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

Introduction

Goals

- To detect social signals robustly on the event-level rather than the frame-level

What's social signals ?

Speech cues (this study)

- Laughter
- Filler
- Backchannel
- Disfluency

Visual cues

- Facial expressions
- Gestures
- Postures
- Gaze

Social signal detection [Schuller+ '13]

✓ Useful for understanding speakers

✓ Informative for dialog systems to behave like human

✓ Rich annotation

Related works: frame-wise classifiers

☹ Does not directly lead to the event-unit detection [Gosztolya+ '15]

☹ Frame-level target labels are required (|inputs| = |outputs|)

☹ Post-processing are required (threshold or HMM etc.)

➡ CTC can solve all these problems!

Approach

Bidirectional Long-Short Term Memory (BLSTM)

✓ Aim for the accurate detection by using the future context

Connectionist Temporal Classification (CTC) [Graves+ '06]

✓ A loss function which can optimize sequence labeling where the input and the target label sequence have different lengths

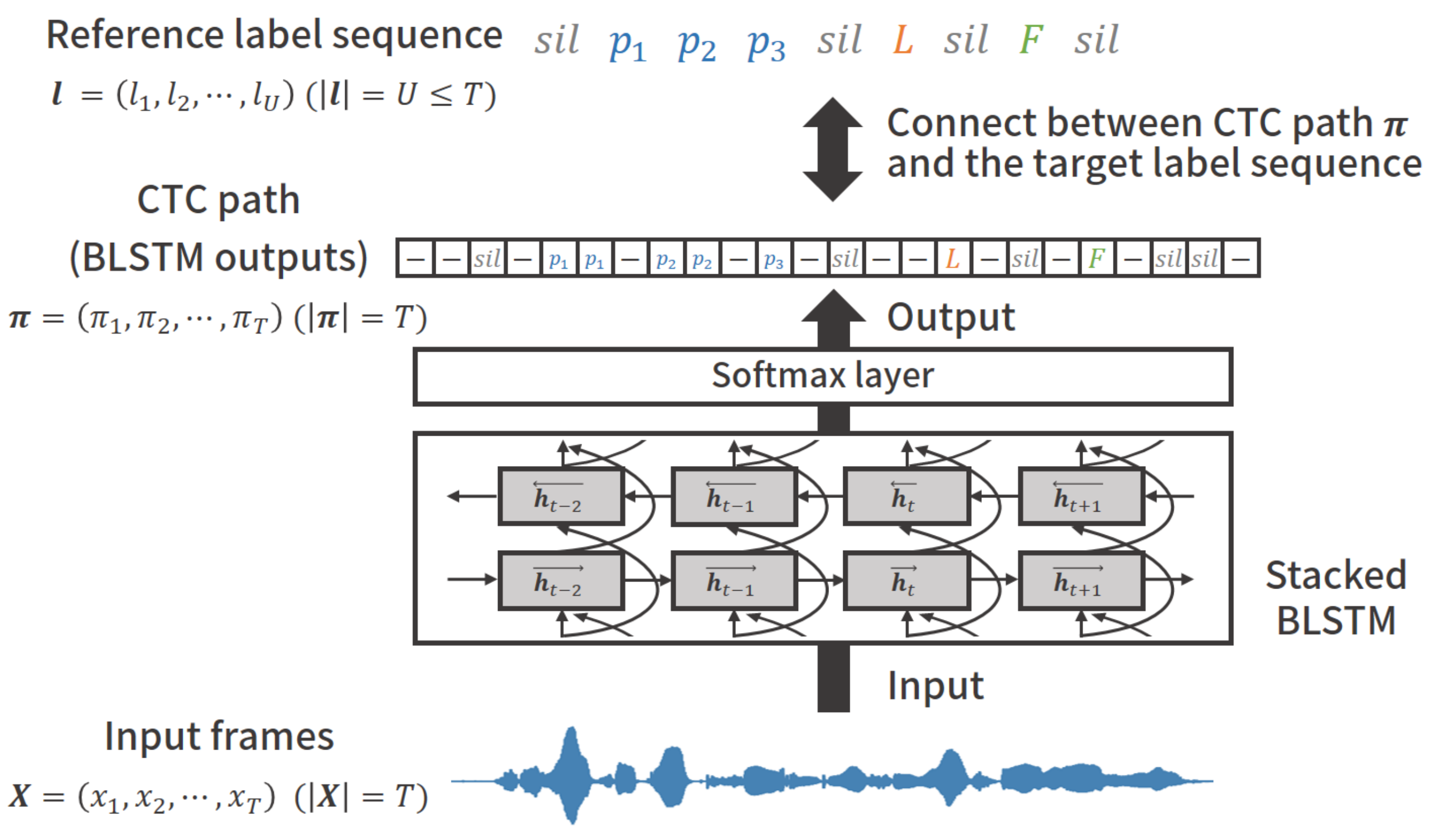
