INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Social Signal Detection in Spontaneous Dialogue Using Bidirectional LSTM-CTC

Hirofumi Inaguma[1], *Koji Inoue*[1], *Masato Mimura*[1], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics Kyoto University, Japan
{inaguma, inoue, mimura, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Non-verbal speech cues such as laughter and fillers, which are collectively called social signals, play an important role in human communication. Therefore, detection of them would be useful for dialogue systems to infer speaker's intentions, emotions and engagements. The conventional approaches are based on frame-wise classifiers, which require precise time-alignment of these events for training. This work investigates the Connectionist Temporal Classification (CTC) approach which can learn an alignment between the input and its target label sequence. This allows for robust detection of the events and efficient training without precise time information. Experimental evaluations with various settings demonstrate that CTC based on bidirectional LSTM outperforms the conventional DNN and HMM based methods.

Index Terms: Social signals, connectionist temporal classification, long-short term memory, human-computer interaction, computational paralinguistics

1. Introduction

Non-verbal speech cues such as laughter and fillers, which are called social signals [1, 2, 3], play an important role in human-to-human communication. Detecting these events is useful for understanding speakers and informative for dialogue systems to behave like human. Moreover, it is expected that removing them leads to improving the performance of automatic speech recognition (ASR) in natural conversation.

Recently, the importance of detecting social signals attracts more attention and a number of conventional machine learning approaches such as Gaussian Mixture Model (GMM) [4], Genetic Algorithm (GA) [5], AdaBoost [6], and Hidden Marcov Model (HMM) [7] were used in previous studies. Motivated by the impressive success of neural networks in ASR, neural networks based approaches such as Deep Neural Network (DNN) [8], Convolutional Neural Network (CNN) [9], and Bidirectional Long-Short Term Memory (BLSTM) [10] have been introduced and shown to outperform other models. In these approaches, models are generally trained as frame-wise classifiers. However, they are not appropriate for this task from two points of view.

Firstly, in terms of information retrieval, our main purpose is to detect the occurrence of social signal events in a huge number of candidates [11], so we want to detect these events robustly on the event unit rather than the frame unit. Secondly, for the classification training, the frame-level target labels are required. It is expensive to make frame-level annotation manually, especially for social signals because their boundaries are unclear compared with utterance boundaries, and in the case of conducting forced alignment using the pre-trained classifier, the quality of the target labels depends on the model.

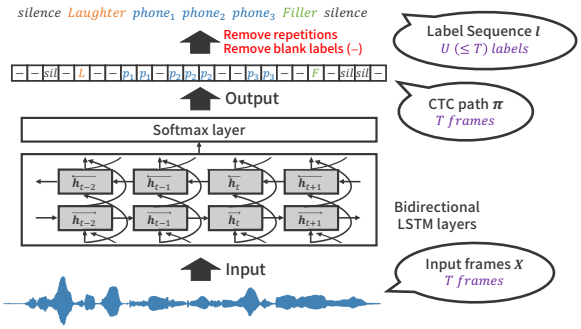

Figure 1: *Decoding in the BLSTM-CTC network*

To address these issues, in this study, we investigate direct detection of the occurrence of social signals in natural conversation using Connectionist Temporal Classification (CTC) [12] (see Figure 1), which conducts optimization over all possible frame-level sequences and has been recently successful in end-to-end speech recognition systems [13, 14, 15, 16]. The CTC approach removes the need to conduct segmentation in the training set and has potential of improving robustness of detection. As another end-to-end approach, the attention mechanism is also successful in machine translation [17], speech recognition [18], and image captioning [19]. But it is difficult to identify the occurrence timing of events by this approach, unlike CTC.

In this study, we confirm that the proposed CTC model outperforms the conventional frame-wise classifiers combined with HMM even without time information during training, and also investigate several methods for generation of the target labels using rough transcripts. We will show that the BLSTM-CTC model can detect social signal events in generally actual timing although the CTC algorithm does not guarantee the alignment of label spikes with the corresponding input frames. Furthermore, an additional experiment using more kinds of social signals is also conducted.

The remaining part of this paper is organized as follows. In Section 2, we describe the purpose of detecting social signals and related works. In Section 3, we describe the BLSTM-CTC model and how to generate the target labels for the CTC network. Section 4 describes the corpus and Section 5 details the experimental results. We conclude this paper in Section 6.

