Incremental Machine Speech Chain
Towards Enabling Listening while Speaking in Real-time

Sashi Novitasari\textsuperscript{1}, Andros Tjandra\textsuperscript{1}, Tomoya Yanagita\textsuperscript{1}, Sakriani Sakti\textsuperscript{1,2}, Satoshi Nakamura\textsuperscript{1,2}

\textsuperscript{1}NAIST, Japan
\textsuperscript{2}RIKEN-AIP, Japan
Outline

I. Introduction

II. Incremental Machine Speech Chain

III. Experiments

IV. Conclusion
I. Introduction

II. Incremental Machine Speech Chain

III. Experiments

IV. Conclusion
Background

ASR and TTS

- Spoken language technologies:
  - Automatic speech recognition (ASR)
  - Text-to-speech synthesis (TTS)

- Crucial for human-machine interaction

- Remarkable performance
  
  \[\text{requires a lot of speech-text paired data}\]

ASR and TTS systems

“Hello”

“Hello”

Copyright © 2020 NAIST Japan, All rights reserved
Background

Machine Speech Chain
[Tjandra et al., 2017]

• Semi-supervised ASR and TTS training via closed feedback loop
• ASR/TTS : standard attention-based seq2seq network
• 2 training phases:
  1) ASR/TTS supervised independent training
  2) ASR/TTS unsupervised joint training with feedback loop
• Full-utterance-based ASR and TTS → High delay
Background

Human Speech Chain

Human speech chain [Denes, 1993]

- Feedback loop between speech production and hearing systems
- **Real-time** process → immediate adaptation
- Feedback delay causes a disturbance during speaking

**Challenge in mimicking human speech chain for machine**
Speech generation or recognition and feedback generation based on incomplete sequence information with **minimum delay**

Propose: Incremental Machine Speech Chain
II. Incremental Machine Speech Chain
Propose

Incremental Machine Speech Chain

Closed short-term feedback loop between incremental ASR (ISR) and incremental TTS (ITTS)

- Reduce feedback delay within machine speech chain training
- Improve ISR and ITTS learning quality
- Enable immediate feedback generation during inference

Move a step closer for ASR and TTS that can adapt to real-time environment unsupervisedly

⇒ Similar to human
Incremental Machine Speech Chain

Components

**Incremental ASR (ISR):** Low delay ASR
- Hidden Markov model ASR
- End-to-end ISR with attention-based seq2seq model
  - Neural transducer [Jaitly et al., 2016]
  - Attention-transfer ISR [Novitasari et al., 2019]

**Incremental (ITTS):** Low delay TTS
- Hidden Markov model TTS
- End-to-end ITTS with attention-based seq2seq model
  - Neural ITTS [Yanagita et al., 2019]
  - ITTS based on prefix-to-prefix framework [Ma et al., 2019]

- Performance limitation due to short-input-based processing
- Previous: independent development
Incremental Machine Speech Chain

Training Mechanism

2 training phases:

1. ISR and ITTS supervised-independent training

2. ISR and ITTS joint training via short-term feedback loop
Incremental Machine Speech Chain Training

1. ISR and ITTS Independent Training

- Incremental: Predict a complete output sequence in $N$ steps.
  - For each step $n$:
    1. Encode a segment of input from input window
    2. Decode and predict a segment of output
    3. Shift the input windows

- ISR and ITTS training by attention transfer from standard non-incremental ASR [Novitasari et al., 2019] → same alignment for ISR and ITTS
 Incremental Machine Speech Chain Training

2. ISR and ITTS Joint Training

• Short-term feedback loop between the components
• Segment-based output passing
• Unrolled processes
  a. **ISR-to-ITTS**
     For each step $n$, ISR predicts $\hat{Y}_n$ from $X_n$, and then ITTS predicts $\hat{X}_n$ from ISR output $\hat{Y}_n$
  b. **ITTS-to-ISR**

\[ \text{Full speech } = (X) \]

\[ \text{Loss}_{TTS_n=1}(x_{n=1}, \hat{x}_{n=1}) \]
\[ \text{Loss}_{TTS_n=2}(x_{n=2}, \hat{x}_{n=2}) \]

\[ Y_{n=1} = \text{“a b c”} \]
\[ Y_{n=2} = \text{“d e”} \]

\[ X_{n=1} = \]
\[ X_{n=2} = \]
Incremental Machine Speech Chain Training
2. ISR and ITTS Joint Training

