Customization of IBM Intu’s Voice by Connecting Text-to-Speech Services with a Voice Conversion Network

Jongyoon Song\textsuperscript{1}  
Jaekoo Lee\textsuperscript{1}  
Hyunjae Kim\textsuperscript{1}  
Euishin Choi\textsuperscript{2}  
Minseok Kim\textsuperscript{2}  
Sungroh Yoon\textsuperscript{1}

\textsuperscript{1}ECE, Seoul National University  
\textsuperscript{2}IBM Korea

Outline

- Introduction
- Related Work
  - IBM Watson and Project Intu
  - Text-to-speech (TTS)
  - Voice conversion
- Model Description
  - Voice conversion network (VCN)
  - Intu
- Experiments and Discussion
- Conclusion
Outline

- Introduction
- Related Work
  - IBM Watson and Project Intu
  - Text-to-speech (TTS)
  - Voice conversion
- Model Description
  - Voice conversion network (VCN)
  - Intu
- Experiments and Discussion
- Conclusion
Introduction

- Intelligent Personal Assistant
- Artificial Intelligence Speaker

- Amazon Echo [1]
- Google Home [2]
- Apple Siri [3]
Preset voice
Customized voice
Customized voice

Voice customization = easy for users to train
Customized voice

Voice customization
= easy for users to train

Our design
- Small target speech data
- No parallel data
Introduction

1. Gather 10~ min of speech of preferred voice

2. Send to Intu to train voice customization module

3. Users can talk with AI speaker with customized voice
Introduction

- **Text-to-speech (TTS)**
  - Text: *linguistic* & phonetic feature
  - Speech: phonetic & *acoustic* feature
  - Requires relatively complex model
  - Needs around 30 min of speech per voice [4]

- **Voice conversion**
  - Inputs and outputs have same feature domain
  - Requires relatively simple model
  - Needs around 10 min of speech per voice [5]
Introduction

Pre-trained using public speech data + Trained using 10~ min of target speech
• **Source speaker/speech**
The speaker of voice before conversion / the speaker’s speech

• **Target speaker/speech**
The speaker whom the user prefers / the speaker’s speech
Introduction

Contribution

• Voice customization for ML-as-a-Service design

• Methods for inference time optimization

• Analysis for proper amount of target speech
Outline

- Introduction
- Related Work
  - IBM Watson and Project Intu
  - Text-to-speech (TTS)
  - Voice conversion
- Model Description
  - Voice conversion network (VCN)
  - Intu
- Experiments and Discussion
- Conclusion
Related Work – IBM Watson and Project Intu

- **IBM Watson**: API service for cognitive task
  - Conversation
  - Text-to-speech
  - Language translator

- **Project Intu**: A platform for intelligent personal assistant service
Related Work – Text-to-speech

Unit selection

Decision tree [7]

Recurrent Neural Network (RNN)
- IBM Watson’s text-to-speech [8]

Hidden Markov model [9]

Dilated convolution neural network
- WaveNet [4]
- Recurrent Neural Network (RNN)
- Deep Voice [10]
Related Work – Voice conversion

Training data

One-to-one w/ parallel data

Many-to-one w/ parallel data

Many-to-one w/o parallel data

Model

Codebook [11]

Gaussian mixture model [12]

Boltzmann machine [13]

Deep belief network [14]

RNN [5]

Lifa Sun et al., 2016 [5]

- Multi-layer bidirectional RNN is used
- Both input & output are from target speech
  ⇒ No parallel data and alignment issue
Outline

• Introduction
• Related Work
  • IBM Watson and Project Intu
  • Text-to-speech (TTS)
  • Voice conversion
• Model Description
  • Voice conversion network (VCN)
  • Intu
• Experiments and Discussion
• Conclusion
Model Description – Voice conversion network (VCN)

- Overall structure of VCN

- Training step
  - Stage I : raw wave $\rightarrow$ linguistic feature
  - Stage II : linguistic feature $\rightarrow$ target speaker’s acoustic feature
Model Description – Voice conversion network (VCN)

- Overall structure of VCN

- Inferring step
  - Stage III: source speech → target speech
Model Description – Voice conversion network (VCN)

- **Stage I**
  
  Raw wave $\rightarrow$ speaker-independent linguistic feature

- **Mel-frequency cepstral coefficients (MFCCs)**
  
  - A kind of speech’s acoustic feature representation
  
  - Energy of each filter bank on mel-scale
Model Description – Voice conversion network (VCN)

• Stage I

   Raw wave → speaker-independent linguistic feature

• Feature-based Maximum Likelihood Linear Regression (fMLLR)
  • Speaker adaptation method transforming speech’s feature vector $x$ [15]
  • Finds affine transformation weight $W$ maximizing likelihood of the speech
Model Description – Voice conversion network (VCN)

- **Stage I**
  - Raw wave → speaker-independent linguistic feature

- **Phonetic Class Posterior Probabilities (PPPs)**
  - Probabilities of phonetic class for each piece of speech
  - *Phoneme*’s representation is limited
  - The number of class of *triphone* is too large
  - *Senone* is cluster of triphones which are similar

