AN END-TO-END APPROACH TO JOINT SOCIAL SIGNAL DETECTION AND AUTOMATIC SPEECH RECOGNITION

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

[1]Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

Social signals such as laughter and fillers are often observed in natural conversation, and they play various roles in human-to-human communication. Detecting these events is useful for transcription systems to generate rich transcription and for dialogue systems to behave as we do such as synchronized laughing or attentive listening. We have studied an end-to-end approach to directly detect social signals from speech by using connectionist temporal classification (CTC), which is one of the end-to-end sequence labelling models. In this work, we propose a unified framework that integrates social signal detection (SSD) and automatic speech recognition (ASR). We investigate several reference labelling methods regarding social signals. Experimental evaluations demonstrate that our end-to-end framework significantly outperforms the conventional DNN-HMM system with regard to SSD performance as well as the character error rate (CER).

Index Terms— Automatic speech recognition, social signals, connectionist temporal classification, end-to-end training

1. INTRODUCTION

Social signals [1–3] which are used to accompany linguistic content in utterances suggest mental states such as reactions and thinking. They include laughter, fillers and backchannels. These play an important role in human communication, and thus will be useful for intelligent machines to interpret the user's mental states and generate appropriate responses.

Recently, the importance of detecting social signals has attracted more attention and various conventional machine learning approaches have been investigated [4–11]. These models are generally trained as frame-wise classifiers, and pre- and post-processing are required. On the other hand, as mentioned in [12], we need to detect the occurrence of social signal events robustly on the event-level metric rather than frame-level metric for real-world applications. We have studied an end-to-end approach to directly detect the occurrence of social signals from speech [13] using bidirectional long-short term memory (BLSTM) [14] with connectionist temporal classification (CTC) loss [15], which has been successful in end-to-end speech recognition [16–19] and allows for optimizing model parameters without pre-segmentation of target labels by marginalizing probabilities of all possible frame-level alignments.

Although social signal detection (SSD) and automatic speech recognition (ASR) have complementary relationships, they have been dealt with independently. With regard to SSD, it is expected that rich information of not only occurrences but also types of fillers or disfluencies and transcriptions of laughing utterances are acquired by the joint modeling of social signals together with other phonetic or morphological information, which would lead to the improvement of detection performance. At the same time, with regard to ASR, it is also expected that auxiliary information of utterances such as social signals helps ASR performance improve.

In this study, we propose a unified framework that integrates SSD and ASR by utilizing the potentials of CTC. Modeling this framework in an end-to-end manner leads to a simple architecture without any components such as special language models for social signals. We expect that joint modeling makes it possible to capture boundaries of social signals while recognizing subword units, which leads to the improvement of performance of both tasks. Thus, we investigate several reference labelling methods regarding social signals. Our end-to-end framework significantly outperforms the conventional cascaded model, where social signals are detected by a language model following a DNN-HMM acoustic model, in both the SSD and character-level ASR performance. We also show that the joint framework of SSD and ASR leads to rich transcription including social signals without degradation of ASR performance.

2. SOCIAL SIGNAL DETECTION (SSD)

2.1. Roles of social signals

Social signals [1–3] are useful for estimating a speaker's mental states, such as emotions, engagements, personalities, and intention. They are informative for dialogue systems to generate human-like behaviors such as synchronized laughing and attentive listening. Social signals are composed of behaviors of various modalities and classified into audio (vocalizations) and visual information (postures, gestures, facial expressions, and gaze etc.). In this study, we focus on four vocal social signal events: laughter, fillers, backchannels, and disfluencies because these are easy to observe in natural conversation and familiar to us.

Each vocalization has some important roles. Laughter relieves the meaning of the preceding utterance and helps speakers express their emotions and personalities [20–23]. Fillers (vocalizations like "uhm", "eh", and "ah" etc.) are used to hold the floor for recollecting thoughts or preventing listeners from breaking the speaking turn [24]. Backchannels (vocalizations like "yeah", "right", and "okay" etc.) are used to express that listeners are paying attention, understanding, or showing agreement, and to encourage the speaker to continue [25]. Disfluencies [26] suggest a trouble in utterance generation, and have several forms such as repetitions, repairs and false starts.

