Vol.2018-SLP-125 No.5
2018/12/10

マルチゲートGRUユニットを用いた2D-RNNによるEnd-to-End始終端フリー単語検出

田中 智宏[1]　篠崎 隆宏[1]

概要：単語境界が未知である条件で単語検出を行う始終端フリー DP マッチングを 2D-RNN を用いて End-to-End 型のニューラルネットとして実現する手法を提案する．従来 2D-RNN を用いた DP マッチングはテキスト処理において提案されていたが，本研究では連続 DP に拡張した上で，音声に適用する．さらに，正例と負例のバランスが大きく異なったタスクにおいて有効な評価尺度である F 値を学習時の目的関数として用いる手法を提案する．提案法はニューラルネットとしての拡張性を備えるとともに，単語検索における一般的な評価尺度を直接目的関数として利用することで，単語検出性能を改善する効果が期待できる．WSJ コーパスを用いた実験において，従来法である連続 DP マッチングおよび LSTM を用いた単語埋め込み手法と比較して単語検出性能が向上することを示す．

1. はじめに

音声単語検出タスクに対し，従来は動的計画法を用いたマッチング手法（連続 DP マッチング） [1] が使われてきた．我々はこれを 2D-RNN [2] を用いて一般化することで性能向上を図る．また，単語埋め込みを用いた手法の研究が近年盛んに行われている [3], [4], [5], [6], [7]．このうち，単語境界を仮定していない手法 [8] をもとにフレームワイズな出力を行うよう変更したモデルと提案法を比較し，2D-RNN を用いた提案法の方が単語検出性能に優れていることを示す．さらに，正例と負例の数が大きく異なるタスクにおいて有効な評価尺度である F 値を直接に学習時の目的関数として用いる手法を提案し，その有効性を示す．

2. 提案モデル

ニューラルネットによる End-to-End 始終端フリー単語検出器として我々が提案するモデルを図 2 に示す．検索対象の音声 $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_T\}$ とキーワード音声 $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_M\}$ を入力とし，キーワード終端である予測事後確率 $\mathbf{y} = \{y_1, y_2, \cdots, y_T\}$ を出力とする．この構造は連続 DP をもとに各ユニットをニューラルネット化することで連続 DP を一般化したものである．つまり，位置 (t,m) のユニットの計算を式 1 とおくと，f が式 2，$dist$ が式 3 で定義される場合は連続 DP であるが，我々はそれぞれ図 3，4 に示すニューラルネットを代わりに用いることで一般化する．

$$h_{t,m} = f\left(h_{t-1,m}, h_{t-1,m-1}, h_{t,m-1}, dist\left(\mathbf{q}_m, \mathbf{s}_t\right)\right). \quad (1)$$

$$f\left(h1, h2, h3, d\right) = min\left(h1, h2, h3\right) + d, \quad (2)$$

$$dist\left(\mathbf{a}, \mathbf{b}\right) = \left(\mathbf{a} - \mathbf{b}\right)^{\mathrm{T}}\left(\mathbf{a} - \mathbf{b}\right). \quad (3)$$

GRU を用いたこのような構造は 2D-GRU といわれ，これを用いた連続 DP マッチングはすでにテキスト処理 [9] やパーキンソン病患者の症状マッチング [10] などで用いられているが，音声単語検索の分野ではまだ試されていない．

3. ソフト F 値目的関数

正例と負例の数が大きく異なるデータを用いて学習するための目的関数として，式 4 に示すソフト F 値を用いることを提案する．ソフト F 値は再現率 R(式 5) や適合率 P(式 6) に基づいた定義自体は F 値と同じである．違いは真陽性 TP(式 7)，偽陽性 FP(式 8)，偽陰性 FN(式 9) のデータ数算出方法であり，ソフト F 値は検索キーワード終点の予測事後確率 $\mathbf{y}$ をもとにソフトな算出がされている．ここで l_t は時刻 t における 2 値の正解ラベルである．このため，図 1 のようにニューラルネットとして実装することが可能であり，これを単語検出器と連結することで End-to-End な最適化が可能となる．