2. Detection of Social Signals

2.1. Social Signals

Social signals [1, 2, 3] are non-verbal speech cues, which carry information on speakers' mental states in natural conversation. In this study, we adopt laughter, filler, backchannel, and disfluency as social signal events because these events are easy to observe and express, and familiar to us. Each vocalization

http://dx.doi.org/10.21437/Interspeech.2017-457

has some important roles. Laughter relieves the meaning of the preceding utterance and helps speakers express their emotions and personalities [20, 21, 22]. Fillers (vocalizations like "uhm", "eh", and "ah" etc.) are used to hold the floor in order to recollect thoughts or prevent listeners from breaking the speaking turn [23]. Backchannel feedback (vocalizations like "yeah", "right", and "okay" etc.) is used to express that listers are paying attention and understanding, and encourage the speaker to continue [24]. Disfluency [25] has several forms such as repetitions, repairs and false starts.

Detecting these events is useful for inferring speakers' emotion states, intentions, personalities, and engagements. In addition, it can be informative for dialogue systems to behave as we do. Moreover, it is expected that removing them leads to improving of the performance of ASR in natural conversation. There is a "Chicken and Egg" problem between detection of social signals and the improvement of performance in ASR. When detecting social signals, it is expected that the detection performance is improved by separating social signals from other phonemes. For example, in order to recognize laughing utterances, it is necessary to detect laughter and recognize what is spoken. On the other hand, when recognizing speech in natural conversation, it is known that recognition errors often occur around social signals. Therefore, training them all together can lead to the improvement of both performances. We formulate a unified framework in Section 5.2.

2.2. Related Works

In the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) [26], social signal detection was one of main tasks and many approaches were proposed [4, 6]. They focused on frame-wise classification of social signal events based on Unweighted Average Area Under the Curve (UAAUC) metric. The best system of the challenge was based on DNN [8], and social signal detection receives more attention and various models have been proposed afterwards [5, 7, 27], especially neural network based models [9, 10]. These models are generally trained as frame-wise classifiers, and often show additional improvement of UAAUC by posterior smoothing. However, Gosztolya [11] mentions that smoothing posteriors drastically degrades the performance when evaluating them on the event unit using HMM. This is because likelihoods of laughter and filler events become relatively lower than the "garbage" class, which leads to high precision and low recall. Therefore we need to evaluate with precision, recall, and F-measure for event-level detection.

3. Bidirectional LSTM-CTC

3.1. Bidirectional Long-Short Term Memory

Recurrent Neural Networks (RNNs) can exploit context information, and among them Long-Short Term Memory (LSTM) [28], which can access long-range context by introducing memory blocks to regulate the flow of information, has become successful. By assuming that ASR transcribes an IPU unit, we leverage future context in the IPU as well. In this case bidirectional LSTM (BLSTM), where two separate forward and backward layers are fed into the same next output layer [13], can detect social signal events more accurately than unidirectional models.

3.2. Connectionist Temporal Classification (CTC)

In the conventional neural network training such as cross entropy or mean square error criterion, where the length of input frames is the same as that of its target label sequence, the training data must be aligned frame by frame to the input. The Connectionist Temporal Classification (CTC) [12] uses a loss function for sequence labeling where the input and the target label sequence have different lengths without pre-segmentation. CTC works together with RNNs and is applied after a softmax layer following RNNs. The key idea of CTC is to introduce a $blank$ label, which means the network emits no labels, and to suppress frame-wise outputs including repetitions of the same labels into the sequence of target outputs (e.g., phonemes or characters). Given an input sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, CTC trains the model to maximize the probability distribution $P(\boldsymbol{l}|\boldsymbol{X})$ for the corresponding target label sequence $\boldsymbol{l}$ of length $U(\leq T)$. This distribution is represented by a summation of all possible frame-level intermediate representations $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ (here after, CTC path):

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

where $\boldsymbol{\Phi}(\boldsymbol{l})$ is the set of CTC paths allowing insertion of blank labels and repetition of non-blank labels to $\boldsymbol{l}$, i.e., $\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}) = \boldsymbol{l}$. Then, if $l_u \in L = \{1, \ldots, K\}$, the softmax layer is composed of $|L \cup \{blank\}| = K + 1$ units. Based on the conditional independence assumption, the posterior $P(\boldsymbol{\pi}|\boldsymbol{X})$ is decomposed as follows:

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

where y_k^t is the k-th output of the softmax layer at time t, which denotes the occurrence probability of the corresponding label. The probability distribution $P(\boldsymbol{l}|\boldsymbol{X})$ is computed efficiently with the forward-backward algorithm.

