社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

重みベクトルの適応的正則化に基づく発音推定

久保　慶伍†　サクティ　サクリアニ†　グラム　ニュービッグ†　戸田　智基†　中村　哲†

† 奈良先端科学技術大学院大学情報科学研究科　〒 630–0192　奈良県生駒市高山町 8916–5
E-mail: †{keigo-k,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

あらまし　音声認識や音声合成の重要な課題の一つである発音推定の分野において現在最も高精度な手法として，多値分類のオンライン識別学習法である MIRA（Margin Infused Relaxed Algorithm）に基づく構造学習が提案されている．これは精度良く発音を推定する一方，学習データを過学習する傾向にあり，Web 上の辞書などノイズを多く含んだ学習データでは性能の劣化が著しいと考えられる．そこで，我々は過学習に強い二値分類手法である重みベクトルの適応的正則化手法（AROW：Adaptive Regularization of Weight Vectors）を構造学習に拡張する方法を提案し，学習データにノイズを含む発音推定においてその手法を評価した結果，5.3%の誤り削減率を得ることができた．

キーワード　発音推定，未知語，識別学習，構造学習，重みベクトルの適応的正則化手法

Grapheme-to-phoneme Conversion based on Adaptive Regularization of Weight Vectors

Keigo KUBO†, Sakriani SAKTI†, Graham NEUBIG†, Tomoki TODA†, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology
8916–5 Takayama-cho, Ikoma, Nara 630–0192, Japan
E-mail: †{keigo-k,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

Abstract　The current state-of-the-art approach in grapheme-to-phoneme (g2p) conversion is structured learning based on the Margin Infused Relaxed Algorithm (MIRA). However, it is known that the aggressive weight update method of MIRA is prone to overfitting, even if the current example is an outlier or noisy. Adaptive Regularization of Weight Vectors (AROW) has been proposed to resolve this problem for binary classification. In this paper, we first apply AROW to g2p conversion which is structured learning problem. In an evaluation, our proposed approach achieves a 5.3% error reduction rate compared to MIRA implemented in DirecTL+ in terms of phoneme error rate.

Key words　g2p conversion, out-of-vocabulary word, online discriminative training, structured learning, AROW

1. はじめに

文字列の発音推定は，書記素列（Graphemes）から音素列（Phonemes）へと変換することから g2p（grapheme-to-phoneme）変換と呼ばれる（以後，発音推定を g2p 変換と書く）．この技術は未知語の発音を推定することに使われ，大規模音声認識システム [1] やテキスト音声合成システム [2] において重要な役割を果たす．現在まで，ルールに基づくアプローチ [3] や，最大エントロピー法 [4]，決定木 [5]，ニューラルネットワーク [6] といった統計的アプローチが試みられてきた．最近このタスクで用いられている手法として結合系列モデル [7], [8] と Margin Infused Relaxed Algorithm（MIRA）[9] に基づく構造学習手法が挙げられる．結合系列モデルは，書記素列と音素列の断片を合わせて一つの単位とした結合 N-gram を用いる生成モデルである．MIRA は，現在対象としているデータの正解クラスのスコアが誤りのクラスのスコアよりも十分な差で高くなるように特徴量の重みを学習する多値分類のオンライン識別学習手法である．MIRA は g2p 変換のようなクラスの候補数が極端に多い構造学習問題にも拡張されており，先行研究では g2p 変換のタスクにおいて結合系列モデルよりも低い単語誤り率を実現している [10], [11]．しかしながら，MIRA は，もし現在対象としているデータが外れ値または正解ラベルが間違っているデータ（以後，このようなデータをノイズデータと書く）であっても，それを正確に分類できるように特徴量の重みを大きく動かしてしまうため，過学習を引き起こす傾向がある．

最近，Web 上に存在する言語の専門家のクロスチェックなしで構築された膨大な発音のデータを g2p 変換の学習データとして用いることが提案されている [12]．このようなデータセットは不特定多数のユーザにより構築されているため，ノイズデータ（g2p 変換においてノイズデータとは誤った発音が付与された単語のことである）が多く含まれていると考えられ，実際に文献 [12] において，そのようなデータセットを用いた場合，辞

