Vol.2017-SLP-117 No.7
2017/7/28

End-to-End モデルによる Social Signals 検出および音声認識との統合

稲熊 寛文[1]　井上 昂治[1]　三村 正人[1]　河原 達也[1]

概要：人間同士の対話では，言語的情報だけでなく，笑いやフィラー，相槌，言い淀みなどの Social Signals と呼ばれる非言語的振る舞いがしばしば観測される．Social Signals を検出することは話者の感情状態やエンゲージメントなどを推定するのに有効であり，対話システムがより人間らしく振る舞うための情報源にもなり得る．著者らは，学習データ中の正解ラベル系列の区間分割が不要な End-to-End モデルである Connectionist Temporal Classification (CTC) を損失関数とする BLSTM-CTC を用いることで，音声対話中に表出する Social Signal の頑健で直接的な検出を行っている．本稿では，従来の文字単位の End-to-End 音声認識と Social Signals の検出を同時に行い，これらが統一的な枠組みで扱えることを示す．さらに，通常の発話特別して Social Signals を検出し，それらを除去することによって，大規模コーパスにおいて認識精度が改善することを確認する．

キーワード：音声認識，End-to-End モデル，Social Singals，Connectionist Temporal Classification

1. はじめに

音声対話中に観測される笑いやフィラー，相槌などの非言語的振る舞いは Social Signals [1–3] と呼ばれ，人間同士のコミュニケーションにおいて重要な役割を果たす．これらの振る舞いを検出することは，話者の感情状態，エンゲージメント，意図，欲求，個性などを推定するのに有効であり，対話システムがより人間らしく振る舞うための情報源にもなり得る．さらに，これらの振る舞いを検出・除去することで，自然会話の音声認識精度の向上にもつながることが期待される．

近年, Social Signals の検出は注目を集めており，これまでに GMM (Gaussian Mixture Model) [4] や AdaBoost [5]，遺伝的アルゴリズム [6]，HMM(Hidden Marcov Model) [7] など，様々な従来の機械学習の手法が提案されている．また，音声認識のタスクと同様に，DNN (Deep Neural Network) [8,9] や CNN (Convolutional Neural Network) [10]，BLSTM (Bidirectional Long-Short Term Momory) [11] などのニューラルネットワークを用いた手法がより高い精度を示している．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が，以下の観点から Social Signals の検出というタスクには適していない．

まず，情報検索の観点から，膨大な Social Signals の発生候補の中から実際の振る舞い自体の発生を検出することが目的であり [12]，フレーム単位よりも振る舞い単位で頑健に検出するのが望ましい．次に，モデルの学習において，従来の手法ではフレーム単位の正解ラベルが必要となる．Social Signals は通常発話に比べてその境界が曖昧なので，フレーム単位で人手によるアノテーションを行うことはコストが高く，また事前学習された識別器を用いて強制アライメントを行う場合，正解ラベルの品質はそのモデルの精度に依存する．さらに，検出段階では，フレーム単位でラベルを予測した後，振る舞い単位で検出するために閾値処理や HMM でのスムージングなどの後処理が必要となる．

そこで著者らは，近年の End-to-End 音声認識で成功しており [13–16]，あらゆる可能なフレーム単位のアライメントを周辺化することで，あらかじめ学習データセット中の正

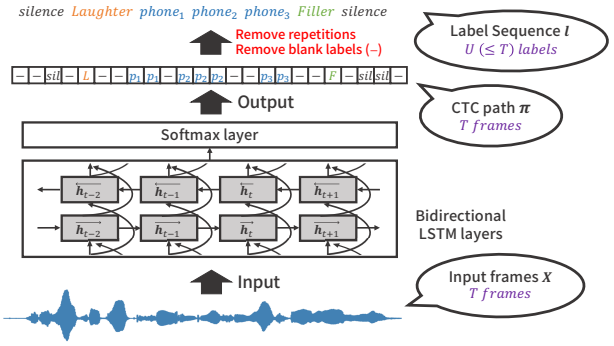

図 1: BLSTM-CTC によるデコーディング

[1] 京都大学 大学院情報学研究科

解ラベル系列の区間分割をすることなくモデルを直接学習できる CTC (Connectionist Temporal Classification) [17] を用いることで，Social Signals の振る舞い単位での検出を行う（図 1）．CTC を用いて直接的な検出を行うことで，頑健性の向上も期待できる．

