

論文 / 著書情報
Article / Book Information

論題(和文)	尤度最大化原理によるHMM適応化法
Title(English)	
著者(和文)	南 泰浩, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	1995年日本音響学会春季講演論文集, , No. 2-5-16, pp. 61-62
Citation(English)	, , No. 2-5-16, pp. 61-62
発行日 / Pub. date	1995, 3

尤度最大化原理によるHMM適応化法*

○南 泰浩 古井貞熙 (NTTヒューマンインタフェース研究所)

1. はじめに

音声認識を実際の環境で使用する場合、背景雑音や回線雑音、回線ひずみなどが大きな問題となる。これらの雑音は通常、加算性の雑音と乗算性のひずみとの線形モデルでモデル化することが多い。もし、このような加算性雑音と乗算性の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。

我々はこれまで、加算性の雑音に極めて有効なHMM合成 (HMM補正) 手法を提案した^{[1][2]}。本論文では加算性の雑音だけを対象にしていたHMM合成を拡張し、乗算性のひずみも考慮したアルゴリズムを提案する。このアルゴリズムはHMM合成において、入力音声に対する合成されたHMMの尤度が最大になるように、乗算性ひずみを推定するというものである。

2. 尤度最大化によるHMM適応

2.1 雑音を含んだ音声のモデル化

雑音を含んだ音声の生成するモデルを図1のようにモデル化する。音声信号Sは音声HMMによって生成される。また、雑音信号Nは雑音HMMに

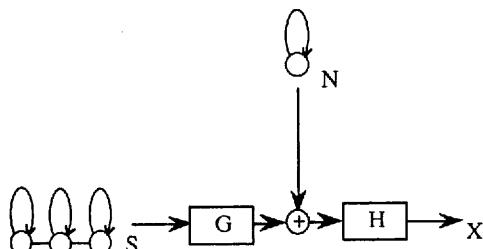


図1 雑音を含む音声のモデル化

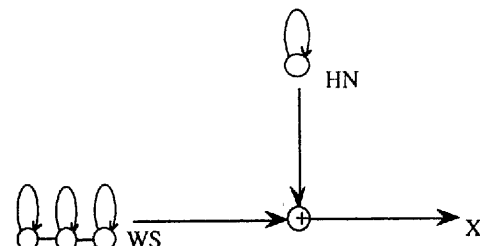


図2 雑音を含む音声モデルの変形 ($W=HG$)

*A maximum likelihood procedure for an HMM adaptation method By Yasuhiro Minami and Sadaoki Furui (NTT Human Interface Laboratories)

よって生成される。ここではSもNも線形スペクトル領域で定義される。このモデルではSは初め話者の特徴やロンバート効果による発声のひずみなどを示す乗算性のひずみGが掛けられる^[3]。その後加算性の雑音Nが音声信号GS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 $X = H(GS + N) = HGS + HN$ となる。ここで $W = HG$ と置くと、 $X = WS + HN$ となる。このため図1のモデルは図2のように書き直すことができる。

HNを生成するHMMは、音声が発声されていない状態で、回線やマイクを通して収録された雑音から学習する。また、Sを生成するHMMとして不特定話者のHMMを用いる。ここで、Wが推定できれば、Xを生成するHMMはHMM合成により求めることができる。Wを推定するために、パワースペクトル領域で以下のようにWを定義する。

$$W = \{w_0, w_1, w_2, \dots, w_p\} \quad (1)$$

ここで、 $p+1$ はパワースペクトルの次数である (定式化の簡単化のため、ケプストラムの次数も $p+1$ とした)。

2.2 尤度最大化

Xを生成するHMMがHNを生成するHMMとWSを生成するHMMからHMM合成によって合成できると仮定する。その合成によって生成されるHMMのセットを $M(W)$ とする。ここで、 $M(W)$ はWの関数である。Wを推定するために $P(O|M(W))$ が $P(O, \Lambda|M(W))$ を最大にするWを求める。ここで $P(O|M(W))$ はトレリス計算による尤度であり、 $P(O, \Lambda|M(W))$ はビタービ計算による尤度である。この $O = \{x_1, x_2, \dots, x_T\}$ は入力音声の時系列であり、 $\Lambda = \{s_1, s_2, \dots, s_T\}$ はビタービアルゴリズムによって得られる状態の時系列である。