✓ Works together with RNNs

✓ Removes the need to conduct segmentation

✓ Has potential of improving robustness of detection (spike prediction)

Key idea of CTC

1. Introduction of a *blank* label (—) (the network emits no labels)
2. Allow repetitions of the same labels


Reference label sequence sil p_1 p_2 p_3 sil L sil F sil
$\boldsymbol{l} = (l_1, l_2, \cdots, l_U)$ $(|\boldsymbol{l}| = U \leq T)$
Connect between CTC path $\boldsymbol{\pi}$ and the target label sequence
CTC path (BLSTM outputs)
— — sil — p₁ p₁ — p₂ p₂ — p₃ — sil — — L — sil — F — sil sil —
$\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_T)$ $(|\boldsymbol{\pi}| = T)$
Output
Softmax layer
$\overleftarrow{h}_{t-2}$ $\overleftarrow{h}_{t-1}$ $\overleftarrow{h}_{t}$ $\overleftarrow{h}_{t+1}$
$\overrightarrow{h}_{t-2}$ $\overrightarrow{h}_{t-1}$ $\overrightarrow{h}_{t}$ $\overrightarrow{h}_{t+1}$
Stacked BLSTM
Input
Input frames
$\boldsymbol{X} = (x_1, x_2, \cdots, x_T)$ $(|\boldsymbol{X}| = T)$

CTC Training

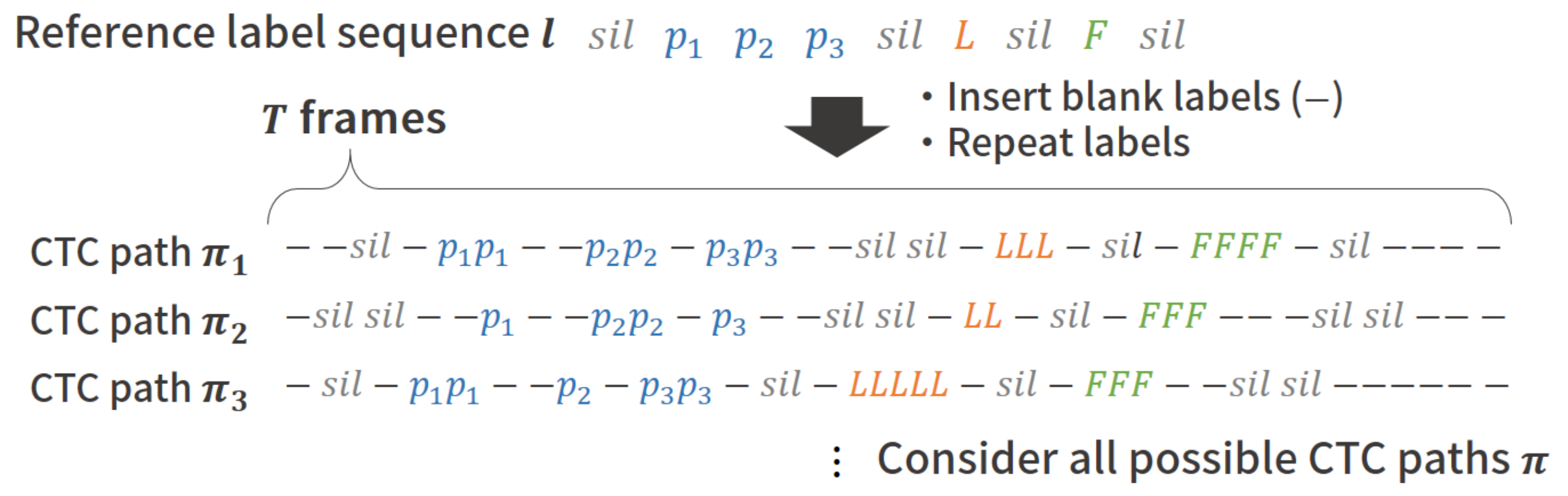
✓ Minimize $L_{CTC} = -\ln p(\boldsymbol{l}|\boldsymbol{X})$

