

強化学習に基づくオンライン逐次環境適応による雑音下音声認識

On-line Incremental Adaptation Based on Reinforcement Learning for Robust Speech Recognition

西田 昌史[†]
Masafumi Nishida

間宮 嘉誉[†]
Yoshitaka Mamiya

堀内 靖雄[†]
Yasuo Horiuchi

市川 熹[†]
Akira Ichikawa

1. はじめに

本研究では、音声対話システムといった実環境での利用を想定した頑健な音声認識の実現を目指している。これらのシステムでは、話者と雑音の両方の変動に対応する必要がある。そこで本研究では、様々な環境に対応することができるシステムの構築を目指して、話者や雑音といった事前知識なしに環境適応を行い、システムを繰り返し使用し続けることで認識精度が改善されるような音声認識の枠組みを実現することを目的とする。

従来、学習時と評価時での音響的なミスマッチを解消する方法が数多く研究されてきている。これらの研究には、信号処理とモデル補間の大きく分けて二つのアプローチがある。まず信号処理によるアプローチでは、近年 Missing feature theory [1] や Multi-band speech recognition [2] といった方法が提案されている。しかし、これらの方法ではある程度雑音の影響を抑えることができるが、話者の変動を適応することはできない。一方モデル補間によるアプローチでは、初期モデル（一般的にはクリーン音声により学習されたモデル）を環境に適応することで行われる。主に、Parallel Model Combination (PMC) [3] や Maximum A Posteriori (MAP) 推定 [4], Maximum Likelihood Linear Regression (MLLR) 法 [5] といった方法が提案されている。PMC 法は SN 比に基づいてクリーンな音声で学習した HMM と雑音により学習した HMM を合成する手法である。しかし、この方法ではあらかじめ雑音のデータが必要であり、ケプストラム平均値正規化 (CMN) といった他の補間法とを組み合わせるのが一般的に困難である。また、MAP 推定や MLLR 法は話者と環境の適応を行うことが可能であるが、教師なしで適応を行う際、認識結果のすべてに対して同じ重みで適応を行うため、ベースラインの認識精度が悪いと適応の効果が低くなってしまふ。

そこで、本研究では話者と雑音の両方の変動を適応する手法として、新たに強化学習に基づく逐次教師なし適応手法を提案する。強化学習 [6] は、教師データの代わりに出力の正しさを表す報酬に基づく学習手法であり、与えられた報酬に基づいて学習が徐々に進んでいく。

提案手法の有効性を示すために、雑音下音声認識評価ワーキンググループから提供されている AURORA-2J コーパス [7] を用いて、雑音環境下における音声認識実験を行った。

2. 強化学習

強化学習 [6] はロボット工学などの分野で用いられており、試行錯誤を通じて環境に適応する学習手法であり、教師データを必要としない代わりに出力の正しさを示す

報酬を必要とする。強化学習では、エージェントと呼ばれる動作の主体が、環境の状態を観測し、ある動作を行ったとき、環境から返ってくる反応によって報酬を受け取る。行った動作が正しいものであれば、エージェントは次も同じ動作を行うように確率を変動させる。行った行動が誤ったものであれば、エージェントは次には同じ動作を行わないように確率を変動させる。これを繰り返すことによって確率の変動が蓄積して行き、次第に学習が進んでいく。

従来の教師ありの学習とは異なり、強化学習では報酬を最大化するために、状況に基づいた動作選択の方法を学習する。エージェントはどの行動を取るべきか教えられず、どの行動を取ればより一層の報酬に結びつくかを見つけ出す必要がある。行動は直接的な報酬だけでなく、その次の状況に影響を与えるため、それ以降の全ての報酬にも影響を与える。これらの二つの特徴、つまり試行錯誤的な探索と遅延報酬は強化学習にとって重要な特徴となっている。そのため、教師あり学習のように外部から与えられる事例からの学習とは異なり、エージェントの行動に対して正確かつ典型的な望ましい反応が返ってくるとは限らない現実的な環境において、相互作用的に学習するのに適している。