For example, in the case that a laughter ($Laughter$) and a filler event ($Filler$) occur across an utterance ($phone_1$ $phone_2$ $phone_3$) in this time order, the target label sequence is given as follows:

$$silence\ Laughter\ phone_1\ phone_2\ phone_3\ Filler\ silence$$

Note that $silence$ means silence, and $phone_i$ means the corresponding sequential unit in the utterance.

3.3. Generation of Training Labels for CTC

There are options on how to represent segments other than social signal events. They are usually subword segments such as phonemes or silence, but in the SVC corpus they are collectively annotated as garbage. Though the CTC network can directly learn the alignments between input and target label sequences, target labels corresponding to acoustic events in the input are required for each utterance. Thus, it is necessary to classify garbage into the corresponding subword units, at least speech or silence. Since there are not any transcripts in the SVC corpus, we try to estimate approximate subword units by the following method. At first, we conduct speech recognition on each audio clip using Kaldi Toolkit [29] and get sequences of 41 kinds of subword labels (40 phones and silence) with time information. We use the acoustic model and language model trained with the

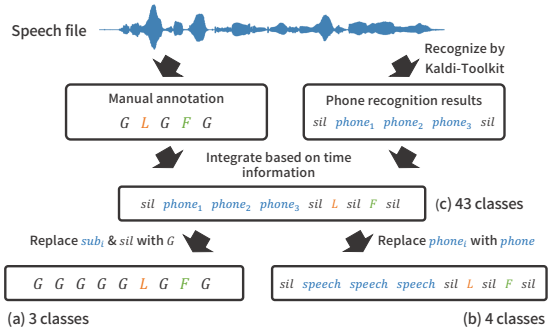

Figure 2: *How to generate the target labels to the CTC network*

WSJ (Wall Street Journal) corpus. Next, these resulting subword labels and manually annotated laughter and filler labels are integrated along the time sequence, which makes 43 kinds of labels in total (40 phones + silence + laughter + filler). Finally, three replacement patterns are considered as follows (see Figure 2):

(a) replacing all phone labels and silence label with garbage labels,
- 3 classes: garbage + laughter + filler

(b) replacing all phone labels with single speech labels,
- 4 classes: speech + silence + laughter + filler

(c) using all subword labels as is.
- 43 classes: 40 phones + silence + laughter + filler

In this study, we used the time annotation of laughter and filler for alignment with ASR results in order to generate the final label sequence, but these processes are not needed if we have transcripts. In fact, we do not use time information at all in the experiment in Section 5.2.

4. Corpus

4.1. The SSPNet Vocalization Corpus (SVC)

The SSPNet Vocalization Corpus (SVC) was used in the Social Signals Sub-Challenge of the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) [26]. The corpus was made from a collection of 60 phone calls involving 120 subjects (63 female, 57 male), where they were fully unacquainted, and composed of 2763 audio clips (total 8.4h). Each clip lasts for 11 seconds and contains at least one laughter or filler event between $t = 1.5$ and $t = 9.5$ seconds (the voice of one speaker only). Overall, the corpus contains 1158 laughter events (3.6%) and 2988 filler events (4.9%), and garbage (including speech and silence) with time information, which were manually annotated. Both types of vocalization can be considered fully spontaneous. The data were divided into speaker disjoint subsets for training (70 speakers, 4.8h), development (20 speakers, 1.5h), and testing (30 speakers, 2.1h), respectively.

4.2. ERATO Human-Robot Interaction Corpus

This corpus is a collection of Japanese face-to-face spontaneous dialogue with an android ERICA [30], which was remotely operated by 6 amateur actresses. 91 sessions were recorded and each session lasts about ten minutes (total 16.8h). There are 91 subjects who talked freely with ERICA. ERICA had a various social roles and subjects were engaged in dialogue in the corresponding social situation. The utterances of operators were recorded using a stand microphone on the table, and those of

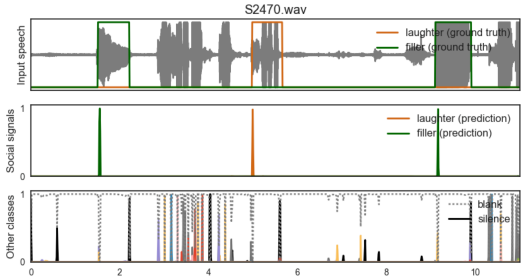

Figure 3: *The output of softmax layer of the BLSTM-CTC model in the SVC corpus. The x-axis shows the time and y-axis shows the label posteriors by softmax layers. This model was trained using 43 class labels (40 phones + silence + laughter + filler) in rough transcripts.*

subjects were recorded using a directional microphone. Transcripts and four kinds of labels of social signal events are manually annotated: 984 laughters, 8609 fillers, 6293 backchannels, and 1204 disfluencies. The data were divided into training (13.4h), development (1.3h), and testing (2.1h) subsets, respectively.