✓ Marginalize $p(\boldsymbol{l}|\boldsymbol{X})$ by a summation of probability distribution of all possible frame-level alignments

$$p(\boldsymbol{l}|\boldsymbol{X}) = \sum_{\boldsymbol{\pi}\in\Phi^{-1}(\boldsymbol{l})} p(\boldsymbol{\pi}|\boldsymbol{X}) = \sum_{\boldsymbol{\pi}\in\Phi^{-1}(\boldsymbol{l})} \prod_{t=1}^{T} y^{t}_{\pi_t}$$

✓ Decompose $p(\boldsymbol{\pi}|\boldsymbol{X})$ based on the conditional independence assumption

✓ Compute $p(\boldsymbol{l}|\boldsymbol{X})$ efficiently with the forward-backward algorithm


Reference label sequence $\boldsymbol{l}$ sil p_1 p_2 p_3 sil L sil F sil
T frames
• Insert blank labels (–)
• Repeat labels
CTC path $\boldsymbol{\pi}_1$ $- - sil - p_1 p_1 - - p_2 p_2 - p_3 p_3 - - sil\ sil - LLL - sil - FFFF - sil - - - -$
CTC path $\boldsymbol{\pi}_2$ $- sil\ sil - - p_1 - - p_2 p_2 - p_3 - - sil\ sil - LL - sil - FFF - - - sil\ sil - - -$
CTC path $\boldsymbol{\pi}_3$ $- sil - p_1 p_1 - - p_2 - p_3 p_3 - sil - LLLLL - sil - FFF - - sil\ sil - - - - - -$
⋮ Consider all possible CTC paths $\boldsymbol{\pi}$

CTC Decoding

1. Remove repetitions
2. Remove all blank labels

Experiments

The SSPNet Vocalization Corpus (SVC)

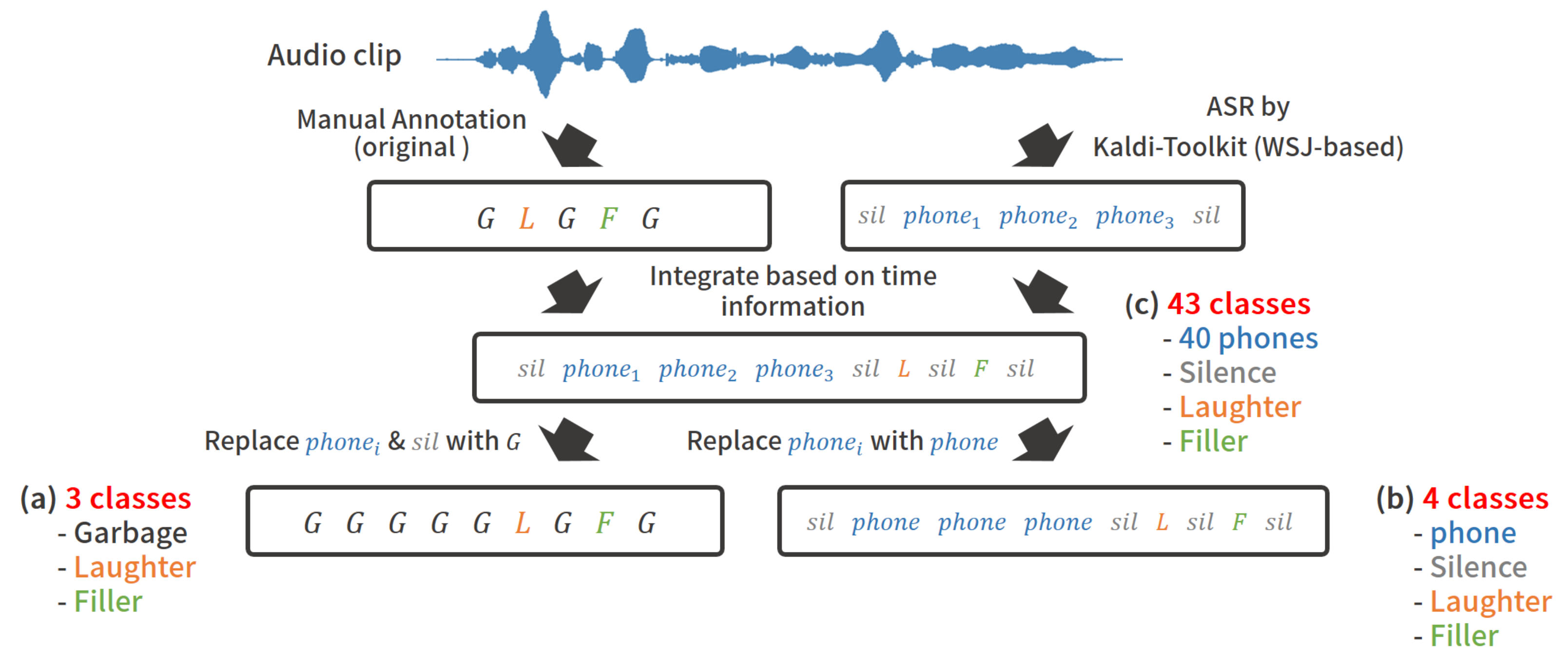
✓ Used in Interspeech 2013 ComParE (total 8.4h) [Schuller+ '13]

