一次線形予測による残響下音声認識の検討*

○滝口 哲也　西村 雅史（日本アイ・ビー・エム 東京基礎研究所）

1 はじめに

ユーザがマイクロフォンから離れて音声を入力するハンズフリー音声認識では、残響などの影響を受けて認識精度が劣化する。従来、音声の伝達経路による影響に対処する方法として、ケプストラム平均減算法 (Cepstrum Mean Subtraction: CMS) が使われている。この手法は、例えば電話回線の影響などのように、伝達特性のインパルス応答が比較的短い場合には有効であるが、マイクロフォンから離れて発話した際には、残響の影響を受けて十分な性能が得られない。これは、一般に室内の残響の伝達特性の長さが、音声認識に用いられる短区間分析の窓幅よりも長くなるためである。このような残響成分を除去する方法として、長い分析窓と短い分析窓を組み合わせる方法 [1] 等が提案されているが、残響成分の除去と同時に音声がひずむ可能性がある。本稿では、短区間分析における残響成分の影響に対処するために、フレーム外影響成分を過去の観測系列から一次の線形予測で表現する方法を検討する。

2 一次線形予測を用いたモデル適応化

残響下での観測音声信号の短区間スペクトルを、残響成分の影響が同一フレーム内におさまるものと、それ以外のものとに大別し、以下の式により近似する。

$$O(\omega;t) \approx S(\omega;t)\cdot H_0(\omega) + \sum_{d=1} S(\omega;t-d)\cdot H_d(\omega) \quad (1)$$

ここで、ω は周波数、t はフレーム番号を表し、クリーン音声を S、音響伝達特性を H で表す。また d はいくつ前の時間フレームからの反射であるかを表している。式 (1) の第二項目のフレーム外影響成分を、次式のように、実際の過去の観測系列からの一次の線形予測にて近似する。

$$\begin{aligned}\hat{O}(\omega;t) &\approx S(\omega;t)\cdot H(\omega;t) + \alpha\cdot O(\omega;t-1)\\ &= \exp\{\cos\{S(c;t)+H(c;t)\}\}\\ &\quad +\alpha\cdot O(\omega;t-1)\end{aligned} \quad (2)$$

ここでの $S(c;t)$、$H(c;t)$ は各々、t フレームにおけるクリーン音声信号と音響伝達特性のケプストラム c 次の値を表している。式 (2) に従い、音声 HMM の出力確率分布を一時刻前の観測信号を使い残響音に適応させ、音声認識を行う。図 1 に、出力確率分布の適応化アルゴリズムを示す。まず、ケプストラム領域にて、クリーン音声 HMM ($\lambda_{S_{cep}}$) とフレーム内伝達特性 HMM ($\lambda_{H_{cep}}$) の合成を行う。

$$\mu^{(SH)}_{p,j,k} = \mu^{(S)}_{p,j,k} + \mu^{(H)}, \quad \Sigma^{(SH)}_{p,j,k} = \Sigma^{(S)}_{p,j,k} + \Sigma^{(H)}$$

ここで、音韻 p の状態 j の分布番号 k の平均ベクトルを $\mu_{p,j,k}$、分散を $\Sigma_{p,j,k}$ とする。次にケプストラムからスペクトラム領域へ変換し、モデル合成を行う。ここでの適応は分布の平均ベクトルのみに対し行い、分散はそのままとした。

$$\hat{\mu}^{(O)}_{p,j,k} = \mu^{(SH)}_{p,j,k} + \alpha\cdot O(t-1), \quad \hat{\Sigma}^{(O)}_{p,j,k} = \Sigma^{(SH)}_{p,j,k}$$

得られた分布をケプストラム領域まで変換し、尤度計算を行う。

3 EM アルゴリズムによる予測係数の推定

次に、式 (2) のフレーム内伝達特性 H と予測係数 α の推定方法について述べる。パラメータ H と α は、観測信号に対する音響モデルの尤度が最大になるようにして求める。まず、ケプストラム領域にて H の推定、次にスペクトラム領域にて α の推定を行う。今回は、まず H の推定は $\alpha = 0$ として、文献 [2] の方法を使った。それからフレーム内伝達特性 H で推定しきれない要素を、過去の観測系列から一次の線形予測にて推定する。推定アルゴリズムを以下に示す。

1) ケプストラム領域にて H の推定を行う [2]。
2) ケプストラム領域にて $\lambda_{S_{cep}}$ と $\hat{\lambda}_{H_{cep}}$ の合成 HMM ($\hat{\lambda}_{SH_{cep}}$) を求める。

図 1: 一次線形予測を用いた出力確率分布の合成

* "Reverberant speech recognition using a first-order linear prediction," by T. Takiguchi and M. Nishimura (IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.)

