

論文 / 著書情報
Article / Book Information

論題(和文)	ゆう度最大化による雑音ひとずみへのHMM適応化法
Title(English)	HMM Adaptation Method for Noise and Distortion by Maximizing Likelihood
著者(和文)	南泰浩, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会論文誌A, Vol. J80-A, No. 7, pp. 1179-1186
Citation(English)	, Vol. J80-A, No. 7, pp. 1179-1186
発行日 / Pub. date	1997, 7
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1997 Institute of Electronics, Information and Communication Engineers.

ゆう度最大化による雑音とひずみへのHMM適応化法

南 泰浩[†] 古井 貞熙^{†*}

HMM Adaptation Method for Noise and Distortion by Maximizing Likelihood

Yasuhiro MINAMI[†] and Sadaoki FURUI^{†*}

あらまし 本論文では、加算性雑音と乗算性ひずみにHMMを適応化させる新しいHMM合成法について述べる。従来のHMM合成法では、加算性の雑音にしか適応できなかった。本手法では、適応音声に対する合成HMMのゆう度を最大化することによって、乗算性ひずみを推定し、乗算性ひずみへの適応を行う。本手法の枠組みでは従来のHMM合成で、問題となっていたSN比の変動に対する適応も乗算性ひずみへの適応の一部として定式化できる。本手法を用いて音韻認識率で評価した結果、本手法が加算性雑音と乗算性ひずみを含む音声の認識率の改善に効果があることが確認された。

キーワード 音声認識、雑音適応、回線適応、HMM

1. ま え が き

実際の環境で音声認識システムを使用する場合、背景雑音や回線雑音、回線ひずみやマイクロフォンのひずみなどが大きな問題となる。これまでに、背景雑音のような加算性雑音または回線ひずみなどの乗算性ひずみに対して、特徴量を補償する簡単でかつ有効な手法が提案されている。加算性の雑音に対しては、線形パワースペクトル領域で雑音を補償するスペクトルサブトラクション法がある[1]。また、乗算性のひずみに対しては、ケプストラム領域で補償するcepstral mean normalization法が提案されている[2]。これらの手法は線形パワースペクトル領域あるいはケプストラム領域で、入力された特徴量を平均化することによって、加算性雑音あるいは乗算性ひずみを推定し、特徴量の補償を行うものである。これらの手法は非常に単純で強力な手法である反面、加算性雑音や乗算性ひずみのモデル化を行わないで、特徴量のただ単純な平均を加算性雑音や乗算性ひずみの推定値としている。このため、加算性雑音と乗算性ひずみを同時に含むような音声を対象にする場合には、理論上、雑音やひずみを正確に推定することができない。

Vargaらは、加算性の雑音が付加した音声モデル化して、HMMを適応させる手法を提案した[3]。この手

法は、対数パワースペクトル領域で、加算性雑音の重畳した音声を加算性の雑音か音声のどちらかの最大値で近似し、モデル化を行う。このモデルを使って雑音のHMMと音声のHMMから、加算性雑音の重畳した音声のHMMの出力確率を計算する。この手法の近似の精度をより高めて、かつ認識時の計算時間を削減する手法が、NTTやCambridge大学で提案された[4]、[5]。この手法は、雑音のHMMと音声のHMMから、雑音の重畳したHMMを合成する手法であり、HMM合成法、HMM補償法、PMCなどと呼ばれている。

一方、乗算性のひずみに対しては、最ゆう推定法を用いてケプストラムのバイアスを計算することによって補償を行う方法が提案されている[6]、[7]。

ここまでで述べた手法は、加算性雑音かあるいは乗算性ひずみのどちらかだけを対象にしたものである。しかし、実環境では両者が同時に存在するのが普通である。このことを考慮して、加算性雑音と乗算性ひずみの両方をモデルに組み込むことができれば、極めて有効な音声認識が実現できる。しかし、ケプストラムと線形パワースペクトラムの間には非線形な変換があるため、加算性雑音と乗算性ひずみを同時に推定することは難しかった。

本論文では加算性の雑音だけを対象にしていたHMM合成法を拡張し、乗算性のひずみも考慮した手法を提案する。本手法は入力音声に対する合成されたHMMのゆう度が最大になるように、加算性雑音と乗算性ひずみを推定するというものである。

[†] NTTヒューマンインタフェース研究所、武蔵野市
Human Interface Laboratories, NTT, 3-9-11 Midori-cho, Musashino-shi,
180 Japan

* 現在、東京工業大学

2. HMM合成法

本章ではHMM合成法の基本的な概念について述べる。HMM合成法は図1に示すように音声HMMと雑音HMMから雑音の重畳した音声のHMMを合成する手法である。このとき合成されたHMMの構造は図に示すように二つのHMMの構造を掛け合わせた形となる。状態遷移確率もそれぞれの遷移確率の積になる。この

