音声言語情報処理 14－10
(1996. 12. 13)

Taylor 展開による音響モデルの適応

山口 義和　高橋 淳一†　高橋 敏　嵯峨山 茂樹

NTT ヒューマンインタフェース研究所　†NTT システムエレクトロニクス研究所

〒239 神奈川県横須賀市光の丘 1-1
yamaguch@nttspch.hil.ntt.co.jp

あらまし　本研究では，新しい雑音耐性のアプローチとして Taylor 展開に基づいた音響モデルの適応手法を提案する．雑音が時間的に変化する場合，HMM 合成法では雑音の変化の度に音響モデルを合成しなければならない．しかし，「ある雑音から別の雑音へ」音響モデルを修正する「雑音適応」というアプローチに立てば，少ない計算量で音響モデルを更新することが可能となる．本手法は，HMM 合成法で行っているモデルパラメータの非線形変換を，Taylor 展開の 1 次項を用いて線形近似することで高速な音響モデル適応を実現している．実験より，適応に用いるデータを観測してからモデルを更新するまでの処理時間が少なく，適応に用いるデータも少量で済むことを確認した．

キーワード　音響モデル，環境雑音，雑音適応，Taylor 展開

Acoustic Model Adaptation by Taylor Series

Yoshikazu YAMAGUCHI　Jun-ichi TAKAHASHI†
Satoshi TAKAHASHI　Shigeki SAGAYAMA

NTT Human Interface Laboratories
†NTT System Electronics Laboratories

1-1, Hikarinooka, Yokosuka-Shi, Kanagawa, 239 Japan

Abstract　This paper presents a new approach against environmental noise, an acoustic model adaptation technique based on the Taylor series. Using an HMM composition, a model containing noise should be composed from a speech model and a noise model each time when environmental noise changes. From the viewpoint of a "noise adaptation" approach, we modify an acoustic model "from one noise to another". This makes it possible to adapt the acoustic model at a low computation cost. In an HMM composition, non-linear conversion from cepstrum parameters is required. To accomplish a fast model adaptation, this paper uses first-order terms of the Taylor series. Based on the experimental results, we confirm a low computation cost for the adaptation processes, and the need for a small amount of noise data for the model adaptation.

key words　acoustic model, environmental noise, noise adaptation, Taylor series

1 はじめに

音声認識を実環境で利用すると，発声環境により認識性能が大きく劣化することがある．これは，発声環境と音響モデル作成時の環境とのミスマッチが認識性能を劣化させるためである．このミスマッチをいかに少なくするかが雑音耐性技術の重要課題である．発声環境と音響モデルとのミスマッチを少なくする技術には NOVO 合成法 [1, 2] や PMC 法 [3, 4] といった HMM 合成法があり，比較的定常な雑音の環境下での認識に効果があると報告されている．しかし，自動車走行中で路面状況により雑音が時々刻々と変化する場合や，駅構内や展示ホール，交差点など環境雑音が頻繁に変化する場合など，音声認識の利用環境によっては背景雑音が時々刻々と変化する場合も多い．このような場合，認識性能を劣化させないためには背景雑音の変化に応じてモデルを動的に更新する必要がある．HMM 合成法を用いると，雑音が変動するたびに HMM 合成法によってモデルを合成して作り直すことになる．しかし，雑音の変動に動的に音響モデルを更新するには，モデルを作り直すよりも「ある雑音から別の雑音へ」音響モデルを修正する「雑音適応」の方がより高速に背景雑音と音響モデルとのミスマッチを軽減できると思われる．このような「雑音適応」の観点に立った雑音耐性対策はこれまでされていなかった．

本研究では，この「雑音適応」という新しい雑音耐性のアプローチを考案した [5]．雑音環境に応じて雑音重畳音声 HMM を予め作成しておき，背景雑音の変化を観測した場合にその変化に応じて雑音重畳音声 HMM を更新する，いわゆる音響モデルの雑音適応を行うものである．この音響モデルの雑音適応のために，クリーン音声，背景雑音から雑音重畳音声へのケプストラム領域での非線形変換を Taylor 展開によって線形近似することで本アプローチを実現している．このようなアプローチを実現することにより，適応に用いるデータを観測してからモデルを更新するまでの処理時間が少なく，適応に用いるデータも少量で適応することができる．これらの特徴をもつ本手法を用いれば，時間変動する雑音に対して高速なモデル適応が可能となる．

2 従来の雑音耐性技術

認識時の発声環境に最も適合した音響モデルを作るには，実際の背景雑音を含む音声を用いて音響モデルを再学習するのがよい．しかし，モデルの再学習は必要なデータ量と時間の面から，決して現実的な雑音耐性対策ではない．ここでは実利用において実現可能な雑音耐性対策を 2 種類に分類して説明する．