2.2. Related work

Social signal detection (SSD) has attracted much recent attention, including being selected as one of the main tasks in the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) [27], and a number of approaches have been investigated such as Gaussian mixture model (GMM) [4,5], genetic algorithm (GA) [6], AdaBoost

[7], and hidden Marcov model (HMM) [8]. As in the fields of ASR, recently deep learning approaches have been impressively successful [9–11, 28]. These models are generally trained as frame-wise classifiers and evaluated by the frame-level metric such as Area Under the Curve (AUC), but detecting social signals frame by frame is not effective from three viewpoints.

Firstly, in terms of information retrieval, our main purpose is to detect the occurrence of social signal events robustly among various utterances [12]. Secondly, in the training stage, frame-level target labels are required because the length of input speech frames must be the same as that of its target label sequence. It is expensive to make frame-level annotation especially for social signals because their boundaries are unclear compared with utterance boundaries, so it is prone to subjective factors to decide their boundaries. If we conduct forced alignment, the quality of the target labels depends on the pre-trained classifier. Thirdly, in the detection stage, post-processing such as threshold processing and smoothing using HMM is also needed after frame-level classification.

Therefore, we have investigated an end-to-end approach to directly detect social signals from speech using BLSTM-CTC [13]. We have confirmed that this approach leads to robust detection of social signals on the event-level metric such as F_1 score.

3. JOINT SOCIAL SIGNAL DETECTION AND AUTOMATIC SPEECH RECOGNITION

3.1. System overview

Social signal detection (SSD) and automatic speech recognition (ASR) are considered to be in a complementary relationship. However, they have been treated as separate problems conventionally.

As mentioned above, some studies with regard to SSD addressed direct detection of laughter and fillers from speech [4–11], but they have dealt with only the occurrences of social signals, and their types or transcription have not been considered. Other studies [26, 29] investigated extraction of fillers and disfluencies from ASR results in a cascaded manner, which depends on ASR performance and their processing is complicated. By the joint modeling of social signals together with other phonetic or morphological information, it is expected that detection performance would be improved.

On the other hand, from the standpoint of ASR, it is generally difficult to recognize utterances around ambiguous speech frames such as social signals. Some of them such as laughter can be modeled with non-speech classes, but many of them such as fillers and disfluencies have segmental information that can be transcribed but hardly modeled except for typical fillers and backchannels. Thus, it is also expected that auxiliary information of social signals helps improve ASR performance.

In this work, we focus on the complementary relationship between SSD and ASR, and propose a unified framework where social signals are directly detected from speech while recognizing subword units based on BLSTM-CTC. We expect this framework leads to the improvement of both performances. In addition, a more simplified architecture without any special components for social signals can be realized by an end-to-end modeling. If we can distinguish between social signals and subword units, rich information will be obtained regarding not only occurrences but also types or transcription of the social signals.

The overall system architecture is shown in Figure 1. A label sequence, which includes both social signals and subword units, is decoded from outputs of the softmax layer following stacked BLSTM layers. This is used for the SSD task. For the ASR task, the final transcript is obtained by removing all social signal labels from the label sequence.