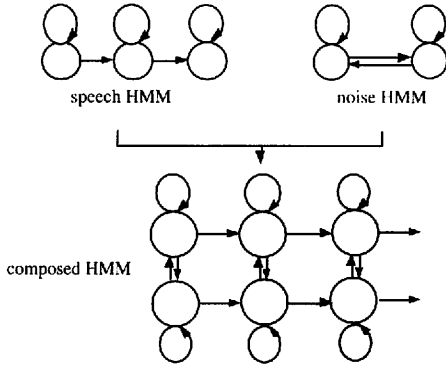


図1 HMM合成法
Fig. 1 HMM composition.

合成で最も問題となるのが出力確率分布の合成である。出力確率分布の合成法を図2を用いて説明する。音声のケプストラム領域での確率変数を $S_{cep}=(S_{cep0}, S_{cep1}, \dots, S_{cepP})$ 、雑音の確率変数を $N_{cep}=(N_{cep0}, N_{cep1}, \dots, N_{cepP})$ とし、雑音が重畳した音声の確率変数を $X_{cep}=(X_{cep0}, X_{cep1}, \dots, X_{cepP})$ とする。 $p+1$ はケプストラムの次元数を表す。

ここで、本論文の中の確率変数の表記法について定義をしておく。 c 領域での確率変数 R を R_c と表す。 $R \in \{N$ (雑音), S (音声), X (雑音が重畳しひずんだ音声)}であり、 $c \in \{cep$ (ケプストラム領域), lg (対数パワースペクトル領域), lin (線形パワースペクトル領域)}である。例えば、 X_{lin} は線形パワースペクトル領域の雑音が重畳してひずんだ音声の確率変数である。また、各領域での正規分布の平均値と分散を (μ^{Rc}, Σ^{Rc}) で表す。またその成分を $(\mu_{w_i}^{Rc}, \sigma_{w_i}^{Rc})$ と表す。

HMMの各状態で、 S_{cep} と N_{cep} の出力確率分布はそれぞれ図2の左のように正規分布をしている。この二つの分布から X_{cep} の分布が合成される。この方法を以下に示す。

音声と雑音の間に相関がないと仮定すると、雑音を含んだ音声の線形パワースペクトルは、雑音の線形パワー

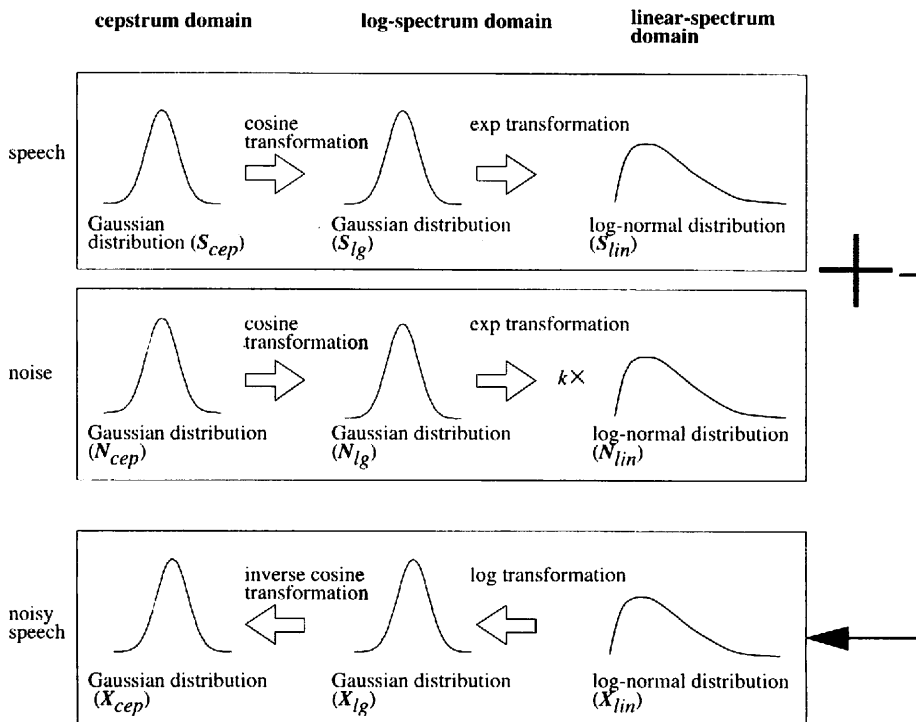


図2 HMM合成法での出力確率分布の合成法
Fig. 2 Method for creating output probability distribution in HMM composition.