(1) 音響モデル合成手法

いわゆる HMM 合成法であり，NOVO 合成法 [1, 2] や PMC 法 [3, 4] などがある．雑音 HMM，クリーン音声 HMM をそれぞれ学習し，これらの HMM を合成することで，モデルを再学習したのと同等の雑音重畳音声 HMM を得る方式である．処理時間のかかるクリーン音声 HMM の学習を事前に行えるので，モデルを再学習するよりも圧倒的に処理量が少ない．しかし，モデル合成の際にフーリエ変換，指数変換，対数変換，逆フーリエ変換を行うために実時間処理が可能な処理量とは言い難い．

(2) 音声強調手法

代表的な手法としてスペクトルサブトラクション法がある [6]．入力された雑音重畳音声から，これとは別に観測した雑音をスペクトル領域上で差し引くことで，クリーンな音声を求める手法である．非常に処理量の少ない手法であるが，雑音を差し引くことによる引き残り，引き過ぎによるスペクトル歪みが生じるといった問題があり，HMM 合成法を用いる場合より認識性能は劣る．

3 Taylor 展開による音響モデルの適応

3.1 雑音耐性の新たなアプローチ — 音響モデルの雑音適応

本研究では，高い雑音耐性と実時間処理可能な雑音耐性技術の実現を目的とした新たな雑音適応のアプローチを提案する．HMM 合成法は，雑音環境の変化に応じて雑音 HMM とクリーン音声 HMM から雑音重畳音声に対する HMM をその都度，作成するアプローチである．これに対して本手法は，初期の雑音環境に応じて雑音重畳音声 HMM をあらかじめ作成しておき，環境雑音の変化を観測した場合に，その変化に応じて雑音重畳音声 HMM を更新することを考える．存在する雑音重畳音声 HMM と実環境とのずれを補正するアプローチであり，計算量の大幅な削減が期待できる．

このアプローチを実現するために，我々が着目したのは Taylor 展開である．背景雑音の変動が音響モデルパラメータに及ぼす影響を計算するには非線形演算が必要である．これを Taylor 展開を用いて近似的な線形演算をすることで，適応処理の高速化を行う．なお本手法の適応対象であるモデルパラメータは，HMM のケプストラムパラメータの出力確率分布の平均値ベクトルである．

3.2 基本概念

雑音重畳音声のスペクトル $\boldsymbol{S}_{S+N}$ (ベクトルで表す) は，クリーンな音声のスペクトル $\boldsymbol{S}_S$ と背景雑音のスペクトル $\boldsymbol{S}_N$ の線形和で表される．

$$\boldsymbol{S}_{S+N} = \boldsymbol{S}_S + \boldsymbol{S}_N \tag{1}$$

上記の関係をケプストラムパラメータに変換した場合，雑音重畳音声ケプストラム $\boldsymbol{C}_{S+N}$ とクリーン音声ケプストラム $\boldsymbol{C}_S$，雑音ケプストラム $\boldsymbol{C}_N$ との関係は以下のような

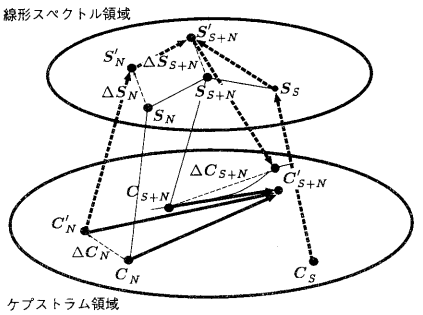


図 1　背景雑音の変化に対するモデルパラメータの更新

非線形な関係となる．

$$\boldsymbol{C}_{S+N} = \mathcal{F}^{-1}\left(\log\left(\exp\left(\mathcal{F}\left(\boldsymbol{C}_S\right)\right) + \exp\left(\mathcal{F}\left(\boldsymbol{C}_N\right)\right)\right)\right) \tag{2}$$

ここで，$\mathcal{F}(\cdot)$, $\mathcal{F}^{-1}(\cdot)$, $\log(\cdot)$, $\exp(\cdot)$ はそれぞれフーリエ変換，逆フーリエ変換，対数変換，指数変換を表す．背景雑音が変化した場合，HMM 合成法 (図 1中の太い波線) のように音声および雑音ケプストラムから雑音重畳音声ケプストラムを求めようとすると，一度，線形スペクトル領域に変換しなければならず，その処理は複雑で演算量が多い．

ここで，音声と雑音のケプストラムが微小に変動した場合の雑音重畳音声の変動分を Taylor 展開を用いて考えてみる．関数 $f(x)$ の x に関する一般的な Taylor 展開を以下に示す．

$$\begin{aligned} f(x+\Delta x) &= f(x) + \frac{f'(x)}{1!}\Delta x + \frac{f''(x)}{2!}(\Delta x)^2 + \cdots \\ &\quad + \frac{f^{(n-1)}(x)}{(n-1)!}(\Delta x)^{n-1} + \frac{f^{(n)}(x+\theta\Delta x)}{n!}(\Delta x)^n \\ &\qquad (0<\theta<1) \end{aligned} \tag{3}$$

