F-045

強化学習によるエコーキャンセラの制御戦略の獲得

Acquirement of a Control Strategy of Echo Cancellers by Reinforcement Learning

○箱石 直士, 西山 清*

○ Naohito HAKOISHI and Kiyoshi NISHIYAMA*

1 はじめに

携帯電話などにおけるエコーキャンセラ [1] の実用化においては、適応アルゴリズムの高性能化だけでなく、ダブルトーク検出が不可欠である。残留エコーを増大させるダブルトーク発生時においては、フィルタ係数の更新を休止すべきであり、その制御の成否はエコーキャンセラの性能を大きく左右する。本研究では、エコーキャンセラにおけるダブルトーク時の適応フィルタの制御戦略を我々が考案した遅延 Q-learning を用いて学習し、その性能をシミュレーションにより評価する。

2 エコーキャンセラ

エコーを $\{z_k\}$、背景雑音を $\{v_k\}$ とし、遠端話者信号 $\{u_k\}$ がエコーパスへの入力信号となることを考慮すれば、エコーパスのインパルス応答 $\{h_i\}$ により、エコーの観測値 $\{y_k\}$ は次式で表される。

$$y_k = z_k + v_k = \sum_{i=0}^{M} h_i u_{k-i} + v_k, \quad k \geq 0 \qquad (1)$$

ただし、近端話者信号 s_k は無視し、M はタップ数 (インパルス応答長) とする。このとき、エコーパスのインパルス応答の推定値 $\{\hat{h}_i\}$ が得られれば、これより疑似エコー $\hat{z}_k$ が次のように得られる。

$$\hat{z}_k = \sum_{i=0}^{M} \hat{h}_i u_{k-i}, \quad k = 0, 1, 2, \cdots \qquad (2)$$

この疑似エコーを送信信号から差し引くことによってエコーをキャンセルすることができる。ただし、$k - i < 0$ のとき $u_{k-1} = 0$ とする。

以上より、エコーキャンセラは直接観測可能な遠端話者信号 $\{u_k\}$ とエコーの観測値 $\{y_k\}$ から適応フィルタを用いてインパルス応答を推定することによって実現できることがわかる。

*岩手大学 工学部 情報システム工学科, 〒 020-8551 盛岡市上田 4-3-5, e-mail: nisiyama@cis.iwate-u.ac.jp

このとき、エコーパスの推定を妨害する要素は、背景雑音と近端話者からの信号 (音声) である。一般に、話者 2 人が同時に話し始めた (ダブルトーク) ときはインパルス応答の推定を中断する。

3 Q-learning

Q-learning は強化学習の一種であり、最適な行動価値関数 $Q^*(s, a)$ を試行錯誤により推定するものである。特に、状態が離散的な場合、Q 関数は Q テーブルとして表すことができる。以下に $Q^*(s, a)$ の推定値である $Q_k(s_k, a_k)$ の更新式を示す [2] 。

$$\begin{aligned} Q_{k+1}(s_k, a_k) &= (1-\alpha) Q_k(s_k, a_k) \\ &\quad + \alpha \{ r_k + \gamma \max_{a' \in A} Q_k(s_{k+1}, a') \} \end{aligned} \qquad (3)$$

ここで、α は学習率、γ は割引率であり、s_{k+1} は状態 s_k で行動 a_k をとったときの遷移先の状態を表す。なお、本研究では行動選択法として次式のボルツマン選択を用いる。

$$\pi(s, a) = \frac{e^{Q(s,a)/T}}{\sum_{b \in A} e^{Q(s,b)/T}} \qquad (4)$$

4 制御戦略の学習

ダブルトークは適応フィルタの係数を大きく乱す。そのため、エコーキャンセラにおいて、ダブルトークの発生時には適応フィルタの係数の更新を休止する必要がある。

4.1 環境

エージェントの置かれる環境の時刻 k における状態は、エコーの観測値 y_k、残留エコー e_k の関数より得られる。本論文では、環境の状態として分散 $\mathrm{E}\{y_k^2\}$,$\mathrm{E}\{e_k^2\}$ を用い、次のように計算される $\sigma^2_{y|k}$,$\sigma^2_{e|k}$ によってそれぞれ近似される。

$$\sigma^2_{y|k} = \lambda_y \sigma^2_{y|k-1} + (1-\lambda_y) y_k y_k \qquad (5)$$

$$\sigma^2_{e|k} = \lambda_e \sigma^2_{e|k-1} + (1-\lambda_e) e_k e_k \qquad (6)$$

ここで、$0 \leq \lambda_y, \lambda_e < 1$ である。さらに、これらの値は常用対数を取った上で量子化される。これにより、広い範囲の $\sigma^2_{y|k}$ 及び $\sigma^2_{e|k}$ の値を扱うことが可能となる。このとき、状態は

$$s_k = \begin{bmatrix} q_1(\log_{10}(\sigma^2_{y|k})) \\ q_2(\log_{10}(\sigma^2_{e|k})) \end{bmatrix} \quad (7)$$