スペクトルと音声の線形パワースペクトルの和になる。また、ケプストラムはcosine変換によって対数パワースペクトルに変換できる。これらのことから、確率変数の間には、次の式が成り立つ。

$$X_{cep} = \Gamma^{-1} \log(\exp(\Gamma S_{cep}) + k \cdot \exp(\Gamma N_{cep})) \quad (1)$$

ここで Γ はcosine変換であり、 $\log$ 、 $\exp$ はそれぞれベクトルの各要素の対数、指数をとる演算を表している。 k は雑音と音声との間のS/N比に対応する定数である。

cosine変換は線形変換なので、正規分布はこの変換の後でも正規分布となる。 $\exp$ 変換は正規分布を対数正規分布に変換する。これらの変換によって $\exp(\Gamma S_{cep})$ と $\exp(\Gamma N_{cep})$ は線形パワースペクトル領域で対数正規分布となる。線形パワースペクトル領域での雑音の重畳した音声の確率変数 $X_{lin} = S_{lin} + k \cdot N_{lin} = \exp(\Gamma S_{cep}) + k \cdot \exp(\Gamma N_{cep})$ は、雑音と音声の確率変数の和の形であるので、この確率変数の分布は、雑音と音声の分布の畳込みとなる。しかし、二つの対数正規分布の畳込みは非常に複雑な分布なので、図中下段の右のように、この分布を対数正規分布で近似する。この対数正規分布は、対数パワースペクトル領域では正規分布に変換される。最後に、逆cosine変換によって、ケプストラム領域で図中下段左のような正規分布が得られる。以上の説明では正規分布を各状態で一つとしているが、混合分布の場合でも、雑音の正規分布と音声の正規分布の各々の組合せについて上記の演算を順次していけばよい。

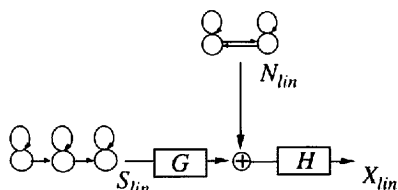


図3 雑音とひずみを含む音声の生成モデル
Fig. 3 Model for producing noisy and distorted speech.

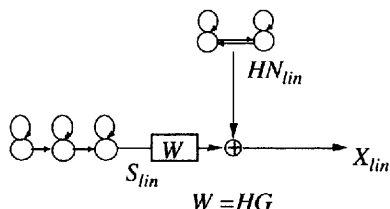


図4 雑音とひずみを含む音声の生成のモデル変形
Fig. 4 Model for producing noisy and distorted speech setting $W=HG$.

本論文で扱うHMMは分散に対角成分以外の要素を含まないようなものである。これは、ここで展開する式そのものは正規分布の分散に対角成分以外の要素も含むようなHMMに対しても適応できるが、HMM合成法において正規分布の分散に対角成分以外の成分を含むようなHMMでの評価実験を行っていないからである。

いくつかの状態で正規分布を共有するようなtied mixtureの場合でも、共有関係に注意すれば、上記手法を各分布に適応することができる。

3. 雑音が重畳しひずんだ音声のモデル化

雑音が重畳しひずんだ音声を生成するモデルを図3のようにモデル化する。音声信号 S_{lin} は音声HMMによって生成される。また、雑音信号 N_{lin} は雑音HMMによって生成される。この確率変数はそれぞれ、 $S_{lin} = (S_{lin0}, S_{lin1}, \dots, S_{linp})$, $N_{lin} = (N_{lin0}, N_{lin1}, \dots, N_{linp})$, $X_{lin} = (X_{lin0}, X_{lin1}, \dots, X_{linp})$ である。 $p+1$ は線形パワースペクトルの次元数を表す。このモデルでは音声信号 S_{lin} には、はじめに乗算性のひずみ G が掛けられる。ここでの乗算は各次元の成分ごとのかけ算を意味している。その後、加算性の雑音 N_{lin} が音声信号 GS_{lin} に加えられる。最後に回線ひずみやマイクのひずみなどを表す乗算性ひずみ H が掛けられる。最終的な信号は $X_{lin} = (H(GS_{lin} + N_{lin})) = HGS_{lin} + HN_{lin}$ となる。雑音とひずみはこの領域で、線形であると仮定している。また、 H 、 G は確率変数ではなく、定数であると仮定する。ここで $W = HG$ とおくと、 X_{lin} は $X_{lin} = WS_{lin} + HN_{lin}$ となる。このため図3のモデルは図4のように書き直すことができる。ここで、 W を以下のように表記する。

$$W = (w_0, w_1, \dots, w_p) \quad (2)$$

4. ゆう度最大化によるHMM適応

3.での式 $X_{lin} = WS_{lin} + HN_{lin}$ に対応するモデルの作成を WS_{lin} と HN_{lin} とのHMM合成で実現する。 S_{lin} を生成するHMMとしては不特定話者のHMMを用いる。ここでの未知数は W と HN_{lin} である。合成によって生成される雑音とひずみの重畳したHMMのモデルセットを $M(W, HN_{lin})$ とする。 $M(W, HN_{lin})$ は W と HN_{lin} の関数である。本論文では、 M を推定するために $P(O|M(W, HN_{lin}))$ を最大にする W と HN_{lin} を求める。ここで $P(O|M(W, HN_{lin}))$ はトレリス計算によるゆう度であり、 $O = \{o_1, o_2, \dots, o_T\}$ は入力音声の時系列である。