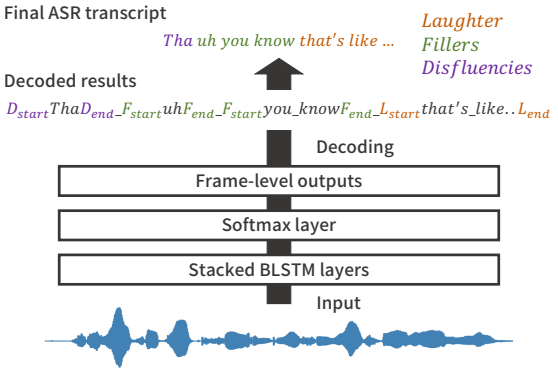

Fig. 1: System overview

3.2. Connectionist temporal classification (CTC)

Connectionist temporal classification (CTC) [15] is an objective function for sequence labelling problems where the input and target label sequence have different lengths, and allows for learning an alignment between them without pre-segmentation. To bridge the gap between input and target label sequence lengths, CTC introduces an extra *blank* label, which means the network emits no label at a given time, and also allows repetitions of the same labels possibly interleaved with blank labels. A class corresponding to blank labels is added to nodes in the softmax layer. Therefore, the output nodes are composed of subword units, social signals, and a blank label. Given an input sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and the corresponding target labels $\boldsymbol{l} = (l_1, \ldots, l_U)$ $(U \leq T)$, the CTC network can be trained to optimize the negative log probability using the probability distribution $P(\boldsymbol{l}|\boldsymbol{X})$ (network outputs) based on the maximum likelihood criterion. Thus, the objective function is formulated as follows:

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

Here, $P(\boldsymbol{l}|\boldsymbol{X})$ is marginalized by a summation of probabilities of all possible frame-level alignments.

$$P(\boldsymbol{l}|\boldsymbol{X}) = \sum_{\boldsymbol{\pi} \in \boldsymbol{\Phi}^{-1}(\boldsymbol{l})} P(\boldsymbol{\pi}|\boldsymbol{X})$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ is output labels of the softmax layer, and this intermediate representation including blank labels is called a *CTC path*. $\boldsymbol{\Phi}^{-1}(\boldsymbol{l})$ is the set of all CTC paths, and $\boldsymbol{\Phi}$ is a many-to-one collapsing function to suppress repeated labels and then remove blank labels, i.e., $\boldsymbol{\Phi}(\boldsymbol{\pi}) = \boldsymbol{l}$. Assuming the conditional independence of outputs at each time step, $P(\boldsymbol{\pi}|\boldsymbol{X})$, the probability distribution of a CTC path, 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. $P(\boldsymbol{l}|\boldsymbol{X})$ is computed efficiently with the forward-backward algorithm as in HMM.

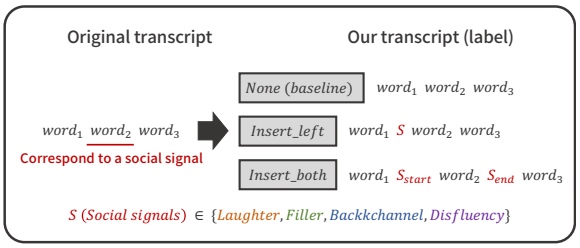

Fig. 2: Reference label generation regarding social singals. S means a social signal label, which is used in $Insert\_left$. S_{start} and S_{end} mean start and end labels of the corresponding social signal, which are used in $Insert\_both$.

3.3. Generation of reference labels

This section describes how to generate reference labels regarding social signals for both SSD and ASR tasks. We consider three labelling methods as follows (see Figure 2).

None (baseline) The same reference labelling as the conventional end-to-end speech recognition systems, where social signals are not considered.

Insert_left Each social signal label is inserted on the left side of the corresponding subword units. This is intended to mark the beginning of the social signals. At least one acoustic frame must correspond to these labels in the CTC framework. We presume some acoustic cues exist around the social signals.

Insert_both The start and end labels of each social signal are inserted on both sides of the corresponding subword units. By assuming that there is some acoustic cues at the end as well, we expect the model to learn rough segmentation of the social signals.

4. EXPERIMENTAL EVALUATION

4.1. System settings

The input features were 40-channel log-mel filterbank outputs plus energy and their delta and acceleration coefficients, computed every 10 ms. Thus, each input frame was a 123-dimensional vector. The features were normalized by the mean and the standard deviation on the speaker basis. In addition, each set of 3 frames (total 30ms) was stacked and concatenated for reducing the arbitrariness of the alignment of CTC training [30], which resulted in reducing the input frame length by a factor of 3. The network consisted of 5 stacked BLSTM layers with 256 memory cells (320 in Section 4.3) per direction and the softmax layer. Optimization was performed on mini-batches of 64 utterances (32 in Section 4.3) using Adam [31] with a learning rate 1.0×10^{-3}. For stable training, all utterances in the training set were sorted by their lengths in the early training stage [17, 19]. All weights were initialized with random values drawn from a uniform distribution with a range $[-0.1, 0.1]$. Bias vectors of the forget gates in each LSTM layer were initialized with 1.0 [32]. We also clipped the norms of gradients and cell activations so that they have maximum absolute values 5 and 50, respectively. The dropout ratio was 0.5 (0.8 in Section 4.3). Beam search decoding was performed with a beam width 20. All networks were implemented with a TensorFlow framework [33]. Note that we did not use any language models.