式 (3) に音声，雑音および雑音重畳音声をあてはめ，音声と雑音の微小変動に対する雑音重畳音声の変動分を以下のように求める．

$$\Delta\boldsymbol{C}_{S+N} = \frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_S}\Delta\boldsymbol{C}_S + \frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}\Delta\boldsymbol{C}_N \tag{4}$$

なお，上式では Taylor 展開の 1 次微分項までを考えている．2 次項以降も含めると近似精度を向上できるが，今回は扱わない．式 (4) は，雑音重畳音声ケプストラムの変動分が，音声ケプストラムの変動分と雑音ケプストラムの変動分から線形結合により求められることを示している．つまり，音声ケプストラムの変動分と雑音ケプストラムの変動分を線形スペクトル領域に変換せずに，図 1中の太い実線に示すように近似的に求めることができる．

音響モデル作成時に収録した音声と雑音の条件が，実際の認識時の条件と異なる場合，音響モデルの出力確率分布の平均値ベクトルを式 (4) に基づいて適応すれば，条件のミスマッチを少なくすることができる．本稿では，式 (4) にみられる $\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_S}$ や $\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}$ がヤコビアン (ヤコビ行列) と呼ばれることから，本手法をヤコビアン適応法と呼ぶことにする．

3.3　ヤコビアン適応法

音響モデルと実環境との間の背景雑音に関するミスマッチのみを考慮する ($\Delta\boldsymbol{C}_S = 0$) 場合，式 (4) は次のようになる．

$$\boldsymbol{C}'_{S+N} = \boldsymbol{C}_{S+N} + \frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}\left(\boldsymbol{C}'_N - \boldsymbol{C}_N\right) \tag{5}$$

ただし，プライム ($'$) は雑音変動後のパラメータを意味する．式 (5) より，雑音変動後の雑音重畳音声ケプストラム $\boldsymbol{C}'_{S+N}$ は，雑音変動前の雑音 (初期雑音) のケプストラム $\boldsymbol{C}_N$ と雑音変動前の雑音重畳音声 (初期雑音重畳音声) のケプストラム $\boldsymbol{C}_{S+N}$，雑音変動後の雑音 (適応対象雑音) のケプストラム $\boldsymbol{C}'_N$，そしてヤコビ行列 $\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}$ から求めることができる．

次に，式 (5) 中のヤコビ行列の計算法を述べる．ヤコビ行列 $\boldsymbol{J}_N$ は以下のように展開できる．

$$\begin{aligned} \boldsymbol{J}_N &= \frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N} \\ &= \frac{\partial \boldsymbol{C}_{S+N}}{\partial(\log \boldsymbol{S}_{S+N})}\frac{\partial(\log \boldsymbol{S}_{S+N})}{\partial \boldsymbol{S}_{S+N}}\frac{\partial \boldsymbol{S}_{S+N}}{\partial \boldsymbol{S}_N} \\ &\quad \times \frac{\partial \boldsymbol{S}_N}{\partial(\log \boldsymbol{S}_N)}\frac{\partial(\log \boldsymbol{S}_N)}{\partial \boldsymbol{C}_N} \end{aligned} \tag{6}$$

コサイン変換，逆コサイン変換，対数変換，指数変換より，式 (6) に含まれる各偏微分項を求める．

- $\boldsymbol{C}_{S+N} = \mathcal{F}^{-1}(\log \boldsymbol{S}_{S+N}) = \boldsymbol{F}^{-1}(\log \boldsymbol{S}_{S+N})$

$$\left[\frac{\partial \boldsymbol{C}_{S+N}}{\partial(\log \boldsymbol{S}_{S+N})}\right]_{ij} = F^{-1}_{ij} \tag{7}$$

- $(\log \boldsymbol{S}_{S+N}) = \log(\boldsymbol{S}_{S+N})$

$$\left[\frac{\partial(\log \boldsymbol{S}_{S+N})}{\partial \boldsymbol{S}_{S+N}}\right]_{ij} = \delta_{ij}\frac{1}{S_{S+N_i}} \tag{8}$$

- $\boldsymbol{S}_{S+N} = \boldsymbol{S}_S + \boldsymbol{S}_N$

$$\left[\frac{\partial \boldsymbol{S}_{S+N}}{\partial \boldsymbol{S}_N}\right]_{ij} = \delta_{ij} \tag{9}$$

- $\boldsymbol{S}_N = \exp(\log \boldsymbol{S}_N)$

$$\left[\frac{\partial \boldsymbol{S}_N}{\partial(\log \boldsymbol{S}_N)}\right]_{ij} = \delta_{ij}S_{N_i} \tag{10}$$

- $(\log \boldsymbol{S}_N) = \mathcal{F}(\boldsymbol{C}_N) = \boldsymbol{F}\boldsymbol{C}_N$

$$\left[\frac{\partial(\log \boldsymbol{S}_N)}{\partial \boldsymbol{C}_N}\right]_{ij} = F_{ij} \tag{11}$$