3) $\hat{\lambda}_{SH_{cep}}$ をスペクトラム領域まで変換する。

$$\hat{\lambda}_{SH_{lin}} = \mathrm{Exp}\{\mathrm{Cos}(\hat{\lambda}_{SH_{cep}})\} \quad (3)$$

4) α の推定を尤度最大化基準により行う。

$$\hat{\alpha} = \underset{\alpha}{\mathrm{argmax}} \Pr(O_{lin}|\lambda_{SH_{lin}}, \alpha) \quad (4)$$

式 (4) を EM アルゴリズムにより求める。まず E-step (Expectation step) では、以下の Q 関数を計算する。

$$Q(\hat{\alpha}|\alpha) = E[\log \Pr(O|\lambda_{SH_{lin}}, \hat{\alpha})|\lambda_{SH_{lin}}, \alpha] \quad (5)$$

上記 Q 関数は、出力確率分布のみに関する項に注目すると以下のようになる。

$$\begin{aligned} &Q(\hat{\alpha}|\alpha) \\ &= -\sum_p \sum_j \sum_k \sum_t \gamma_{p,j,k}(t) \cdot \Big[\tfrac{1}{2} \log(2\pi)^D \Sigma_{p,j,k}^{(SH)} \\ &\quad + \tfrac{\{O_p(t) - \mu_{p,j,k}^{(SH)} - \hat{\alpha} \cdot O_p(t-1)\}^T \cdot}{} \\ &\quad \tfrac{\{O_p(t) - \mu_{p,j,k}^{(SH)} - \hat{\alpha} \cdot O_p(t-1)\}}{2\Sigma_{p,j,k}{}^{(SH)}} \Big] \end{aligned} \quad (6)$$

ここで、O_p は音韻 p に関連する観測系列とする。また、$\gamma_{p,j,k,t} = \Pr(s_p(t) = j, m_p(t) = k|O_p, \lambda_{SH_{lin}}, \alpha)$ とする。($s(t)$ は状態系列、$m(t)$ は混合要素系列を表す。)

従って、Q 関数を最大にする $\hat{\alpha}$ は、$\partial Q(\hat{\alpha}|\alpha)/\partial\hat{\alpha} = 0$ より、

$$\hat{\alpha} = \frac{\sum_p \sum_j \sum_k \sum_t \gamma_{p,j,k}(t) \frac{O_p(t-1)\{O_p(t) - \mu_{p,j,k}^{(SH)}\}}{\Sigma_{p,j,k}^{(SH)}}}{\sum_p \sum_j \sum_k \sum_t \gamma_{p,j,k}(t) \frac{O_p^2(t-1)}{\Sigma_{p,j,k}^{(SH)}}}$$

4 認識実験

4.1 実験条件

残響下音声を作成するために、RWCP 実環境音声・音響データベースより残響時間 300ms と 120ms のインパルス応答を使用した（マイクまでの距離は 2m)。このインパルス応答と ATR 音声データベースのクリーン音声と畳み込みを行い、テストデータと適応データを作成した。タスクは語彙 500 単語として、テストデータは男性話者一人が対象語彙を一回発声したものである。また適応データ数は 10 単語とした。特定話者 HMM（54 音韻、学習用データ数 2620 単語）を使用して認識実験を行う。信号の分析条件は、サンプリング周波数 12kHz、窓幅 32ms である。分析周期を 8ms としたので、窓間で重なりが生じないように、式 (2) において 4 フレーム分ずらした（過去の）観測信号を用いた。

表 1: 認識結果

残響時間	CMS	モデル適応	
		従来法	提案法
300ms	86.0%	89.8%	93.4%
120ms	92.6%	95.4%	95.8%

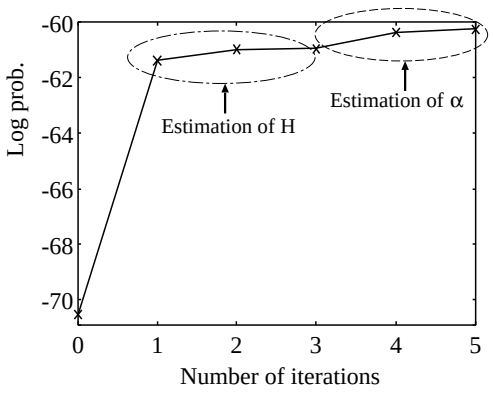

図 2: パラメータ推定における収束性

4.2 実験結果

実験結果を表 1 に示す。残響時間が 300ms の時、フレーム内成分 H のみを使った場合（従来法 [2]）の認識率は 89.8%であった。さらに過去の観測系列の一次線形予測により、フレーム外影響成分を補正した提案法では、認識率は 93.4%まで改善された。比較手法として CMS を使った場合では 86%であった。また matched model (音響モデルを評価環境と同じ学習データ 2620 単語で再学習) での認識率は 96.4%である。この結果と比べるとまだ認識精度は劣っているので、今後さらに改善が必要である。残響時間が比較的短い 120ms の時は、従来法と提案法との差はほとんどなかった。また図 2 に提案手法の収束性を示す（残響時間：300ms)。本実験では EM アルゴリズムは数回で収束しているのが分かる。

5 おわりに

本稿では、残響下における音声認識精度を改善するために、フレーム外からの残響成分の影響を観測系列の一次線形により近似する方法を検討した。また線形予測係数を EM アルゴリズムを用いて、シングルマイクロフォンにて観測される音声信号のみから推定する方法を提案した。今後は、さらに背景雑音が存在する環境下において評価をすすめていく予定である。

参考文献

[1] C. Avendano, S. Tivrewala, and H. Hermansky, “Multiresolution channel normalization for ASR in reverberant environments,” Eurospeech, pp. 1107-1110, 1997.

[2] T. Takiguchi, S. Nakamura, and K. Shikano, “HMM-Separation-Based Speech Recognition for a Distant Moving Speaker,” IEEE Trans. on SAP, Vol. 9, pp. 127-140, Feb. 2001.