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.

フレームごとのモデル合成による残響下音声認識

山本　仁　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: {yamamoto,nishi,sagayama}@hil.t.u-tokyo.ac.jp

あらまし　本稿は，残響に頑健な音声認識を実現するための，音響モデルを残響音声に逐次的に適応させる方法について報告する．残響のある室内では，音源からの直接音だけでなく壁などからの反射音（残響成分）が観測される．このとき，ある音声分析フレームにおける残響成分は，それより前のフレームの音源信号に応じて時々刻々変動する．したがって，残響による特徴量のひずみを，各フレーム同一の乗法性ひずみとして単純に扱うことはできない．そこで本稿では，残響によるひずみを 2 種類にわけ，直接音の伝搬特性としての乗法性の成分と反射音としての加法性の残響成分として扱う．あるフレームに重畳する残響成分をそれ以前の音響信号から推定し，これを加法性雑音とみなして音響モデルに合成することによって，残響下音声に対する動的なモデル適応化を行なう．評価として残響下音声の特定話者孤立単語音声認識実験を行ない，効果を確認した．さらに，フレームごとのモデル適応化にヤコビ適応法を用いることについての検討を行なった．

キーワード　残響，フレーム同期，モデル適応

Reverberant Speech Recognition Using Frame-Synchronous Model Composition

Hitoshi YAMAMOTO, Takuya NISHIMOTO, and Shigeki SAGAYAMA

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

Abstract　This paper describes a technique for robust speech recognition under a reverberant environment. Acoustic signals are distorted by its reflection inside a room, and this reverberation degrades the performance of automatic speech recognition system. When the effect of reverberation spans for several frames, dealing such the distortion due to reverberation simply as convolutive distortion is not sufficient. It is necessary to consider the reverberation effect of other frames over current frame. In this paper, the frame-synchronous estimation of reverberation component and HMM combination are proposed. The proposed technique was evaluated with an isolated word reverberant speech recognition task, and the result showed improvement in recognition performance. Moreover, using Jacobian adaptation as a frame-by-frame model adaptation technique is investigated.

Key words　Reverberation, Frame-Synchronous, Model Adaptation

1. はじめに

本稿は，残響に頑健な音声認識の実現するための，音響モデルを残響音声に逐次的に適応させる方法について報告する．

実環境における音声認識では，周囲雑音や残響などの存在によってモデル学習時と音声認識時の収音環境に不整合が生じるため，認識性能が劣化する [1]．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音声強調（雑音抑圧）手法や，モデル合成法などの環境適応化手法が用いられる．回線特性やマイクロホン特性などの乗法性ひずみに対しては，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ひずみは，各フレームで同一の乗法性ひずみとして単純に扱うことはできない．なぜなら，残響のある室内環境では音源からの直接音に加えて壁などからの反射音（残響成分）が観測されるが，このときある音声分析フレームにおける残響成分は過去の（自らとは異なる）フレームの音源信号に依存し

て時々刻々変動するためである．

残響の影響を受けた信号から残響のない信号を得るため，さまざまな音響処理手法が用いられているが，短時間スペクトルに基づく特徴量の領域ではひずみを生じる可能性がある．特徴量の補正方法 [2], [3] でも，十分なひずみの補正が得られていないことが報告されている．また，音響モデル側で対処する手法 [4], [5] では，長い残響のときには十分な性能向上を得るのが難しい．これに対しては，フレーム外からの影響を線形予測で推定する [6] が提案されている．

本稿では，音声の音素モデルと室内の音響伝達特性が与えられたときの，残響音声の認識問題をあつかう．マイクロホンは一つとする．この状況で，常にある時点の残響環境と発話に適応した音響モデルを得ることができれば，音声認識の頑健性の向上が期待できる．われわれは，残響成分を逐次的に推定し，これをモデル合成することによって残響環境下の音声に動的に適応する方法について検討した [7]．本稿では，新たに直接音のひずみも考慮した方法について評価を行ない，さらに，モデル適応としてヤコビ適応法を用いることについて検討する．

以下，2 章では残響のある環境における音声認識問題について，3 章で提案手法であるフレームごとのモデル合成について述べる．4 章では音声認識実験による評価結果を報告する．5 章ではフレームごとのヤコビ適応について検討する．