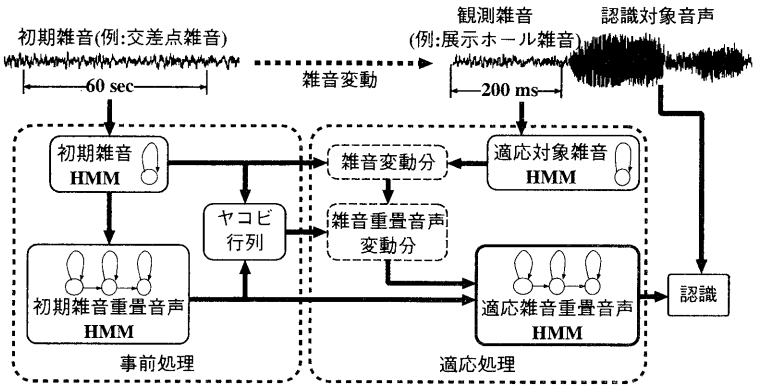


図 2　ヤコビアン適応法を用いた背景雑音への適応

ここで $0 \leq i,j \leq p$ であり, p はケプストラムの次数で 0 次には残差パワーを用いている. また, $\boldsymbol{F}$ はコサイン変換行列 (F_{ij} は行列 $\boldsymbol{F}$ の i 行 j 列目の要素), δ_{ij} は

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

である. 式 (7)~(11) を式 (6) に代入して整理すると, ヤコビ行列は以下のような簡潔な式で求まる.

$$[\boldsymbol{J}_N]_{ij} = \sum_k F_{ik}^{-1} \frac{S_{N_k}}{S_{S+N_k}} F_{kj} \quad 0 \leq i,j,k \leq p \tag{13}$$

すなわち, ヤコビ行列は初期の雑音重畳音声スペクトル $\boldsymbol{S}_{S+N}$ と初期の雑音スペクトル $\boldsymbol{S}_N$ から求めることができる (S_{S+N_k}, S_{N_k} はベクトル $\boldsymbol{S}_{S+N}$, $\boldsymbol{S}_N$ の k 番目の要素). $\boldsymbol{S}_{S+N}$, $\boldsymbol{S}_N$ はそれぞれ初期雑音重畳音声, 初期雑音のケプストラムパラメータを線形スペクトル領域に変換することで求まる. したがって, ヤコビ行列は適応の事前処理としてあらかじめ計算しておくことができる.

3.4　背景雑音の変動に対する音響モデルの雑音適応

本節では, 前節で述べた定式化による音響モデルの雑音適応の手法について述べる. 式 (5) に従い, 各音響モデルの出力確率分布の平均値ベクトル (ケプストラム) を用いて雑音適応後の音響モデルを求める. 以下に, 式 (5) 中のケプストラムパラメータに対応する HMM を示す.

(a) $\boldsymbol{C}_N$: 事前に観測した (環境変動前の) 雑音から計算される HMM (以下, 初期雑音 HMM) の平均値ベクトル

(b) $\boldsymbol{C}_{S+N}$: 事前に用意した雑音重畳音声 HMM (以下, 初期雑音重畳音声 HMM) の平均値ベクトル

(c) $\boldsymbol{C}'_N$: 認識時に観測した (環境変動後の) 適応すべき雑音から計算される HMM (適応対象雑音 HMM) の平均値ベクトル

(d) $\boldsymbol{C}'_{S+N}$: 適応後の雑音重畳音声 HMM の平均値ベクトル

雑音重畳音声 HMM を適応するには, 雑音重畳音声 HMM のすべてのモデル中に存在する出力確率分布に対して式 (5) を計算する. すなわち, 適応した雑音重畳音声 HMM の m 番目の出力確率分布の平均値ベクトル $\boldsymbol{C}^{m\prime}_{S+N}$ を求めるために, 初期雑音重畳音声 HMM の m 番目の出力確率分布の平均値ベクトル $\boldsymbol{C}^m_{S+N}$ を式 (5) に用いる. 一方, 本研究では雑音 HMM として 1 状態 1 混合分布のモデルを用いるため, 出力確率分布の平均値ベクトルは 1 個であり, これを常に式 (5) に用いる. 以上を考慮して式 (5) を書き換える.

$$\boldsymbol{C}^{m\prime}_{S+N} = \boldsymbol{C}^m_{S+N} + \frac{\partial \boldsymbol{C}^m_{S+N}}{\partial \boldsymbol{C}_N}(\boldsymbol{C}'_N - \boldsymbol{C}_N) \tag{14}$$

また, 式 (14) からわかるようにヤコビ行列も雑音重畳音声 HMM の出力確率分布ごとに用意する必要がある. 初期の雑音重畳音声 HMM の m 番目の出力確率分布に対するヤコビ行列 $\boldsymbol{J}^m_N = \dfrac{\partial \boldsymbol{C}^m_{S+N}}{\partial \boldsymbol{C}_N}$ は以下のようになる.

$$[\boldsymbol{J}^m_N]_{ij} = \sum_k F_{ik}^{-1} \frac{S_{N_k}}{S^m_{S+N_k}} F_{kj} \quad 0 \leq i,j,k \leq p \tag{15}$$

$\boldsymbol{S}^m_{S+N}$ は $\boldsymbol{C}^m_{S+N}$ を線形スペクトル領域に変換したものである.