書構築のコストと時間を削減する代わりに，音声認識システムの性能が減少することが示されている．そのため，このようなノイズデータをトレーニングに用いても，ノイズデータを過学習せずに頑健に高い性能を保つことができるアプローチが近年必要となっている．

このようなノイズデータの問題を解決するために，二値分類において，重みベクトルの適応的正則化手法（AROW：Adaptive Regularization of Weight Vectors）[13] というオンライン識別学習が提案されている．以後，これを AROW と書く．現在対象としているデータを正しく分類できる特徴量の重みを求める MIRA とは異なり，AROW は現在のデータを正しく分類できることを保証しない代わりに，学習データを正しく分類できる方向へと特徴量の重みを少しずつ動かす．また，他のデータにおいて良く出現する特徴量の重みは，あまり出現しない特徴量の重みよりも動かさない．これにより AROW はノイズデータを正しく分類するために特徴量の重みを大きく動かすことを防ぎ，ノイズデータの過学習に対して頑健さを持つ．加えて AROW の更新規則は MIRA と比較して単純であり，より短時間で学習することができる．複数の二値分類タスクにおいて，AROW は，MIRA の二値分類手法と見なすことができる Passive-Aggressive (PA) アルゴリズム [14] を超える性能を示した．

そのため，我々は二値分類手法である AROW を構造学習に拡張し，それを構造学習問題である g2p 変換タスクへと初めて適用する．本報告では，まず第 2 節で従来手法である MIRA や提案手法である AROW の構造学習において用いられる線形分類器に基づく g2p 変換を説明し，第 3 節で従来手法である MIRA に基づく構造学習を説明する．第 4 節では二値分類手法の AROW とその前身の二値分類手法である Confidence Weighted Algorithm（CW）[15], [16] を説明し，第 5 節で提案手法である構造学習へと拡張した AROW を説明する．第 6 節では，結合系列モデルと MIRA に基づく構造学習を比較手法として，提案手法である AROW に基づく構造学習を g2p 変換タスクにより評価した結果を報告する．最後に第 7 節でまとめを述べる．

2. 線形分類器に基づく g2p 変換

まず最初に書記素列 $\boldsymbol{x}$ から音素列 $\boldsymbol{y}$ へと変換する g2p 変換を定義する．ある書記素列 $\boldsymbol{x}$ から正しい音素列 $\boldsymbol{y}$ を得るために，以下のように定義される線形分類器を用いる．

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \boldsymbol{w} \cdot \Phi(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

ここで $\boldsymbol{w}$ は分類器の特徴量の重みベクトルを意味しており，$\Phi(\boldsymbol{x}, \boldsymbol{y})$ は，$\boldsymbol{x}$ と $\boldsymbol{y}$ に出現するある結合 N-gram の頻度 [11] といった特徴量から構成される特徴量ベクトルを意味している．式（1）において，$\hat{\boldsymbol{y}}$ は動的計画法を用いることにより効率的に得ることができる．構造学習はこの線形分類器の枠組みにおいて，正しい音素列を推定することができる $\boldsymbol{w}$ を得るために用いられる．この後の節において，MIRA に基づく構造学習手法と AROW に基づく構造学習手法を説明する．

3. MIRA によるオンライン構造学習

本報告において，構造学習へと拡張されたオンライン識別学習をオンライン構造学習と定義する．g2p 変換における MIRA に基づくオンライン構造学習は Jiampojamarn ら [10] により提案された．i 番目のデータ $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ と $\boldsymbol{w}_{t-1} \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})$ により選ばれた N-best の仮説 $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ が与えられた時，現在の重みベクトル $\boldsymbol{w}_{t-1}$ を以下の制約付き最適化問題を解くことにより更新する．

$$\begin{aligned} &\min_{\Delta \boldsymbol{w}} \frac{1}{2} \|\Delta \boldsymbol{w}\|^2 \\ &\text{s.t.} \quad \forall n \\ &(\boldsymbol{w}_{t-1} + \Delta \boldsymbol{w}) \cdot \boldsymbol{u}_{in} \geqq d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n) \end{aligned} \tag{2}$$

ここで $\Delta \boldsymbol{w}$ は重みの更新量ベクトルであり，更新後の重みベクトル $\boldsymbol{w}_t$ は以下のように定義される．

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \Delta \boldsymbol{w} \tag{3}$$