2. 残響が音声認識におよぼす影響

本章では，残響のある環境での音声認識の問題について述べる．ここでは，特徴量が短時間スペクトルに基づくものであり，音響モデルがクリーン音声（残響を含まない音声）の音素モデルであるとする．

2.1 残響下の音声認識

残響の存在する室内では，図 1 に示すように，音源からの直接音に加えて壁などからの反射音（残響成分）が観測される．これは残響や反響とよばれる現象である．ある区間の音声の反射音が後続区間の音声に重なることによって，音声認識システムの認識性能は低下する．連続音声認識では，残響時間が 0.4 秒をこえると，その残響下の音声で学習したモデル（matched モデル）を用いた場合でも，十分な認識性能が得られないことが報告されている [4]．

従来，残響の影響を受けた信号から残響のない信号を得るために，逆フィルタやマイクロホンアレーなどの音響処理手法が用いられてきた．しかし，多くの手法は主に聴覚的な音質や明瞭度の改善に焦点をあてている [2] ため，特徴量の領域では得られた音声がクリーン音声に近いものであるとは限らない．このことから，これらの手法を音声認識システムの前処理部として単純につけ加えても，十分な認識性能向上は期待できず，音声認識のための残響対策が必要であると考えられる．

残響下音声の特徴量を補正する手法のうち，残響時間を考慮した複数のフレーム長の分析を用いたもの [2] では，残響下の音声については改善が見られるものの残響のない音声でかえってひずんでしまう．特徴量系列からのフィルタによって音声成分を推定する方法 [3] では，その認識率向上が小さいことが報告されている．また，音響モデル側で対処する手法としては，複数の環境適応モデルからのモデル選択法 [4] の効果が大きいが，多くのモデルを持たなければならず，また選択を誤る可能性もある．音響伝達特性のモデル化に基づく HMM 分離法 [5] では，フレーム間の関係を扱っておらず，長い残響のときには十分な性能向上を得るのが難しい．

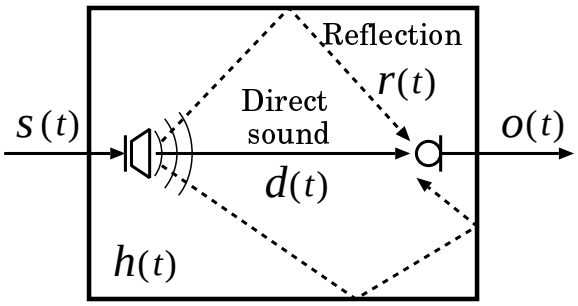


図 1　残響を含む室内の音響伝達経路
Fig. 1　Acoustical path inside a room.

2.2 残響の長さと音声認識

残響下の音声認識問題は，室内の音源（発声者）から観測点（マイクロホン）までの音響伝達特性によって扱いをわけて考えることができる．ここではそれを線形系と仮定し，インパルス応答の長さ（残響時間と考えてもよい）と，音声分析フレーム長／ HMM の状態継続時間／音素継続長との関係によって，次のように分類する．

（1） 残響時間が短いとき: 室内の伝達特性の影響は各フレームの観測スペクトルの変化として現れる．ひずみはチャネル適応などの静的なモデル適応化によって扱える．

（2） 残響時間が長いとき: 室内の伝達特性はフレームから別のフレームに加算的に影響をおよぼす．先行の音声の残響成分が後続の音声（別のフレーム）に重なり，入力音声とモデルとの間にミスマッチが発生するため，フレーム内だけでの適応処理は難しい．このことは,「フレーム」を「HMM の状態」や「音素」と置き換えても同様である．

音声認識では通常フレーム長が数十 ms であるのに対し，室内の残響時間は数百 ms である（例えば会議室では 0.7 秒前後，コンサートホールでは 1.5 秒前後，など）ので，多くの残響下音声認識問題は後者にあてはまると考えられる．

