7M-04

ニューラルネットによる音声対話における非言語的振る舞いの検出

稲熊 寛文[1]　井上 昂治[1]　河原 達也[1]
[1]京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

人間同士の対話で観測される笑いやフィラーなどの非言語的振る舞いはSocail Signalsと呼ばれ，コミュニケーションにおいて重要な役割を果たす．笑いは対話を和ませる役割がある．一方，フィラーには発話内容を整理するときに自分の発話ターンを保持する役割がある．これらの振る舞いを検出することは，話者の感情や意図，個性，エンゲージメントなどを推定するのに有効であると考えられる．そこで本稿では，ニューラルネットの一種であるLSTM-CTCモデル[1]を用いて音声対話中に表出される笑いとフィラーの振る舞い単位での検出を試みる．

2. Social Signalsの検出

近年，音声対話におけるSocial Signalsのコンペディション[2]が開催されるなど注目が集まっており，様々な手法が提案されている．中でもDNN (Deep Neural Network)[3]やCNN (Convolutional Neural Network)[4]などの順伝搬型やBLSTM (Bidirectional LSTM)[5]などの再帰型のニューラルネットが高い精度を示している．しかし，これらはいずれもフレーム単位の識別器として学習されるためフレーム単位の正解ラベルが必要となる．しかし，膨大なコーパスに対して人手によるアノテーションを行うことはコストの面から非現実的である．さらに，Social Signalsのフレーム単位での正確な検出は冗長であると言える．そこで，CTC (Connectionist Temporal Classification)[6]を用いて振る舞いのラベル系列を直接学習することで，学習データ中のSocial signalsの区間分割をすることなくモデルの学習を行う（図1）．

3. CTCによる学習

クロスエントロピーや最小二乗誤差などを目的関数とする従来のニューラルネットの学習では，入力系列に対する同じ長さの正解ラベル系列をフレーム単位で学習するため，フレーム単位での正解データのアライメントが必要であった．CTCでは長さの異なる入出力系列を扱う目的関数を用いて学習することができ，正解データのアライメントが不要となる．例えば，音声認識への応用[7]では入力の音声特徴量系列に対して音素ラベル系列「arayurugeN...(あらゆる現実を...)」を正解ラベルとして直接学習できる．

次に，CTCのニューラルネットへの適用方法について述べる．まず，目的のラベルの集合をL，Lに空白ラベル$'blank'$を追加した集合を$L'(=L\cup\{'blank'\})$とすると，CTCを適用するネットワークの出力層は$|L'|(|L|+1)$個のユニットからなる．時刻tの出力$\boldsymbol{y}_t$の第k成分y_t^kは時刻tにおけるラベルkの生起確率を表すとすると，系列ラベルの生起が互いに独立であると仮定したとき，入力系列$\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$に対する，$L'$から生成される長さ$T$の任意のラベル系列$\boldsymbol{\pi}=(\pi_1,\ldots,\pi_T)$（パスと呼ぶ）の事後確率は以下のように表される．

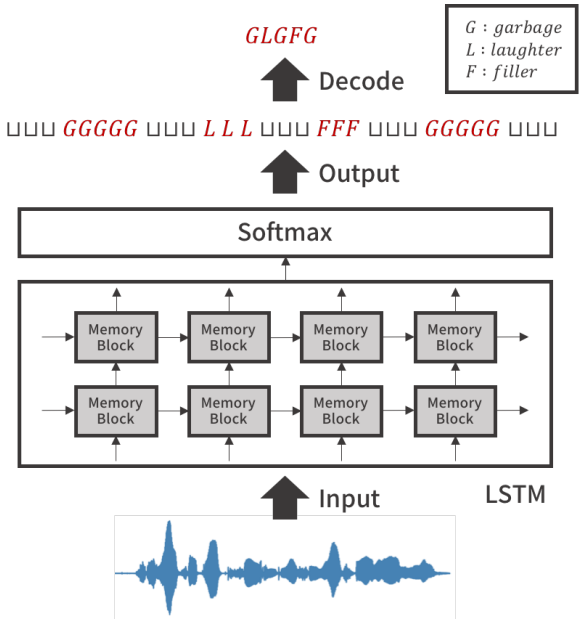

図 1: LSTM-CTCの処理の流れ

$$p(\boldsymbol{\pi}|\boldsymbol{X})=\prod_{t=1}^{T}y_{\pi_t}^t,\ \forall\boldsymbol{\pi}\in L'^T$$