本研究では，Social Signals の正解ラベルの与え方を調査すると同時に，検出した Social Signals を除去することで音声認識精度が改善することを示す．評価は比較的小さな対話コーパスおよび大規模な講演コーパス（CSJ）の両方を用いて行う．

本稿の構成を以下に示す．2 章では，Social Signals の検出の意義と関連研究について述べる．3 章では，提案手法である CTC について述べる．4 章では，CTC での Social Signals を考慮した正解ラベルの与え方について述べる．5 章では Social Signals の検出実験を行う．最後に，本研究のまとめと今後の展望について 6 章で述べる．

2. Social Signals

2.1 Social Signals の役割

Social Signlas [1–3] とは，自然会話中に話者の感情状態などを理解する目安となる情報源であり，その検出のコンペティション [18] が開催されるなど近年多くの注目を集めている．観測可能な Social Signals として，音声，姿勢，ジェスチャ，視線，表情などの様々なモダリティの振る舞いがあるが，本研究では Social Signals として，笑い，フィラー，相槌，言い淀みを対象とする．これらの 4 つの振る舞いを採用したのは，これらが音声対話において比較的観測しやすく，また我々が日常的に表現しているものであるからである．

これらの振る舞いにはそれぞれ役割がある．笑いには，先行発話を和ませたり話者の感情を表す役割がある [19–21]．一方，フィラー（“うーん”，“えー” など）には発話内容を整理したり相手が自分のターンに割り込まないように自分の発話権を保持する役割がある [22]．また，相槌（“はい”，“うん” など）は話者に注意を払っていることや発話内容を理解していることを示すフィードバックであり，発話の継続を促進する役割がある [23]．言い淀みには発話の繰り返し，訂正などのいくつかの役割がある [24]．

2.2 Social Signals と音声認識

Social Signals の検出と音声認識は相補的な関係にある．Social Signals の検出では，笑いやフィラーなどの Social Signals と音素を区別することで，検出精度が向上することが期待される．一方音声認識では，自然会話の音声を認識する際，特に笑いやフィラーなどの周りで認識誤りが発生する傾向がある．したがって，Social Signals を通常発話と区別できれば，認識精度の向上にもつながると考えられる．そこで本研究では，これまで別々の枠組みとして扱われてきた Social Signals の検出と音声認識を統一的な枠組みで扱い，同時学習することで双方の精度向上を目指す．また，これらを同時に扱うことで，発話笑いなどを検出・認識できることも期待される．

2.3 関連研究

Interspeech において毎年のように開催されている非言語情報に関するコンペティション Computational Paralinguistics Challenge (ComParE) の中で，2013 年に Social Signals の検出のコンペティションが開催され [18]，様々な手法が提案された [4,5,8]．このコンペティションの目標は，Unweighted Average Area Under the Curve (UAAUC) を評価基準としたフレーム単位の分類であった．

DNN ベースの手法 [8] が最高精度を示し，その後も様々な手法が提案されたが [6,7]，特にニューラルネットワークによる手法が成果を挙げ [9–11]，さらに事後確率のスムージングによって UAAUC の精度向上を示している．

しかし，Gosztolya ら [12] によると，これらのフレーム単位で識別学習されたモデルの事後確率を音声認識と同様に HMM によってスムージングし，振る舞い単位の F 値による検出精度を評価すると，スムージングにより精度が低下した．したがって本研究では，Gosztolya らの提案に従い，Social Signals の検出の評価基準として，適合率，再現率，F 値を採用して評価を行う．