$\boldsymbol{u}_{in}$ は正解の特徴量ベクトルと仮説の特徴量ベクトルの差ベクトル $(\Phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_n))$ として定義される．$d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n)$ は $\boldsymbol{y}_i$ を $\hat{\boldsymbol{y}}_n$ と推定した際の損失値である．g2p 変換において，損失値 $d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n)$ は $\boldsymbol{y}_i$ と $\hat{\boldsymbol{y}}_n$ 間の音素誤り率が用いられる．式（2）の制約式は正解のスコア $\boldsymbol{w}_t \cdot \Phi(\boldsymbol{x}_i, \boldsymbol{y}_i)$ が仮説のスコア $\boldsymbol{w}_t \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_n)$ よりも損失値以上高くなるよう $\boldsymbol{w}_t$ を制約することを意味する．式（2）の最適化問題をラグランジュの未定乗数法で解くと，式（2）の双対問題が以下のように得られる．

$$\begin{aligned} &\max_{\alpha_1, ..., \alpha_N} \frac{1}{2} \sum_n \sum_m \alpha_n \alpha_m \boldsymbol{u}_{in} \cdot \boldsymbol{u}_{im} \\ &+ \sum_n \alpha_n (d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n) - \boldsymbol{w}_{t-1} \cdot \boldsymbol{u}_{in}) \end{aligned} \tag{4}$$

ここで，$\alpha_1, ..., \alpha_N$ はラグランジュの未定乗数であり，最適化されるパラメータである．また，$\Delta \boldsymbol{w}$ は $\alpha_1, ..., \alpha_N$ を用いて以下のように定義できる．

$$\Delta \boldsymbol{w} = \sum_n \alpha_n \boldsymbol{u}_{in} \tag{5}$$

式（4）は $\alpha_1, ..., \alpha_N$ のパラメータを持つ 2 次計画問題である．これを 2 次計画問題の解法 [17] を用いて解き，最適な $\alpha_1, ..., \alpha_N$ を得て，式（3）と式（5）から更新後の重みベクトル $\boldsymbol{w}_t$ を求める．式（3）と式（5）から求まる $\boldsymbol{w}_t$ は，現在のデータ $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ の正解のスコアを各仮説 $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ のスコアよりもそれぞれの仮説が持つ損失値以上高くし，かつ更新前の重みベクトル $\boldsymbol{w}_{t-1}$ に最も近い重みベクトルである．

もし求めなければならないパラメータが多い場合，2 次計画問題は計算コストの観点から解くことが非常に難しくなる．MIRA に関して，求めなければならないパラメータは $\alpha_1, ..., \alpha_N$ であり，これは更新において用いられる仮説の数に等しい．もし，バッチ学習を考えた場合，求めなければならないパラメータの数は学習データ数に仮説数を掛けた数であり，パラメータ数が多くなる．それゆえ，MIRA では計算コストを削減するためにバッチ学習ではなくオンライン学習が用いられる．

Algorithm 1 Online structured learning based on MIRA

Input: Training dataset $\boldsymbol{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_{|\boldsymbol{D}|}, \boldsymbol{y}_{|\boldsymbol{D}|})\}$
Output:$\boldsymbol{w}$
$\boldsymbol{w} = \boldsymbol{0}$
repeat
  for $i = 1$ **to** $|\boldsymbol{D}|$ **do**
    Predict N-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ by $\boldsymbol{w} \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})$
    Update $\boldsymbol{w}$ by solving the constrained optimization problem of Eq.(2)
  end for
until Stop condition is met

MIRA に基づくオンライン構造学習の手続きを **Algorithm 1** に示す．文献 [10] において，N-best 仮説 $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ はフレーズ単位デコーダ [18] に基づくビームサーチにより近似的に推定される．

MIRA の欠点として過学習しやすいことが挙げられる．もし現在対象としているデータが外れ値またはノイズデータであっても，MIRA は必ずそれを正解ラベル通りに正しく分類できる $\boldsymbol{w}_t$ を求める．その際，それを実現するために他のデータの分類において重要な特徴量の重みを性能が劣化する方向へと大きく動かし過学習する可能性がある．これによりシステムの性能が劣化する恐れがある．この問題を解決するため，過学習に対して頑健性を持つ AROW に基づくオンライン構造学習を提案する．次の節では，提案手法である AROW に基づくオンライン構造学習を説明する前段階として，二値分類手法である AROW とその前身となった CW について説明する．