5. Experiments

5.1. Evaluation in SVC

5.1.1. Experimental setup

At first, we conduct experiments using the SVC corpus described in Section 4.1. The input features for the BLSTM-CTC models are 40-channel log-mel filterbank outputs and log energy ($+\Delta, \Delta\Delta$), computed every 10 ms. Each input frame is a 123-dimensional vector. The features are normalized by the mean and the standard deviation over the training set. The BLSTM network consists of 5 bidirectional LSTM layers with 256 memory cells per layer and the softmax layer. Optimization was performed on minibatches of 64 utterances using RMSProp with learning rate 10^{-3}. All the weights were initialized with the range $[-0.1, 0.1]$ of uniform distribution. Note that bias vectors of the forget gates were initialized with 1.0 as in [31]. The dropout ratio of input-hidden and hidden-hidden layers were 0.8 and 0.5, respectively. All the networks are implemented with TensorFlow [32].

5.1.2. Results

Results in the SVC corpus are shown in Table 1. We adopt precision, recall, F-measure and their average over all social signal events as evaluation metrics. Note that we predict social signal events in case that the label spikes of CTC outputs (posteriors) exceed 0.5 and regard them as correct detection only when they are included in the corresponding true intervals, for severe comparison with other frame-wise classifiers. We compare the CTC model with DNN and AdaBoost models combined with HMM whose state transitional probability values are uniform for the sake of simplicity because they are evaluated on the event unit in [11], Although the CTC model does not use time information for the training stage, it outperforms both AdaBoost and DNN models, which use time information. When comparing the three label generation method (a), (b), and (c) mentioned in Section 3.3, contrary to our expectation, using more classes do not lead to improving the accuracy. This is because the target subword labels are obtained by ASR, thus not so accurate. As shown in Figure 3, however, we can confirm that the CTC network can

Table 1: *Accuracy for the event unit detection in the SVC corpus (3 class classification)*

Model	Laughter			Filler			F-measure Average
	Precision	Recall	F-measure	Precision	Recall	F-measure	
AdaBoost-HMM [11]	0.58	0.74	0.65	0.65	0.71	0.68	0.66
DNN-HMM [11]	0.58	0.72	0.64	0.71	0.60	0.65	0.65
BLSTM-CTC ((a) 3 classes)	**0.65**	**0.66**	**0.66**	**0.66**	**0.80**	**0.72**	**0.69**
BLSTM-CTC ((b) 4 classes)	0.60	0.49	0.54	0.59	0.78	0.67	0.61
BLSTM-CTC ((c) 41 classes)	0.79	0.51	0.62	0.71	0.78	0.74	0.68

$word_1$ ($Laughing$ $word_2$) $word_3$ → $word_1$ L $word_2$ $word_3$

$word_1$ ($Backchannel$ $word_2$) $word_3$ → $word_1$ B $word_2$ $word_3$

$word_1$ ($Filler$ $word_2$) $word_3$ → $word_1$ F $word_2$ $word_3$

$word_1$ ($Disfluecy$ $word_2$) $word_3$ → $word_1$ D $word_2$ $word_3$

Figure 4: *How to insert social signal labels to the corresponding words*

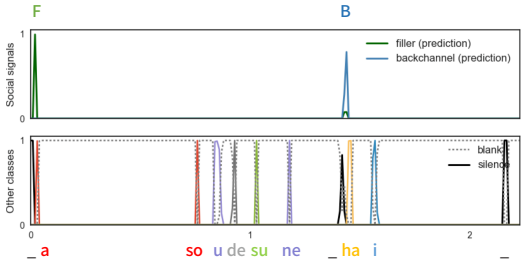

Figure 5: *The output of softmax layer of the BLSTM-CTC model in the ERATO corpus. The x-axis shows the time and y-axis shows the label posteriors by softmax layers. True labeling is "_ F ha so u de su ne _ B ha i _". "_" means a silence label.*

learn the alignment between input and target labels although it is trained using rough transcripts.