3. Connectionist Temporal Classification (CTC)

クロスエントロピーや平均二乗誤差などを損失関数とする従来のニューラルネットワークの学習では，入力系列と正解ラベル系列の系列長は等しくなければならず，フレーム単位での正解ラベルのアライメントが必要であった．Connectionist Temporal Classification (CTC) [17] は正解データの事前のアライメントが不要であり，系列長の異なる入出力系列を直接学習できる損失関数を用いる．CTC は理論上任意のニューラルネットワークに適用できるが，Recurrent Neural Networks (RNN) との相性がよく，RNN のソフトマックス層の後に接続される [13]．CTC では，ネットワークが何も出力しないことを意味する空白ラベル *blank* の導入，および同じラベルの繰り返しの許容により，入出力の系列長の違いに対応する．空白ラベルは，ネットワークのソフトマックス層に新たなクラスとして追加される．この空白ラベルを含んだ CTC の中間表現（ソフトマックス層の出力）は CTC パスと呼ばれる．

入力音声 $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ とそれに対応する L 種類のラベル集合からなる長さ $U(\leq T)$ の正解ラベル系列 $\boldsymbol{l}$ が与えられた時，CTC を適用するネットワークは以下の $\boldsymbol{X}$ に対する $\boldsymbol{l}$ の負の対数尤度 $L_{CTC}(\boldsymbol{X})$ を最小化するように最

尤推定によりパラメータを学習する．

$$L_{CTC}(\boldsymbol{X}) = -\ln P(\boldsymbol{l}|\boldsymbol{X}) \quad (1)$$

事後確率 $P(\boldsymbol{l}|\boldsymbol{X})$ は，あらゆる CTC パス $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ の事後確率の総和として周辺化される．

$$P(\boldsymbol{l}|\boldsymbol{X}) = \sum_{\boldsymbol{\pi} \in \boldsymbol{\Phi}(\boldsymbol{l})} P(\boldsymbol{\pi}|\boldsymbol{X}) \quad (2)$$

ただし，$\boldsymbol{\Phi}(\boldsymbol{l})$ は CTC パスの集合であり，$\boldsymbol{\Phi}$ は CTC パスから連続したラベルを除去し，空白ラベルを取り除く操作を表す関数の逆関数である（$\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) = \boldsymbol{l}$）．

CTC パスの事後確率 $P(\boldsymbol{\pi}|\boldsymbol{X})$ は各時刻の出力ラベルの条件付き独立性の仮定により，以下のように分解される．

$$P(\boldsymbol{\pi}|\boldsymbol{X}) = \prod_{t=1}^{T} y_{\pi_t}^t \quad (3)$$

ただし，y_k^t は時刻 t におけるソフトマックス層のクラス k に対応するノードの出力であり，対応するラベルの生起確率を表す．これらより，$P(\boldsymbol{l}|\boldsymbol{X})$ 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効率的に計算される．CTC は明示的に時刻情報を正解として与えないため，正確な区間検出は保証していないものの，RNN として双方向の入力を使用する Bidirectional Long-Short Term Memory (BLSTM) を使用することで，単方向の LSTM よりも正確な Social Signals の検出を目指す．

4. 正解ラベルの生成

本研究における Social Signals に関する正解ラベルの与え方について，*Remove*(*Baseline*)，*Insert*，*Insert*2，*Replace* の 4 通りを考える（**図 2**）．*Remove*(*Baseline*) は Social Signals を考慮しない従来の正解ラベルで，通常の音声認識と同様のラベルの与え方である．

Insert は Social Signals ラベルに対応する仮名のラベルの前に挿入する．分かち書きのスペースラベルを介さずに対応する仮名とつなげることで，CTC の持つ言語モデルに似た性質を利用することを狙いとしている．

*Insert*2 はさらに対応する仮名のラベルの後ろに Social Signals の終了ラベルを挿入する．これにより，*Insert* ではできなかった Social Signals に対応する区間情報を正解として与えることを狙いとしている．

Replace は Social Signals に対応する仮名のラベルを Social Signals で置換する．フィラーや相槌などは仮名自体を認識することにはあまり意味を持たないので，特定のイベントとして検出することが目的である．Social Signals の検出は *Insert*，*Insert*2，*Replace* の 3 通り，音声認識は *Remove*(*Baseline*)，*Insert*，*Insert*2 の 3 通りで評価・比較する．

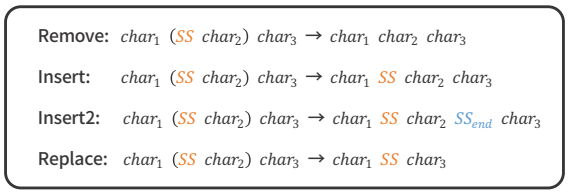

図 2: Social Singals ラベルを考慮した正解ラベルの生成方法．SS は任意の Social Signals ラベル，SS_{end} は対応する Social Signals の終了ラベルを表す．