となる。ここで、q_i は量子化器である。

エージェントの行動集合 A は、適応フィルタの係数の更新を行うか休止するかの二通りとなる。

$$A = \{a_{\mathrm{S}}, a_{\mathrm{C}}\} \quad (8)$$

時刻 k においてエージェントへ与える報酬 r_k は、時刻 $k+1$ におけるタップ誤差 E^2_{k+1} から時刻 k におけるタップ誤差 E^2_k を引いたものを用いる。

$$r_k = -\Delta E^2_k \quad (9)$$

ここで、

$$\Delta E^2_k = E^2_{k+1} - E^2_k \quad (10)$$

$$E^2_k = \|\hat{\boldsymbol{h}}_k - \boldsymbol{h}_k\|^2 = \sum_{i=0}^{M} (\hat{h}_i - h_i)^2 \quad (11)$$

4.2 遅延 Q-learning

更新を継続したときにタップ誤差 E^2_k が急増大する場合を考える。増大開始点で $E^2_k = \epsilon$ とすると次の時点で $E^2_{k+1} >> \epsilon$ となり、報酬が $r_k << 0$ となる。このとき、従来の Q-learning では $E^2_k = \epsilon$ の状態 s^{ϵ}_k で更新継続 (a_{C}) すると大きなペナルティが付くこととなる。これはタップ誤差が小さいときに更新を休止する戦略をとる結果となる。よって、本研究では遷移後 (1 ステップ後) の状態にペナルティを付けるように改良した。

$$Q_{k+1}(s_{k+1}, a_k) = (1-\alpha) Q_k(s_{k+1}, a_k) + \alpha \left\{ r_k + \gamma \max_{a' \in A} Q_k(s_{k+1}, a') \right\} \quad (12)$$

これを遅延 Q-learning と呼ぶことにする。

4.3 学習と制御の流れ

遠端話者信号 u_k とエコーの観測値 y_k がエコーキャンセラ (EC) に入力され、疑似エコー $\hat{z}_k$ と残留エコー e_k が出力される。式 (7) を用いて u_k,y_k,e_k よりエージェントは状態 s_k を得る。次に Q 関数の値から得られる政策 π により s_k に対する行動 a を決定する。エージェントの行動が決定すると、その行動を行い、環境の状態が変化する。遷移後の状態を s' とし、環境の変化によって、タップ誤差より報酬 r を得る。以上のようにして得られた s, a, r, s' と現在の Q 関数の値を用いて、Q 関数の値をより最適な値へと更新していくことで制御戦略を学習する。

各エピソード毎にダブルトークの発生区間をずらしながら学習を行う。1 エピソードあたりの学習過程を図 1 に示す。

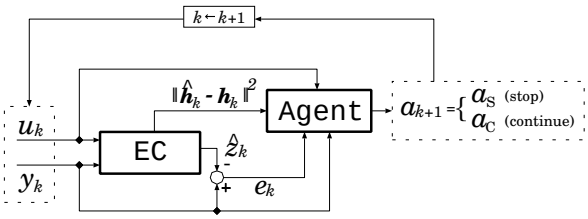


図 1: 1 エピソードあたりの学習過程

学習後、エージェントは学習によって得られた政策 π に基づいて EC の制御を行う。

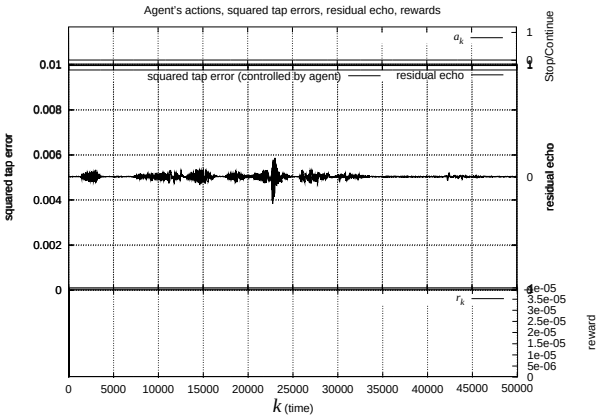
5 シミュレーション

適応フィルタアルゴリズムとして高速 H_∞ フィルタ ($\gamma_f = 42$)[3] を用い、式 (3) を用いて Q 関数を更新した場合と式 (12) を用いて Q 関数を更新した場合のエージェントによる戦略学習を行った。シミュレーション時に用いたパラメータは $\alpha = 0.1$, $\gamma = 0.1$, $T = 5.0 \times 10^{-6}$, $\lambda_u = 0.99$, $\lambda_y = 0.99$, $\lambda_e = 0.99$, 乱数の種 12345 とした。