本論文では $P(O, \Lambda|M(W))$ を以下のような最急降下法を使うことによって最大化を行なう。

$$W(k) = W(k-1) + \epsilon \frac{\partial \log(P(O, \Lambda|M(W)))}{\partial W} \quad (2)$$

ここで、 ϵ はステップサイズであり、 k は更新回数を示す。

式(2)を混合ガウス分布に対して求めるのは複雑であるので、状態 s_t から s_t へ遷移するときの混合ガウス分布の代わりに混合ガウス分布の中の最大のガウス分布(式(3))を使う。

$$b_{s_t s_t}(x_t) = \frac{1}{(2\pi)^P | \Sigma_{s_t s_t} |} e^{-\frac{1}{2} (x_t - \mu_{s_t})^T \Sigma_{s_t s_t}^{-1} (x_t - \mu_{s_t})} \quad (3)$$

このガウス分布を使ったビタービアルゴリズムに

より、 $P(O, \Lambda(M(W)), \mu = \{\mu_1, \mu_2, \dots, \mu_T\}, \Sigma = \{\Sigma_1, \Sigma_2, \dots, \Sigma_T\})$ を計算する。ここで、 μ と Σ はビタビアルゴリズムによって得られる平均値と分散の時系列を表している。

c領域での確率変数Rを R_c と表す。ここで $R \in \{HN$ (雑音), S (音声), X (雑音を含む音声)}であり、 $c \in \{cep$ (ケプストラム領域), lg (対数スペクトル領域), lin (線形スペクトル領域)}である。例えば、 X_{lin} は線形スペクトル領域の雑音を含む音声の確率変数である。この確率変数のガウス分布の平均値と分散を $(\mu_{lin}^{X_{lin}}, \Sigma_{lin}^{X_{lin}})$ で表す。

HMM合成から次式が得られる。

$$\mu_{lg}^S = \Gamma \mu_{cep}^S \quad (4)$$

$$\Sigma_{lg}^S = \Gamma \Sigma_{cep}^S \Gamma^T \quad (5)$$

$$\mu_{lin}^{S_{lin}} = \exp(\mu_{lg}^{S_{lin}} \frac{\sigma_{lin}^{S_{lin}}}{2}) \quad (6)$$

$$\sigma_{lin}^{S_{lin}} = \mu_{lin}^{S_{lin}} \mu_{lin}^{S_{lin}} (\exp(\sigma_{lin}^{S_{lin}}) - 1) \quad (7)$$

$$\mu_{lin}^{X_{lin}} = \log(w_u \mu_{lin}^{S_{lin}} + \mu_{lin}^{HN_{lin}}) - \log(\frac{w_u \mu_{lin}^{S_{lin}} \sigma_{lin}^{S_{lin}} + \mu_{lin}^{HN_{lin}}}{(w_u \mu_{lin}^{S_{lin}} + \mu_{lin}^{HN_{lin}})^2}) \quad (8)$$

$$\sigma_{lin}^{X_{lin}} = \log(\frac{w_u \mu_{lin}^{S_{lin}} \sigma_{lin}^{S_{lin}} + \mu_{lin}^{HN_{lin}}}{(w_u \mu_{lin}^{S_{lin}} + \mu_{lin}^{HN_{lin}})(w_u \mu_{lin}^{S_{lin}} + \mu_{lin}^{HN_{lin}})} + 1) \quad (9)$$