式 (14), (15) に従って雑音適応後のモデルを求めるためには, 初期雑音 HMM, 初期雑音重畳音声 HMM (例えば NOVO 合成法により作成する), 適応対象雑音 HMM, そしてヤコビ行列が必要である. このうち, 初期雑音 HMM, 初期雑音重畳音声 HMM, ヤコビ行列は環境変動前に事前処理として, 計算しておくことが可能である. し

たがって環境変動後の適応処理として行なうことは，環境変動後の雑音を観測して適応対象雑音 HMM を計算し，式 (14) に従って適応後の雑音重畳音声 HMM を求めることだけである．この式 (14) の計算は p 次のベクトルおよび行列の簡単な計算であり，雑音重畳音声 HMM のそれぞれの出力確率分布に対して加減算 $p(p+1)$ 回，乗算 p^2 回で済むので，高速な適応処理が可能である．

ヤコビアン適応法を用いた雑音適応の実施方法の一例を説明する (図 2)．

[事前処理]

1) 事前に仮定した雑音から初期雑音 HMM を求める．(実際は平均と分散を計算)

2) HMM 合成法により初期雑音重畳音声 HMM を求める．

3) 式 (15) によりヤコビ行列を計算し記憶する．

[適応処理]

4) 発声直前の区間から適応すべき雑音を観測し，適応対象雑音 HMM を求める．(平均を計算)

5) 式 (14) の行列計算により，適応した雑音重畳音声 HMM を近似計算する．(平均を更新)

4 スペクトルサブトラクション法の導入

ヤコビアン適応法や HMM 合成法などの手法は，雑音が重畳した入力信号に音響モデルを適合させることでミスマッチを軽減している．しかし，S/N 比が悪くなるにしたがい，識別性能が劣化することは避けられない．そこで，入力信号自体の S/N 比を改善するために，音声強調手法を併用することが有効であると考えられる．本章では，ヤコビアン適応法および NOVO 合成法のフロントエンドとしてスペクトルサブトラクション法 (以下，SS 法) の導入を検討した [7]．

4.1 スペクトルサブトラクション法

本研究では以下の定式化に基づく SS 法を用いる [8]．

$$\hat{\boldsymbol{S}}_S = \boldsymbol{S}_{S+N} - \alpha \hat{\boldsymbol{S}}_N$$
$$\tilde{\boldsymbol{S}}_S = \begin{cases} \hat{\boldsymbol{S}}_S & \text{if } \hat{\boldsymbol{S}}_S > \beta \boldsymbol{S}_{S+N} \\ \beta \boldsymbol{S}_{S+N} & \text{otherwise} \end{cases} \quad (16)$$

雑音重畳音声のスペクトル $\boldsymbol{S}_{S+N}$ から雑音の平均スペクトル $\hat{\boldsymbol{S}}_N$ を差し引くことで，消し残り雑音を含む音声のスペクトル $\tilde{\boldsymbol{S}}_S$ を求め，これを出力とする．この際，雑音を α 倍に強調して差し引き，その結果として雑音による引き過ぎが生じた場合は係数 β によってフロアリングする．

4.2 SS- ヤコビアン適応法

雑音が重畳した音声に SS 法を適用して得られる「消し残り雑音を含む音声」に音響モデルをマッチングさせることを考える．そのためには，従来の NOVO 合成法や前章で述べたヤコビアン適応法で用いる雑音 HMM を「消し残り雑音」を用いて学習した HMM に置き換えればよい．「消し残り雑音」は，式 (16) に従い，観測した雑音データの各フレームのスペクトルから平均スペクトルを差し引いて求める．

SS 法をフロントエンドとした NOVO 合成法 (以下，SS-NOVO 合成法) では，クリーン音声 HMM と「消し残り雑音」HMM から「消し残り雑音を含む音声」HMM を求める．一方，SS 法をフロントエンドとしたヤコビアン適応法 (以下，SS- ヤコビアン適応法) では，NOVO 合成法で得られた「消し残り雑音を含む音声」HMM を初期モデルとして，「消し残り雑音」HMM の変動を考慮してモデルを更新する．