Table 1: Accuracy for the social signal detection in the ERATO corpus. Prec., Rec., and F_1 stand for precision, recall, and F_1 score (F-measure), respectively.

Labelling	Social Signals	Prec.	Rec.	F_1	Ave. F_1
$Insert\_left$	Laughter	0.88	0.44	0.59	0.59
	Filler	0.73	0.79	**0.76**	
	Backchannel	0.90	0.75	**0.82**	
	Disfluency	0.42	0.12	0.19	
$Insert\_both$	Laughter	0.80	0.53	**0.64**	**0.60**
	Filler	0.70	0.83	**0.76**	
	Backchannel	0.88	0.70	0.78	
	Disfluency	0.41	0.17	**0.24**	

Table 2: Character error rate (CER) in the ERATO corpus. CERs in $Insert\_left$ and $Insert\_both$ are computed by removing all social signal labels from decoded results.

Labelling	CER (%)
$None\ (baseline)$	19.41
$Insert\_left$	19.80
$Insert\_both$	19.69

4.2. Evaluation on dialogue corpus

4.2.1. Experimental conditions

In this section, we conduct experiments using the ERATO Human-Robot Interaction Corpus (ERATO corpus), which is a collection of Japanese face-to-face spontaneous dialogue corpus recorded with an autonomous android ERICA [34]. ERICA was remotely operated by 6 amateur actresses. There are 91 sessions and each session lasts about 10 minutes (total 14.7 hours). The subjects talked freely with remotely operated ERICA. ERICA had various social roles and each subject was engaged in conversation in the corresponding situation. Recording of operators and subjects were conducted by using a stand microphone on the table and a directional microphone, respectively. Transcripts and four kinds of social signal events were manually annotated: laughter, fillers, backchannels, and disfluencies[1]. Reference labels were composed of 145 kinds of Japanese kana characters, 4 kinds of social signals, space, and noise. We adopted Japanese kana characters as targets instead of phones because Japanese kana characters include both phonetic and morphological information. The whole corpus was divided into training (11.8 hours), development (1.3 hours), and testing subsets (1.6 hours).

4.2.2. Experimental results

Results of SSD experiments using the ERATO corpus are shown in Table 1. We followed [12] and adopted precision, recall, F_1 score (F-measure), and their average over all social signal events as evaluation metrics. Fillers and backchannels were detected with comparable high accuracy, and laughter was also detected to some extent although recall was slightly lower. It is very difficult to detect laughing utterances, where laughing voices and utterances are overlapped. The detection performance of disfluencies is also very low due to the insufficient training data to cover a large variation of them. However, $Insert\_both$ works better than $Insert\_left$ in detecting laughter and disfluencies. These social signals are longer and usually followed by a short pause, so the ending mark is effective.

[1]The number of social signals contained in the ERATO corpus is laughter: 1131/183, fillers: 8348/741, backchannels: 3424/687, and disfluencies: 1231/163, respectively (train/test).

