重みベクトルの適応的正則化手法の発音推定における評価*

☆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 (奈良先端大)

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やテキスト音声合成システムにおいて重要な役割を果たす．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 [1, 2] と Margin Infused Relaxed Algorithm（MIRA）[3] 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 [4, 5]．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．

このような過学習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）[6] 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，過学習に対して頑健さを持つ．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 [7] 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した [8]．本報告では様々なデータセットを用いた g2p 変換タスクによる AROW に基づく構造学習の評価実験について報告する．

2 線形分類器に基づく g2p 変換

まず最初に線形分類器に基づく g2p 変換について定義する．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 $\boldsymbol{y}$ に出現する結合 N-gram の頻度 [5] 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．

3 AROW に基づくオンライン構造学習

ここでは線形分類器において用いられる重みベクトル $\boldsymbol{w}$ を得るための AROW に基づくオンライン構造学習について説明する．AROW は 2 値分類のオンライン識別学習として提案された．AROW は重みベクトルが多次元ガウシアン分布 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ に従うと仮定することで，各重みに関する更新量を以下のように制御する．頻繁に出現した特徴量（他の多くのデータに出てきた特徴量）の重みは現在の位置に信頼性があるので大きく動かさない．逆に，今まであまり出現しなかった特徴量の重みは現在の位置に信頼性がないため大きく動かす．これにより，AROW はノイズデータを学習しても，他のデータに影響を与える重要な特徴量の重みをシステムの性能が落ちる方向へと大きく動かすことを防ぐ．これが，従来手法の MIRA と比べて，AROW が過学習に頑健な理由である．また推定の間，AROW は重みベクトルの期待値 $\mathrm{E}[\boldsymbol{w}_t] = \boldsymbol{\mu}_t$ を線形分類器の重みベクトルとして用いる．

AROW を構造学習へと拡張した我々の提案手法は，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，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．

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

* Evaluation of Adaptive Regularization of Weight Vectors on Grapheme-to-Phoneme Conversion. by Kubo Keigo, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura (Nara Institute of Science and Technology)

ここで $N(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$ は現在の重みベクトルの分布，$\boldsymbol{u}_{in}$ は正解と仮説の特徴量ベクトルの差ベクトル $\Phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_n)$，$r > 0$ はパラメータの更新量を制御するための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．

$$\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_n, \boldsymbol{\mu}_t) = (\max\{0, d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n) - \boldsymbol{\mu}_t \cdot \boldsymbol{u}_{in}\})^2 \quad (3)$$

ここで $d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n)$ は損失値であり，g2p 変換では音素誤り率などが用いられる．

式（2）を $\boldsymbol{\mu}_t$ で偏微分し，0 と置くことで，以下に定義される 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．

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \frac{\max\{0, d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_n) - \boldsymbol{\mu}_t \cdot \boldsymbol{u}_{in}\}}{\boldsymbol{u}_{in}^{\mathrm{T}} \Sigma_{t-1} \boldsymbol{u}_{in} + r} \Sigma_{t-1} \boldsymbol{u}_{in} \quad (4)$$

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（2）の目的関数を Σ_t の p 番目の対角行列の要素 $(\Sigma_t)_{p,p}$ 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．

$$\frac{\partial}{\partial (\Sigma_t)_{p,p}} L(\boldsymbol{\mu}_t, \Sigma_t) = \frac{1}{2}\left(\frac{1}{(\Sigma_{t-1})_{p,p}} - \frac{1}{(\Sigma_t)_{p,p}} + \frac{(\boldsymbol{u}_{in})_p^2}{r} \right) = 0 \quad (5)$$

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．

$$(\Sigma_t)_{p,p} = \frac{r(\Sigma_{t-1})_{p,p}}{r + (\boldsymbol{u}_{in})_p^2 (\Sigma_{t-1})_{p,p}} \quad (6)$$

$p = 1, ..., d$ の各対角要素 $(\Sigma_t)_{p,p}$ は式（6）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 1** 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（6）から，$(\Sigma_{t-1})_{p,p} \geq (\Sigma_t)_{p,p}$ 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 1** の収束は保証される．**Algorithm 1** において，N-best 仮説 $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N$ は文献 [4] と同様にフレーズ単位デコーダ [9] に基づくビームサーチにより近似的に推定される．

Algorithm 1 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.(4) and Eq.(6) respectively
      end if
    end for
  end for
until Stop condition is met

4 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（English），Brulex（French），Beep（English）は，Pascal Letter-to-Phoneme Conversion Challenge[1] から得た単語の発音辞書である．また，CMUdict（English）[2]，Celex（English，German，Dutch）[3] もまた単語の発音辞書である．文献 [2] の実験で用いられているデータセット（NETtalk，Brulex，Beep，CMUdict）において，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 [2] の実験の再現を試みた．また，AROW に基づくオンライン構造学習が過学習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，過学習に対して頑健性を持たない手法の性能は，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 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[4] と Di-

[1] http://pascallin.ecs.soton.ac.uk/Challenges/PRONALSYL/Datasets
[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[3] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC96L14
[4] http://sequitur.info/

Table 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	1000	1000	10
Noisy NETtalk	26/50	17595 (1760)	1000	1000	10
Brulex	40/39	23353	1373	2747	1
CMUdict	27/39	100886	5941	12000	1
Beep	26/44	169823	8938	19862	1
CELEX English	26/53	39995	15000	5000	1
CELEX German	30/59	206807	25851	77552	1
CELEX Dutch	41/44	196587	24573	73721	1

Table 2　各手法において設定が必要な特徴量とパラメータ.