5 実験

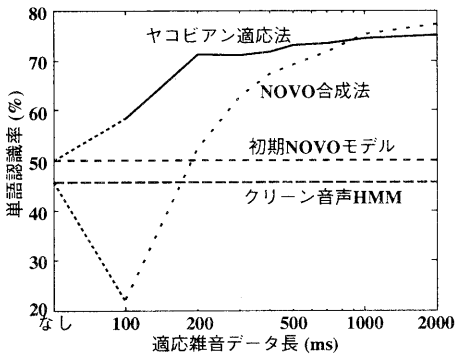
5.1 実験 I：ヤコビアン適応法の評価実験

ヤコビアン適応法の有効性を示すために，NOVO 合成法との比較を行った．認識用音声データの直前の雑音データを用いて適応する場合について検討した．音声の開始点には視察によるラベル付けがされており，雑音区間および音声区間の切り出し誤りはないとする．13 名のオープン話者による 100 都市名単語音声に，雑音を計算機上で重畳させたものを評価データとした．認識語彙サイズは 400 単語である．クリーン音声 HMM には 523 状態 4 混合分布の HMnet，雑音 HMM には 1 状態 1 混合分布の HMM を用いた．特徴量には，16 次 LPC ケプストラム，16 次 Δ ケプストラム，Δ パワー，パワー (認識時を除く) を用い，うちケプストラムとパワーのみを適応対象とした．

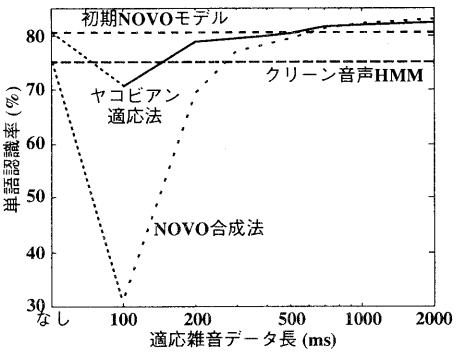
5.1.1 雑音のスペクトル形状が変化した場合

ここでは雑音の変化として，雑音のスペクトル形状が変化した場合の実験を行った．S/N 比は 10 dB とした．データ長 60 sec の《事前に観測した初期雑音》を用いて NOVO 合成法により合成したモデルを初期雑音重畳音声 HMM として用いた．適応雑音データ長を 100 ms ～ 2000 ms に変化させたときの，雑音の変化 (《事前に観測した初期雑音》→《認識時の観測雑音》) に対する平均単語認識率を図 3に示す．図 3中の「初期 NOVO モデル」は，《事前に観測した初期雑音》を用いて NOVO 合成したモデルで，《認識時の観測雑音》の重畳した評価データを認識した場合の結果である．

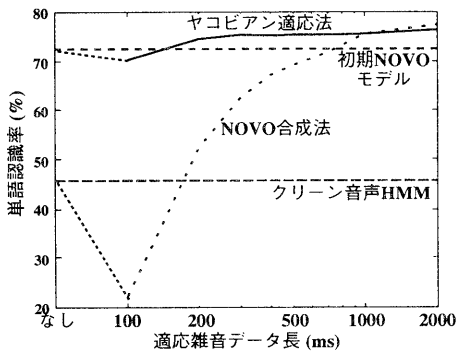
図 3より，本手法によってモデル適応することで初期モデルより認識性能が向上し，適応の効果が大きいことがわかった．また，NOVO 合成法は適応データ量が多い場