4. 二値分類手法 CW・AROW

CW と AROW は二値分類手法に関するオンライン識別学習アルゴリズムである．AROW は CW の改善手法として提案された．両手法は，平均 $\boldsymbol{\mu} \in \mathbb{R}^d$ と分散共分散行列 $\Sigma \in \mathbb{R}^{d \times d}$ をパラメータとしてもつ多次元正規分布 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ に重みベクトル $\boldsymbol{w}$ が従うと仮定する．ここで d はモデルにおける特徴量の数である．CW と AROW は重みベクトルの期待値 $\mathrm{E}[\boldsymbol{w}] = \boldsymbol{\mu}$ をクラスの推定時に重みベクトルとして用いる．また，分散共分散行列を考慮することにより，CW と AROW は各特徴量の重みの更新量を以下のように制御する．過去において何度も出現し，更新された特徴量の重みは現在の位置が最適である可能性（CW と AROW ではこれを信頼度という言葉を用いて説明している）が高いため，そのような特徴量の重みは各更新において極端に動かさない．その一方で，過去において稀にしか出現せず，あまり更新されなかった特徴量の重みは現在の位置が最適である可能性（信頼度）が低いため，そのような特徴量の重みは更新において大きく動かす．MIRA が持たないこの特性により，CW と AROW は外れ値やノイズデータが出現した際に，他のデータにも頻繁に出現する信頼度の高い特徴量の重みをシステムの性能が劣化する方向へと大きく動かすことを防ぐことができる．この節の残りでは，CW と AROW に関してそれぞれ説明する．

4.1 CW

i 番目のデータ $(\boldsymbol{x}_i, y_i)$ が与えられた時，CW は以下の制約付き最適化問題を解くことにより重みベクトルに関する更新された分布 $\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)$ を得る．

$$(\boldsymbol{\mu}_t, \Sigma_t) = \min_{\boldsymbol{\mu}_t, \Sigma_t} \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \quad \text{s.t.} \quad \Pr_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)}[y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) \geqq 0] \geqq \eta \qquad (6)$$

ここで $\mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$ は現在の重みベクトルに関する分布であり，$\mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}))$ は更新後の分布と現在の分布間のカルバック・ライブラ擬距離である．$\eta \in (0.5, 1]$ は $\boldsymbol{\mu}$ と Σ に関する更新量を制御するハイパーパラメータである．また，第 2 節，第 3 節では扱うデータが書記素列 $\boldsymbol{x}_i$ と音素列 $\boldsymbol{y}_i$ であったのに対し，ここでは二値分類のため入力ベクトル $\boldsymbol{x}_i$ と正解ラベルのスカラー値 $y_i \in \{-1, +1\}$ をデータとして扱っていることに注意する．式（6）において CW は，現在のデータ $(\boldsymbol{x}_i, y_i)$ が少なくとも確率 $\eta \in (0.5, 1]$ で正しく分類されるという制約を満たし，かつ前回の分布に最も近い分布を求める．CW の学習は現在のデータを必ず正しく分類するという制約を満たすために積極的に重みを動かし，すぐに収束することが実験 [16] により知られている．しかしながら MIRA のように，その積極的な学習は制約を必ず満たすために信頼度が高い重みでさえ大きく動かすため，過学習を引き起こしやすい．

4.2 AROW

MIRA と CW の問題を避けるために，AROW は正則化項として CW の制約を目的関数に置く．AROW により発見される重みの分布は現在のデータ $(\boldsymbol{x}_i, y_i)$ を正確に分類することを保証しない．しかしながら，学習データは分布が更新されるたびに正しく分類されるようになる．一方で，外れ値が現れた時でさえ AROW は信頼できる重みをシステムの性能が落ちる方向へと大きく動かさない．

AROW は以下に定義される制約なし最適化問題を解くことにより重みベクトルに関する更新された分布を得る．

$$(\boldsymbol{\mu}_t, \Sigma_t) = \min_{\boldsymbol{\mu}_t, \Sigma_t} \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) + \tfrac{1}{2r} \ell_{h^2}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_t) + \tfrac{1}{2r} \boldsymbol{x}_i^{\mathrm{T}} \Sigma_t \boldsymbol{x}_i \qquad (7)$$

