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

Background

Social signals

✓ Laughter
✓ Filler (”uh-huh”, “eh” etc.)
✓ Backchannel (”yeah”, “right” etc.)
✓ Disfluency

➡

◆ Useful for estimating speaker’s mental states
  ✓ emotions, engagements, personalities, intention
◆ Informative for dialogue systems to generate human-like behaviors
  ✓ attentive listening, synchronous laughing
◆ Rich annotation for subsequent tasks
  ✓ text normalization, spoken language understanding (SLU), syntax parsing

Motivation

✓ SSD and ASR have been treated as separate problems conventionally
✓ However, they are in the complementary relationship

SSD (social signal detection)

◆ Detection from speech [Schuller+ 2013]
  ✓ Types of social signals or transcription have not been considered (occurrence only)
◆ Detection from ASR results in a cascaded manner
  ✓ Depend on ASR performance
  ✓ Complicated systems

→ The joint modeling with phonetic or morphological information would lead to the improvement of SSD performance

⬌

ASR (automatic speech recognition)

◆ Difficult to recognize utterances around social signals
◆ Fillers and disfluencies have countless forms
  ✓ Difficult to model all of them

→ Auxiliary information of social signals would help improve ASR performance

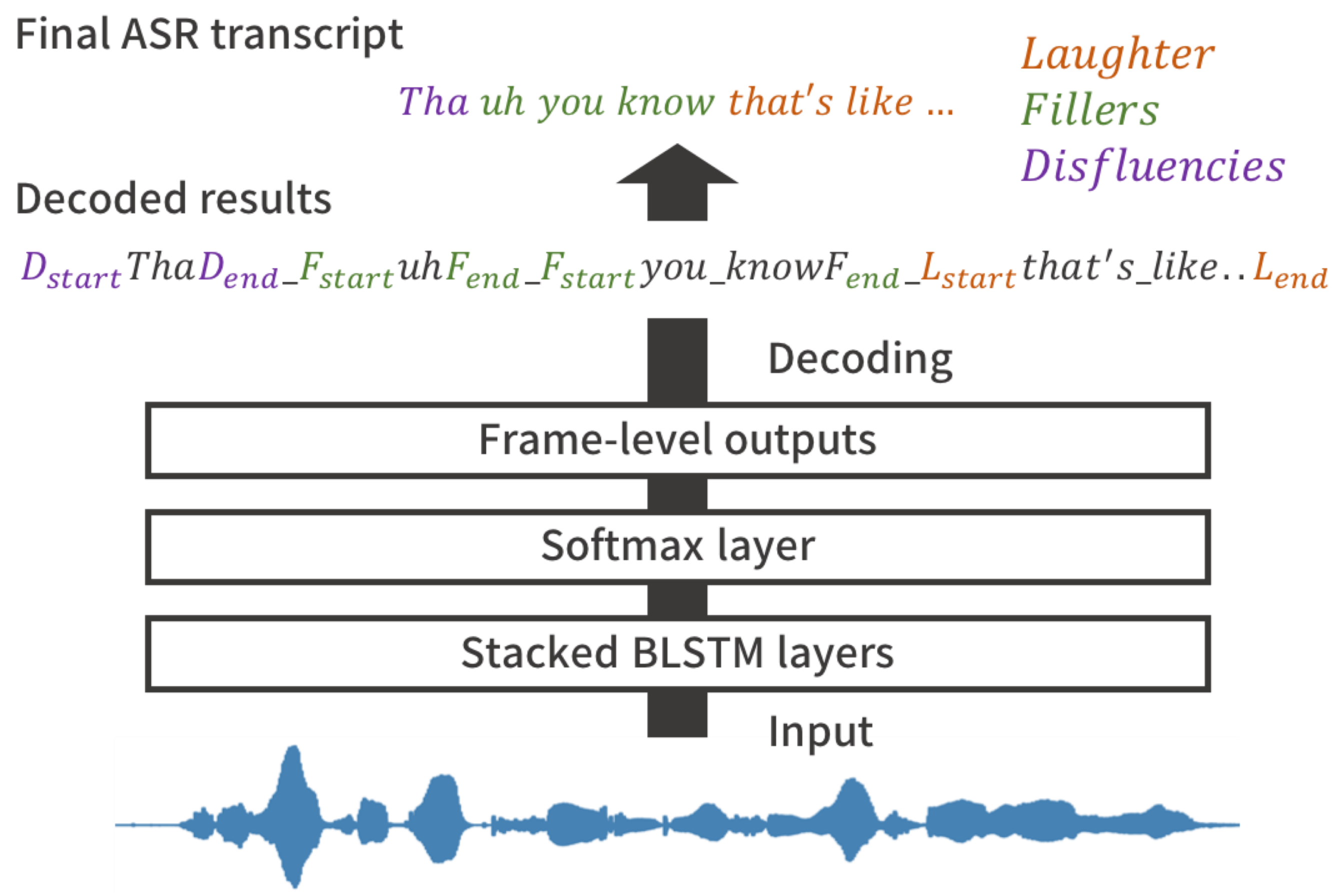
Joint SSD-ASR framework

We propose a unified framework where
  ✓ social signals are directly detected from speech
  ✓ while recognizing sub-word units
based on BLSTM-CTC [Graves+ 2006]

