LSTM を用いた本心でない発話の自動検出
Automatic Detection of Insincere Utterances with LSTM

見尾　和哉†
Kazuya Mio

段原　優和†
Yuto Danbara

大道　博文‡
Hirofumi Omichi

† 広島経済大学　石野研究室
Ishino Laboratory, Hiroshima University of Economics

‡ 広島市立大学　言語音声メディア工学研究室
Language and Speech Research Laboratory,
Hiroshima City University

概要 本研究では、発話中の音声や表情の情報を用いて、本心でない発話を自動検出するというタスクにおいて、LSTM を利用する手法を提案した。また、SVM を利用した既存手法との比較実験を行い、台詞固定の場合で F 値を 0.13 ポイント、台詞自由の場合で F 値を 0.14 ポイント向上させることに成功し、提案手法の有効性を確認した。

1　はじめに

近年、人間とコミュニケーションを行う対話システムが盛んにサービス化している。対話システムがユーザとより円滑なコミュニケーションを行うためには、ユーザの感情を理解する必要がある。しかし、表出された感情が常に本心であるとは限らないため、発話が本心か否かというような抑圧された感情まで推定する技術が必要となってくる。

そこで、本研究では、抑圧された感情まで推定するシステムの構築を目的に、本心でない発話を自動検出する手法を提案する。提案手法では、機械学習には LSTM を使用し、特徴量として発話中の音声や表情の情報を利用する。

2　先行研究

本研究の先行研究として、上村[1]の研究がある。上村は、発話中の音声や表情から得られる特徴量を使用して、本心でない発話を自動検出する手法を提案している。音声特徴量としては、音声の時系列データから、openSMILE[2]により算出された音量の最大値や最小値などの 384 個の特徴を利用している。表情特徴量としては、発話が終わった瞬間の静止顔画像から、オムロンの OKAO Vision[3]を用いて算出された 5 感情（無、喜、驚、怒、悲）分の推定感情の尤度を利用している。このように、上村らの手法では、音声、表情ともに事前に定義された特徴量を使用して、SVM によって本心でない発話を検出している。本研究では、より細かい音声や表情の時系列での変化を把握するため、LSTM（Long short-term memory）を用いた手法を提案する。

3　LSTM を用いた本心でない発話の検出手法

本研究では、音声と表情の時系列での変化を考慮し、本心でない発話の検出を行うため、LSTM を使用した手法を提案する。LSTM とは、RNN（Recurrent Neural Network）を改良した時系列データを扱うことができるモデルである。

LSTM を用いた本心でない発話の自動検出手法のイメージ図を図 1 に示す。図 1 の縦長の長方形は、LSTM の入出力となる特徴ベクトルを模式的に表したものである。

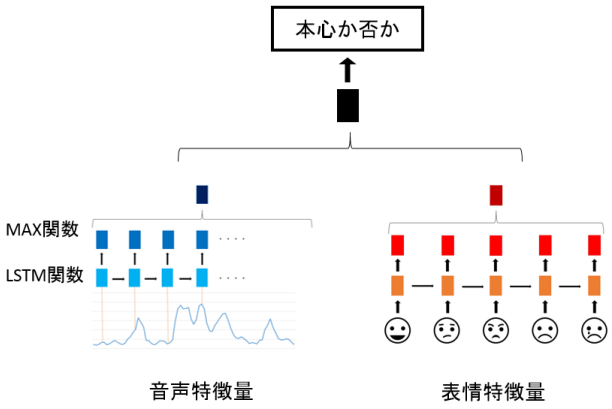

図 1： LSTM を利用した手法のイメージ図

LSTM に与える特徴量について説明する。まず、音声特徴量について説明する。発話中にマイクにより録音したデータから、openSMILE を使って、25msec の窓に対して「音圧」、「基本周波数（F0）」、「自己相関関数から算出される声らしさ」の値を算出する。この窓を 10msec ごとにシフトさせ時系列の音声特徴量を得る。

次に、表情特徴量について説明する。発話中にビデオカメラにより撮影された動画から、OKAO Vision により顔部位の特徴点（左目頭、左目尻、右目頭、右目尻、鼻左、鼻口、口上、口元左、口元右）を検出し、その座標を時系列に並べたデータを表情特徴量として使用する。表情特徴量として、OKAO Vision により推定された感情を利用せず、顔部位の特徴点の座標を利用するのは、Ekman[4] の squelched expression（抑圧された表情）のように顔の部位ごとの感情表出やその程度を考慮するためである。

