TOSHIBA
Leading Innovation

SIMULTANEOUS SPEECH RECOGNITION AND ACOUSTIC EVENT DETECTION USING AN LSTM-CTC ACOUSTIC MODEL AND A WFST DECODER

*Hiroshi Fujimura*1, Manabu Nagao*1*2, Takashi Masuko*
Corporate Research & Development Center, Toshiba Corporation

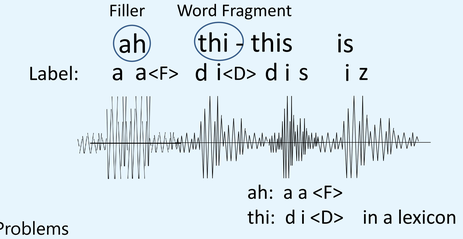
*1:Equal contribution *2: Currently with Toshiba Digital Solutions Corporation

Summary

● Detection of filler and word fragment

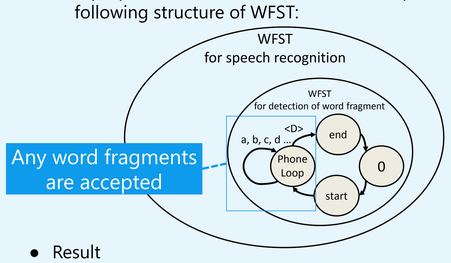
- Filler and word fragment are distinctly modeled

Previously proposed method



- Problems
  - How do both the detection and decoding work by one-pass decoding?
  - Prepare all word fragments in a lexicon? → not realistic

- Proposed method
  - A proposed WFST decoder can solve the problems using the following structure of WFST:



- Result
  - The proposed WFST decoder could simultaneously achieve speech recognition and detection of fillers and word fragments using a conventional lexicon

Background

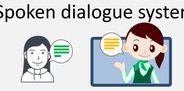
● Motivation

- **Influence of Filler and Word fragment**

→Difficult to recognize it and read it

- **Requirement of real-time computation for speech recognition**

Spoken dialogue system



Translation

Japanese English

→ Develop real-time filler and word fragment detection

● Related works

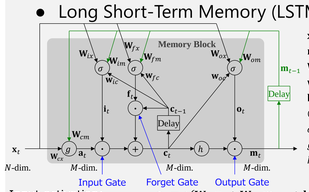
- **Many studies of Language-aspect methods for detection of word fragment by second pass decoding**
- **Direct modeling of word fragment using acoustic feature**
  - Word fragments are directly modeled by adding a fragment tag to phonemes composing word fragments
  - Phoneme models with the fragment tag are used as garbage models

→ Develop detection of fillers and word fragments using acoustic feature by one-pass decoding

Acoustic Model with detection of fillers and word fragments

● LSTM-CTC

- Long Short-Term Memory (LSTM)



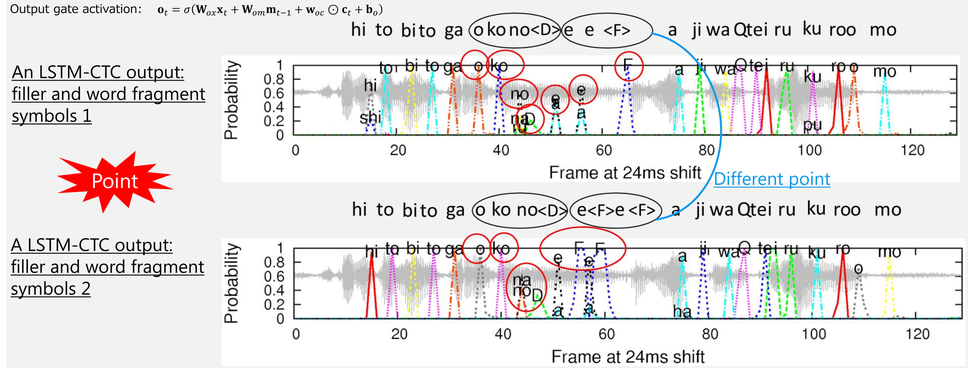
$\mathbf{x}_t$: Input vector (N-dim.)
$\mathbf{m}_{t-1}$: Hidden state vector (M-dim.)
$\mathbf{W}_{\cdot\cdot}$: Weight matrix
$\mathbf{w}_{\cdot\cdot}$: Weight vector
$\mathbf{b}_{\cdot}$: Bias vector
$\odot$: Element-wise multiplication
$\sigma(\cdot)$: Element-wise sigmoid function
$g(\cdot)$: Element-wise tanh function
$h(\cdot)$: Element-wise tanh function