$$F_{soft} = 2\frac{sR \cdot sP}{sR + sP}, \quad (4)$$

$$sR = \frac{sTP}{sTP + sFN}, \quad (5)$$

[1] 東京工業大学
Tokyo Institute of Technology, Tokyo, Japan
www.ts.ip.titech.ac.jp

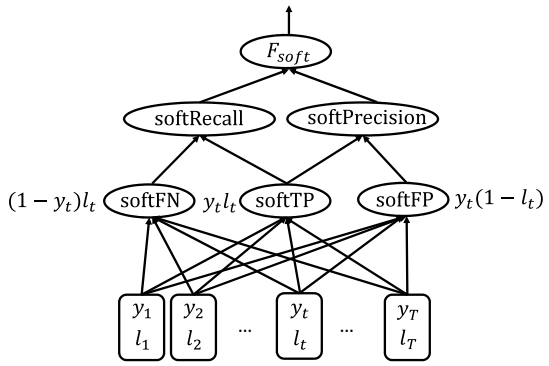

図 1 ソフト F 値を計算するニューラルネット

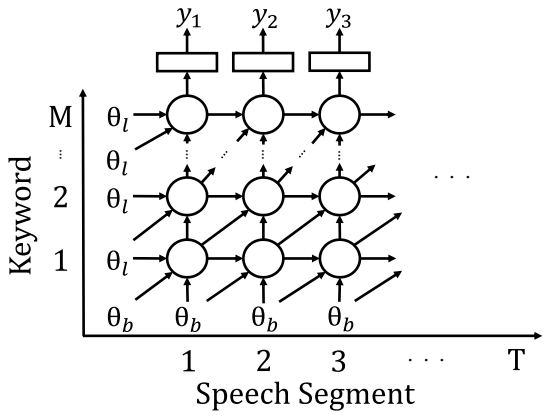

図 2 2D-GRU 単語検出モデル

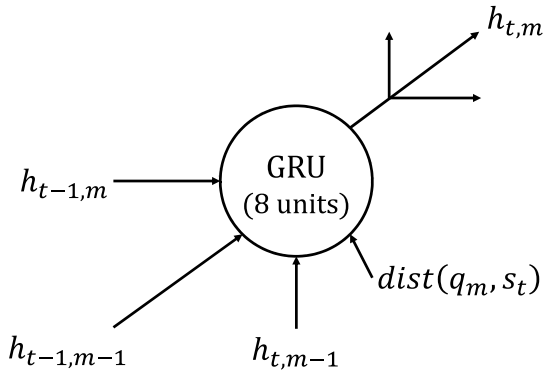

図 3 マルチゲート GRU ユニット

$$sP = \frac{sTP}{sTP + sFP}. \tag{6}$$

$$sTP = \sum_t y_t l_t, \tag{7}$$

$$sFP = \sum_t y_t \left(1 - l_t\right), \tag{8}$$

$$sFN = \sum_t \left(1 - y_t\right) l_t. \tag{9}$$

4. 実験条件

実験には WSJ コーパスを用いた．特徴量は 13 次元の MFCC を用いた．MFCC の抽出は Kaldi ツールキットを用いて行い，25ms 幅 10ms ステップ幅のウィンドウを用いた．検索対象の音声は 5 秒とした．キーワードの種類は，