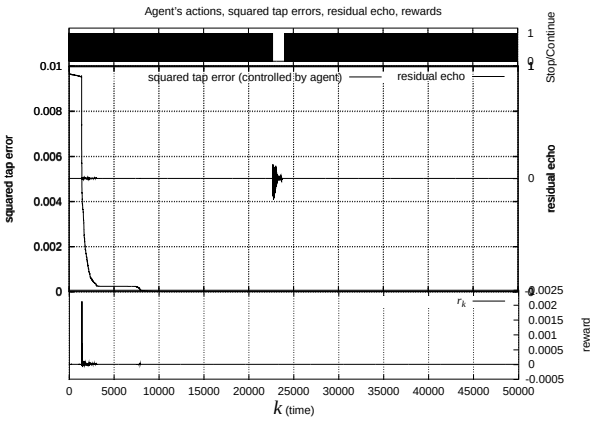
ダブルトークの長さは 1000 ステップとし、ダブルトークの発生位置を $k = 10000$ から 1 学習エピソード毎に 100 ステップずつずらしながら学習を行った。以下に 128 エピソードの学習を行った結果を示す。

図 2(a) に従来の Q-learning により学習を行うエージェントにより EC を制御した場合、図 2(b) に本研究で提案した遅延 Q-learning により学習を行うエージェントにより EC を制御した場合の行動 a_k、タップ誤差、残留エコー、エージェントに与えられた報酬 r_k をそれぞれ示す。

従来の Q-learning を用いた場合は、図 3 より、エコーの観測値及び残留エコーのパワーのレベルが小さい状態においてもフィルタ係数の更新を休止する確率が高くなっていることがわかる。本研究で提案した遅延 Q-learning の場合は、図 4 より、エコーの観測値及び残留エコーのパワーのレベルが比較的大



(a) Q-learning を用いたエージェントによる制御; 行動、タップ誤差、残留エコー、報酬



(b) 遅延 Q-learning を用いたエージェントによる制御; 行動、タップ誤差、残留エコー、報酬

図 2: 各手法による制御; 行動、タップ誤差、残留エコー、報酬

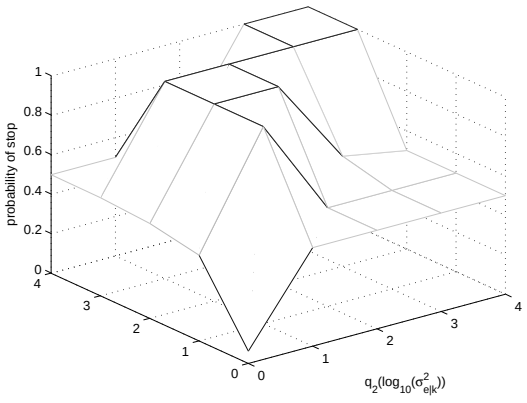


図 3: 従来の Q-learning における更新の休止行動を選択する確率

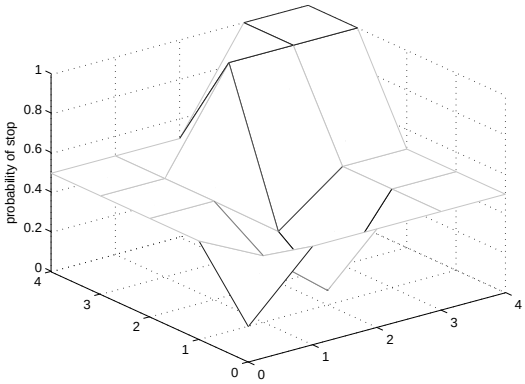


図 4: 遅延 Q-learning における更新の休止行動を選択する確率

きい時に係数更新を休止するような戦略を学習し、残留エコーを抑えていることがわかる。

6 おわりに

強化学習によるエコーキャンセラの制御を実装し、ダブルトークを含むエコー信号に対して良好なエコーキャンセル性能があることを示した。エージェントはタップ誤差を最小にするような制御戦略を自律的に学習する。そのため、エコーキャンセラの置かれる環境が変化した場合に閾値法において必要となる閾値の再設定を行うことなく、置かれた環境に対して適切な制御戦略を自動的に得ることが可能となる。

今後の課題としては、初期の収束性を改善するような戦略を学習するための報酬や環境の検討、エコー経路が変動する場合の制御などが挙げられる。

参考文献

[1] 西山 清 : 最適フィルタリング, 培風館, 2001.

[2] 長行 康男, 伊藤 実 : “2 体エージェント確率ゲームにおける他エージェントの政策推定を利用した強化学習法,” 電子情報通信学会論文誌, Vol.J86-D-I, No.11, pp.821-829, 2003.

[3] K. Nishiyama : “An H ∞ Optimization and Its Fast Algorithm for Time-Variant System Identification,” IEEE Transactions on Signal Processing, 52, 5, pp.1335-1342, 2004.