CTCを適用するネットワークでは，$\boldsymbol{\pi}$から冗長性を取り除いた，すべての空白ラベルおよび連続したラベルを除去（e.g. $\boldsymbol{B}(aa\sqcup abb\sqcup\sqcup c)=abc$ ($\sqcup:'blank'$)）した系列$\boldsymbol{l}$を実現するあらゆるパス$\boldsymbol{\pi}$の事後確率の総和を考え，正解系列$\boldsymbol{l}$の事後確率

$$p(\boldsymbol{l}|\boldsymbol{X})=\sum_{\boldsymbol{\pi}\in\boldsymbol{B}^{-1}(\boldsymbol{l})}p(\boldsymbol{\pi}|\boldsymbol{X}) \tag{1}$$

を最大化するようなネットワークパラメータを推定する．また，$\boldsymbol{B}:L'^T\mapsto L^{\leq T}$はこの冗長性を取り除く操作を表す多対一の写像($\boldsymbol{B}(\boldsymbol{\pi})=\boldsymbol{l}$)である．式(1)はHMMなどと同様にフォワード・バックワードアルゴリズムによって計算する．例えば，ある1つのファイルで笑いとフィラーがこの順番で発生したとすると，正解ラベルは以下のようになる．

$$garbage\ \ laughter\ \ garbage\ \ filler\ \ garbage$$

CTCは正確な区間検出はできないものの，スパイク状の事後確率が対応する音素の中心付近に出現するのでため（図3），フレーム単位ではなく振る舞い単位で検出を行うことで頑健性の向上が期待できる．

4. データセット

実験に用いたデータセットはInterspeech 2013 Computational Paralinguistics Challenge (ComParE)の中のSocial Signals Sub-Challenge[2]で使用されたSSPNet Vocalication Corpus (SVC)である．収録音声は120

Detection of Social Signals in Dialogue by Neural Network: *Hirofumi Inaguma, Koji Inoue, Tatsuya Kawahara (Kyoto University)*

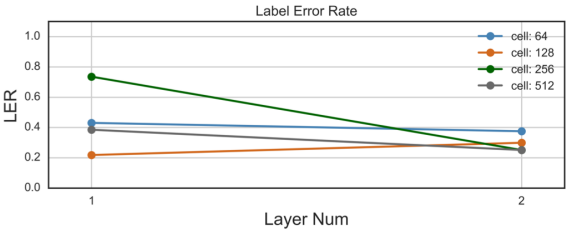

図 2: 実験結果：Label Error Rate

表 1: 実験結果：振る舞い単位での検出精度

セル	層数	笑い			フィラー		
		適合率	再現率	F 値	適合率	再現率	F 値
64	1	0.34	0.18	0.23	0.76	0.45	0.56
	2	0.40	0.27	0.32	0.69	0.64	0.66
128	1	0.32	0.21	0.26	0.67	0.66	0.66
	2	0.42	0.34	0.38	0.83	0.50	0.28
256	1	0.25	0.00	0.00	0.67	0.01	0.02
	2	0.29	0.06	0.10	0.79	0.72	0.75
512	1	0.38	0.11	0.17	0.64	0.75	0.69
	2	0.31	0.04	0.07	0.81	0.73	0.77

名の初対面の被験者（男性 63 名，女性 57 名）の 1 対 1 の電話対話であり，長さが約 11 秒の 2,763 個の WAV ファイルからなる（計 8.4 時間）．各ファイルには 1.5〜9.5 秒の長さの笑いまたはフィラーの音声が少なくとも 1 つは含まれており，音声は 1 話者のみからなる．アノテーションは人手で行われ，2,988 個の笑い，1,158 個のフィラー，その他の区間（有声・無音区間両方を含む）にそれぞれ “laughter”,“filler”, “garbage” の 3 種類のラベルが 1ms 単位で付与されている．

5. 評価実験

5.1 実験目的・条件

SVC を用いて LSTM-CTC の学習を行った．SVC の 120 名分の全 2763 個のファイルのうち，70 話者分（1583 個）を学習データ，20 話者分（500 個）を開発データ，30 話者分（680 個）をテストデータとして使用した．特徴量は，openSMILE[8] を用いてフレーム幅 25ms で 10ms 毎に抽出した 12 次元の MFCC (+Δ,ΔΔ)，対数エネルギー (+Δ)，発話確率 (+Δ)，Harmonics-to-Noise Ratio (+Δ)，F0 (+Δ)，ゼロ交差率 (+Δ) の 47 次元の特徴量に加え，その前後 4 フレームにおける平均，標準偏差を合わせた計 141 次元とした．これは Social Signals Sub-Challenge[2] で定められた特徴量セットと同様のものである．さらに，これらの特徴量は学習データの平均，標準偏差を用いて正規化した．“laughter”,“filler”,“garbage” の 3 種類のラベルを学習時およびデコード時に使用した．