Table 3: Accuracy for the social signal detection in the CSJ

Model	Labelling	Social Signals	eval1			eval2			Ave. F_1
			Prec.	Rec.	F_1	Prec.	Rec.	F_1	
CTC (w/o LM)	$Insert\_left$	Filler	0.95	0.92	0.93	0.93	0.94	0.93	0.93
		Disfluency	0.75	0.54	0.63	0.63	0.55	0.59	0.61
	$Insert\_both$	Filler	0.95	0.93	0.94	0.93	0.93	0.93	**0.94**
		Disfluency	0.75	0.58	0.65	0.67	0.58	0.62	**0.64**
DNN-HMM (w/ 3-gram)	—	Filler	0.88	0.91	0.89	0.83	0.90	0.86	0.88
		Disfluency	0.54	0.27	0.36	0.45	0.24	0.31	0.34

In addition, we investigated whether the joint modeling of SSD and ASR led to the improvement of ASR performance. Table 2 shows the character-level ASR results. We adopted character error rate (CER) as evaluation metric. Although CER was not improved in this corpus even when considering social signals, there was no significant difference among the three labelling methods. This means that CTC could detect laughter, fillers, and backchannels robustly without degradation of ASR performance.

4.3. Evaluation on large-scale lecture corpus

4.3.1. Experimental conditions

In this section, we conduct experiments using a large-scale spontaneous speech corpus, Corpus of Spontaneous Japanese (CSJ) [35], which is one of the largest Japanese speech corpora. The CSJ consists of about 600 hours of spontaneous speech including academic and simulated lectures, but in this work we focus on the academic lectures which have been the major target of ASR research using this corpus, consisting of about 240 hours of training data in total. There are two evaluation sets (eval1 and eval2), each of which is composed of 10 lectures. We picked up different 19 lectures as the development set. Three types of social signals, laughter, fillers, and disfluencies were annotated in the CSJ, and we used 150 kinds of reference labels including Japanese kana characters, filler, disfluency, space, and noise. Since there are few laughter in academic lectures, we excluded laughter from the SSD[2].

4.3.2. Experimental results

Results of SSD in the CSJ are shown in Table 3. Fillers were detected with high accuracy, and the accuracy of disfluencies also improved thanks to sufficient data compared with Section 4.2. Comparing $Insert\_left$ and $Insert\_both$, $Insert\_both$ slightly outperformed $Insert\_left$ in filler and disfluency detection as in Section 4.2.

In addition, we compared the detection accuracy of fillers and disfluencies with a hybrid system, where social signals were detected by a 3-gram language model and a DNN-HMM acoustic model. Fillers and disfluencies were treated as separate words and added into the dictionary. The DNN-HMM acoustic model was composed of six hidden layers with 2048 nodes and an output layer with 3k nodes, and trained using 240 hour data (the same training data as CTC models). Sequence discriminative training was also performed using sMBR criterion. Word 3-gram model was trained using all data in the CSJ (about 600 hours). We observed that the end-to-end framework by BLSTM-CTC significantly outperformed the hybrid system in both cases of filler and disfluency. It is confirmed that disfluencies are hardly covered by the language model while fillers are easily covered.

[2]The number of social signals contained in the CSJ is filler: 442,930/1,720/1,279 and disfluencies: 98,356/388/355, respectively (train/eval1/eval2).

Table 4: Character error rate (CER) in the CSJ

Model	Labelling	CER (%)		
		eval1	eval2	Ave.
CTC (w/o LM)	$None\ (baseline)$	7.70	6.11	6.90
	$Insert\_left$	8.11	6.36	7.23
	$Insert\_both$	8.18	6.34	7.26
DNN-HMM (w/ 3-gram)	—	8.65	7.44	8.04

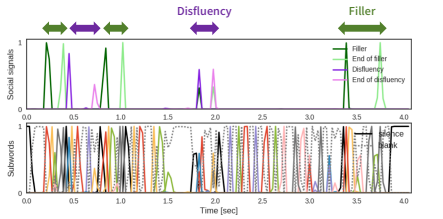

Fig. 3: The output posteriors of CTC ($Insert\_both$) in the CSJ. We can confirm that the CTC could conduct rough segmentation of both fillers and disfluencies.

Next, we evaluated the performance of character-level speech recognition. Table 4 shows ASR results. There was not statistically significant differences among three labelling methods of CTC models, but all of them significantly outperformed the DNN-HMM system in ASR performance as well as SSD performance. Note that BLSTM-CTC did not use a word-level lexicon and language models explicitly while the DNN-HMM system used a word 3-gram language model. We can conclude that our end-to-end framework is suitable for solving both SSD and ASR tasks at the same time, and this leads to robust detection of social signals with ASR performance enhancement.