Input activation: $\mathbf{a}_t = g(\mathbf{W}_{cx}\mathbf{x}_t + \mathbf{W}_{cm}\mathbf{m}_{t-1} + \mathbf{b}_c)$
Input gate activation: $\mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{im}\mathbf{m}_{t-1} + \mathbf{w}_{ic} \odot \mathbf{c}_{t-1} + \mathbf{b}_i)$
Forget gate activation: $\mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fm}\mathbf{m}_{t-1} + \mathbf{w}_{fc} \odot \mathbf{c}_{t-1} + \mathbf{b}_f)$
Cell activation: $\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{a}_t$ Output activation: $\mathbf{m}_t = \mathbf{o}_t \odot h(\mathbf{c}_t)$
Output gate activation: $\mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{om}\mathbf{m}_{t-1} + \mathbf{w}_{oc} \odot \mathbf{c}_t + \mathbf{b}_o)$

- Connectionist Temporal Classification (CTC)
  - Maximize probability for output sequence that is length U from input sequence that is length T ($U < T$)

Label samples: 2 patterns for filler labeling

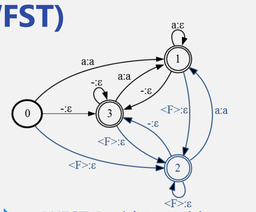
Transcription	Thi-this is ah a pen
Phonetic symbols	di di s Iz aa a pe N
+ filler and word fragment symbols1	di <D>di s Iz aa <F> a pe N
+ filler and word fragment symbols2	di <D>di s Iz a <F> a <F> a pe N



Proposed Weighted Finite State Transducer (WFST)

● WFST for general speech recognition

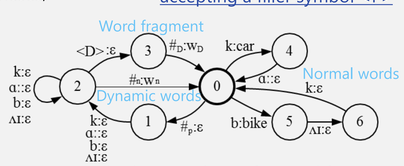
- A WFST for speech recognition consists of R, L and G using LSMT-CTC
  - R: WFST Squashing a label sequence of the acoustic model in a CTC manner
    ex.) "AAAA", "ΦΦAAAΦ"→ "A", "A" by squashing (Φ: blank symbol)
  - L: WFST of a lexicon, G: WFST of a language model



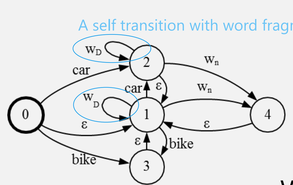
● WFST for filler detection

- Filler symbol can be detected and deleted by a new WFST R
  - By checking input symbols on the transitions of word hypotheses, it is possible to detect words as fillers
- Filler confidence score using filler symbol2 (e <F> e <F>)
  - Confidence score $c = f/p$,
    where f: the number of detected filler symbols
    p: the number of phonetic symbols in a word
    ex.) ah (a <F> a <F>) c = 1.00,
    third (s a <F> a d) c= 0.25

WFST R with transitions accepting a filler symbol <F>



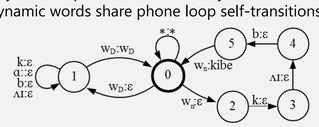
● WFST for word fragment detection



Point

WFST L
Word fragment detection
WFST G
WFST D

- A WFST L for word fragment detection
  - A WFST L should simultaneously accept normal words, dynamic words and arbitrary phonetic symbol sequences terminated by <D>
  - Word fragments and dynamic words share phone loop self-transitions



- A WFST G for word fragment detection
  - A self transition with word fragment w_D is added to a WFST G
  - A dynamic word w_n is regarded as a normal word
  - A combined WFST $RLG = \pi_\varepsilon(opt(R \circ opt(proj_{i\to o}(L \circ G))))$

- A WFST D for word fragment detection
  - Convert a phoneme sequences to dynamic words
  - Convert phonetic symbol sequences to w_D
  - A combined WFST for decoding $RLGD = RLG \circ D$