5. 評価実験

5.1 システムの構成方法

入力特徴量は，HTK [25] を用いてフレーム幅 25ms で 10ms 毎に抽出した 40 次元対数メルフィルタバンク特徴量と対数エネルギー $(+\Delta, \Delta\Delta)$ の計 123 次元とした．これらの特徴量は話者毎の平均，標準偏差を用いて正規化した．さらに，3 フレーム分の入力（30ms）をそれぞれ 1 つの入力に連結して入力フレーム数を削減することで，CTC のアライメントの任意性を減らした [26]．ネットワークは単方向のメモリセルのユニット数 256 の BLSTM5 層とソフトマックス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 64 発話をミニバッチとして，RMSProp によって行った（学習係数 10^{-3}）．学習を安定させるために学習データを発話長でソートしてからミニバッチを構成し [14, 16]，ミニバッチ内で発話をシャッフルした．すべての重みパラメータは $[-0.1, 0.1]$ の一様分布からランダムにサンプリングして初期化した．LSTM の忘却ゲートのバイアスの初期値は 1.0 に初期化した [27]．勾配は $[-5.0, 5.0]$，LSTM のメモリセルのアクティベーションは $[-50, 50]$ でクリッピングした．Weight decay はすべてのパラメータに適用し，その係数は $1e-6$ とした．隠れ層間のドロップアウト率はそれぞれ 0.8 とした．CTC の出力のデコードはビームサーチによって行い，ビーム幅は 20 とした．ネットワークの実装は TensorFlow [28] によって行った．

5.2 ERATO コーパスによる評価

5.2.1 実験条件

ERATO コーパスは自律型アンドロイド ERICA [29] を用いて収録した日本語の対面対話コーパスである．ERICA は 6 人の女性オペレータによって遠隔操作されている．現在までに，91 セッションの初対面対話が収録されており，それぞれ約 10 分程度の対話から構成されている（計 16.8 時間）．被験者は 91 名で，一人ずつオペレータに遠隔操作された ERICA と自由に対話する．ERICA は様々な社会的役割を担い，被験者はそれに応じて会話する．オペレータの音声は卓上に設置されたスタンドマイクで，被験者の音声は被験者の足下に設置したガンマイクでそれぞれ収録し

表 1: ERATO コーパスにおける Social Signals の振る舞い単位での検出結果

ラベル付与	Social signals	適合率	再現率	F 値	F 値平均
Insert	Laughter	0.61	0.35	**0.44**	**0.56**
	Filler	0.72	0.81	0.76	
	Backchannel	0.90	0.70	**0.79**	
	Disfluency	0.31	0.20	0.25	
*Insert*2	Laughter	0.65	0.31	0.42	0.54
	Filler	0.73	0.81	0.76	
	Backchannel	0.91	0.69	**0.79**	
	Disfluency	0.30	0.14	0.19	
Replace	Laughter	0.61	0.16	0.25	0.50
	Filler	0.72	0.83	**0.77**	
	Backchannel	0.90	0.60	0.72	
	Disfluency	0.35	0.20	**0.26**	

表 2: ERATO コーパスにおける文字単位の音声認識結果

ラベル付与	CER
Remove(*Baseline*)	**20.82 %**
Insert	21.14 %
*Insert*2	21.26 %

たものを用いた．書き起こしとアノテーションは人手で行われ，コーパスには 707 個の笑い，9312 個のフィラー，？個の相槌，1459 個の言い淀みの 4 つの振る舞いが含まれている．ラベルは 145 文字の日本語の仮名とスペース，ノイズ，4 種類の Social Signals の計 151 種類を使用した．全データは学習データ（13.4 時間），開発データ（1.3 時間），テストデータ（2.1 時間）にそれぞれ分割した．