2.3 残響によるひずみの分離

以上を踏まえ，あらためて残響下の音声の特徴量のひずみを考える．残響環境では，前述の通り直接音と反射音とが観測されるので，それぞれのひずみについて対処することが効果的であると考えられる．直接音のひずみに対しては，それを時間遅れのみとみなせば特に対処は不要であり，スペクトルのなまけがあるとすればチャネル適応などによって対処することができる．一方，残響成分は現在のフレームに加算的に影響する．また，残響成分は前のフレームの音声が室内の伝達特性にたたみ込まれたものに由来するため，時々刻々変動する．このような状況では，各フレーム同一のひずみとして扱うことはできず，先行の音声に応じて動的にモデル適応化する必要がある．

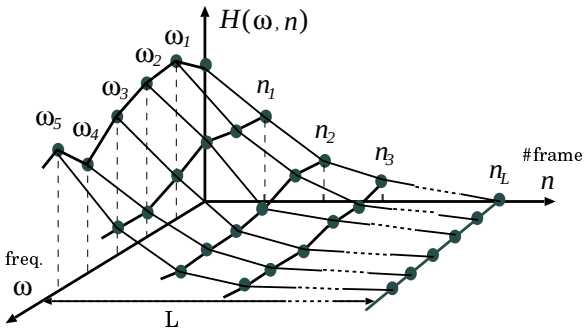


図 2　残響特性の短時間周波数応答による表現

Fig. 2　Characteristics of reverberation expressed by a response of short-time spectra.

本稿では，動的なモデル適応化手法として，過去のフレームに由来する残響成分を加法的雑音とみなしたフレームごとのモデル適応法を検討する．

3.　フレームごとのモデル合成

本章ではフレームごとのモデル合成について述べる．この手法は，各フレームにおいて推定した残響成分を音響モデルに合成するものである．

3.1　室内の残響特性の表現

まず準備として，短時間スペクトル領域での室内の残響特性の表現について述べる．室内の残響現象は，波形レベルでは室内のインパルス応答を用いて表現できる．図 1 で，時刻 t における観測信号 $o(t)$ は，$s(t)$ は音源信号，$h(t)$ はインパルス応答として次式にしたがう．

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

短時間スペクトル領域で，まず残響時間がフレーム長に比べて短い場合を考える．フレーム n における，観測信号 $o(t)$ の短時間スペクトル $O(\omega, n)$ は次式であらわされる．

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

ここで，$H(\omega)$ と $S(\omega, n)$ はそれぞれ $h(t)$ と $s(t)$ の短時間スペクトルである．また，ω は周波数であるが，たとえば特徴量が MFCC の場合はフィルタバンク（番号や中心周波数）に相当する．多くの従来手法では，残響も含めた乗法性のひずみを，式 (2) のように n によらない各フレーム同一のものとして扱っているといえる．

しかし一方，残響時間が長い場合は上式は成り立たない．前述のとおり，フレームごとの短時間スペクトル領域では，あるフレーム m のスペクトル $S(\omega, m)$ に，それ以前のフレームのスペクトル $S(\omega, m-1), S(\omega, m-2), \cdots,$ がそれぞれ伝達特性 $H(\omega, n)$ にしたがって重なっている．これはフレームを単位とした短時間スペクトルの系列のたたみ込みによって次式のように表わせる．

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

なおこの式は，H と S が無相関であるときの，同一周波数成分についての両者のランダムな位相関係の期待値として成り立つ．

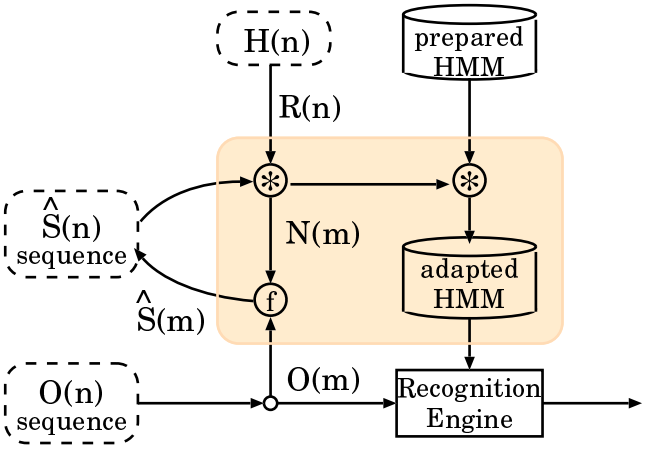


図 3　提案する手法の枠組み

Fig. 3　Framework of proposed technique.

