、また単純に残響時間との関連はみられない。このことから、認識率が低いのは考慮に入れた残響が短かったというよりも、残響推定の精度や 3.2 節と 3.3 節で行ったパターン削減による偏りが原因であると考えられる。これを解決するために、例えばコンテキスト依存 HMM に対して本手法を適用することが考えられる。コンテキスト依存 HMM では、ある音素に対してその直前にある音素が決定できるため、3.3 節で直前音

表 2　評価音声に畳み込んだものと若干異なるインパルス応答を与えて、提案手法を行った場合の残響音声の単語認識率（%）

Table 2　Word recognition rate（%）for reverberant speech under slightly different impulse response from the one convolved to evaluate speech.

	baseline	proposed1		proposed2	
		matched	mismatched	matched	mismatched
E2A	74.81	86.87	85.34	85.95	86.11
E1B	63.36	75.42	72.37	74.05	72.82
JR1	38.02	60.76	55.88	59.39	57.86

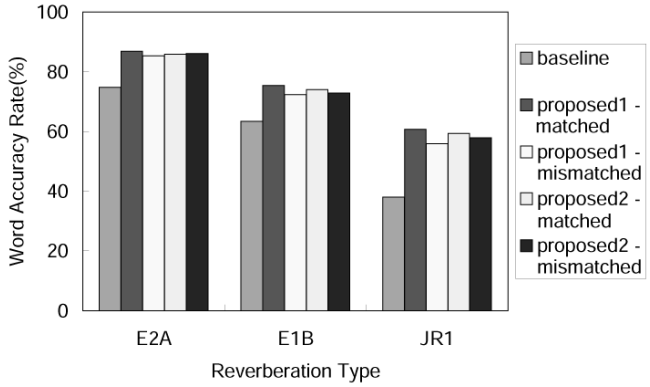


図 4　評価音声に畳み込んだものと若干異なるインパルス応答を与えて、提案手法を行った場合の残響音声の単語認識率（%）

Fig. 4　Word recognition rate（%）for reverberant speech under slightly different impulse response from the one convolved to evaluate speech.

素の内容を統計的に扱う必要はない。また、proposed2 ではフレーム間の動的特徴量 (ΔMFCC) を考慮しないでよいと仮定して残響の推定を行った。動的特徴量を考慮した出力ベクトル列の導出は既に音声合成の分野で解決されているので [12]、これを導入することでより歪みの少ない残響モデルの推定が期待できる。

表 2 では、各環境において mismatched でも十分な認識率向上が得られた。このことにより、本手法がインパルス応答の変動に対して頑健であることが確認された。今後はさらに困難な状況として、異なる環境で計測されたインパルス応答を与えたときの認識率の変化を調べる必要がある。仮に一つのインパルス応答である程度多くの残響環境がカバーできるならば、未知の環境でも事前に用意したインパルス応答の中から類似した環境を選ぶことで、ある程度良好なモデル変換ができると考えられる。さらに、いくつかのインパルス応答による残響音声モデルを統計的に組み合わせれば、多くの残響環境でクリーン音響モデルよりも良い結果を与える汎用的なモデル変換ができる可能性がある。

5. おわりに

本稿では、残響に頑健な音声認識を実現するため、クリーン音声のモデルと残響特性を与えて、残響環境に適応させた音響モデルを作成する手法を提案した。変換する音素に対してその直前にある音素列の可能性を場合分けし、それぞれの場合で残響モデルを求めた。そして残響モデルを音素列の出現確率によって重ね合わせて、変換結果とした。残響モデルの求め方としては、各フレームの残響成分を独立した分布とみなしてモデル合成をする方法と、HMM から MFCC の出力系列を構成し、直接計算した残響を残響分布の平均とする 2 通りの方法を提案した。残響下音声の特定話者孤立単語音声認識実験により認識率の向上が確認できたが、マッチドモデルに比べて十分な性能ではなくまだ改善の余地がある。

今後の課題としては、残響分布推定の精度向上やコンテキスト依存 HMM の導入などが挙げられる。また、いくつかのインパルス応答による変換結果を統計的に組み合わせることで、多くの残響環境で汎用的に使用可能な残響モデルへの変換を行うことが考えられる。

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] H. Hermansky, N. Morgan: “RASTA Processing of Speech,” IEEE Trans. on Speech and Audio Process. 2, pp. 578-589, 1994.

[3] R. Mukai, S. Araki, S. Makino: “Separation and Dereverberation Performance of Frequency Domain Blind Source Separation for Speech in a Reverberant Environment,” Proc. Eurospeech 2001, pp.2599-2602, 2001.

[4] C. Avendano, S. Tibrewala and H. Hermansky: “Multiresolution channel normalization for ASR in reverberant environments,” Proc. Eurospeech 97, pp. 1107-1110, 1997.

[5] 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.

[6] 滝口哲也, 中村哲, 鹿野清宏: “加法性雑音、伝達特性による歪みを受けた音声の HMM 合成による認識,” 日本音響学会 1995 年秋季研究発表会講演論文集 1-2-2, 1995.

[7] 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 年秋季研究発表会講演論文集 1-6-7, 2003.

[8] 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音源波形回復,” 電子情報通信学会論文誌, Vol. J81-A, no. 10, pp. 1323-1330, 1998.

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

[10] S. Sagayama, Y. Yamaguchi, S. Takahashi and J. Takahashi: “Jacobian approach to fast acoustic model adaptation,” Proc. ICASSP97, pp. 835-838, 1997.

[11] C. K. Raut, T. Nishimoto, S. Sagayama: “Model Composition by Lagrange Polynomial Approximation for Robust Speech Recognition in Noisy Environment,” Proc. ICSLP2004, 2004.

[12] 徳田恵一, 益子貴史, 小林隆夫, 今井聖: “動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム,” 日本音響学会誌, vol. 53, no. 3, pp. 192-200, 1997.

[13] http://julius.sourceforge.jp/