5.2.2 実験結果

Social Signals の検出結果を**表 1** に示す．評価尺度として適合率，再現率，F 値を採用した [12]．フィラーと相槌に関しては高い精度で検出できたが，笑いと言い淀みの検出精度は低い結果となった．笑いに関しては，発話笑いのような通常の音声に笑いが重なる場合を認識できなかったことが大きな原因であると推測できる．言い淀みに関しては，従来言語モデルによって検出されているように [24]，言語情報を特徴量として使用しなかったことに起因すると考えられる．ラベルの付与の仕方で比較すると，F 値の平均から *Insert* と *Insert*2 がラベルの与え方として適切であることがわかった．

次に，Social Signals を検出し，それらを除去することによって音声認識精度が向上するか調査した．結果を**表 2** に示す．評価尺度は仮名単位の Character Error Rate (CER) とした．*Insert* の方が *Insert*2 よりクラス数が少なくコストが低いので，5.3 節の CSJ での実験では *Remove*(*Baseline*) と *Insert* のみを比較対象する．

CTC の 4 通りの正解ラベルの与え方ごとのデコード結果は以下のようになる．

***Remove*(*Baseline*)**
正解: __ソ__ソノ__リユウ__ハイ__ハイ__
出力: __ソ__ソノ__リユウカイ__ハイ__

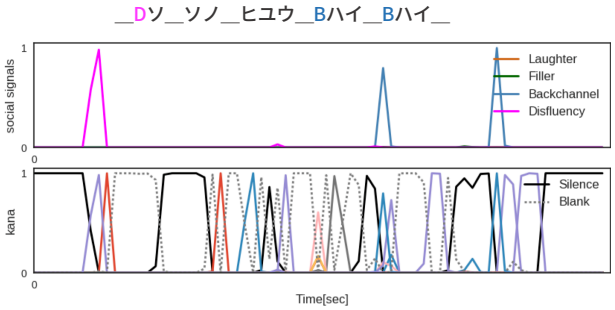

図 3: ERATO コーパスにおける BLSTM-CTC の出力．ただし，Social Signals の正解ラベルの与え方は *Insert* である．

Insert
正解: __ D ソ__ソノ__リユウ__ B ハイ__ B ハイ__
出力: __ D ソ__ソノ__ヒユウ__ B ハイ__ B ハイ__

***Insert*2**
正解: __ D ソ d __ソノ__リユウ__ B ハイ b __ B ハイ b __
出力: __ D ソ d __ソノ__イウ__ B ハイ b __ B ハイ b __

Replace
正解: __ D __ソノ__リユウ__ B __ B __
出力: __ D __ソノ__リユウ D __ B __

ただし，__はスペースラベルであり，空白ラベルとは異なることに注意する．英語などの同様に，単語間の区切りを明確にするために与えた．

また，BLSTM-CTC のソフトマックス層の出力を**図 3** に示す．これより，CTC によって仮名だけでなくそれが Social Signals であるかどうかも検出できていることがわかる．

5.3 講演コーパス（CSJ）による評価

5.3.1 実験条件

次に，日本語話し言葉コーパス（CSJ）[30] の講演を使用して評価を行った．学習データは学会講演のみの 240 時間と，模擬講演なども含めた 586 時間の 2 種類用意した．評価データは Eval1，Eval2，Eval3 の 3 種類あり，それぞれ 10 講演ずつから構成されている．また，開発データとして学習データから 19 講演を選択した．CSJ は独話であるが，笑い，フィラー，言い淀みの 3 種類の Social Signals がアノテーションされており，ラベルは 145 文字の仮名とスペース，ノイズ，上記 3 種類の Social Signals の計 150 種類を使用した．アーキテクチャは，BLSTM のメモリセルのユニット数を 320 にしたことを除き，5.2 節と同じである．

5.4 実験結果

Social Signals の検出結果を**表 3** に示す．5.2 節の ERATO コーパスの結果と比較すると，フィラーと言い淀みに関しては高い精度で検出できた．これは CSJ は ERATO コーパスに比べて学習データ量が大きく，CTC の持つ言