これまで我々は、この強化学習の考え方を音韻認識に適用することを試みた [8]。これを踏まえて本研究では、システムが使用される環境について事前知識がなかったとしても、教師なしで環境に適応する手法について検討を行う。

3. 強化学習に基づく環境適応

強化学習に基づいて音響モデルを話者と雑音に適応する方法について述べる。本研究では、オンライン逐次適応による環境への適応を目的としており、音響モデルの適応は適応データ量の割合と事前分布を考慮した適応手法である MAP 推定により行う。提案手法では、強化学習における報酬として MAP 推定の事前分布の重みを制御することで、環境への適応を行うことにする。

MAP 推定は、観測データ x を用いて事後確率 $P(\theta|x)$ が最大となるモデルパラメータ θ を求める手法である。事後確率はベイズの定理により次のように表される。

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \quad (1)$$

ここで、観測データ x を用いてパラメータ θ を求める際、パラメータ θ は事前分布 $P(x)$ と無関係である。そこで、パラメータ θ は事前分布に従う確率変数であると仮定すると、次のように推定される。

$$\bar{\theta} = \arg \max_{\theta} P(\theta|x) = \arg \max_{\theta} P(x|\theta)P(\theta) \quad (2)$$

[†]千葉大学 大学院自然科学研究科

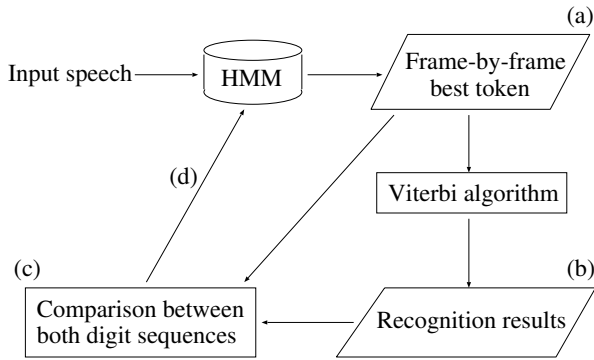


図 1: 提案手法の概念図

事前分布が正規分布 $N(\mu_0, \sigma_0)$ に従うと仮定すると、HMM の混合分布の平均ベクトルは MAP 推定により、次のように推定される。

$$\bar{\mu} = \frac{1}{\alpha + T} \left\{ \alpha \mu_0 + \sum_{i=1}^T x_i \right\} \quad (3)$$

ここで、 x_i は学習データの i フレーム目の特徴ベクトル、 T は学習データのフレーム数、 α は更新係数を表している。式 (3) により平均ベクトルが推定されると、推定された平均ベクトル $\bar{\mu}$ は推定前の平均ベクトル μ_0 よりも学習データの方向に近づくことになる。また、更新係数 α は事前分布の割合を制御することができ、逐次的に適応することで過去の適応の重みは減少することになる。

次に、強化学習に基づいた音響モデルの適応手法について述べる。本研究では、連続数字列の認識を行うため、入力と出力は数字列となっている。強化学習により HMM を適応するには、入力に対する出力の正しさを表す報酬が必要となる。そこで、本研究ではフレーム毎に最尤の HMM の状態をもつ数字列とビタビアルゴリズムにより得られた数字列を比較し、これらの比較結果に基づいて報酬を決定することにする。この報酬は、適応処理において式 (3) の更新係数 α に与えることにする。

提案手法の流れを図 1 に示す。また、詳細な処理手順を以下に示す。

1. 入力音声に対して各 HMM の尤度が計算され、フレーム毎に最尤の HMM の状態をもつ数字列が得られる (図 1 の (a)).
2. その後、ビタビアルゴリズムにより発話全体での最尤な数字列が得られる (図 1 の (b)).
3. フレーム毎に最尤の HMM の状態をもつ数字列と、ビタビアルゴリズムにより得られた数字列を比較する (図 1 の (c)).
4. 比較結果により得られた報酬に従って、MAP 推定に基づいて音響モデルを適応する (図 1 の (d)).