Joint Social Signal Detection (SSD) and Automatic Speech Recognition (ASR)

System overview

1. Both subword units and social signal labels are recognized by BLSTM-CTC for the SSD task
2. The final transcription for the ASR task is obtained by removing all social signal labels



Final ASR transcript

Tha uh you know that's like ...

Laughter
Fillers
Disfluencies

Decoded results

$D_{start}ThaD_{end}\_F_{start}uhF_{end}\_F_{start}you\_knowF_{end}\_L_{start}that's\_like..L_{end}$

Decoding

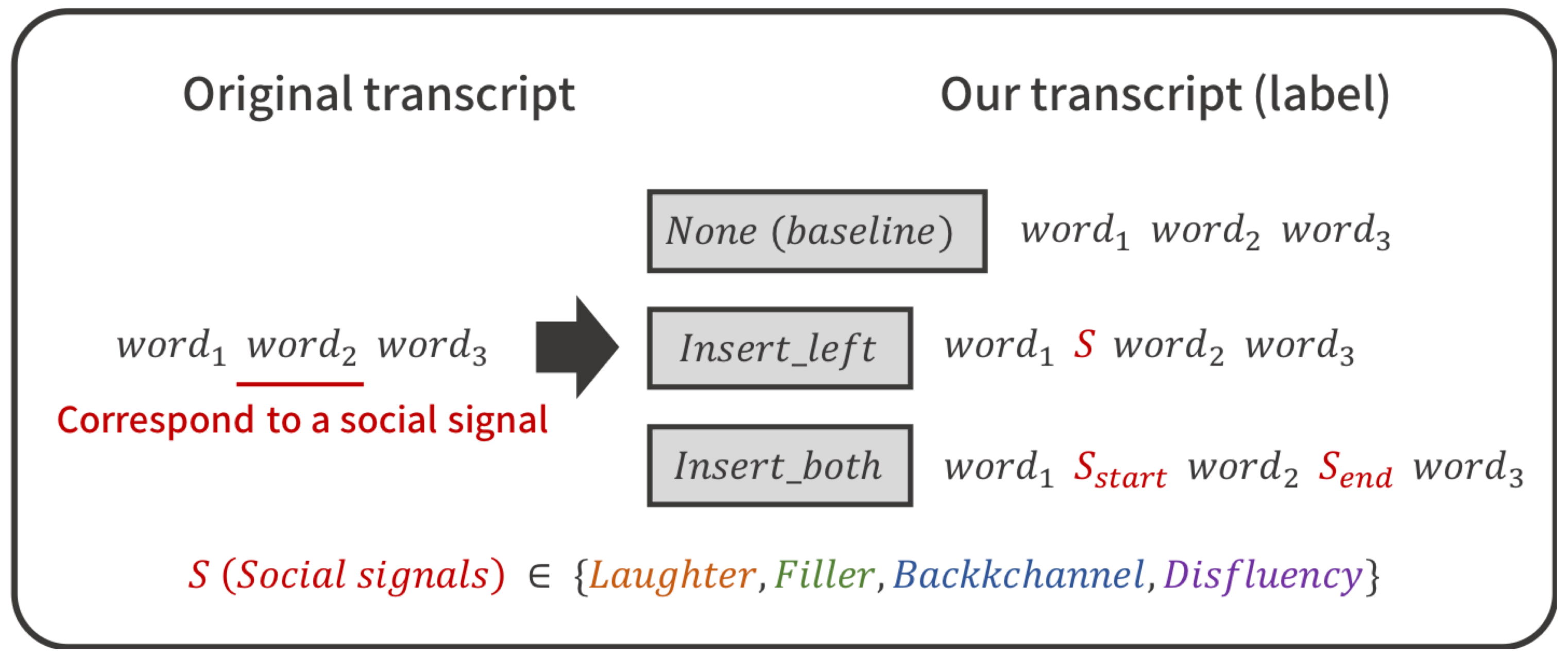
Frame-level outputs

Softmax layer

Stacked BLSTM layers

Input

Generation of reference labels



Original transcript

Our transcript (label)