ここで r は $\boldsymbol{\mu}$ と Σ に関する更新量を制御するハイパーパラメータであり，$r > 0$ と制約する．$\ell_{h^2}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_t)$ は以下のように定義される損失関数である．

$$\ell_{h^2}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_t) = (\max\{0, 1 - y_i(\boldsymbol{\mu}_t \cdot \boldsymbol{x}_i)\})^2 \qquad (8)$$

式（7）の最適化問題を解くことは，損失関数の値と出現した各特徴量の分散を減らしながら可能な限り前の分布に近い分布を求めることと同じである．複数の二値分類問題において AROW は CW と PA の性能を超えることが示された [13]．我々は次の節において AROW を構造学習へと拡張した手法を提案する．

5. AROW に基づくオンライン構造学習

i 番目のデータ $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ と n 番目の仮説 $\hat{\boldsymbol{y}}_n$ が与えられた時，我々の提案手法である AROW に基づくオンライン構造学習は

Algorithm 2 AROW に基づくオンライン構造学習

Input: Training dataset $\boldsymbol{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_{|\boldsymbol{D}|}, \boldsymbol{y}_{|\boldsymbol{D}|})\}$
Output:$\boldsymbol{\mu}$ as weight vector $\boldsymbol{w}$
$\boldsymbol{\mu} = \boldsymbol{0}, \Sigma = \boldsymbol{I}$
repeat
  for $i = 1$ **to** $|\boldsymbol{D}|$ **do**
    Predict N-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ by $\boldsymbol{\mu} \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})$
    for $n = 1$ **to** N **do**
      if $\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_n, \boldsymbol{\mu}) > 0$ **then**
        Update $\boldsymbol{\mu}$ and Σ by Eq.(11) and Eq.(13) respectively
      end if
    end for
  end for
until Stop condition is met

以下の目的関数を最小化する分布 $\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)$ を求める．

$$L(\boldsymbol{\mu}_t, \Sigma_t) = \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)||\mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) + \tfrac{1}{2r}\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_n, \boldsymbol{\mu}_t) + \tfrac{1}{2r}\boldsymbol{u}_{in}^{\mathrm{T}}\Sigma_t\boldsymbol{u}_{in} \quad (9)$$

ここで $\boldsymbol{u}_{in}$ は前述の通り $\Phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_n)$ として定義され，r もまた前述の通り $r > 0$ の制約を持つハイパーパラメータである．$\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_n, \boldsymbol{\mu}_t)$ は以下のように定義される損失関数である．

$$\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_n, \boldsymbol{\mu}_t) = (\max\{0, d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n) - \boldsymbol{\mu}_t \cdot \boldsymbol{u}_{in}\})^2 \quad (10)$$

式（9）を $\boldsymbol{\mu}_t$ で偏微分し，0 と置くことで，以下のように定義される AROW に基づくオンライン構造学習の $\boldsymbol{\mu}_t$ に関する更新式を得る．

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \frac{\max\{0, d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n) - \boldsymbol{\mu}_t \cdot \boldsymbol{u}_{in}\}}{\boldsymbol{u}_{in}^{\mathrm{T}}\Sigma_{t-1}\boldsymbol{u}_{in} + r}\Sigma_{t-1}\boldsymbol{u}_{in} \quad (11)$$

g2p 変換における特徴の数は巨大であるため，それらの共分散関係を扱うことは困難である．そのため，我々は Σ_t を対角行列であると仮定する．式（9）の目的関数を Σ_t の p 番目の対角行列の要素で偏微分し，0 と置くと，以下のように Σ_t に関する更新式を得る．

$$\frac{\partial}{\partial(\Sigma_t)_{p,p}}L(\boldsymbol{\mu}_t, \Sigma_t) = \frac{1}{2}\left(\frac{1}{(\Sigma_{t-1})_{p,p}} - \frac{1}{(\Sigma_t)_{p,p}} + \frac{(\boldsymbol{u}_{in})_p^2}{r}\right) = 0 \quad (12)$$