話者が音声を発声していない場合は $X_{lin} = WS_{lin} + HN_{lin}$

$=HN_{lin}$ となる。音声を含んでいない雑音データを使って、EMアルゴリズムにより、 HN_{lin} を生成するHMMが学習できる (H と N_{lin} を別々に学習する必要はない)。本論文ではこの統計量が音声発声部分でも変化しないと仮定して定式化を行う。以上のようにして HN_{lin} を推定することにより、未知数は W だけとなり $P(O|M(W), HN_{lin})$ は、 W だけの関数になる。これを、 $P(O|M(W))$ と表記する。この式を最大化する W が推定できれば、 X_{lin} を生成するHMMは、HMM合成法により求めることができる。

4.1 繰返し演算による方法

計算の簡単化のために、ここでは、 $P(O, A|M(W))$ を $P(O|M(W))$ のゆう度の代わりに用いる。 $P(O, A|M(W))$ はビタビ計算によるゆう度である。 $A = \{\lambda_1, \lambda_2, \dots, \lambda_T\}$ はビタビアルゴリズムによって得られる状態の時系列である。このゆう度を使って、以下のような最急降下法による繰返し演算により、ゆう度の最大化を行う。

$$w_{u,i+1} = w_{u,i} + \varepsilon \frac{\partial \log(P(O, A|M(W)))}{\partial w_u} \quad (3)$$

ここで、 ε はステップサイズであり、 i は更新回数を示す。式(3)を混合正規分布に対して求めるのは複雑であるので、混合正規分布の代わりに混合正規分布の中の最大の正規分布を使う(ビタビアルゴリズムによって求める)。

式(3)を計算するために、各状態での出力確率分布が2.で示したHMM合成法のあとで、どのような W の関数になっているかを示す。まず、 $X_{lin} = WS_{lin} + HN_{lin}$ という式から式(1)は $X_{cep} = \Gamma^{-1} \log(W \exp(\Gamma S_{cep}) + \exp(\Gamma HN_{cep}))$ となる。まず、 $\exp(\Gamma S_{cep})$ による分布の変換を考える。cosine変換は線形変換なので、正規分布はこの変換の後でも正規分布に変換される。変換される分布の平均値は式(4)のようになる。分散は(5)のようになる。またexp変換によって分布は正規分布から対数正規分布になる。確率分布における平均値と分散の定義に従って平均値と分散を計算すると、式(6)、(7)のようになる。

$$\mu^{S_{lin}} = \Gamma \mu^{S_{cep}} \quad (4)$$

$$\Sigma^{S_{lin}} = \Gamma \Sigma^{S_{cep}} \Gamma^T \quad (5)$$

$$\mu_u^{S_{lin}} = \exp(\mu_u^{S_{cep}} + \frac{\sigma_{uu}^{S_{cep}}}{2}) \quad (6)$$

$$\sigma_{uv}^{S_{lin}} = \mu_u^{S_{cep}} \mu_v^{S_{cep}} (\exp(\sigma_{uv}^{S_{cep}}) - 1) \quad (7)$$

以上の式は HN に対しても同様に成り立つ。 W は定数であり、確率変数の和の分布はそれぞれの分布の畳込

みになるので、 $W \exp(\Gamma S_{cep}) + \exp(\Gamma HN_{cep})$ に対する平均値と分散は(8)、(9)のようになる。

$$\mu_u^{X_{lin}} = w_u \mu_u^{S_{cep}} + \mu_u^{HN_{lin}} \quad (8)$$

$$\sigma_{uv}^{X_{lin}} = w_u w_v \sigma_{uv}^{S_{cep}} + \sigma_{uv}^{HN_{lin}} \quad (9)$$

この対数正規分布の平均値と分散はlog変換によって式(10)、(11)のように変換される。これは式(6)、(7)の逆変換になっている。

$$\mu_u^{X_{lin}} = \log(\mu_u^{X_{lin}}) - \frac{1}{2} \log\left(\frac{\sigma_{uu}^{X_{lin}}}{\mu_u^{X_{lin}^2}} + 1\right) \quad (10)$$

$$\sigma_{uv}^{X_{lin}} = \log\left(\frac{\sigma_{uv}^{X_{lin}}}{\mu_u^{X_{lin}} \mu_v^{X_{lin}}} + 1\right) \quad (11)$$