数字列の比較手法を図 2 に示す。図中の上段はフレーム毎に最尤の HMM の状態をもつ数字列、下段はビタビアルゴリズムにより得られた数字列を表している。フ

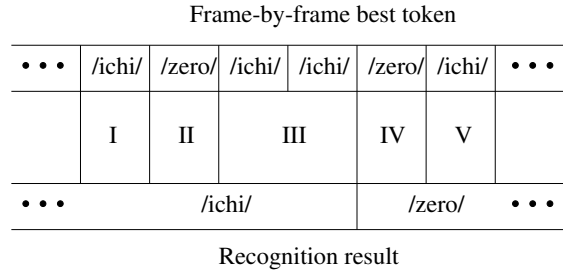


図 2: 数字列の比較方法

レーム毎に最尤の HMM の状態をもつ数字列は HMM からの直接の出力であり、ビタビアルゴリズムにより得られた数字列は前後のフレームを考慮した結果となっている。そこで、フレーム毎に最尤の HMM の状態をもつ数字列と、ビタビアルゴリズムにより得られた数字列が同じ場合、その音声区間の認識結果は信頼度が高いと考えられる。したがって、フレーム毎に最尤の HMM の状態をもつ数字列と、ビタビアルゴリズムにより得られた数字列が同じ場合 (図 2 中の I, III, IV), 更新係数 α を小さくして MAP 推定を行う。また、数字列を比較した結果が異なっている場合 (図 2 中の II, V), 更新係数 α を大きくして MAP 推定を行う。それぞれの場合における MAP 推定は、対応する数字列の HMM の各状態に属する全混合分布の平均ベクトルに対して行う。

このように、状況に応じた報酬を与えることで、話者と雑音の変動に対して逐次的に教師なしで適応することができると考えられる。

4. 音声認識実験

4.1 実験条件

本研究では、雑音環境下連続数字認識コーパスである AURORA-2J[7] を用いて、音声認識実験を行った。

音声データは、サンプリング周波数 16kHz、フレーム長 25msec、フレーム周期 10msec で分析されている。特徴量には、12 次 MFCC、パワーのそれぞれの 1 次、2 次回帰の計 39 次元のパラメータを用いた。各発話には、ケプストラム平均値正規化 (CMN) を行った。音響モデルは、16 状態、20 混合分布の HMM により各数字をモデル化した。初期モデルは、110 名の話者による 8440 発話のクリーンな音声により学習を行った。各発話には、1 から 7 桁の数字列が含まれている。

評価には、8 種類の雑音 (Subway, Babble, Car, Exhibition, Restaurant, Street, Airport, Station) を 3 種類の SN 比 (10, 15, 20 dB) で重畳させたデータのうち、50 名の話者による 500 発話のデータを用いた。これらのデータは、10 発話ごとに話者が異なっている。

4.2 実験結果と考察

各 SN 比に対する認識結果を表 1 に示す。表中の “Baseline” は適応を行わずに初期モデルでの結果、“MAP” は従来の MAP 適応を逐次的に教師なしで行ったときの結果、“Reinforcement learning” は強化学習に基づいて逐次的に教師なしで適応を行ったときの結果を示している。ここで、従来の MAP 適応の更新係数 α は、予備実験に

より 30 に設定した。また、強化学習においては、フレーム毎に最尤の HMM の状態をもつ数字列と、ビタビアルゴリズムにより得られた数字列が同じ場合は更新係数 α を 30, 数字列が異なる場合は 70 にそれぞれ予備実験により設定した。

表 1: 各 SN 比に対する認識結果

	Baseline	MAP	Reinforcement learning
20 dB	89.4	96.8	97.0
15 dB	65.8	90.1	91.7
10 dB	43.0	48.3	61.3

従来の MAP 適応は、ベースラインに比べて認識精度が改善されている。しかし、10 dB のときの認識精度の改善率が小さかった。これは、SN 比が小さくなれば認識誤りが増加するにもかかわらず、従来の MAP 適応ではすべての認識結果に対して同じ重みで適応を行っているため、正しい方向に適応が進まなかったからであると考えられる。