ネットワークは 1〜2 層の LSTM (Long-Short Term Memory) 層と softmax 関数を活性化関数とする全結合層から構成した．パラメータの更新は 20 ファイルをミニバッチとして，学習係数が 10^{-2} の Adam[9] によって行った．学習係数はエポック毎に開発データでの精度が改善しなかった場合に半減させ，6 回半減したところで十分小さいものであるとみなして固定し，最大 50 エポックを学習した．ただし，10 エポック精度の改善が見られなかった場合，そこで学習を打ち止めた．重みパラメータは $[-0.1, 0.1]$ からランダムにサンプリングして初期化した．ネットワークの実装は TensorFlow[10] を用いた．

5.2 評価尺度

評価尺度として LER (Label Error Rate)[6] を使用した．S_{eval} を評価データ，S_{eval} から得られる入出力系列をそれぞれ $\boldsymbol{x}$,$\boldsymbol{z}$ としたとき，検出モデル M の LER は以下のように算出される．

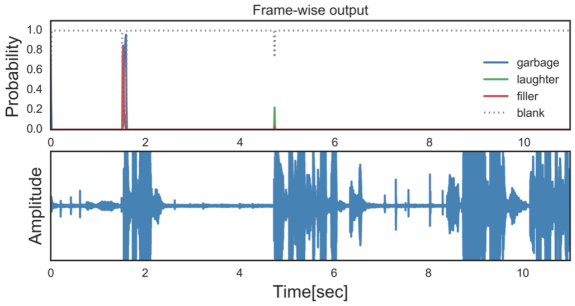

図 3: CTC の出力例

$$LER(M, S_{eval}) = \frac{1}{|S_{eval}|} \sum_{(\boldsymbol{x},\boldsymbol{z})\in S_{eval}} \frac{ED(M(\boldsymbol{x}), \boldsymbol{z})}{|\boldsymbol{z}|}$$

ただし，$ED(\boldsymbol{p}, \boldsymbol{q})$ は 2 つの系列 $\boldsymbol{p}$,$\boldsymbol{q}$ 間の編集距離を表す．

5.3 実験結果

実験結果を図 2 に示す．最も高いもので，約 20%程度の誤り率を得た．メモリセル数で比較するとあまり有意な差は見られなかったが，層数で比較すると 1 層よりも 2 層の方が精度が向上した．また，振る舞い単位での検出精度を表 1 に示す．フィラーに関しては比較的高い精度で検出できているが，笑いの検出はほとんどできておらず，これが誤り率の主な原因であると考えられる．これは図 3 の CTC のフレーム単位の出力を見てもわかる．SVC では “garbage” に発話が含まれているため，有声・無声区間を分け，さらに Socail Signals 以外の音声にも音素ラベルを対応付けて学習させることで，さらなる精度の向上が期待できる．

6. おわりに

本稿では，LSTM-CTC を用いて笑いとフィラーの振る舞い単位での検出を行った．結果より笑いの検出精度は低いがフィラーは高い精度で検出できることを確認した．今後は，Bidirectional LSTM でも同様の実験を行い，また音声認識で使用される音素ラベルに Socail Signals の音素ラベルを追加し，音声認識と統合して検出を試みる予定である．

謝辞 本研究は，JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクトの支援を受けて実施された．

参考文献

[1] 那須悠 *et al.* LSTM-CTC を用いた音響イベント検出・除去音声認識システムの検討. **研究報告コンピュータビジョンとイメージメディア** *(CVIM)*, 2016(22):1–6, 2016.

[2] Björn Schuller *et al.* The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. *in Proc. Interspeech*, 2013.

[3] Rahul Gupta *et al.* Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. *in Proc. Interspeech*, 173–177, 2013.

[4] Lakshmish Kaushik *et al.* Laughter and filler detection in naturalistic audio. *in Proc. Interspeech*, 2015.

[5] Raymond Brueckner *et al.* Social signal classification using deep blstm recurrent neural networks. *in Proc. ICASSP*, 4823–4827. IEEE, 2014.

[6] Alex Graves *et al.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *in Proc. ICML*, 369–376, 2006.

[7] Alex Graves *et al.* Speech recognition with deep recurrent neural networks. *in Proc. ICASSP*, 6645–6649. IEEE, 2013.

[8] Florian Eyben *et al.* Opensmile: the munich versatile and fast open-source audio feature extractor. *in Proc. ACM Multimedia*, 1459–1462. ACM, 2010.

[9] Diederik Kingma *et al.* Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Martın Abadi *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.