最後の逆コサイン変換によって平均値と分散は、以下のように変換される。

$$\mu^{X_{cep}} = \Gamma^{-1} \mu^{X_{lin}} \quad (12)$$

$$\Sigma^{X_{cep}} = \Gamma^{-1} \Sigma^{X_{lin}} \Gamma^{-1T} \quad (13)$$

Γ : コサイン変換

t : 転置

u, v : パラメータのインデックス ($0 \leq u, v \leq p$)

上記の式で、(4)、(5)、(12)、(13)ではコサイン変換の行列が対称になるような変数変換を行っている[8]。 $\mu^{X_{cep}}$ 、 $\Sigma^{X_{cep}}$ は雑音が重畳してひずんだ音声の正規分布の平均値と分散である。この二つの値は以上のように W の関数となっていることがわかる。

式(3)を展開すると $0 \leq u \leq p$ に対して次のような式が得られる。このとき、 $\Sigma^{X_{cep}}$ の W に対する微係数の項は無視した。

$$w_{u,i+1} = w_{u,i}$$

$$+ \varepsilon \frac{\partial \sum_t \left[\log(a_{\lambda_{t-1}, \lambda_t}) - \frac{1}{2} \log((2\pi)^{p+1} |\Sigma_{\lambda_{t-1}, \lambda_t}^{X_{cep}}|) - \frac{1}{2} (x_t - \mu_{\lambda_{t-1}, \lambda_t}^{X_{cep}})^T \Sigma_{\lambda_{t-1}, \lambda_t}^{X_{cep}^{-1}} (x_t - \mu_{\lambda_{t-1}, \lambda_t}^{X_{cep}}) \right]}{\partial w_u} \\ = w_{u,i} + \varepsilon \frac{\partial \sum_t \left[-\frac{1}{2} (x_t - \mu_{\lambda_{t-1}, \lambda_t}^{X_{cep}})^T \Sigma_{\lambda_{t-1}, \lambda_t}^{X_{cep}^{-1}} (x_t - \mu_{\lambda_{t-1}, \lambda_t}^{X_{cep}}) \right]}{\partial w_u} \\ = w_{u,i} + \varepsilon \frac{\partial \sum_t \left[-\frac{1}{2} (x_t - \Gamma^{-1} \mu_{\lambda_{t-1}, \lambda_t}^{X_{lin}})^T \Sigma_{\lambda_{t-1}, \lambda_t}^{X_{lin}^{-1}} (x_t - \Gamma^{-1} \mu_{\lambda_{t-1}, \lambda_t}^{X_{lin}}) \right]}{\partial w_u} \quad (14)$$

この式で $\mu_{\lambda_{t-1}, \lambda_t}^{X_{lin}}$ 、 $\Sigma_{\lambda_{t-1}, \lambda_t}^{X_{lin}}$ はビタビアルゴリズムによって選択された状態 λ_{t-1} から λ_t への遷移での最大の正規分布の平均値と分散を表す。また、 $a_{\lambda_{t-1}, \lambda_t}$ は状態 λ_{t-1} から

λ_c への遷移確率を表す。

式(14)を展開するのに必要な微係数を以下に示す。

$$\frac{\partial \mu_u^{x_{iu}}}{\partial w_u} = \frac{\mu_u^{S_{iu}}}{\mu_u^{x_{iu}}} - \frac{w_u \mu_u^{x_{iu}} \sigma_{uu}^{S_{iu}} - \mu_u^{S_{iu}} \sigma_{uu}^{x_{iu}}}{\mu_u^{x_{iu}^2} + \mu_u^{x_{iu}} \cdot \sigma_{uu}^{x_{iu}}} \quad (15)$$

$$\frac{\partial \mu_v^{x_{iv}}}{\partial w_u} = 0 \quad (u \neq v) \quad (16)$$

式(15), (16)を式(14)に代入することによって目的の式を得る。

図4のモデル化では雑音と音声のSN比の相違の項をモデルの中に陽に記述していなかった。このことを考察するために図5のようなモデルを考える。このモデルでは、雑音のHMMと音声のHMMとの間のSN比の整合性をとるため k という定数を雑音HMMの方に乗算している。この場合、式(8), (9)は次のようになる。

$$\mu_u^{x_{iu}} = w_u \mu_u^{S_{iu}} + k \mu_u^{HN_{iu}} \quad (17)$$

$$\sigma_{uv}^{x_{iu}} = w_u w_v \sigma_{uv}^{S_{iu}} + k^2 \sigma_{uv}^{HN_{iu}} \quad (18)$$