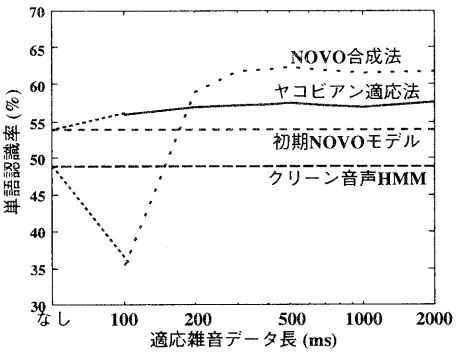
(a) 初期雑音: 交差点 — 観測雑音: 展示ホール



(b) 初期雑音: 交差点 — 観測雑音: 人混み

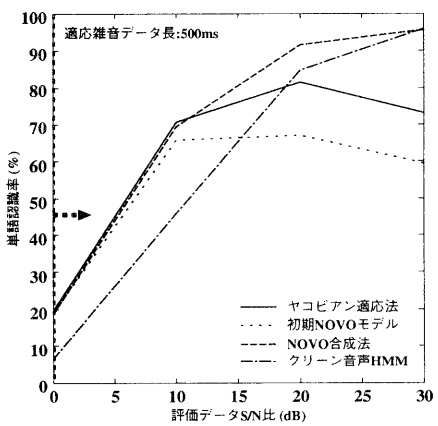


(c) 初期雑音: 人混み — 観測雑音: 展示ホール

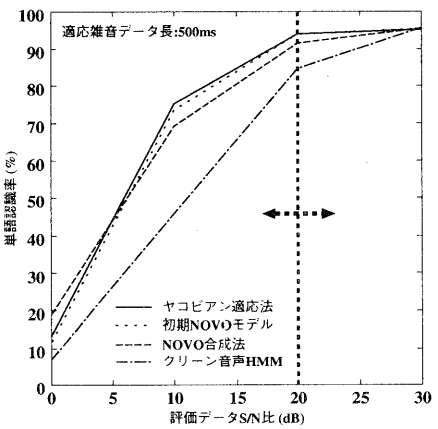


(d) 初期雑音: 人混み — 観測雑音: 在来線

図 **3** 雑音観測時間に対する音声認識率向上の比較



(a) 初期モデル 0 dB



(b) 初期モデル 20 dB

図 **4** S/N 比変動に対する音声認識率

表1 適応処理に要する CPU time (音響分析を除く)

	ヤコビアン	NOVO	ヤコビアン/NOVO
事前処理	2,216 ms	4,416 ms	1/2
適応処理	149 ms	5,066 ms	1/34

(Sun SPARCstation20 で測定)

合では性能が高いが, 適応データが少量の場合に性能が急激に低下することがわかった. 一方, ヤコビアン適応法は適応データが少量の場合でも性能が大幅に低下することがなかった. これは, NOVO 合成法が出力確率分布の平均値に加えて分散も適応するため, ヤコビアン適応法よりも多量のデータが必要となるからである. 十分なデータが得られれば, NOVO 合成法がヤコビアン適応法より認識性能が高くなることは図3よりわかる.

ただし, 図3(b) では適応データが少量の場合 (500 ms 以下) に初期 NOVO モデルが本手法を上回り, 適応の効果が見えなかった. これは, 「人混み」雑音が「交差点」雑音とスペクトル形状 (含まれている音の種類) が比較的似ていること, 無音部分を多く含んでいること, そして適応データ少量であることなどが要因と考えられる. 本稿で示した 4 種類の雑音変動を含めて, 電子協騒音データベースより「交差点」, 「展示ホール」, 「人混み」, 「駅構内」, 「在来線」, 「自動車走行音 (2000cc)」を用いて現時点で 15 種類の雑音変動について実験を行ったが, 図3(b) のように適応データが 200 ms 以上で初期 NOVO モデルが本手法を上回る傾向が見られたのは 3 種類であった.

5.1.2 雑音の S/N 比が変化した場合

次に雑音のスペクトル形状がさほど変化せず, S/N 比が変化した場合の実験を行った. 《事前に観測した初期雑音》, 《認識時の観測雑音》ともに「展示ホール」雑音を用いた. 初期雑音のデータ長が 60 sec, 適応雑音データ長が 500 ms のときの平均単語認識率を図4に示す. 図4(a),(b) は, 初期モデルを作成する際の S/N 比をそれぞれ 0, 20 dB としたときに, 横軸が示す S/N 比の雑音データに適応した場合の認識率を示している. 初期モデルの作成時の S/N 比が低い場合の方が特に本手法の効果が高いことがわかった.

5.1.3 適応処理量

適応処理に要する処理量 (CPU time: Sun SPARCstation20) を表1に示す. 表1では, 適応に用いる雑音データを観測するまでに行うことができる事前処理と, 観測後に行う適応処理に分割して示している. ヤコビアン適応法の場合, ヤコビ行列の計算が事前処理, 適応対象雑音 HMM の作成, モデルパラメータのベクトル計算が適応処理である. NOVO 合成法の場合, クリーン音声 HMM の出力確率分布のケプストラムパラメータの線形スペクトルへの変換が事前処理, 雑音 HMM を作成, 出力確率分布のケプストラムパラメータの線形スペクトルへの変換, クリーン音声と雑音の合成, 線形スペクトラムからケプストラムへの変換が適応処理である.

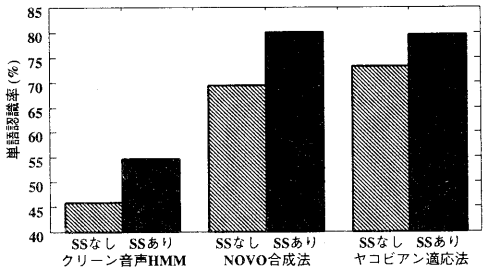


図5 SS 法の導入による音声認識率向上

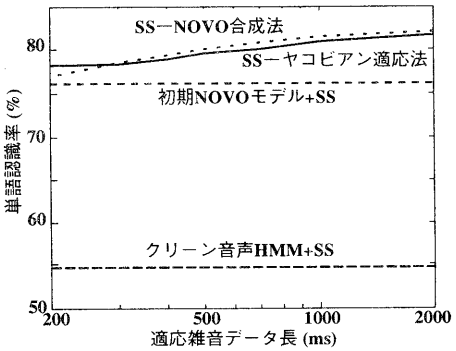


図6 SS 法を導入した場合の雑音観測時間に対する音声認識率向上の比較

高速に音響モデルを更新して背景雑音とモデルとのミスマッチをなくすには, 雑音を観測してから音響モデルを適応するまでの時間つまり適応処理の処理量が重要である. 表1よりヤコビアン適応法は NOVO 合成法に比べて適応時に必要な処理が NOVO 合成法の 1/34 で済む. 以上より, ヤコビアン適応法は少量の適応データで高速な適応処理が可能であり, 変動する背景雑音に音響モデルを実時間で適応するのに適したアルゴリズムであると言える.