$word_1\ word_2\ word_3$

Correspond to a social signal

None (baseline) $word_1\ word_2\ word_3$

Insert_left $word_1\ S\ word_2\ word_3$

Insert_both $word_1\ S_{start}\ word_2\ S_{end}\ word_3$

$S\ (Social\ signals) \in \{Laughter, Filler, Backkchannel, Disfluency\}$

Baseline
The same method as the conventional end-to-end ASR

Insert_left
Start label is inserted on the left side of subword units to detect acoustic cues such as short pauses before social signals

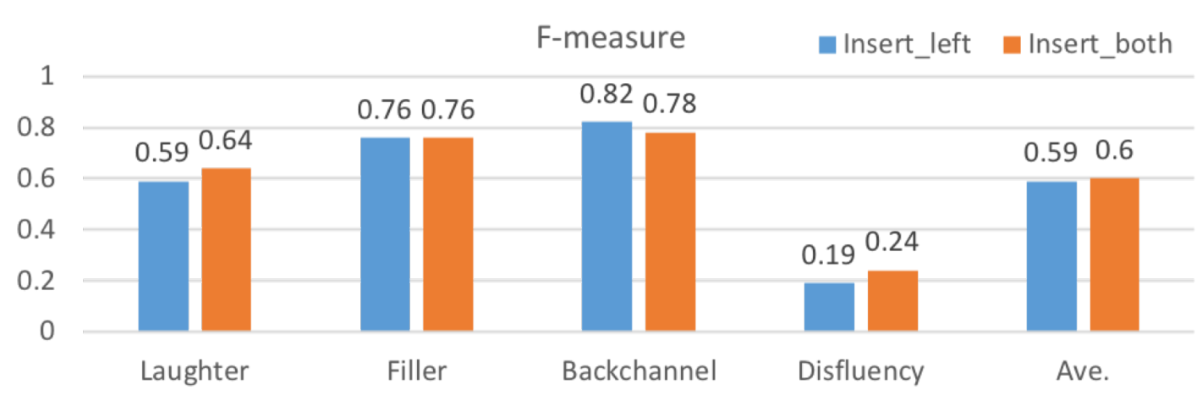
Insert_both
The end label is also inserted on the right side to learn rough segmentation of the social signals

Experimental Evaluations

Evaluation on ERATO Human-Robot Interaction corpus

✓ Dialogue corpus recorded with an autonomous android ERICA via Wizard-of-Oz (11.8h)
✓ Social signals: laughter, filler, backchannel, disfluency
✓ Vocabulary: 145 kinds of characters

SSD



✓ Fillers and backchannels were easy to detect
✓ Disfluencies were hard to detect due to the insufficient training data
✓ The end label was effective in disfluencies (0.24 vs. 0.19)