Figure 3 shows examples of the CTC outputs (posteriors) with $Insert\_both$ labelling method. It was observed that the BLSTM-CTC could recognize not only subword units but also detect boundaries of social signals simultaneously. We also found that the CTC could conduct rough segmentation of social signals by marking both the beginning and end points of them.

5. CONCLUSION

In this paper, we have proposed a unified framework that integrates social signal detection (SSD) and automatic speech recognition (ASR) based on connectionist temporal classification (CTC), which is one of the end-to-end models. We also investigated several reference labelling methods regarding social signals and confirmed that our end-to-end framework by BLSTM-CTC significantly outperformed the conventional DNN-HMM system with a language model in both SSD and ASR performance. CTC could identify rough locations of social signals. We also found that this framework leads to rich transcription including social signal information without degradation of ASR performance with two speech corpora. For future work, we will implement the attention-based model [36, 37], which is another end-to-end model, to capture relationships more explicitly between subword units and social signals.

6. REFERENCES

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

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

[3] Paul Brunet and Roderick 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] Teun F Krikke and Khiet 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] Artur Janicki, "Non-linguistic vocalisation recognition based on hybrid GMM-SVM approach," in *Proceedings of Interspeech*, 2013, pp. 153–157.

[6] Gábor 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.

[7] Gábor Gosztolya, Róbert Busa-Fekete, and László Tóth, "Detecting autism, emotions and social signals using adaboost," in *Proceedings of Interspeech*, 2013, pp. 220–224.

[8] Hugues Salamin, Anna Polychroniou, and Alessandro 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.

[9] Rahul Gupta, Kartik Audhkhasi, Sungbok Lee, and Shrikanth Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking," in *Proceedings of Interspeech*, 2013, pp. 173–177.

[10] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen, "Laughter and filler detection in naturalistic audio," in *Proceedings of Interspeech*, 2015, pp. 2509–2513.

[11] Raymond Brueckner and Björn Schulter, "Social signal classification using deep BLSTM recurrent neural networks," in *Proceedings of ICASSP*, 2014, pp. 4823–4827.

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

[13] Hirofumi Inaguma, Koji Inoue, Masato Mimura, and Tatsuya Kawahara, "Social signal detection in spontaneous dialogue using bidirectional LSTM-CTC," in *Proceedings of Interspeech*, 2017, pp. 1691–1695.

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

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of ICML*, 2006, pp. 369–376.

[16] Alex Graves, Abdelrahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of ICASSP*, 2013, pp. 6645–6649.

[17] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *Proceedings of ASRU*. IEEE, 2015, pp. 167–174.

[18] Kanishka Rao, Andrew Senior, and Haşim Sak, "Flat start training of CD-CTC-SMBR LSTM RNN acoustic models," in *Proceedings of ICASSP*. IEEE, 2016, pp. 5405–5409.

[19] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

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

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

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

[23] Willibald Ruch and Paul Ekman, "The expressive pattern of laughter," *Emotion, qualia, and consciousness*, pp. 426–443, 2001.

[24] Herbert Clark and Jean Fox Tree, "Using "uh" and "um" in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

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

[26] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi, "Disfluency detection using a bidirectional LSTM," in *Proceedings of Interspeech*, 2016, pp. 2523–2527.

[27] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013, pp. 148–152.

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

[29] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[30] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proceedings of Interspeech*, 2015, pp. 1468–1472.

[31] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] Yajie Miao, Mohammad Gowayyed, Xingyu Na, Tom Ko, Florian Metze, and Alexander Waibel, "An empirical exploration of CTC acoustic models," in *Proceedings of ICASSP*, 2016, pp. 2623–2627.

[33] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[34] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara, "Talking with erica, an autonomous android," in *Proceedings of SIGDIAL*, 2016, pp. 212–215.

[35] Kikuo Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[36] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proceedings of ICASSP*, 2016, pp. 4945–4949.

[37] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.