$$\mu_{cep}^X = \Gamma^{-1} \mu_{lg}^X \quad (10)$$

$$\Sigma_{cep}^X = \Gamma^{-1} \Sigma_{lg}^X \Gamma^{-1T} \quad (11)$$

Γ : コサイン変換

T : 転置

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

式(2)の微係数に対する共分散行列の効果を無視すると $0 \leq u \leq p$ に対して次のような式が得られる。

$$w_u(k) = w_u(k-1)$$

$$\begin{aligned} & \frac{\partial \sum_i \{-\log((2\pi)^p |\Sigma_i^{X_{cep}}|) - \frac{1}{2}(x_i - \mu_i^{X_{cep}})^T \Sigma_i^{X_{cep}-1} (x_i - \mu_i^{X_{cep}})\}}{\partial w_u} \\ &= w_u(k-1) + \\ & \varepsilon \sum_i \left\{ \frac{1}{2} \left(\frac{\partial \mu_i^{X_{cep}}}{\partial w_u} \right)^T \Sigma_i^{X_{cep}-1} (x_i - \mu_i^{X_{cep}}) + \frac{1}{2} (x_i - \mu_i^{X_{cep}})^T \Sigma_i^{X_{cep}-1} \frac{\partial \mu_i^{X_{cep}}}{\partial w_u} \right\} \\ &= w_u(k-1) + \\ & \varepsilon \sum_i \left\{ \frac{1}{2} \left(\Gamma^{-1} \frac{\partial \mu_{lg}^{X_{lin}}}{\partial w_u} \right)^T \Sigma_i^{X_{cep}-1} (x_i - \mu_i^{X_{cep}}) + \frac{1}{2} (x_i - \mu_i^{X_{cep}})^T \Sigma_i^{X_{cep}-1} \Gamma^{-1} \frac{\partial \mu_{lg}^{X_{lin}}}{\partial w_u} \right\} \end{aligned} \quad (12)$$

また、式(8)の微分により次式が得られる。

$$\begin{aligned} \frac{\partial \mu_{lin}^{X_{lin}}}{\partial w_u} &= \frac{\mu_{lin}^{S_{lin}}}{\mu_{lin}^{HN_{lin}} + w_u \mu_{lin}^{S_{lin}}} \\ & - \frac{w_u \mu_{lin}^{HN_{lin}} \sigma_{lin}^{S_{lin}} - \mu_{lin}^{S_{lin}} \sigma_{lin}^{HN_{lin}}}{(\mu_{lin}^{HN_{lin}} + w_u \mu_{lin}^{S_{lin}})^2 + (\mu_{lin}^{HN_{lin}} + w_u \mu_{lin}^{S_{lin}})(\sigma_{lin}^{HN_{lin}} + w_u^2 \sigma_{lin}^{S_{lin}})} \end{aligned} \quad (13)$$

この式と式(10)、(11)を用いて、式(12)の計算が実現できる。(12)を収束するまで行なうことによって

最終的に、加算性雑音と乗算性ひずみを考慮したHMMを生成する。

3. 認識実験

提案する手法の効果を調べるために音韻認識実験を行なった。不特定話者のHMMをASJの連続音声認識データベースから作成した。テストデータには話者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S/N比が12dBになるように計算機室雑音を加え、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)を計算機上で掛けた。適応化データとして同じ話者の発声したATR音韻バランス文の先頭の1文を用いた。適応化データにもテストデータと同様な処理を行なった。評価ではラベルなし評価法¹⁰⁾を用いて音韻認識率を求めた。

実験結果を表1に示す。"HMM合成なし"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、6dB、9dB、12dB、15dB、18dBの条件でHMM合成を行ない、認識率が最大になるHMMを使った。表から提案する手法を使った場合、HMM合成のみの場合に比べ認識率が11%改善されていることが分かる。

表1 音韻認識実験結果

HMM合成なし	HMM合成のみ 加算性雑音のみ	HMM合成 乗算性ひずみ推定
44.7%	56.7%	67.7%

4. まとめ

本論文ではHMM合成を拡張し、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。実験によりこの手法の有効性を確認した。

文献

- [1] 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., March 1992, pp. 233-236.
- [2] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models", Proc. Eurospeech, September 1993, pp. 1031-1034.
- [3] Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, July 1994, pp. 380-394.
- [4] 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、1994、A、Vol. J77-A、No. 2、pp. 267-273.