5.2. Evaluation in ERATO HRI corpus

5.2.1. Experimental setup

In addition, we conduct experiments using the ERATO Corpus described in Section 4.2. The experimental setting of the BLSTM-CTC model is the same as in Section 5.1. In this corpus, we annotate additional social signal events of backchannel and disfluency. The target labels are composed of 83 kinds of kana characters, silence, and 4 kinds of social signals, 88 in total. There is a large variation in the utterance length in the ERATO corpus, so we sort all utterances in the training set by length to stabilize the training. In regard with label generation of social signals, we insert each social signal label in front of the corresponding word (see Figure 4). This label generation can be suitable for both social signal detection and speech recognition.

5.2.2. Results

We evaluate detection accuracies in the entire utterance because the ERATO corpus does not have frame-level annotation. We regard as correct detection when the predicted social signal labels are included in the corresponding target label sequence. Table 2 shows detection results in the ERATO corpus. Compared with in Section 5.1.2, fillers were detected with comparable high accuracy, but the accuracy of laughter is slightly lower. This is because that the ERATO corpus has more laughing utterances than the SVC corpus. Backchannels were detected more accurately than fillers. However, disfluency could hardly be detected. Disfluency should be detected by considering not only acoustic but also linguistic features with a language model [25]. Figure 5 show an example of the CTC outputs. From this, it is observed that the CTC network can learn not only subwords but also whether they are social signals or not.

Table 2: *Accuracy for the event unit detection in the ERATO corpus*

Social signals	Precision	Recall	F-measure
Laughter	0.89	0.35	0.50
Filler	0.75	0.75	0.75
Backchannel	0.86	0.87	0.86
Disfluency	0.44	0.15	0.22

Moreover, in order to see that considering social signals when constructing the acoustic model leads to the improvement of the ASR performance, we evaluate two BLSTM-CTC models based on Character Error Rate (CER). When evaluating the model considering social signals ($Insert$), we acquire the output label sequences of the CTC model at first, then remove social signal labels, before compute CER. From Table 3, CER by the CTC acoustic model considering social signals outperforms the conventional CTC acoustic model which does not consider them. Therefore, we conclude that acoustic models should be constructed considering social signals in the case that many non-verbal cues are observed such as in spontaneous dialogue. We will compare the CTC model with the hybrid model such as DNN-HMM and BLSTM-HMM in the future.

Table 3: *Character Error Rate (CER). Insert means the CTC model which is trained considering social signals events.*

	CER
BLSTM-CTC	19.1 %
BLSTM-CTC ($Insert$)	18.6 %

6. Conclusions

In this paper, we address detection of social signals of the event unit by using the BLSTM-CTC model. We confirmed that the proposed CTC model outperformed several conventional models combined with HMM even without time information for the training stage. The CTC approach has the advantages of not only removing the requirement of pre-alignment between input and its target label sequence but also making detection more robustly. In addition, although the CTC algorithm does not guarantee the alignment of label spikes with the corresponding input sequence, the alignments are generally matched with the actual timing of the occurrence of social signal events. Finally, another experiment using the corpus more rich annotation of social signals was conducted and consistent results was confirmed. For future work, we will construct a language model considering the statistical nature of the occurrence of social signals.

7. References

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

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

[3] P. Brunet and R. Cowie, "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] T. F. Krikke and K. P. Truong, "Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech," in *Proceedings of Interspeech*, 2013, pp. 163–167.

[5] G. Gosztolya, "Detecting laughter and filler events by time series smoothing with genetic algorithms," in *Proceedings of International Conference on Speech and Computer*, 2016, pp. 232–239.

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

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

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

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

[10] R. Brueckner and B. Schulter, "Social signal classification using deep blstm recurrent neural networks," in *Proceedings of ICASSP*, 2014, pp. 4823–4827.

[11] G. Gosztolya, "On evaluation metrics for social signal detection," in *Proceedings of Interspeech*, 2015, pp. 2504–2508.

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

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

[14] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *Proceedings of ASRU*. IEEE, 2015, pp. 167–174.

[15] K. Rao, A. Senior, and H. Sak, "Flat start training of cd-ctc-smbr lstm rnn acoustic models," in *Proceedings of ICASSP*. IEEE, 2016, pp. 5405–5409.

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

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

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

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[20] J. Bachorowski, M. Smoski, and M. Owren, "The acoustic features of human laughter," *Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[21] J. Vettin and D. Todt, "Laughter in conversation: Features of occurrence and acoustic structure," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 93–115, 2004.

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

[23] H. Clark and J. F. Tree, "Using "uh" and "um "in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[24] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[25] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional lstm," in *Proceedings of Interspeech*, 2016, pp. 2523–2527.

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

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

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584, 2011.

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

[31] Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, "An empirical exploration of ctc acoustic models," in *Proceedings of ICASSP*, 2016, pp. 2623–2627.

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