表 3: CSJ における Social Singals の振る舞い単位での検出結果

学習データ	Social Signals	Eval1			Eval2			Eval3			F 値平均
		適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値	
240h	Laughter	0.66	0.11	0.20	0.00	0.00	0.00	1.0	0.04	0.08	0.09
	Filler	0.96	0.91	0.93	0.92	0.91	0.92	0.84	0.88	0.86	0.90
	Disfluency	0.80	0.34	0.48	0.72	0.38	0.50	0.64	0.29	0.40	0.46
	Mean (L+F+D)	0.81	0.45	0.54	0.55	0.43	0.47	0.83	0.40	0.44	0.48
586h	Laughter	1.00	0.35	0.53	0.00	0.00	0.00	1.00	0.12	0.22	0.25
	Filler	0.96	0.89	0.92	0.94	0.90	0.92	0.87	0.88	0.88	0.90
	Disfluency	0.83	0.32	0.46	0.76	0.34	0.47	0.77	0.23	0.36	0.43
	Mean (L+F+D)	0.93	0.52	0.63	0.56	0.41	0.46	0.88	0.41	0.49	0.52

表 4: CSJ における文字単位の音声認識結果

学習データ	ラベル付与	CER			平均
		Eval1	Eval2	Eval3	
240h	Remove	9.88%	8.45%	15.62%	11.32%
	Insert	**9.56%**	**8.27%**	**14.21%**	**10.68%**
586h	Remove	10.32%	7.98%	9.69%	9.33%
	Insert	**10.21%**	**7.59%**	9.84%	**9.22%**

語モデルに似た性質がいかされたためだと考えられる．一方，笑いは適合率は高いもののほとんど検出できなかった．そもそもテストデータに笑いがほとんど含まれていないのが原因である．

次に，検出した Social Signals を除去することによる音声認識精度を評価した．結果を**表 4** に示す．ERATO コーパスに比べて CSJ では学習データが大規模なので，学習時に Social Signals を考慮した正解ラベルを与えた効果が現れ，検出した Social Signals を除去することで音声認識精度が向上することを確認した．これにより，Social Signals の検出と音声認識を統一的な枠組みで扱える示唆を得た．

6. おわりに

本研究では，系列長の異なる正解ラベル系列を直接学習することができる CTC を損失関数とした BLSTM-CTC モデルを用いることで，事前に学習データ中の区間分割を行うことなくをパラメータを学習し，振る舞い単位での Social Signals の検出を試みた．自律型アンドロイドを遠隔操作して収録した音声対話コーパスと CSJ を用いた実験結果より，フィラーと相槌が高い精度で検出できることを確認した．また，Social Signals を考慮した正解ラベルの与え方を工夫することで，大規模コーパスである CSJ において，検出した Social Signals を除去することで音声認識精度が向上することを確認した．今後は，発話笑いなどに対応するため，Attention Mechanism [31, 32] を用いて Social Signals の検出を行うほか，DNN-HMM などのハイブリッドモデルと比較し，さらに英語などの多言語でも評価する予定である．

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（課題番号：JPMJER1401）の一環で行われた．

参考文献

[1] Vinciarelli, A., Pantic, M. and Bourlard, H.: Social signal processing: Survey of an emerging domain, *Image and Vision Computing Journal*, Vol. 27, No. 12, pp. 1743–1759 (2009).

[2] Poggi, I. and D'Errico, F.: Social signals: a framework in terms of goals and beliefs, *Cognitive Processing*, Vol. 13, No. 2, pp. 427–445 (2012).

[3] Brunet, P. and Cowie, R.: Towards a conceptual framework of research on social signal processing, *Journal on Multimodal User Interfaces*, Vol. 6, No. 3-4, pp. 101–115 (2012).

[4] Krikke, T. F. and Truong, K. P.: Detection of nonverbal vocalizations using Gaussian Mixture Models: looking for fillers and laughter in conversational speech, *Proceedings of Interspeech*, pp. 163–167 (2013).

[5] Gosztolya, G., Busa-Fekete, R. and Tóth, L.: Detecting autism, emotions and social signals using adaboost, *Proceedings of Interspeech*, pp. 220–224 (2013).

[6] Gosztolya, G.: Detecting Laughter and Filler Events by Time Series Smoothing with Genetic Algorithms, *Proceedings of International Conference on Speech and Computer*, pp. 232–239 (2016).