ここで $(\boldsymbol{u}_{in})_p$ は $\boldsymbol{u}_{in}$ における p 番目の特徴量を意味する．上記の式を $(\Sigma_t)_{p,p}$ に関する式に以下のように変形する．

$$(\Sigma_t)_{p,p} = \frac{r(\Sigma_{t-1})_{p,p}}{r + (\boldsymbol{u}_{in})_p^2(\Sigma_{t-1})_{p,p}} \quad (13)$$

$p = 1, ..., d$ の各対角要素 $(\Sigma_t)_{p,p}$ は式（13）により更新する．また，$\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_n, \boldsymbol{\mu}_{t-1})$ が 0 の時，$\boldsymbol{\mu}_{t-1}$ と Σ_{t-1} は更新しない．

AROW に基づくオンライン構造学習の手続きを **Algorithm 2** に示す．$\boldsymbol{\mu}$ と Σ は 0 ベクトルと単位行列により各々初期化される．$(\Sigma_0)_{p,p} = 1$ と $r > 0$，式（13）から，$(\Sigma_{t-1})_{p,p} \geqq (\Sigma_t)_{p,p}$ が全ての t において成り立つ．$(\Sigma_t)_{p,p} = 0$ の時，$\boldsymbol{\mu}$ の p 番目の特徴量の重みは固定される．故に **Algorithm 2** の収束は保証される．**Algorithm 2** において，N-best 仮説 $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ は文献 [10] と同様にフレーズ単位デコーダ [18] に基づくビームサーチにより近似的に推定される．**Algorithm 2** における $\boldsymbol{\mu}$ と Σ に関する逐次的な更新処理は多値分類 CW [19] における逐次更新に似ている．その違いは，CW の逐次更新が各仮説に対して制約付き最適化問題を解くことに対し，提案手法は制約なし最適化問題を解くことである．また，**Algorithm 2** は MIRA と同様にオンライン構造学習であるが，提案手法は 2 次計画問題を解く必要がないため簡単にバッチ学習を行うことができる．

表 1　g2p 変換タスクの評価実験で使用するデータセット

Dataset	g/p	Vocabulary size			
		Train (Noisy)	Dev	Test	K-fold
NETtalk	26/50	17595 (0)	1000	1000	10
Noisy NETtalk	26/50	17595 (1760)	1000	1000	10

6. 評価実験

提案手法である AROW に基づくオンライン構造学習を g2p 変換タスクにおいて評価する．表 1 はこの実験において用いたデータセットのデータ名（Dataset），出現する書記素と音素の種類数（g/p: g が書記素，p が音素の種類数に対応），学習データ数（Train），開発データ数（Dev），テストデータ数（Test），交差検定の回数（K-fold）を示している．データセットの NETtalk は，Pascal Letter-to-Phoneme Conversion Challenge(注1) から得た英語の辞書である．我々は，学習データから開発データをランダムに選んだことを除いて，書記素列が 1 文字で構成されるといった例外データの取り除き方，学習データ数（＋開発データ数）とテストデータ数の割合に関して，文献 [8] の NETtalk に関する実験の再現を試みた．AROW に基づくオンライン構造学習がノイズデータに対して頑健であることを確かめるため，我々は学習データの 10%の書記素列に対して辞書内の音素列をランダムに選択し，それに付与することでノイズデータを人工的に作り出し，新しく Noisy NETtalk データセットを作成した．Noisy NETtalk において，ノイズデータに頑健性を持たない手法の推定性能は，ノイズデータを過学習することにより劣化すると考えられる．表 1 内の Noisy は人工的に作りだしたノイズデータの数を示している．Noisy NETtalk は 17595 個の語彙のうち，1760 個のノイズデータを含んでいる．また，開発データ（Dev）は，ハイパーパラメータなどといった学習により決定できないパラメータを決定するためのデータ数を意味している．

比較手法の g2p 変換ツールとして，Sequitur(注2) と DirecTL+(注3) を用いた．Sequitur は書記素列と音素列に関する結合 N-gram の生成モデルである結合系列モデルが実装され

(注1)：http://pascallin.ecs.soton.ac.uk/Challenges/PRONALSYL/Datasets
(注2)：http://sequitur.info/
(注3)：http://code.google.com/p/directl-p/