✓ Laughter , Filler , Garbage (speech and silence)

✓ Not transcripts available in SVC

✓ Target labels corresponding to acoustic events in the input are required (speech or silence)

◆ Generation of training labels for CTC

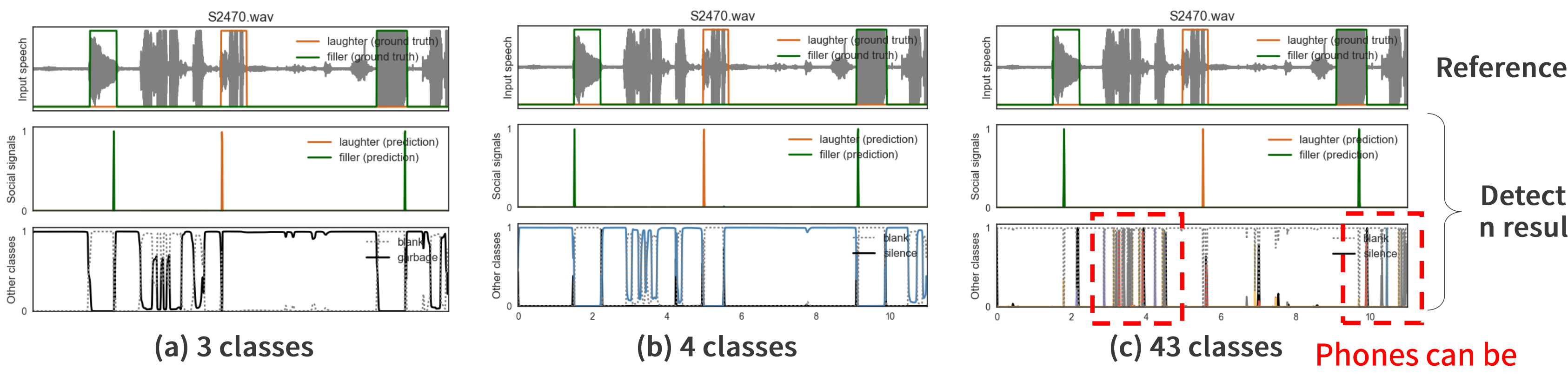

Audio clip
Manual Annotation (original)
ASR by Kaldi-Toolkit (WSJ-based)
G L G F G
sil $phone_1$ $phone_2$ $phone_3$ sil
Integrate based on time information
sil $phone_1$ $phone_2$ $phone_3$ sil L sil F sil
(c) 43 classes
- 40 phones
- Silence
- Laughter
- Filler
Replace $phone_i$ & sil with G
Replace $phone_i$ with $phone$
(a) 3 classes
- Garbage
- Laughter
- Filler
G G G G G L G F G
sil $phone$ $phone$ $phone$ sil L sil F sil
(b) 4 classes
- phone
- Silence
- Laughter
- Filler