	Sequitur	DirecTL+	Proposed
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**

recTL+[5] を用いた．Sequitur は書記素列と音素列の結合 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 [5] 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．表 2 はそれらの特徴量や設定が必要なパラメータの詳細を示している．文献 [5] 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．NETtalk において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，各交差検定において開発データの音素誤り率が最小になるように決定した．太字は 10 回の NETtalk の交差検定中，一度でも用いられた値を示している．また，NETtalk 以外の他のデータセットに関して，特徴量とパラメータは NETtalk の実験において多く採用された値を使用し，学習回数とハイパーパラメータ r は NETtalk と同じ方法で決定した．他のデータセットで用いた特徴量とパラメータは文脈窓サイズが 6，結合 N-gram サイズが 5，学習時における N-best 仮説が 5 である．また，NETtalk の実験においてビーム幅を 50 にしても性能の劣化が見られなかったため，他のデータセットでは探索のビーム幅を 50 にした．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[6] に実装されている文献 [10] 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 の損失値は音素誤り率を用いた．

表 3 は評価実験の結果を示している．PER と WER は各々音素誤り率と単語誤り率を意味し，“±” は 90% 信頼区間を示している．NETtalk，CELEX の German と Dutch を除いて，提案手法は Sequitur と DirecTL+ の音素誤り率と単語誤り率を改善している．NETtalk において，提案手法は DirecTL+ と同等の性能であるのに対し，Noisy NETtalk では提案手法が DirecTL+ を上回る性能を示している．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．そのため，他のデータセットにおいても性能の改善が見られたと考えられる．

Sequitur が他の手法と比べて CELEX の German と Dutch において性能を改善したのは，極端に低い誤り率と Sequitur のバッチ学習によるものだと考えられる．提案手法や MIRA で採用されるオンライン学習では，個々のデータを使って重みベクトルを更新するたびに，過去に学習したデータの識別精度が薄れていく．そのため，極端に低い誤り率を持つ CELEX の German と Dutch において，提案手法や MIRA は過去に学習したデータの識別精度が薄れる影響により，全てのデータを同時に学習する Sequitur のバッチ学習よりも高い性能を示すことができなかったと考えられる．

また，NETtalk 以外のデータセットにおいて，Sequitur よりも DirecTL+ の性能が劣っていた．文献 [4, 5] では Sequitur に実装された結合系列モデルよりも DirecTL+ に実装された MIRA の方が高い性能を示している．これは今回ランダムに選択した学習データやテストデータの影響によるものだと考えられる．より正確な評価のため，NETtalk のようにクロスバリデーションによる評価が必要だと考えられる．

5　まとめ

我々は AROW をオンライン構造学習へと拡張し，様々なデータセットを用いた 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 MIRA に基

[5] http://code.google.com/p/directl-p/

[6] http://sourceforge.jp/projects/mpaligner/

Table 3　g2p 変換タスクにおける評価実験の結果.

Dataset	Measure	Sequitur	DirecTL+	Proposed
NETtlak	PER	7.63%±0.24	**6.75%±0.22**	**6.75%±0.20**
	WER	31.54%±0.80	**28.15%±0.76**	**28.56%±0.62**
Noisy	PER	**9.78%±0.23**	10.33%±0.27	**9.79%±0.45**
NETtlak	WER	34.01%±0.85	**33.52%±0.46**	**33.02%±0.95**
Brulex	PER	1.30%	1.97%	**1.12%**
	WER	6.70%	8.26%	**5.75%**
CMUdict	PER	6.80%	7.25%	**6.09%**
	WER	28.83%	28.99%	**26.38%**
Beep	PER	2.85%	4.45%	**2.22%**
	WER	15.37%	19.58%	**12.00%**
CELEX	PER	2.83%	4.23%	**2.51%**
English	WER	13.33%	16.76%	**11.83%**
CELEX	PER	**0.08%**	0.31%	0.13%
German	WER	**0.67%**	1.86%	1.04%
CELEX	PER	**0.08%**	1.05%	0.27%
Dutch	WER	**0.67%**	5.28%	1.85%

づくオンライン構造学習よりも過学習に頑健で，g2p 変換の性能を改善することを示した．今後の課題として，NETtalk 以外のデータセットをクロスバリデーションにより評価することや，提案手法の性能をさらに改善するために，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。

参考文献

[1] 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.

[2] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” Speech Communication, vol.50, no.5, pp.434–451, 2008.

[3] K. Crammer and Y. Singer, “Ultraconservative online algorithms for multiclass problems,” Journal of Machine Learning Research, vol.3, pp.951–991, 2003.

[4] S. Jiampojamarn and G. Kondrak, “Online discriminative training for grapheme-to-phoneme conversion,” Proc. INTERSPEECH, pp.1303–1306, 2009.

[5] S. Jiampojamarn, C. Cherry, and G. Kondrak, “Integrating joint n-gram features into a discriminative training framework,” Proc. NAACL-HLT, pp.697–700, 2010.

[6] K. Crammer, A. Kulesza, and M. Dredze, “Adaptive regularization of weight vectors,” Advances In Neural Information Processing Systems, vol.23, pp.414–422, 2009.

[7] 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.

[8] 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，“重みベクトルの適応的正則化に基づく発音推定,” 信学技報，第 113 巻，pp.25–30，2013.

[9] R. Zens and H. Ney., “Improvements in phrase-based statistical machine translation,” Proc. NAACL HLT, pp.257–264, 2004.

[10] K. Kubo, H. Kawanami, H. Saruwatari, and K. Shikano, “Unconstrained many-to-many alignment for automatic pronunciation annotation,” Proc. APSIPA, pp.1–4, 2011.