表 2　各手法において設定が必要なパラメータは開発セットを用いて最適化された．太字は 10 回の交差検定中，一度でもパラメータとして用いられた値である．

	Sequitur	DirecTL+	This work
joint n-gram	**5**,**6**,**7**,8,**9**,10	Follow Sequitur	Follow Sequitur
context window	-	**4**,**5**,**6**	Follow DirecTL+
n-best hypotheses	-	1,3,**5**	Follow DirecTL+
hyperparameter r	-	-	**500**,**1000**,**1500**
beam width	-	**150**	**150**

表 3　NETtalk における評価実験の結果．PER と WER は各々音素誤り率と単語誤り率を意味する．Time は各手法の学習時間である．“±” は 90% 信頼区間を示している．

	PER(%)	WER(%)	Time(hr.)
Sequitur	7.63%±0.24	31.54%±0.80	**1.1h±0.3**
DirecTL+	**6.75%±0.22**	**28.15%±0.76**	8.6h±1.5
This work	**6.75%±0.20**	**28.56%±0.62**	4.7h±1.0

ている．DirecTL+ は MIRA に基づくオンライン構造学習が実装されている．提案手法と DirecTL+ は文献 [11] に従い，文脈特徴量（Context features），連鎖特徴量（Chain features），結合 N-gram 特徴量（Joint n-gram features）を用いている．文献 [11] の遷移特徴量（transition features）は NETtalk において性能の劣化が見られたため用いなかった．書記素列と音素列の最小単位を決めるアライメントに関して，我々は mpaligner(注4) に実装されている文献 [20] の制約なし多対多アライメント手法を用いた．提案手法と MIRA の損失関数は音素誤り率を用いた．文脈窓サイズと結合 N-gram サイズ，ハイパーパラメータ r，学習時における N-best 仮説，ビームサーチのビーム幅，学習の繰り返し回数は開発データにおける音素誤り率が最小になるように決定した．表 2 はそれらの詳細を示している．また，この実験は CPU に Intel Xeon E5649 2.53GHz を持つクラスターマシンで行った．

表 3 は NETtalk における評価実験の結果を示している．提案手法は Sequitur の音素誤り率と単語誤り率を十分な差で改善している．DirecTL+ と比較すると，音素誤り率と単語誤り率においては十分な差が見られなかった一方で，学習時間の観点から見ると，提案手法の学習速度は DirecTL+ よりも速かった．これは提案手法の更新は N-best に含まれている各仮説に対してたった一回更新式を解くだけなのに対し，MIRA の更新は 2 次計画問題を解く手法により式（2）の制約を満たす $\boldsymbol{w}$ を繰り返し探索するからである．この結果は提案手法が MIRA に基づくオンライン構造学習よりも，巨大なデータの学習により適していることを示している．

表 4 は Noisy NETtalk の評価実験の結果を示している．表 4 から，ノイズデータによる提案手法の性能劣化は DirecTL+ の劣化より少ないことが分かる．提案手法と DirecTL+ の音素誤り率における差は有意水準 0.05 の t-検定において有意な差である．この結果は AROW に基づくオンライン構造学習が，二値分類の場合と同様に，MIRA の過学習問題を解決していることを示している．

(注4)：http://sourceforge.jp/projects/mpaligner/

表 4　Noisy NETtalk における評価実験の結果

	PER(%)	WER(%)	Time(hr.)
Sequitur	**9.78%±0.23**	34.01%±0.85	**3.3h±1.0**
DirecTL+	10.33%±0.27	**33.52%±0.46**	100.5h±12.1
This work	**9.79%±0.45**	**33.02%±0.95**	78.1h±15.9

また NETtalk と比べて，Noisy NETtalk の識別学習の学習時間が大幅に長くなっていることが分かる．これは Noisy NETtalk において含まれている人工的なノイズデータが書記素列と音素列のアライメントステップにおいて，新しい，誤った書記素列と音素列の対応付けを生み出すことが原因である．その新しい，誤った対応付けが推定可能な音素列の仮説数を増加させ，MIRA と AROW に基づく識別学習の N-best 仮説の推定時間を大きく長引かせるからである．その対応付けは Sequitur において実装されている結合系列モデルにおけるバックオフスムージングの計算時間にも悪影響を与えるが，それは識別学習の問題と比べて深刻な問題ではない．識別学習における計算時間の問題は文献 [16] において提案されているように分散学習により解決するか，ビームサーチのビーム幅を小さくすることによりある程度解決可能である．

