Real-time Neural Machine Speech Chain

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
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

• ASR and TTS closely related to each other
  → Research trends: independent development

Machine Speech Chain [Tjandra et al., 2017]

• Semi-supervised ASR and TTS training via closed feedback loop
• Inspired from human speech chain [Denes, 1993]
  ○ Listening while speaking
Background

ASR and TTS

Machine speech chain

- 2 training phases:
  1) ASR/TTS supervised independent training
  2) ASR/TTS unsupervised joint training with feedback loop
    - 2 unrolled processes inside the feedback loop:
      
      **A) ASR-to-TTS (speech only)**
      
      **B) TTS-to- ASR (text only)**
      
      Current framework is for the full utterance-based ASR/TTS
      \rightarrow High delay
Human Speech Chain

Human speech chain [Denes, 1993]

- Feedback loop between speech production and hearing systems
- **Real-time** process $\rightarrow$ 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

Unrolled processes in machine speech chain loop
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**
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 $\hat{X}_n$
III. Experiments
Experiments

Dataset

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

• Language : English

  ❖ Training sets:
    o \textit{SI-84} : 16 hours of speech, 83 speakers
    o \textit{SI-200} : 66 hours of speech, 200 speakers
    o \textit{SI-284} : \textit{si84} + \textit{si200}

  ❖ Dev. set : \textit{dev93}

  ❖ Eval. set : \textit{eval92}

• Character-level

• Speech features: 80-dims log Mel spectrogram (window: 50 msec, shift: 12.5 msec)
Experiments

Model Configuration

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

**ASR**

- **Transcription Decoder**
  - LSTM
  - Char Emb.
  - Attention

- **Encoder**
  - BiLSTM
  - FNN

- **Speech features**
  - $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_{T-1}, x_T$

**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]**

- **Encoder**
  - BiLSTM
  - Char Emb.
  - Transcription

- **Speaker Emb.**

**Input/step**
- **ITTS** : max. 30 chars
- **Std. TTS** : full-sentence (avg. 103 chars)
Experiments
Learning Approach

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)
### 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 (delay: 7.88 sec)</td>
<td>Incremental (delay: 0.84 sec)</td>
</tr>
<tr>
<td></td>
<td>nat-sp syn-sp</td>
<td>nat-sp syn-sp</td>
</tr>
<tr>
<td>Independent Training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indep-trn SI-84</td>
<td><img src="image1" alt="Diagram" /></td>
<td></td>
</tr>
<tr>
<td>Indep-trn SI-284</td>
<td><img src="image2" alt="Diagram" /></td>
<td></td>
</tr>
</tbody>
</table>

#### Machine Speech Chain

<table>
<thead>
<tr>
<th>Data</th>
<th>ASR (CER%)</th>
<th>TTS (L2-norm)$^2$</th>
<th>Input type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indep-trn (SI-84) + chain-trn-greedy (SI-200)</td>
<td><img src="image3" alt="Diagram" /></td>
<td><img src="image4" alt="Diagram" /></td>
<td>nat-sp</td>
</tr>
<tr>
<td>Indep-trn (SI-84) + chain-trn-teachforce(SI-200)</td>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
<td>syn-sp</td>
</tr>
</tbody>
</table>

- **Baseline**
  - ISR and ITTS *indep-trn SI-84*
- **Topline**
  - Standard systems *indep-trn SI-284*
- **Proposed**
  - Incremental machine speech chain
- **Input type**
  - [Isolated Speaker Recognition (ISR) 📈]
  - [Text-to-Speech (ITTS) 📈]

---

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

---

Copyright © 2021 NAIST Japan, All rights reserved
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)
Thank you