5.2 実験 II : SS 法の導入による評価実験

次に, SS 法導入の効果について検討した. 平均雑音スペクトルは適応用雑音データの後半の 160 ms の区間から計算した. SS 法で用いる係数は, 認識対象の雑音重畳音声に対して $\alpha = 2.5$, $\beta = 0.3$, 雑音に対して $\alpha = 1.5$, $\beta = 0.1$ と別々の値を設定した. これらの係数値は実験的に求めたものである. 本実験で用いた雑音, および, その

他の実験条件は5.1.1と同様である.

SS 法をフロントエンドとして用いた場合と用いない場合の認識性能の比較を図5に示す. 図5での適応用雑音データ長は 500 ms である. クリーン音声 HMM, NOVO 合成法, ヤコビアン適応法とも SS 法をフロントエンドとして導入することにより認識率の向上が見られ, 特に NOVO 合成法に対する効果が顕著であった.

次に, 適応雑音データ長を 200 ms ～ 2000 ms に変化させた場合の認識結果を図6に示す. 実験 I の実験結果(図3)と比較すると, 適応データ長が短い場合も, 安定して高い認識性能が得られることがわかった. 特に, NOVO 合成法は SS 法を導入することによって適応データ量の減少による認識性能の急激な劣化がおさえられることがわかる. これは SS 法で雑音のバイアス分を引くことでモデルパラメータの移動分を少なくし, ヤコビアン適応法および NOVO 合成法が含む近似の誤差が少なくなったと考えられる.

6 まとめ

本研究では, Taylor 展開に基づく高速な音響モデル適応法を提案し, 背景雑音の時間変動に対する雑音適応を実現した. 評価実験により, 高速な適応処理が可能であり (SS20 で 150 ms, NOVO 合成法の $^1/_{34}$), かつ少量 (200 ms 程度) の適応データでも高い性能が得られることを確認した. また SS 法をフロントエンドとして導入することにより, ごくわずかな計算量の追加で更なる認識率の向上が見られた. 本手法が有するこれらの性質は音響モデルの実時間適応に適している.

しかしながら本手法は,

- 雑音環境の微小変動を前提としている,
- NOVO 合成法では分散も考慮した分布の変換から平均値を導出しているが, 本手法では, 平均値の変換の Taylor 展開で近似している, つまり NOVO 合成法で得られる解の近似解である,
- 平均値のみの適応である,

ことから, 雑音が時間変動するような全ての状況において, 十分なモデル適応ができているとは言い難い. このため, 適応可能な雑音変動の範囲を検討し, 実時間処理可能という利点を考慮した上で本手法の利用場面を考える必要がある. その一方, さらなる性能向上のために Taylor 展開の 2 次微分項の導入や, 分散および Δケプストラムの適応などが今後の課題として挙げられる.

謝辞

日頃から貴重なご意見を頂く音声情報研究部, 古井特別研究室の皆様に感謝いたします. 本研究では電子協の騒音データベースを実験に使用した.

参考文献

[1] F. Martin, K. Shikano, Y. Minami, Y. Okabe, "Recognition of Noisy Speech by Using the Composition of Hidden Markov Models," 音学講論, 1-7-10, pp. 65–66, 1992-10.

[2] Y. Minami and S. Furui, "A Maximum Likelihood Procedure for a Universal Adaptation Method Based on HMM Composition," *Proc. ICASSP95*, pp. 129–132, 1995.

[3] M. J. F. Gales and S. J. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech And Noise," *Proc. ICASSP92*, pp.233–236, 1992.

[4] M. J. F. Gales and S. J. Young, "A Fast And Flexible Implementation of Parallel Model Combination," *Proc. ICASSP95*, pp.133–136, 1995.

[5] 山口 義和, 高橋 淳一, 高橋 敏, 嵯峨山 茂樹, "Taylor 展開に基づく高速な音響モデル適応法", 音学講論, 2-Q-11, pp. 151–152, 1996-9.

[6] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, No. 2, pp. 113–120, 1979.

[7] 高橋 敏, 山口 義和, 嵯峨山 茂樹, "スペクトルサブトラクションと NOVO 合成法を用いた雑音下音声認識", 音学講論, 2-Q-8, pp. 145–146, 1996-9.

[8] J. A. N. Flores and S. J. Young, "Adapting a HMM-Based Recognizer for Noisy Speech Enhanced by Spectral Subtraction," *Proc. Eurospeech93*, pp. 829–832, 1993.