式 (3) の短時間スペクトルの応答 H は，図 2 の模式図のように，周波数帯域ごとのパワーの減衰を表している．本稿では，H は室内の残響時間に対応する有限の応答長 L[フレーム] を持つとし，H によって室内の残響特性を記述できると仮定する．

3.2　残響成分を用いたモデル合成

ある時点の残響成分は，その前の区間の音響信号の反射音に依存すると仮定する．このとき，室内の残響特性 H を

$$H(\omega, n) = D(\omega, n) + R(\omega, n) \tag{4}$$

のように直接音の伝搬特性 D と反射音の伝搬特性 R にわけて考えると，式 (3) は次式のように書ける．

$$O(\omega, n) = D(\omega, n) * S(\omega, n) + R(\omega, n) * S(\omega, n) \tag{5}$$

直接音成分 $D * S$ は乗法性ひずみ D の影響を受けたクリーン音声と見なせる．特に D が平坦な特性（遅れのみ）ならば，クリーン音声のモデルを用いることができる．一方，残響成分 $R * S$ の項は，S の過去のフレームの短時間スペクトル列と R とのたたみ込みである．この成分を N とすると，これは次式によって求まる．

$$N(\omega, n) = \sum_{l=0}^{L} R(\omega, l) \cdot S(\omega, n-l) \tag{6}$$

この N を直接音に対する加法性雑音とみなし，フレームごとに音響モデルと合成すれば，過去の音響信号に応じた，残響下音声に適応したモデルを動的に生成することができる．図 3 に提案手法の枠組みを示す．

この手法では，収音条件や音声分析方法は変更していないので，モデル学習時と比べて特徴量がひずむおそれはない．フレームごとにモデルを適応するので一つのモデルを持てばよい．HMM 分離法 [5] とは，音響伝達特性を限られた状態にモデル化せず，多様な変動を認め，残響特性と先行の音声に応じた動的な適応を行なう点が異なる．線形予測によってフレーム外の影響を表現する方法 [6] とは，残響成分の推定法が異なる．なお，フレームごとにモデル合成を行なうため計算量は多い．

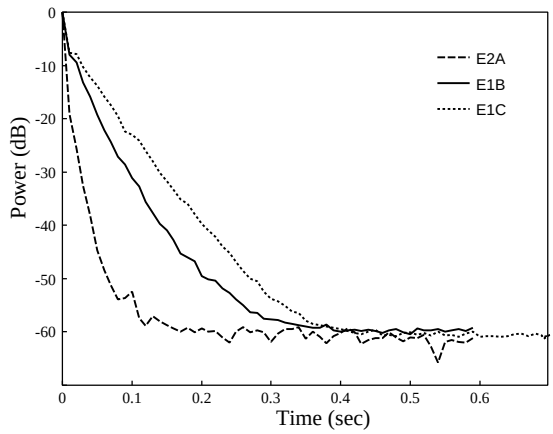


図 4　評価実験に使用した 3 種類のインパルス応答の特性
Fig. 4　Characteristics of impulse response used in evaluation.

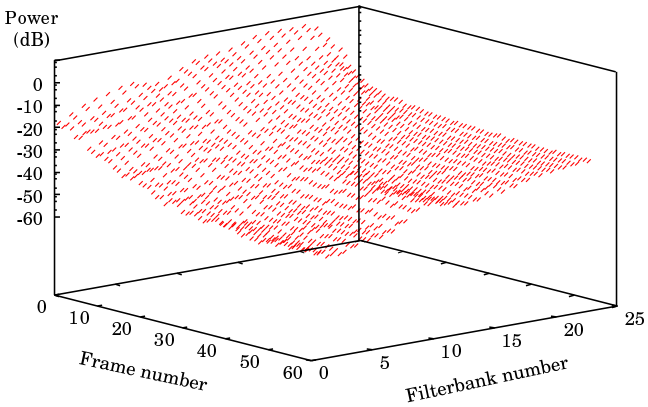


図 5　インパルス応答の短時間スペクトル領域での特性（E1B）
Fig. 5　Characteristics of impulse response in terms of short-time power spectra.