Cat

Phoneme: $k$, $æ$, $t$

Triphone: /$kæt$/
Model Description – Voice conversion network (VCN)

• **Stage I**

Raw wave → speaker-independent linguistic feature

• TIMIT corpus [20] is used
• MFCC, fMLLR and PPP are mapped using Kaldi toolkit [16]
• Speaker-independent auto speech recognizer (SI-ASR) maps MFCC (acoustic) feature to PPP (linguistic) feature
### Model Description – Voice conversion network (VCN)

**Stage II**

- **Deep bidirectional long short-term memory (DBLSTM)**
  - Multi-layer recurrent neural network with LSTM cell
  - It consists of forward and backward directional LSTM

SI-phonetic feature $\rightarrow$ acoustic feature of target speaker
Model Description – Voice conversion network (VCN)

- **Stage II**

  - Mel-cepstral Coefficients (MCEPs)
    - Another feature representation of speech
    - Mel-cepstrum analysis of spectrum $H(z)$ to find coefficients $c_\alpha(m)$ [17]

  - SI-phonetic feature → acoustic feature of target speaker
Model Description – Voice conversion network (VCN)

- **Stage II**

  - Only requires target speech to achieve input and label
  - **Deep bidirectional LSTM model** (DBLSTM) is used to map PPP (linguistic) feature to target speech’s MCEP (acoustic) feature

**SI-phonetic feature → acoustic feature of target speaker**
Model Description – Voice conversion network (VCN)

- **Stage III**

- **Fundamental Frequency (Fo)**
  - Lowest frequency of a periodic waveform [18]
  - It is related with pitch of voice

Acoustic feature → raw wave of target speech
Model Description – Voice conversion network (VCN)

- **Stage III**
  - **Aperiodicity Component (AC)**
    - Non-periodic features of speech
    - It contains details of speech

Acoustic feature → raw wave of target speech
Model Description – Voice conversion network (VCN)

- **Stage III**

  - Whole model is achieved by pipelining the models of previous stages
  - STRAIGHT vocoder [19] is used to convert acoustic features to raw wave

**Acoustic feature → raw wave of target speech**
Model Description – Intu

- Intu structure: echoing model

1. **MIC** input speech
2. (Text extractor) speech to text
3. (Echo agent) change the type
4. (WinSpeech gesture) text to speech
5. (Voice conversion (VCN))
6. (SPK) output speech

**User’s speech** → (Intu voice’s speech) → target voice’s speech
Outline

• Introduction
• Related Work
  • IBM Watson and Project Intu
  • Text-to-speech (TTS)
  • Voice conversion
• Model Description
  • Voice conversion network (VCN)
  • Intu
• Experiments and Discussion
• Conclusion
Experiments and Discussion

Two experiments

1. Additional \textbf{time} measurement

2. Varying \textbf{size} of the target speech samples
Experiments and Discussion

• Additional time measurement

Results - w.r.t step

Results - w.r.t task

Feature extraction is a major factor of time delay
Experiments and Discussion

- Additional time measurement

Main proposal 1 – Parallel processing
← SI-ASR and DBLSTM processes are independent of early process of vocoder
Experiments and Discussion

• Additional time measurement

Main proposal 2 – Extracting feature of Intu’s voice in advance
← IBM Watson TTS follows unit selection method
Experiments and Discussion

- Additional time measurement

80.7% time reduction
Varying the size of target speech samples

- **Mel-cepstral distortion (MCD)**

  MCEP distance between original & reconstructed target speech [5]

  \[
  MCD(dB) = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{D} (c_d - c_d^{converted})^2}
  \]

```
Target speech  ----> VCN  ----> Reconstructed (target) speech
```

\[\oplus\]
Experiments and Discussion

Varying the size of target speech samples

- **Mel-cepstral distortion (MCD)**

  MCEP distance between original & reconstructed target speech [5]

  \[
  MCD(dB) = \frac{10}{\ln 10} \sqrt{\frac{2}{D} \sum_{d=1}^{D} (c_d - c_{d^\text{converted}})^2}
  \]

  **High MCD**

  **Low MCD**
Experiments and Discussion

- Varying the size of target speech samples

MCD for training set

MCD for validation set

\[ \therefore 100+ \text{ of target speech samples avoid overfitting} \]

\[ = 10\sim \text{ min of target speech} \]
Outline

• Introduction
• Related Work
  • IBM Watson and Project Intu
  • Text-to-speech (TTS)
  • Voice conversion
• Model Description
  • Voice conversion network (VCN)
  • Intu
• Experiments and Discussion
• Conclusion
Conclusion

- TTS + VCN is suitable for voice customization service
- Parallel & pre-processing for optimization of inference time
- Our design requires user 10~ min of target speech
Acknowledgements

• Members of our laboratory

A picture with my advisor, his advisor (my academic grandpa), and lab members
Data Science & Artificial Intelligence Laboratory
(http://ailab.snu.ac.kr)

• Researchers in IBM Korea
Reference