◆ Results

Class	Model	Laughter			Filler			Ave.
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1
3	AdaBoost-HMM [Gosztolya+ '15]	0.58	0.74	0.65	0.65	0.71	0.68	0.66
	DNN-HMM	0.58	0.72	0.64	0.71	0.60	0.65	0.65
	(a) BLSTM-CTC	**0.65**	**0.66**	**0.66**	**0.66**	**0.80**	**0.72**	**0.69**
4	(b) BLSTM-CTC	0.60	0.49	0.54	0.59	0.78	0.67	0.61
43	(c) BLSTM-CTC	0.79	0.51	0.62	0.71	0.78	0.74	0.68

CTC outperformed the conventional frame-wise classifiers even without time information in the training stage

◆ CTC outputs (posteriors)



(a) 3 classes (b) 4 classes (c) 43 classes

Phones can be detected jointly

ERATO Human-Robot Interaction Corpus

✓ Japanese face-to-face spontaneous dialog with an android ERICA, which was remotely operated

✓ 91 sessions (about 10 min/session, total 16.8h)

✓ Laughter, Filler, Backchannel, Disfluency

✓ 4 social signals + 83 Japanese kana characters + space

◆ Generation of training labels for CTC

✓ Insert each social signal label in front of the corresponding word

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

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

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

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

◆ Results

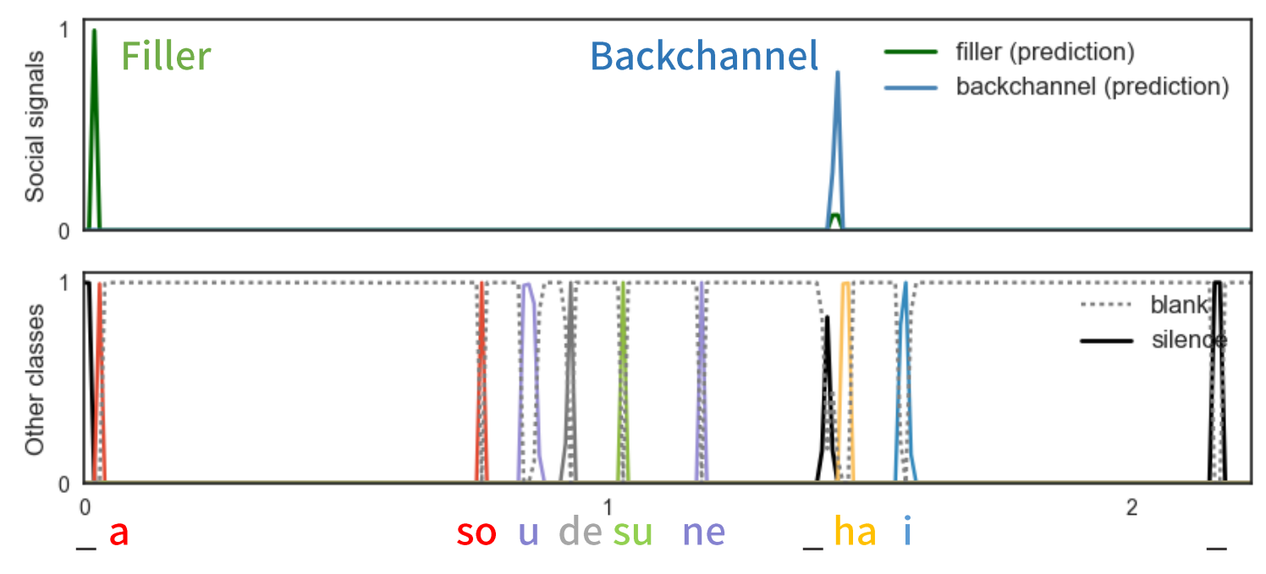
Social signals	Prec.	Rec.	F_1
Laughter	0.89	0.35	0.50
Filler	0.75	0.75	0.75
Backchannel	0.86	0.87	0.86
Disfluency	0.44	0.15	0.22

Social signals	CER (%)
BLSTM-CTC (w/o social signals)	19.1
BLSTM-CTC (w/ social signals)	**18.6**

Joint-training with social signals improved character-level speech recognition accuracy

◆ CTC outputs (posteriors)

CTC could capture relationships between social signals and subwords



Conclusions

Summary

✓ Robust social signal detection on the event-level by BLSTM-CTC

✓ Removed the need of pre-alignment and post-processing

✓ Outperformed the conventional frame-wise classifiers

✓ Alignments are generally matched with the actual timing of the occurrence of social signal events

Future work

✓ Evaluate with large dataset

✓ Attention-based detection