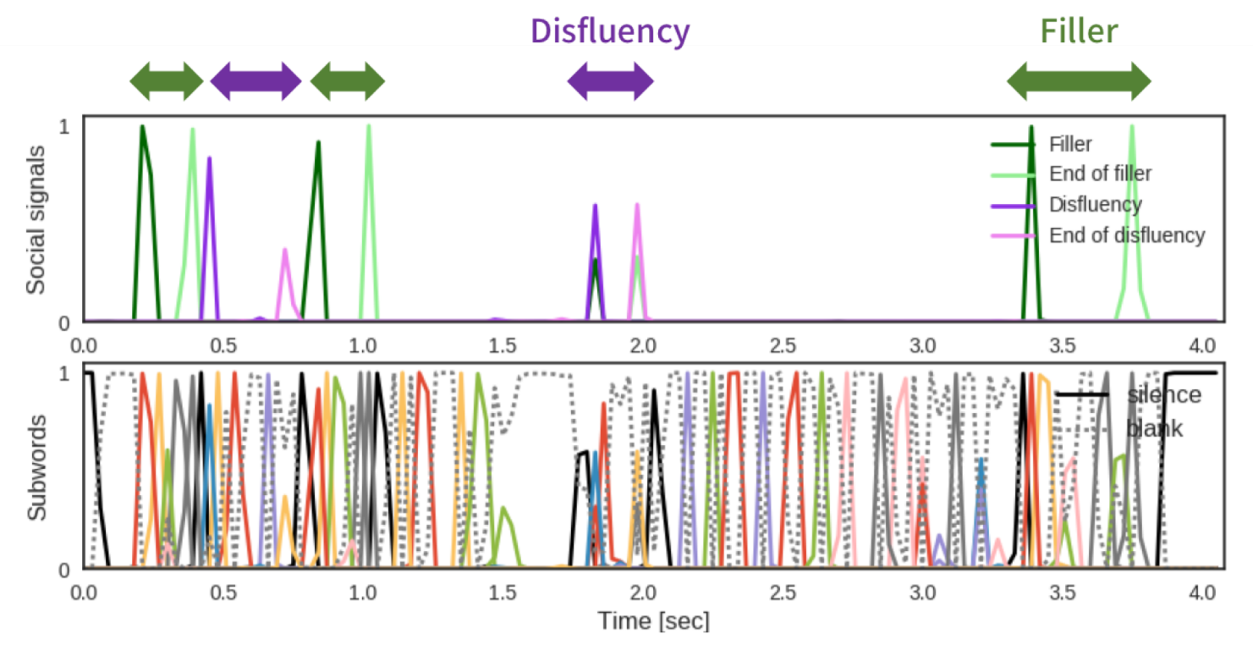
ASR

Labelling	Labelling	CER (%)
CTC (w/o LM)	*Baseline*	19.41
	Insert_left	19.80
	Insert_both	19.69

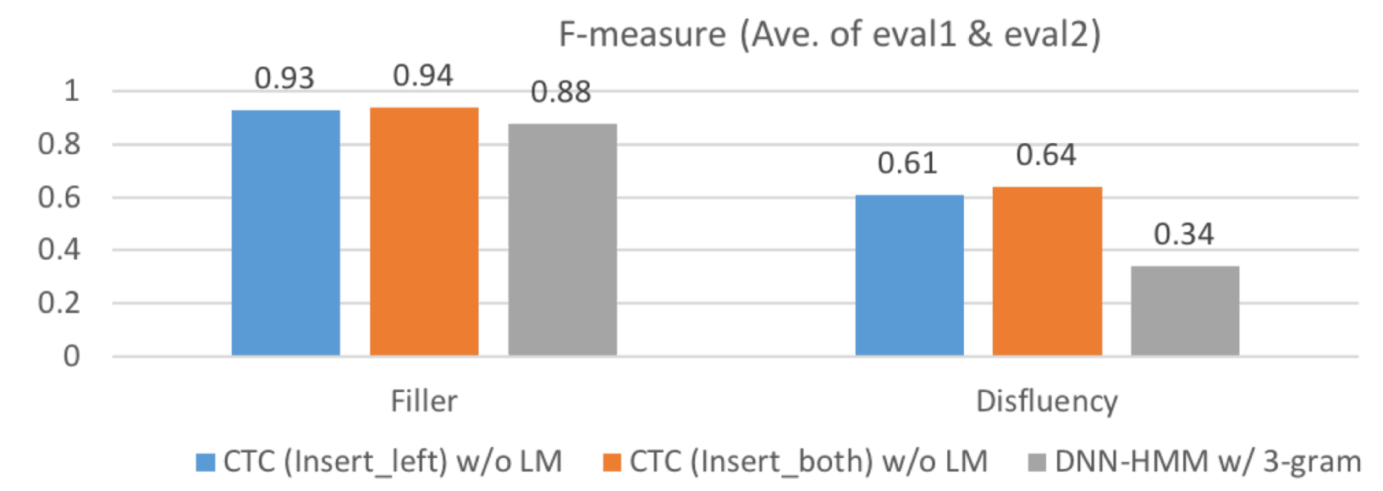
✓ No significant difference
✓ Robust detection of social signals without the degradation of ASR performance

Evaluation on CSJ

✓ Large-scale academic lecture corpus (240h)
✓ Social signals: filler, disfluency
✓ Vocabulary: 150 kinds of characters



SSD



✓ The performance of disfluencies were improved thanks to sufficient data
✓ *Insert_both* outperformed *Insert_left* in disfluencies as in ERATO corpus (0.64 vs. 0.61)
✓ Our framework outperformed DNN-HMM, especially for disfluencies (0.64 vs. 0.34)
✓ It is difficult for the hybrid ASR system to cover disfluencies (0.34)

ASR

Model	Labelling	CER (%)		
		eval1	eval2	Ave.
CTC (w/o LM)	*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

✓ CTC outperformed DNN-HMM
✓ CER of CTC was not improved by the additional social signal labels

Conclusions

✓ We have proposed the unified framework of SSD and ASR by a simplified architecture based on BLSTM-CTC without any special components
✓ Joint SSD-ASR framework outperformed the conventional hybrid system in both SSD and ASR performances
✓ CTC could identify rough locations of social signals
✓ Joint modeling leads to rich transcription including social signal information without the degradation of ASR performance