提案手法では、20 dB のときに 97.0 %, 15 dB のときに 91.7 %, 10 dB のときに 61.3 % の認識精度が得られ、すべての SN 比に対して従来手法よりも高い認識精度が得られた。特に、10 dB のときに従来の MAP 適応に比べて認識精度の改善率が高かった。このことから、認識結果の信頼度に基づいた最適な報酬により適応を行ったことが有効であると考えられる。

各雑音に対する全 SN 比の平均認識率を図 3 に示す。この結果から、提案手法によりほぼすべての雑音に対して認識精度の改善が得られた。

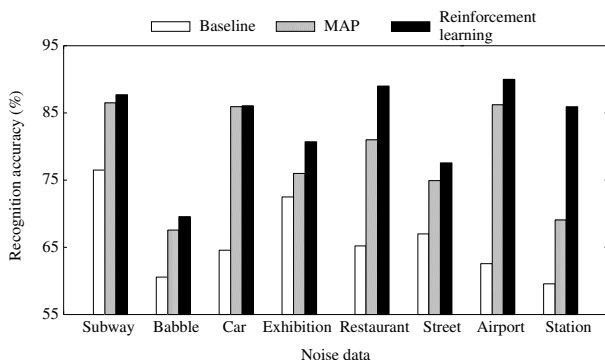


図 3: 各雑音に対する認識結果

次に、SN 比が 10 dB のときの全雑音で平均した 20 発話ごとの認識結果を図 4 に示す。この結果から、提案手法は徐々に認識精度が改善されており、他の手法に比べて認識精度の変動が小さいことがわかる。このことから、強化学習に基づいた逐次教師なし適応が話者と雑音の両方の変動を適応するのに、有効であると考えられる。

5. おわりに

本研究では、雑音に頑健な音声認識手法として、強化学習に基づくオンライン逐次適応手法を提案した。本手法は、話者や雑音といった事前知識を必要とせず、話者

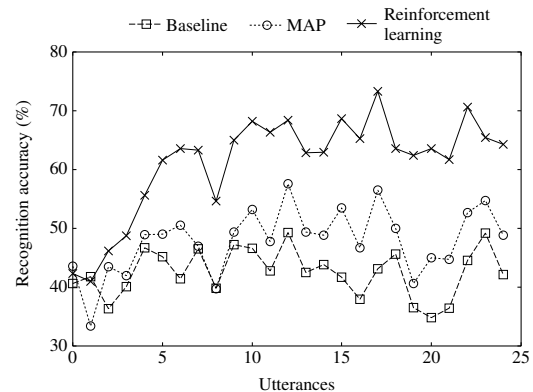


図 4: 20 発話毎の認識結果

と雑音の両方の変動に適応することができる。雑音環境下連続数字認識コーパスである AURORA-2J を用いて音声認識実験を行った結果、提案手法はすべての雑音に対して、従来手法に比べて高い認識精度を得ることができた。

今後は最適な報酬の推定法、提案手法にその他の雑音除去手法を組み合わせることなどについて検討を行う予定である。

謝辞

本研究では、IPSJ SIG-SLP 雑音下音声認識評価 WG の雑音下連続数字認識コーパス AURORA-2J を利用した。

参考文献

- [1] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study," Proc. EUROSPEECH, pp. 2407-2410, 1999.
- [2] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi stream adaptive evidence combination for noise robust ASR," Speech Communication, 34, pp. 25-40, 2000.
- [3] M. J. F. Gales, "Predictive model-based compensation schemes for robust speech recognition," Speech Communication, 25, pp. 49-74, 1998.
- [4] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, 1994.
- [5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," Computer Speech and Language, 9, pp. 171-186, 1995.
- [6] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.
- [7] S. Nakamura, K. Yamamoto, K. Takeda, S. Kuroiwa, and S. Kitaoka et al, "Data collection and evaluation of AURORA-2 Japanese corpus," Proc. ASRU, pp. 619-632, 2003.
- [8] 清水 智之, 堀内 靖雄, 市川 薫, "連続音韻認識への強化学習適応の検討," 人工知能学会 11 回大会論文集, 19-01, pp. 380-383, 1997.