3.3　アルゴリズム

クリーン音声のモデルと観測信号 O のみが与えられるような現実的な問題では，式 (5) や式 (6) を用いるために，直接音の伝搬特性 D と反射音の伝搬特性 R，音源信号成分 S を推定する必要がある．何らかの方法により音源と観測点の間の伝達特性 H が得られた場合でも，D と R を得るのは困難である．そこで，ここでは便宜的に D と R は時間軸で重複しないとし，H の最初の数フレームを D，それ以後を R と分割して扱うことにする．これは D の影響は数フレーム外におよばないと仮定することに相当する．

発話開始時点では，それ以前の区間は無音のため残響成分は存在しない．すなわち $N = 0$ であり $O = D * S$ である．D は，理想的にはある時間の遅れとみなすことができる．本稿では，D を無視する方法（proposed1）と，D を無響室の特性で置き換える方法（proposed2）の 2 通りを行なう．

発話開始後では，(6) 式に基づいて N を求める．このとき音源信号を S を $\hat{S} = O - N$ で近似する．$N = 0$ を初期条件としてフレームごとに再帰的に $\hat{S}$ を推定できる．ここで得られる $\hat{S}$ は，それを音声認識するには精度が低すぎるが，残響成分を推定する程度には利用できるものと考え，上述の方法を適用する．

特徴量を MFCC，HMM の出力確率分布をガウス分布とするときの提案手法のアルゴリズムを以下に示す．MFCC 領域と短時間スペクトル領域の間の変換は PMC 法 [8] と同様の方法で行なうとする．

準備　：
（1）　音声の HMM と音響伝達特性 H を用意
　　（ただし，以降の任意の時点で変更可能）

初期化　：
（1）　D と R を設定（H を分割）
（2）　モデルパラメタをスペクトル領域に変換
（3）　$N(\omega, 0) = 0$，$S(\omega, 0) = O(\omega, 0)$

合成　：各フレーム n において
（1）　各フィルタバンク ω において

a. N の推定（(6) 式）
b. S の計算（$\hat{S}(\omega, n) = O(\omega, n) - N(\omega, n)$）
c. 分布平均に $N(\omega, n)$ を加算

（2）　モデルパラメタを MFCC 領域に再変換
（3）　適応モデルを用いて尤度計算

4.　評　　価

提案手法の評価として，残響をたたみ込んだ音声について特定話者の孤立単語認識実験を行なった．

4.1　実験条件

評価データは ATR 音声データベース A セットの 655 単語とし，残響環境を模擬するために計算機上でインパルス応答をたたみ込んだ．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 3 種類を用いた．音源とマイクロホンの距離は 2m である．使用したインパルス応答の特性を図 4 に示す．

入力音声を標本化周波数 16kHz，量子化ビット数 16bit でディジタル化し，フレーム長 25ms のハミング窓を用いて，周期 10ms で分析した．音響特徴量は 13 次の MFCC（0 次含む）と ΔMFCC の計 26 次元とし，24 個のフィルタバンクを用いて算出した．音響モデルは 3 状態 4 混合の音素環境独立 HMM（音素数 41）とし，評価データと同一話者でかつ評価データを含まない 2620 単語で学習した．デコーダには記述文法用認識エンジン Julian [9] をもとに提案手法を実装したものを用いた．

残響特性 H には，評価データにたたみ込んだインパルス応答を音声同様に短時間分析した短時間スペクトル応答を用いた．D は 3.3 節で述べた 2 通りとした．R は，D を平坦とみなす場合はフレームの重なりを考慮して H の第 4 フレーム目以降とし，D を無響室の特性とみなす場合は D の長さを考慮して H の第 8 フレーム目以降とした．インパルス応答 E1B の短時間スペクトル領域における特性を図 5 に示す．

4.2　実験結果

結果を表 1 および図 6 に示す．2 名の話者 MAU（男性）・FFS（女性）について，図 4 のインパルス応答をたたみ込んだときの単語認識率（%）である．手法は 6 通り行なった．それぞれ，まず，クリーン音声のモデルを使用した場合の，何も