式(10)の右辺第1項の中を k で割って、 k を対数の外に出し、第2項の分子分母を k^2 で割ると式(19)が得られる。また、式(11)の括弧内部の分子分母を k^2 で割ると式(20)が得られる。

$$\mu_u^{x_{iu}} = \log(k) + \log\left(\frac{\mu_u^{x_{iu}}}{k}\right) - \frac{1}{2} \log\left(\frac{\frac{\sigma_{uu}^{x_{iu}}}{k^2}}{\frac{\mu_u^{x_{iu}^2}}{k^2} + 1}\right) \quad (19)$$

$$\sigma_{uv}^{x_{iu}} = \log\left(\frac{\frac{\sigma_{uv}^{x_{iu}}}{k^2}}{\frac{\mu_u^{x_{iu}}}{k} \cdot \frac{\mu_v^{x_{iu}}}{k} + 1}\right) \quad (20)$$

式(19)で、 $\log(k)$ はすべての次元で同じ値なので、直流成分である。この項は逆cosine変換により、ケプスト

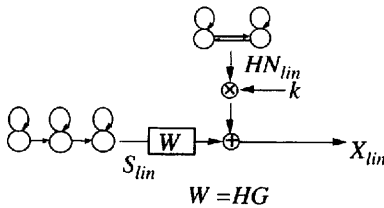


図5 雑音とひずみを含む音声の生成モデル (SN比考慮)
Fig. 5 Model for producing noisy and distorted speech considering S/N ratio.

ラムの0次の項にのみ影響を与える。本論文ではケプストラムの0次を音声認識には利用しない。このため、式(19)の $\log(k)$ は無視できる。このことから、式(10), (11)の $\mu_u^{x_{iu}}, \sigma_{uv}^{x_{iu}}$ を $\frac{\mu_u^{x_{iu}}}{k}, \frac{\sigma_{uv}^{x_{iu}}}{k^2}$ に置き換えても生成されるケプストラムの分布は同じであることがわかる。この結果は、式(17), (18)の右辺を k, k^2 で割っても結果が変わらないことを示している。以上の考察から、図4は図5を包含するようなモデルとなっており、 k を陽に定義しなくてもよいということがわかる。

また、 W を以下のようにケプストラム領域で定義することもできる。

$$W_{cep} = (w_0, w_1, \dots, w_p) \quad (21)$$

このように定義すると、HMM合成の式(4), (8), (9)がそれぞれ以下になる。

$$\mu^{S_{ic}} = \Gamma(\mu^{S_{ref}} + W_{cep}) \quad (22)$$

$$\mu_u^{x_{iu}} = \mu_u^{S_{iu}} + \mu_u^{HN_{iu}} \quad (23)$$

$$\sigma_{uv}^{x_{iu}} = \sigma_{uv}^{S_{iu}} + \sigma_{uv}^{HN_{iu}} \quad (24)$$

これらの式から、式(15), (16)に対応する偏微分係数が計算でき、この結果を式(14)に代入すれば同様の繰返し演算の式が求まる。

4.2 並列モデルによる方法

この節では繰返し演算による方法より簡単な手法について述べる。この方法では図6に示すように、 W の異なるHMMをいくつか並列に用意しておき、その中からゆう度が最大のHMMを選び出す。この方法は、可能な W が無数にある場合は実現不可能であるが、制限された W に対しては極めて有効な手法である。このような制限として、例えば、従来のHMM合成法で問題となっていた学習時と認識時のSN比の違いへの適応 (式

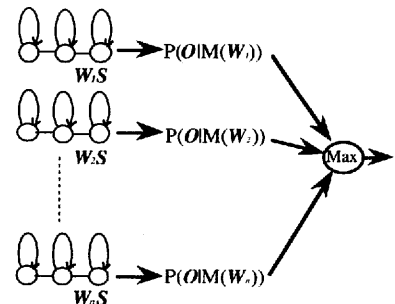


図6 並列法
Fig. 6 Parallel HMM composition.