提案手法では、音声特徴量$v_{aco,1}$と表情特徴量$v_{face,1}$をそれぞれ時系列に並べて LSTM 関数に入力し、300 次元のベクトル$v_{aco,2}$、$v_{face,2}$を得る。$v_{aco,2}$と$v_{face,2}$を合成後、線形関数を適用し、2 次元のベクトル$v_{aco+face}$へ変換する。最後に、$v_{aco+face}$の中で最も値の大きい次元に対応するラベルを予測ラベルとする。

4　実験

4.1　実験に使用したデータ

実験には、上村[1]の研究で使用したデータを利用した。上村のデータ収集手順を説明する。まず、大学生 10 人に、図 2 に示すような、本心を誘発するであろう画像と、本心を誘発しないであろう画像各 20 枚を、1 枚ずつ見せた。

図 2： データ収集に用いた画像
（上村の論文より引用）

画像を見て、例え本意でなくても必ず褒めてもらい、発話した直後に、本心であるか否かを選択させた。その際、音声をマイクで録音、表情をビデオカメラで撮影しておき、後に特徴量として機械学習に使用した。これを、褒め台詞を指定する「台詞固定」と、自由な言葉で褒めてもらう「台詞自由」の 2 パターンでデータを収集し、実験に使用した。

実験に使用したデータを表 1 に示す。台詞固定、台詞自由で、それぞれ 300 件のデータを訓練データに、残りを評価データとして使用した。

表 1： 実験に使用したデータ

	本心でない（正例）	本心である（負例）	合計
台詞固定	201	166	367
台詞自由	167	202	369

4.2　比較手法

提案手法の有効性を確認するため、以下に示す手法で実験を行った。

- SVM 手法（比較手法）：　上村と同様に機械学習に SVM を利用した手法
- LSTM 手法（提案手法）：　機械学習に LSTM を利用した手法

また、機械学習に使用する特徴量は、音声特徴量のみ、表情特徴量のみ、音声特徴量と表情特徴量の両方を利用する 3 パターンで実験を行った。評価尺度には、精度、再現率、F 値を使用した。

4.3　実験結果

実験結果を表 2 に示す。比較手法である SVM 手法では、特徴量に表情のみを使用した場合が、最も F 値が高かった。提案手法である LSTM 手法では、特徴量に音声と表情を使用した場合が、最も F 値が高かった。この 2 つの手法を、台詞固定と台詞自由のそれぞれの場合で比較する。台詞固定の場合、提案手法では、精度は同じ値だったものの、再現率を 0.38 ポイント、F 値を 0.13 ポイント向上させることができた。台詞自由の場合、提案手法では、精度を 0.09 ポイント、再現率を 0.22 ポイント、F 値を 0.14 ポイント向上させることができた。台詞固定と台詞自由の両パターンにおいて、精度を下げることなく、再現率を大幅に向上させることで、F 値を改善することができた。これは、比較手法では、音声、表情ともに事前に定義された特徴量を使用していたが、提案手法では、LSTM を利用することで、より細かい音声や表情の時系列での変化を把握することができたためであると考えらえる。

5　まとめ

本研究では、LSTM を利用し本心でない発話を自動検出する手法を提案した。SVM を利用した既存手法との比較実験を行い、提案手法の有効性を確認した。今後は、Bi-directional LSTM の利用を検討する予定である。

6　参考文献

[1]　上村譲史，口調と表情遷移に基づく抑圧された負の感情の検出，平成29年度広島市立大学修士論文 (2018).
[2]　F. Eyben, et al., openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor, ACM Multimedia Conference – MM,pp.1459-1462 (2010).
[3]　OKAO Vision | オムロン人画像センシングサイト, https://plus-sensing.omron.co.jp/technology/, (参照日 2018/7/18).
[4]　P. Ekman, Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition). W.W. Norton (2009).

表 2： 実験結果

手法	特徴量	台詞固定			台詞自由		
		精度	再現率	F 値	精度	再現率	F 値
SVM 手法（比較手法）	音声	0.58	0.60	0.59	0.52	0.51	0.51
	表情	0.64	0.57	0.60	0.55	0.61	0.58
	音声＋表情	0.57	0.58	0.57	0.54	0.53	0.53
LSTM 手法（提案手法）	音声	0.64	0.95	0.73	0.61	0.83	0.70
	表情	0.62	0.56	0.59	0.83	0.13	0.22
	音声＋表情	0.64	0.95	0.73	0.64	0.83	0.72