[7] Salamin, H., Polychroniou, A. and Vinciarelli, A.: Automatic detection of laughter and fillers in spontaneous mobile phone conversations, *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 4282–4287 (2013).

[8] Gupta, R., Audhkhasi, K., Lee, S. and Narayanan, S.: Paralinguistic event detection from speech using probabilistic time-series smoothing and masking, *Proceedings of Interspeech*, pp. 173–177 (2013).

[9] Brueckner, R. and Schulter, B.: Hierarchical neural networks and enhanced class posteriors for social signal classification, *Proceedings of ASRU*, pp. 362–367 (2013).

[10] Kaushik, L., Sangwan, A. and Hansen, J. H.: Laughter and filler detection in naturalistic audio, *Proceedings of Interspeech*, pp. 2509–2513 (2015).

[11] Brueckner, R. and Schulter, B.: Social signal classification using deep BLSTM recurrent neural networks, *Proceedings of ICASSP*, pp. 4823–4827 (2014).

[12] Gosztolya, G.: On evaluation metrics for social signal detection, *Proceedings of Interspeech*, pp. 2504–2508 (2015).

[13] Graves, A., Mohamed, A. and Hinton, G.: Speech recognition with deep recurrent neural networks, *Proceedings of ICASSP*, pp. 6645–6649 (2013).

[14] Miao, Y., Gowayyed, M. and Metze, F.: EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, *Proceedings of ASRU*, IEEE, pp. 167–174 (2015).

[15] Rao, K., Senior, A. and Sak, H.: Flat start training of CD-CTC-SMBR LSTM RNN acoustic models, *Proceedings of ICASSP*, IEEE, pp. 5405–5409 (2016).

[16] Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G. et al.: Deep speech 2: End-to-end speech recognition in english and mandarin, *arXiv preprint arXiv:1512.02595* (2015).

[17] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *Proceedings of ICML*, pp. 369–376 (2006).

[18] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E. et al.: The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, *Proceedings of Interspeech* (2013).

[19] Bachorowski, J., Smoski, M. and Owren, M.: The acoustic features of human laughter, *Journal of the Acoustical Society of America*, Vol. 110, No. 3, pp. 1581–1597 (2001).

[20] Vettin, J. and Todt, D.: Laughter in conversation: Features of occurrence and acoustic structure, *Journal of Nonverbal Behavior*, Vol. 28, No. 2, pp. 93–115 (2004).

[21] Tanaka, H. and Campbell, N.: Acoustic features of four types of laughter in natural conversational speech, *Proceedings of 17th International Congress of Phonetic Sciences (ICPhS)*, pp. 1958–1961 (2011).

[22] Clark, H. and Tree, J. F.: Using "uh" and "um "in spontaneous speaking, *Cognition*, Vol. 84, No. 1, pp. 73–111 (2002).

[23] Ward, N. and Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese, *Journal of pragmatics*, Vol. 32, No. 8, pp. 1177–1207 (2000).

[24] Zayats, V., Ostendorf, M. and Hajishirzi, H.: Disfluency detection using a bidirectional LSTM, *Proceedings of Interspeech*, pp. 2523–2527 (2016).

[25] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D. et al.: The HTK book (for HTK version 3.4), *Cambridge university engineering department*, Vol. 2, No. 2, pp. 2–3 (2006).

[26] Sak, H., Senior, A., Rao, K. and Beaufays, F.: Fast and accurate recurrent neural network acoustic models for speech recognition, *Proceedings of Interspeech*, pp. 1468–1472 (2015).

[27] Miao, Y., Gowayyed, M., Na, X., Ko, T., Metze, F. and Waibel, A.: An empirical exploration of CTC acoustic models, *Proceedings of ICASSP*, pp. 2623–2627 (2016).

[28] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M. et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems, *arXiv preprint arXiv:1603.04467* (2016).

[29] Inoue, K., Milhorat, P., Lala, D., Zhao, T. and Kawahara, T.: Talking with erica, an autonomous android, *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 212 (2016).

[30] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (2003).

[31] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y.: End-to-end attention-based large vocabulary speech recognition, *Proceedings of ICASSP*, pp. 4945–4949 (2016).

[32] Bahdanau, D., Cho, K. and Bengio, Y.: Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* (2014).