(1)の k の推定)などがあげられる。

5. 認識実験

5.1 実験条件

提案する手法の効果を調べるために音素認識実験を行った。表1に実験に用いたデータを示す。評価用データには話者1名の発声した電話番号案内タスク51文を用いた。

雑音を含んだ音声を作成するために、このデータにSN比が6 dBと12 dBになるように計算機室雑音を加えた。その後、マイクひずみや回線ひずみとして、フィルタ $(1-0.97z^{-1})$ を計算機上で人工的に掛けた。ひずみの特性としては各種のものが考えられるが、マイク特性の違いの最も顕著なものは、高域の周波数特性の傾斜が変化するものである。そこで、今回は $(1-0.97z^{-1})$ のような簡単な高域を強調するフィルタを用いて、この傾斜の変化を実現した。適応用データとして評価用データと同じ話者の発声したATR音素バランス文の先頭の1文を用いた。適応用データの発声内容は既知であり、発声内容のテキストを用いて音素HMMを結合して発声文章全体のHMMを作成した。このHMMを用いてHMMの適応を行った。HMMに用いた特徴量とHMMの構造を表2に示す。ここで用いたHMMは、分布の分散が対角要素のみのものである。特徴量の適応は、ケプストラムについてだけ行った。デルタケプストラムは適応前のHMMの値をそのまま使った。評価ではラベルなし評価法[9]を用いて音素認識率を求めた。

実験は図7に示す手順で行った。最初に、フィルタ $(1-0.97z^{-1})$ を通過させた雑音データのみを使って雑

音モデルを学習した。次に、線形パワースペクトル領域で、 $W=(k, k, \dots, k)$ において、並列モデルによる方法により、 k を推定した。ここでは k の値をSN比に換算して3 dBおきに3 dBから21 dBの間で変えて、ゆう度が最大となるものを選んだ。最後にこの k を初期値とし、繰り返し演算による方法によって、線形パワースペクトル領域とケプストラム領域の両方で W の推定を行った。このような操作を行ったのは次のような理由による。予備実験において、4.1の式をそのまま実行した結果、式(3)が収束をしないことがほとんどであった。これは次のことによると考えられる。最急降下法では、局所的な関数の傾斜方向にパラメータを更新していく。このとき、局所的な傾斜方向と大域的な傾斜方向が違う場合には振動を起こして収束しない場合がある。4.1で述べたように、 W は線形領域で定義されており、学習時と認識時のSN比の違いへの適応も含んでいる。従って、このSN比の違いによって、 W の値はSN比の指数関数的に大きく変化する。設定したSN比が最適となるSN比の値より大きく違う場合には、局所的な傾斜方向と大域的な傾斜方向（最適SN比の方向）に違いがあると考えられる（ W の値があまりにも離れすぎているため）。このために予備実験結果が収束しなかったと考えられる。ここでは、この問題を回避するため、最初に、並列モデルによる方法でSN比に関してゆう度の最大化を行い、この値を用いて、局所的な W の最大化を行った。

5.2 実験結果

実験結果を表3に示す。“適応なし”の実験では、HMM合成を行っていないHMMを用いた。“加算性雑音のみ”の実験では、並列モデルによる方法で求めた k を用いてHMM合成を行った。ゆう度最大となる場合の認識率を表3の“ゆう度最大”の項に記述した。また、比較のため次の列に認識率が最大になるようなときの認識率も示した。

次に、線形パワースペクトル領域とケプストラム領

表1 実験に用いたデータ

Table 1 Experimental data.

不特定話者HMM学習用データ	ASJ連続音声データベース (約9600文)
評価用データ	電話番号案内51文
雑音データ	電子協雑音データベース計算機室雑音
適応用データ	ATR音韻バランス文の最初の1文

表2 HMMの特徴量と構造

Table 2 Feature vectors and structure in HMMs.

	音素HMM	雑音HMM
特徴量	16次元ケプストラム 16次元デルタケプストラム デルタパワー	
状態数	3状態	1状態
混合数	4	1

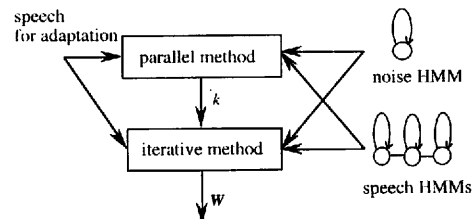


図7 雑音とひずみに対するHMMの生成法
Fig. 7 Procedure for making noisy HMMs.

表3 音素認識結果
Table 3 Phoneme recognition results.

S/N比	適応なし	加算性雑音のみ 尤度最大 (k のみ)	加算性雑音のみ 認識率最大 (k のみ)	加算性雑音と 乗算性ひずみ (線形パワースペクトラム領域)	加算性雑音と 乗算性ひずみ (ケプストラム領域)
6 dB	28.4%	37.0%	41.1%	52.9%	53.4%
12 dB	44.7%	56.5%	56.7%	67.7%	67.3%

域において W を推定したときの音韻認識率を示した。表3から k のみ推定しても6 dB, 12 dBの両者で10%程度認識率が改善することがわかる。ゆう度最大で求めた結果は、事後的に認識率最大で求めたものに近い認識率を示している。このことは、並列モデルによる手法が、HMM合成での、音声HMMと雑音HMMの間の認識時と学習時のSN比の違いを推定できることを示している。更に乗算性ひずみを推定すると、6 dB, 12 dBどちらも更に10%以上認識率が改善することがわかる。このことから、我々のアルゴリズムが加算性雑音だけでなく、乗算性ひずみにも適応できることがわかる。