• Short-term feedback loop between the components
• Segment-based output passing
• Unrolled processes
  a. ISR-to-ITTS
     For each step $n$, ISR predicts $\hat{Y}_n$ from $X_n$, and then ITTS predicts $\hat{X}_n$ from ISR output $\hat{Y}_n$
  b. ITTS-to-ISR
     For each step $n$, ITTS predicts $\hat{X}_n$ from $Y_n$, and then ISR predicts $\hat{Y}_n$ from ITTS output $\hat{X}_n$
Exploration on 2 learning approaches:

A) **Semi-supervised incremental machine speech chain**
   1) ISR/ITTS independent training: supervised
   2) ISR/ITTS joint training: unsupervised (unlabeled data)

B) **Supervised incremental machine speech chain**
   1) ISR/ITTS independent training: supervised
   2) ISR/ITTS joint training: supervised (labeled data)

Unrolled process examples in joint training
(ITTS-to-ISR follows similar mechanism)
III. Experiments
Experiments

Dataset

Wall Street Journal CSR Corpus [Paul and Baker, 1992]

- Language: English
  - Training sets:
    - SI-84: 16 hours of speech, 83 speakers
    - SI-200: 66 hours of speech, 200 speakers
    - SI-284: si84 + si200
  - Dev. set: dev93
  - Eval. set: eval92
- Character-level
- Speech features: 80-dims log Mel spectrogram (window: 50 msec, shift: 12.5 msec)

Copyright © 2020 NAIST Japan, All rights reserved

INTERSPEECH 2020
Experiments
Model Configuration

* Same architecture for standard (non-incremental) and incremental models

**ASR**

Transcription
Decoder

LSTM
Char Emb.

Encoder
BiLSTM
BiLSTM

FNN

Speech features

Hierarchical sub-sampling
8 feature frames into 1 encoder state

Attention

Input/step
- ISR : 0.84 sec
- Std. ASR : full-utterance (avg. 7.88 sec)

**TTS**

Tacotron 2 [Wang et al., 2017] structure with speaker embedding [Tjandra et al., 2018]

Speech features

Decoder
Linear Proj.

2 LSTM
2 Pre-Net

Encoder
BiLSTM

3 Conv
Char Emb.

Transcription

Input/step
- ITTS : avg. 30 chars
- Std. TTS : full-sentence (avg. 103 chars)
Result

ASR (CER%) and TTS (log Mel-spectrogram L2 loss) performances

<table>
<thead>
<tr>
<th>Data</th>
<th>ASR (CER%)</th>
<th>TTS (L2-norm)²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Incremental</td>
</tr>
<tr>
<td></td>
<td>(delay: 7.88 sec)</td>
<td>(delay: 0.84 sec)</td>
</tr>
<tr>
<td>nat-sp</td>
<td>syn-sp</td>
<td>nat-sp</td>
</tr>
</tbody>
</table>

**Independent Training**

- **Indep-trn SI-84**
  - ASR: 17.33, 27.03
  - TTS: 17.81, 44.54
  - Delay: 7.88 sec

- **Indep-trn SI-284**
  - ASR: 7.16, 9.60
  - TTS: 7.97, 19.99
  - Delay: 0.84 sec

**Machine Speech Chain**

- **Indep-trn (SI-84) + chain-trn-greedy (SI-200)**
  - ASR: 11.21, 11.52
  - TTS: 14.23, 32.43
  - Delay: 7.88 sec

- **Indep-trn (SI-84) + chain-trn-teachforce(SI-200)**
  - ASR: 7.27, 6.30
  - TTS: 9.43, 12.78
  - Delay: 0.84 sec

- **Input type:**
  - ISR
  - ITTS

- **Baseline**
  - ISR and ITTS *indep-trn SI-84*

- **Topline**
  - Standard systems *indep-trn SI-284*

- **Proposed**
  - Incremental machine speech chain

- **Incremental machine speech chain**
  - Improved ISR and ITTS
  - Shorter delay with a close performance to the standard system
IV. Conclusion
Conclusion

Incremental machine speech chain

Short-term feedback loop for ISR/ITTS development by mimicking human speech chain

- Reduced the delay with a close performance to the basic framework
- Improve ISR and ITTS (natural/synthetic input)
- Synthetic input processing: demonstration of real-time feedback generation
Thank you