Experimental Setup

Evaluation Datasets

	Style	Speaker	#Filler	#Word Fragment
CSJ testset3	Monologue	5 males 5 females	785	169
Liaison-meeting	Monologue	1 male	238	36

Training labels for acoustic models

Acoustic Model	Training label for "こ, ええと この (ko, eeto kono)"
Normal model (NAM)	ko e e to ko no
Filler + word fragment detection model (FDAM1)	ko <D> e e to <F> ko no
Filler + word fragment detection model (FDAM2)	ko <D> e <F> e <F> to <F> ko no

Lexicons

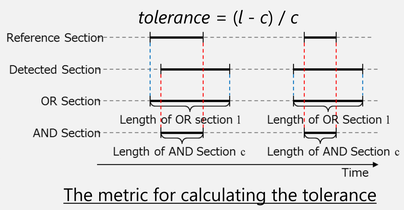
Lexicon	Word and the pronunciation
Normal lexicon (NLex)	kono: ko no, eeto: e e to
Normal lexicon with word fragments (Nlex+D)	kono: ko no, ko: ko, eeto: e e to
Lexicon with filler and word fragment symbols 1 (FDLex1)	kono: ko no, ko: ko <D>, eeto: e e to <F>
Lexicon with filler and word fragment symbols 2 (FDLex2)	kono: ko no, ko: ko <D>, eeto: e <F> e <F> to <F>

LSTM-CTC training

Feature	28-dimentional Mel filterbank
Feature concatenation	1+previous 8 frames
#LSTM layer	3
#Memory unit / layer	1024
Recurrent projection	1024→256
Output class	126 Japanese moras + filler & word fragment
Training data	CSJ training set 580h

Language model training

N-gram	4-gram
Training data	200 million sentences from Web + Toshiba Internal dataset
Vocabulary size	200,000



The metric for calculating the tolerance

Created WFSTs ("WF": Word Fragment)

	WFST name	Acoustic Model	Lexicon	Decoder
	Normal (NWT)	NAM	Nlex+D	Conventional
	Filler & WF (FDWT1)	FDAM1	FDLex1	Conventional
Conventional	Filler & WF (FDWT2)	FDAM2	FDLex2	Conventional
	Filler & WF (FDWT3)	FDAM1	NLex	Proposed
Proposed	Filler & WF (FDWT4)	FDAM2	NLex	Proposed

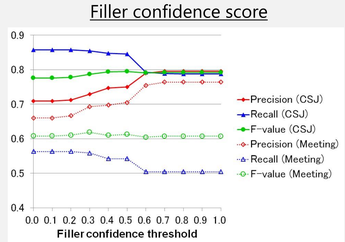
Point

Results

ASR performance for each WFST (CER [%])

WFST name	CSJ testset3	Liaison-meeting	Ave.
NWT	10.34	15.36	12.85
FDWT1	10.35	14.75	12.55
FDWT2	9.83	14.65	12.24
FDWT3	10.53	14.82	12.68
FDWT4	10.16	14.99	12.57

Filler confidence score



Point

Filler Detection performance (F: F-value)

WFST	CSJ testset3 Precision	Recall	F	Liaison-meeting Precision	Recall	F
NWT	0.77	0.61	0.68	0.79	0.45	0.57
FDWT1	0.80	0.77	0.78	0.87	0.57	0.69
FDWT2	0.78	0.81	0.79	0.78	0.58	0.66
FDWT3	0.75	0.85	0.79	0.72	0.58	0.64
FDWT4	0.75	0.85	0.79	0.70	0.54	0.61

Word Fragment Detection performance (F: F-value)

WFST	CSJ testset3 Precision	Recall	F	Liaison-meeting Precision	Recall	F
NWT	0.22	0.04	0.07	0.00	0.00	0.00
FDWT1	0.36	0.02	0.04	1.00	0.03	0.05
FDWT2	0.42	0.03	0.06	1.00	0.06	0.11
FDWT3	0.53	0.25	0.34	0.86	0.33	0.48
FDWT4	0.49	0.25	0.34	0.61	0.31	0.41

- **The proposed decoder can simultaneously recognize speech and detect fillers and word fragments**
- **The proposed decoder outperformed a conventional decoder using lexicons including word fragments in word fragment detection**