線形パワースペクトルで W を推定する場合とケプストラム領域で W を推定する場合を比較すると、乗算性ひずみの推定の効果は線形パワースペクトル領域で行ってもケプストラム領域で行ってもほぼ同程度であることがわかる。

このときの計算量については、以下のような考察ができる。本論文ではケプストラム領域での微係数を計算するために、合成関数の微分を使っている。このため、ケプストラム領域で微係数を求めるためには、一度線形パワースペクトル領域の微係数を求めなければならない。この計算量は線形パワースペクトル領域の微係数を直接求めるのとはほぼ同じである。しかし、この微係数をケプストラム領域まで展開するためには更に、逆cosine変換をしなければならないので、その分の計算量が余計に必要となる。

学習の面から考えると、ケプストラム領域の場合、線形パワースペクトル領域の場合と違いSN比の違いを0次の項への加算で実現できるため、5.で示したような最急降下法を収束させるための制御がより簡単になると考えられる。

6. む す び

本論文ではHMM合成法を拡張し、加算性の雑音と乗算性のひずみを考慮するHMMの合成法を提案した。この手法は、雑音HMMと音声HMMから合成されるHMMを乗算性ひずみの関数として定式化し、このHMMの適

応音声に対するゆう度を最大にするように、乗算性ひずみを推定する方法である。この方法の実現方法として本論文では、繰返し演算による方法と並列モデルによる方法を提案した。二つの方法を使って音素認識率で評価した結果、これらの方法が、加算性雑音と乗算性ひずみの両方に適応できることがわかった。

謝辞 日ごろ、有意義な議論をして下さった古井特別研究室の方々に感謝致します。本実験では、日本音響学会連続音声認識データベースおよび、電子協のデータベースを利用した。

文 献

- [1] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoust. Speech Signal Processing, vol.ASSP-27, pp.11-120, April 1979.
- [2] B.Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," Journal of the Acoustic Society of America, vol.55, pp.1304-1312, June 1974.
- [3] A.P.Varga and R.K.Moore, "Hidden Markov model decomposition of speech and noise," Proc. Internat. Conf. Acoust. Speech Signal Process., pp.845-848, April 1990.
- [4] M.J.F.Gales and S.J.Young, "An improved approach to the hidden Markov model decomposition of speech and noise," Proc. Internat. Conf. Acoust. Speech Signal Process., pp.233-236, March 1992.
- [5] F.Martin, K.Shikano, and Y.Minami, "Recognition of noisy speech by composition of hidden Markov models," Proc. Eurospeech, pp.1031-1034, Sept. 1993.
- [6] M.G.Rahim and B-H Juang, "Signal bias removal for robust telephone based speech recognition in adverse environments," Proc. Internat. Conf. Acoust. Speech Signal Process., pp.1445-1448, April 1994.
- [7] R.C.Rose, E.M.Hofstetter, and D.A.Reynolds, "Integrated models of signal and background with application to speaker identification in noise," IEEE Trans. on Speech and Audio Processing, vol.2, pp.245-257, April 1994.
- [8] F.Martin, "Recognition of noisy speech by composition of hidden Markov models," Master Thesis, University of Tokyo, Jan. 1992.
- [9] 南 泰浩, 松岡達雄, 鹿野清宏, "音韻ラベルを用いないHMM評価法とそれを用いた連続音声認識用HMMの評価," 信学論 (A), vol. J77-A, no. 2, pp. 267-273, 1994. (平成8年8月14日受付, 9年1月13日再受付)



南 泰浩

昭61慶大・理工・電気卒。平3同大大学院博士課程了。同年日本電信電話(株)入社。現在NTTヒューマンインタフェース研究所主任研究員。音声認識の研究に従事。日本音響学会会員。



古井 貞熙 (正員)

昭43東大・工・計数卒。昭45同大大学院修士課程了。同年NTT電気通信研究所入社。以後、同研究所において、音声認識、話者認識、音声知覚などの研究に従事。昭53～54ベル研究所客員研究員。現在東京工業大学教授。工博。昭50米沢賞、昭63、平5論文賞。

平2著述賞、平1科学技術庁長官賞、IEEE ASSP Society Senior Awardなど受賞。著書「ディジタル音声処理」, 「Digital Speech Processing, Synthesis, and Recognition」, 「音響・音声工学」, 編著「Advances in Speech Signal Processing」など。IEEEおよび米国音響学会Fellow。