表 1　提案手法による残響音声の単語認識率（%）

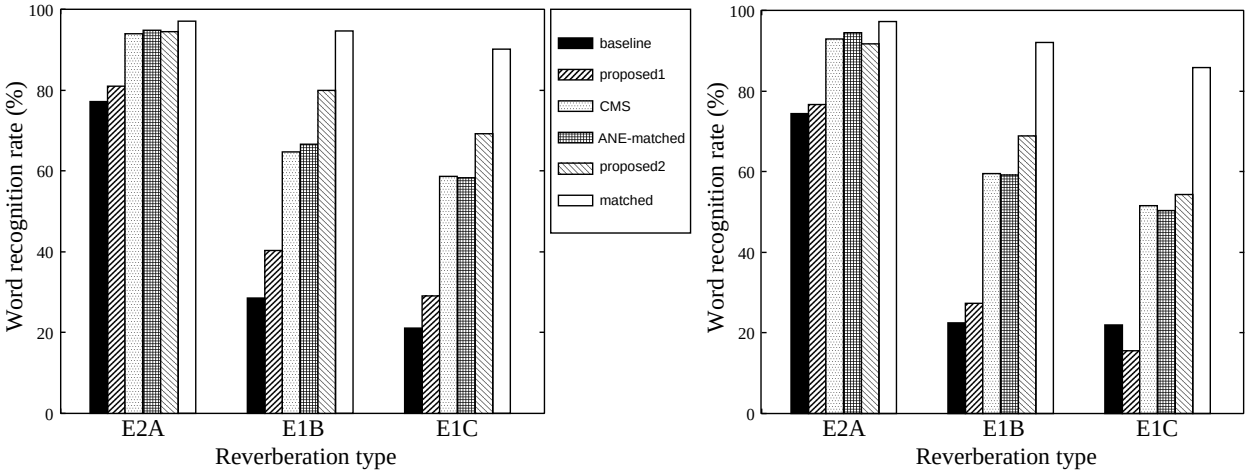
Table 1　Word recognition rate（%）for reverberant speech.

(a) Speaker: FFS(female)

	baseline	proposed1	CMS	ANE	proposed2	matched
E2A	77.1	81.1	94.1	95.0	94.8	97.1
E1B	28.2	39.9	64.9	66.9	79.9	95.4
E1C	20.8	28.4	58.6	58.3	68.6	90.7

(b) Speaker: MAU(male)

baseline	proposed1	CMS	ANE	proposed2	matched
74.5	76.6	93.1	94.5	91.9	97.6
22.3	27.4	59.7	59.5	68.9	92.4
22.1	15.0	51.9	50.7	53.7	86.2



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

図 6　提案手法による残響音声の単語認識率（%）

Fig. 6　Word recognition rate (%) for reverberant speech.

行なわない（baseline），提案手法を施す（proposed1），CMS を行なう（CMS），の 3 通り，次に無響室の特性をたたみ込んだ音声で学習したモデルを使用した場合の，何も行なわない（ANE），提案手法を施す（proposed2），の 2 通り，さらに評価データと同じ残響特性下の音声で学習したモデルを使用した場合（matched），であった．クリーン音声をもとにした場合（proposed1）では，直接音の伝搬特性 D が平坦と近似されている．一方，無響室音声のモデルにもとにした場合（proposed2）は，無響室の特性が D として扱われている．

4.3　考　　察

結果より，提案手法による認識率の向上が確認された．インパルス応答 E2A では，手法 CMS や ANE による効果が大きかった．これは図 4 からもわかるように，残響成分が少なく，直接音のひずみが主なひずみであったためと考えられる．インパルス応答 E1B・E1C の場合では，手法 CMS や ANE の効果は同程度であり，提案手法 proposed2 によってさらに認識率が向上した．以上から，提案手法によってフレーム外からの残響成分に対処できることが確認された．ただし，いずれの条件においても，性能向上は matched 条件に比べて小さなものにとどまった．この点を改善するため，提案法における近似の改良などを今後行なう必要がある．例えば，残響成分推定のための音源信号推定には先行研究 [3] のようにフィルタを用いることができる．あるいは，実際の音素・状態の継続時間を考慮した残響成分の推定や，音声の定常的な部分と遷移部分に留意したモデル合成なども有効である可能性がある．また，今回は第一段階の評価として比較的短い残響時間での孤立単語認識を行なったが，本手法は，残響時間がさらに長い場合や連続音声の認識において有効にはたらく可能性があるため，今後そのような条件で評価する必要がある．