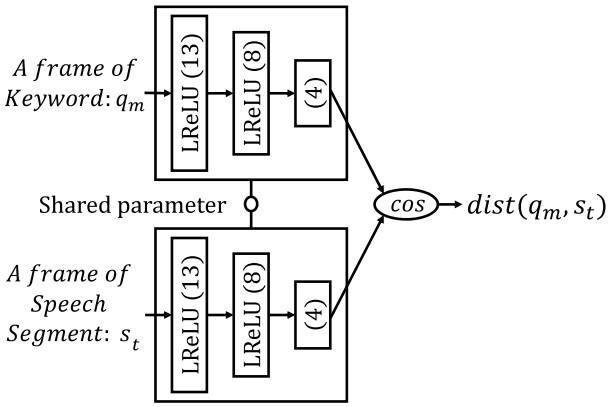

図 4 距離を計算するニューラルネット

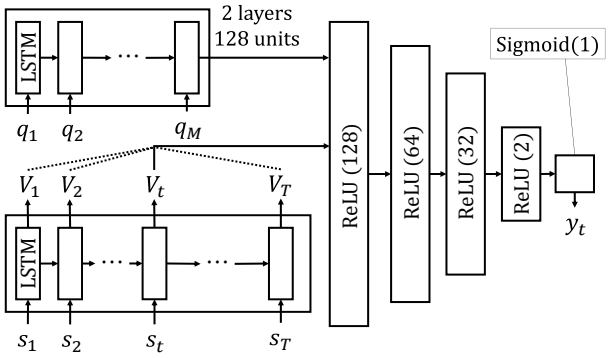

図 5 埋め込み型ニューラルネット単語検出モデル

学習・開発・評価セットについて，それぞれ 480，60，200 とした．各セットはランダムに抽出されたキーワード音声と検索対象音声のペアから構成される．各ペアは異なる話者の音声が用いられる．また，検索対象音声中には対応するキーワードが少なくとも 1 つ含まれる．サンプル数は，学習・開発・評価セットについて，それぞれ 24000，3000，3000 とした．学習・開発セット間はキーワードオープン，学習・評価セット間はキーワード・話者オープンとした．ラベルファイルは Kaldi のトライフォン HMM-GMM モデルを用いた強制アライメントより作成した．ラベル 1 は単語終端を意味し，0 はそうでないことを意味する．アライメントから得られる単語終端を中心に，手前に 9 フレーム，後ろに 10 フレームの 20 フレーム分を 1 に，それ以外を 0 とした．GRU のユニット数は 8 とした．ベースラインには，MFCC を用いた従来の連続 DP マッチング法と，図 5 に示す埋め込み型のニューラルネットモデルを使用した．

5. 実験結果

提案法とベースラインとの比較は図 6 のようになる．従来の MFCC を用いた DP マッチングの 20.3%と比較して，埋め込み型ニューラルネットモデルは 50.3%と大幅に高い性能が得られた．さらに，2D-RNN を用いた提案法では 67.0%とより高い性能が得られた．

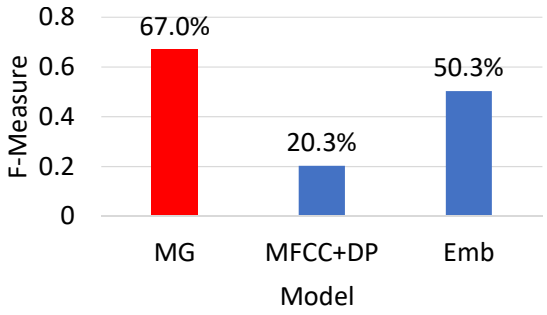

図 6 評価セットを用いた F 値による各モデルの比較．MG はマルチゲート GRU を用いた提案モデル，MFCC+DP は MFCC を用いた DP マッチング，Emb は埋め込み型ニューラルネットのモデルを示す．

6. まとめ

本研究ではマルチゲート GRU ユニットを用いた End-to-End 始終端フリー単語検出モデルを提案し，あわせて連続 DP タスクの End-to-End 学習に適した目的関数としてソフト F 値を用いることを提案した．その結果，従来埋め込み法と比較して 16.7%の性能向上が得られた．今後の課題としては，フレームワイズな評価方法を単語単位に拡張することや，計算量を削減することが挙げられる．

謝辞 本研究は JSPS 科研費 17K20001 および Microsoft Research Core 12 Program の支援を受けたものです．

参考文献

[1] H. Sakoe and S. Chiba: *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*, IEEE Trans Acoust (1978).
[2] Graves, Alex and Fernández, Santiago and Schmidhuber, Jürgen: *Multi-dimensional Recurrent Neural Networks*, in ICANN (2007), pp549–558, Springer Berlin Heidelberg.
[3] H. Kamper and W. Wang and K. Livescu: *Deep convolutional acoustic word embeddings using word-pair side information*, Proc. ICASSP (2016).
[4] Yu-An Chung and Chao-Chung Wu and Chia-Hao Shen and Hung-yi Lee and Lin-Shan Lee: *Audio Word2Vec: Unsupervised Learning of Audio Segment Representations using Sequence-to-sequence Autoencoder*, Proc. Interspeech (2016).
[5] S. Settle and K. Livescu: *Discriminative acoustic word embeddings: Recurrent neural network-based approaches*, Proc. SLT (2016).
[6] Wanjia He and Weiran Wang and Karen Livescu: *Multi-view Recurrent Neural Acoustic Word Embeddings*, Proc. ICLR (2017).
[7] K. Audhkhasi and A. Rosenberg and A. Sethy and B. Ramabhadran and B. Kingsbury: *End-to-End ASR-Free Keyword Search From Speech*, IEEE J Sel Top Signal Process (2017).
[8] Chia-Wei Ao and Hung-yi Lee: *Query-by-example Spoken Term Detection using Attention-based Multi-hop Networks*, Proc. ICASSP (2018).
[9] Pang, Liang and Lan, Yanyan and Guo, Jiafeng and Xu, Jun and Xu, Jingfang and Cheng, Xueqi: *Deeprank: A new deep architecture for relevance ranking in information retrieval*, Proc ACM Int Conf Inf Knowl Manag (2017).
[10] Chao Che and Cao Xiao and Jian Liang and Bo Jin and Jiayu Zho and Fei Wang: *An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease*, Proc SIAM Int Conf Data Min (2017).