7. ま と め

我々は AROW をオンライン構造学習へと拡張し，g2p 変換タスクにおいて評価した．評価実験において，提案手法は音素誤り率と単語誤り率において，DirecTL+ により実装された MIRA に基づくオンライン構造学習と同等の性能を達成した．また，ノイズデータを含むデータセットにおいて，提案手法は MIRA の音素誤り率を改善した．加えて，提案手法の学習速度は DirecTL+ よりも速かった．この結果から提案手法はより大きなデータセットから学習すること，またノイズデータを含んでいるデータセットから学習することに関して，MIRA に基づくオンライン構造学習よりも適していると考えられる．

今後の課題として，より大きな英語のデータセットや他の言語のデータセットを使って，提案手法を評価することが挙げられる．また，提案手法の性能を改善するために，メモリの制限内で Σ における 2 つの特徴量間の共分散関係を近似的に扱う手法を考えることが挙げられる．

謝辞　本研究の一部は，JSPS 科研費 24240032 および（独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」の助成を受けたものである。

文　献

[1] L.R. Bahl, S. Das, P.V. Desouza, M. Epstein, R.L. Mercer, B. Merialdo, D. Nahamoo, M.A. Picheny, and J. Powell, “Automatic phonetic baseform determination,” Proc. ICASSP, pp.173–176, IEEE, 1991.

[2] J. Schroeter, A. Conkie, A. Syrdal, M. Beutnagel, M. Jilka, V. Strom, Y.J. Kim, H.G. Kang, and D. Kapilow, "A perspective on the next challenges for TTS research," 2002.
[3] R.M. Kaplan and M. Kay, "Regular models of phonological rule systems," Computational linguistics, vol.20, pp.331–378, 1994.
[4] S.F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," Proc. EUROSPEECH, pp.2033–2036, 2003.
[5] W. Daelemans and A. van denBosch, "Language-independent data-oriented grapheme-to-phoneme conversion," Progress in Speech Processing, pp.77–89, Springer-Verlag, 1997.
[6] T.J. Sejnowski and C.R. Rosenberg, "Parallel networks that learn to pronounce English text," Complex Syst., vol.1, pp.145–168, 1987.
[7] S. Deligne and F. Bimbot, "Inference of variable-length linguistic and acoustic units by multigrams," Speech Communication, vol.23, no.3, pp.223–241, 1997.
[8] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," Speech Communication, vol.50, no.5, pp.434–451, 2008.
[9] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," Journal of Machine Learning Research, vol.3, pp.951–991, 2003.
[10] S. Jiampojamarn and G. Kondrak, "Online discriminative training for grapheme-to-phoneme conversion," Proc. INTERSPEECH, pp.1303–1306, 2009.
[11] S. Jiampojamarn, C. Cherry, and G. Kondrak, "Integrating joint n-gram features into a discriminative training framework," Proc. NAACL-HLT, pp.697–700, 2010.
[12] T. Schlippe, S. Ochs, and T. Schultz, "Grapheme-to-phoneme model generation for Indo-European languages," Proc. ICASSP, pp.4801–4804, 2012.
[13] K. Crammer, A. Kulesza, and M. Dredze, "Adaptive regularization of weight vectors," Advances In Neural Information Processing Systems, vol.23, pp.414–422, 2009.
[14] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," Journal of Machine Learning Research, vol.7, pp.551–585, 2006.
[15] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," International Conference On Machine Learning (ICML), pp.264–271, 2008.
[16] K. Crammer, M. Dredze, and F. Pereira, "Confidence-weighted linear classification for text categorization," Journal of Machine Learning Research, vol.13, pp.1891–1926, 2012.
[17] R.J. Vanderbei, "Loqo: An interior point code for quadratic programming," Optimization methods and software, vol.11, no.1-4, pp.451–484, 1999.
[18] R. Zens and H. Ney., "Improvements in phrase-based statistical machine translation," Proc. NAACL HLT, pp.257–264, 2004.
[19] K. Crammer, M. Dredze, and A. Kulesza, "Multi-class confidence weighted algorithms," Empirical Methods in Natural Language Processing (EMNLP), vol.2, pp.496–504, 2009.
[20] K. Kubo, H. Kawanami, H. Saruwatari, and K. Shikano, "Unconstrained many-to-many alignment for automatic pronunciation annotation," 2011.