5.　ヤコビ適応法の利用の検討

本章では，提案手法の発展として検討した，フレーム同期ヤコビ適応法による残響下音声認識について述べる．

5.1　ヤコビ適応法

ヤコビ適応法 [10] は音響モデルの高速な環境適応法であり，ある環境（クリーンではない）からに別の環境に適応できるという特徴を持つ．事前に仮定した環境の音声を $\boldsymbol{Y}^c_{ref}$，環境変動要因（たとえば周囲雑音）を $\boldsymbol{N}^c_{ref}$，発話時の環境の音声を $\boldsymbol{Y}^c_{tar}$，雑音を $\boldsymbol{N}^c_{tar}$ とする．また，$\boldsymbol{Y}^c$ と $\boldsymbol{N}^c$ はケプストラム領域，$\boldsymbol{Y}^s$ と $\boldsymbol{N}^s$ はスペクトル領域を表すとする．このとき，ヤコビ適応法では，環境変動要因の変化にともなうモデルの非線形変換を次式のように線形近似し，ケプストラム領域で適応処理を行なう．

$$\boldsymbol{Y}^c_{tar} \simeq \boldsymbol{Y}^c_{ref} + \boldsymbol{J}_N(\boldsymbol{N}^c_{tar} - \boldsymbol{N}^c_{ref}) \tag{7}$$

ここで $\boldsymbol{J}_N$ はヤコビ行列であり，次式から事前に求めておくことができる．

$$\boldsymbol{J}_N \equiv \frac{\partial \boldsymbol{Y}^c}{\partial \boldsymbol{N}^c} \tag{8}$$

$$= \frac{\partial \boldsymbol{Y}^c}{\partial \log \boldsymbol{Y}^s} \frac{\partial \log \boldsymbol{Y}^s}{\partial \boldsymbol{Y}^s} \frac{\partial \boldsymbol{Y}^s}{\partial \boldsymbol{N}^s} \frac{\partial \boldsymbol{N}^s}{\partial \log \boldsymbol{N}^s} \frac{\partial \log \boldsymbol{N}^s}{\partial \boldsymbol{N}^s} \tag{9}$$

$$= \boldsymbol{C} \frac{\boldsymbol{N}^s}{\boldsymbol{Y}^s} \boldsymbol{C}^{-1} \quad (\boldsymbol{C}：\text{コサイン変換行列}) \tag{10}$$

5.2 残響成分を用いたヤコビ適応

3. 章で提案した残響下音声認識の枠組みに，ヤコビ適応法を用いることを検討する．先の提案手法ではクリーン音声のモデルと残響成分のモデル合成を行なっていた．ここにヤコビ適応法を用いる場合は，残響成分を環境変動要因として扱い，ある想定した残響下音声のモデルから適応する．あらかじめ想定した環境における平均的な残響成分（$\boldsymbol{N}_{ref}$）とヤコビ行列（$\boldsymbol{J}_N$）を求めておき，これとフレームごとに推定した残響成分（$\boldsymbol{N}_{tar}$）を用いて (7) 式にしたがって適応処理を行なう．

ヤコビ適応は環境適応モデル（matched）や複数環境で学習したモデル（multi-condition）から適応することができるため，高い認識性能を持つと考えられる．また，残響特性が多少変動した場合にも適応できるという頑健性が期待される．さらに，音声認識時にはケプストラム領域のみで適応処理を行なうため，モデル合成法のようにスペクトル領域に変換する必要がなく，大幅な計算量の削減がなされる．

5.3 事前環境の残響成分

ある残響環境に適応したモデルからヤコビ適応を行なうためには，事前に想定する環境の平均的な残響成分 $\boldsymbol{N}_{ref}$ が必要である．しかし，ここで仮定している $\boldsymbol{N}_{ref}$ は観測によって明らかに求められるものではないため，その推定を行なう必要がある．たとえば，音声のデータベースと残響特性のインパルス応答が与えられているならば，波形レベルのたたみ込みによって反射音成分を模擬し，これらの特徴量から事前の残響成分を求めることができる．また，3. 章で提案した，短時間スペクトル領域での残響成分推定法を用いることもできる．また，それぞれの音素 HMM，あるいはその状態ごとに事前の残響成分を求めれば，より細かな適応が可能となると考えられる．例として，環境適応モデルとクリーン音声のモデルからの推定や，データベースから得られる音声のコンテキストに関する統計量を用いた推定などが挙げられる．

5.4 アルゴリズム

3.3 節と同様の問題において，モデルの環境適応化にヤコビ適応法を用いるときのアルゴリズムを以下に示す．推定残響成分 N がスペクトル領域から MFCC 領域に変換されるが，モデルパラメタの変換に比べればはるかに計算量が少ない．

準備 ：

（1） 環境適応音声の HMM と音響伝達特性 H を用意

（2） 事前の残響成分 $\boldsymbol{N}_{ref}$，ヤコビ行列 $\boldsymbol{J}_N$ の計算

初期化 ：

（1） D と R を設定（H を分割）

（2） $N(\omega, 0) = 0$，$S(\omega, 0) = O(\omega, 0)$

適応 ：各フレーム n において

（1） 各フィルタバンク ω において

a. N の推定（(6) 式）

b. S の計算（$S(\omega, n) = O(\omega, n) - N(\omega, n)$）

（2） N を MFCC 領域に変換

（3） モデル適応処理（(7) 式）

（4） 適応モデルを用いて尤度計算

5.5 予 備 実 験

上述のことがらに基づき，4. 章と同様の条件の音声認識実験を行なった．事前の残響成分はすべての HMM 状態で同一として，フレームごとの残響成分と同じ推定によって求めた．評価データと同じ残響特性下の音声で学習したモデル（matched）から適応処理を行なったとき，matched に比べて認識率の向上は見られなかった．また，HMM の状態ごとに異なる事前の残響成分として，残響適応モデルとクリーン音声モデルの差分ベクトルを用いた場合も同様の結果であった．ただし，計算時間については大幅に削減された．今後，事前の残響成分について十分調査し，その推定方法を検討することにより，有効なモデル適応手法となることが期待される．

6. ま と め

本稿では，残響に頑健な音声認識のための手法として，各フレームにおける残響成分をその過去の音響信号から推定し，これを音響モデルと合成することによって逐次的にモデルを残響下音声に適応化させる方法を提案した．残響下音声の特定話者孤立単語音声認識実験によって提案手法による認識性能向上を確認した．また，モデル合成の代わりにヤコビ適応法を用いることについての検討を行なった．

今後の課題としては，残響成分推定の精度の向上をはじめ，さまざまな残響環境や連続音声での評価，今回は既知として与えた室内の音響特性の推定などが挙げられる．さらに，ヤコビ適応において事前に想定する残響成分の推定についても検討する必要がある．また，フレーム同期モデル適応の手法を周囲雑音などに対しても適用することも考慮中である．

文　　献

[1] 中村哲, “実音響環境に頑健な音声認識を目指して,” 信学技報, SP 2002-12, 2002.

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

[3] 杉村耕司, 滝口哲也, 中村哲, 鹿野清宏, “フレーム間の関係を考慮した残響音声認識の検討,” 音講論, 3-Q-5, Mar. 1999.

[4] 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,” 音講論, 1-9-14, Sep. 2002.

[5] T. Takiguchi, S. Nakamura and K. Shikano, “HMM-Separation-Based Speech Recognition for a Distant Moving Speaker,” IEEE Trans. on SAP, Vol. 9, No. 2, 2001.

[6] 滝口哲也, 西村雅史, “一次線形予測による残響下音声認識の検討,” 音講論, 3-Q-6, Sep. 2003.

[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,” 音講論, 1-6-7, Sep. 2003.

[8] M. J. F. Gales and S. Young, “An improved approach to the hidden Markov model decomposition of speech and noise,” Proc. ICASSP92, pp. 233-236, 1992.

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

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