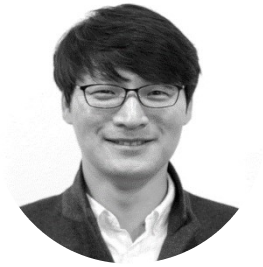
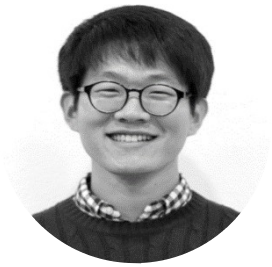
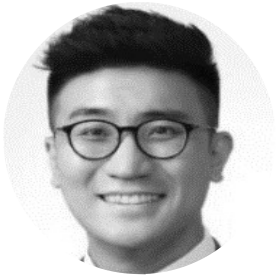
Customization of IBM Intu's Voice by Connecting Text-to-Speech Services with a Voice Conversion Network

Jongyoon Song[1] Jaekoo Lee[1] Hyunjae Kim[1]
Euishin Choi[2] Minseok Kim[2] Sungroh Yoon[1]

[1]ECE, Seoul National University
[2]IBM Korea




HICSS IBM VERI LUX TAS MEA

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]

Introduction

Preset voice

Introduction

Customized voice

Introduction

Customized voice

Voice customization
= easy for users to train

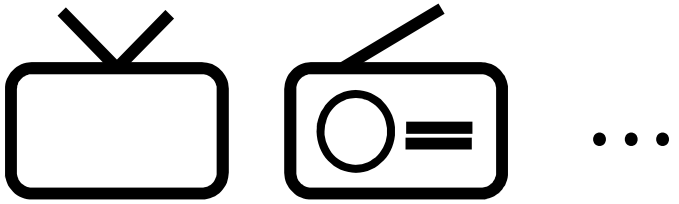
Introduction

Customized voice

Voice customization
= easy for users to train

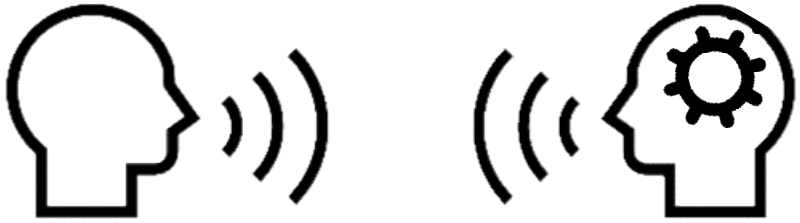
Our design
- Small target speech data
- No parallel data

Introduction



1. Gather 10~ min of speech of preferred voice

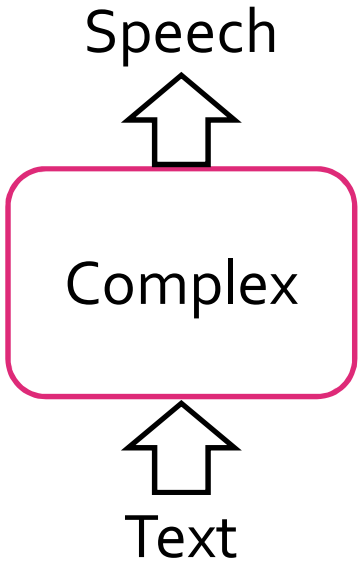
2. Send to Intu to train voice customization module



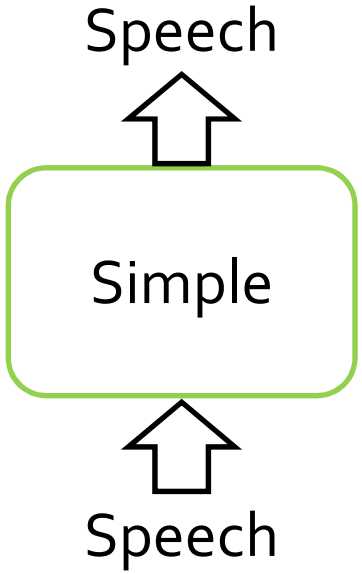
3. Users can talk with AI speaker with customized voice

Introduction

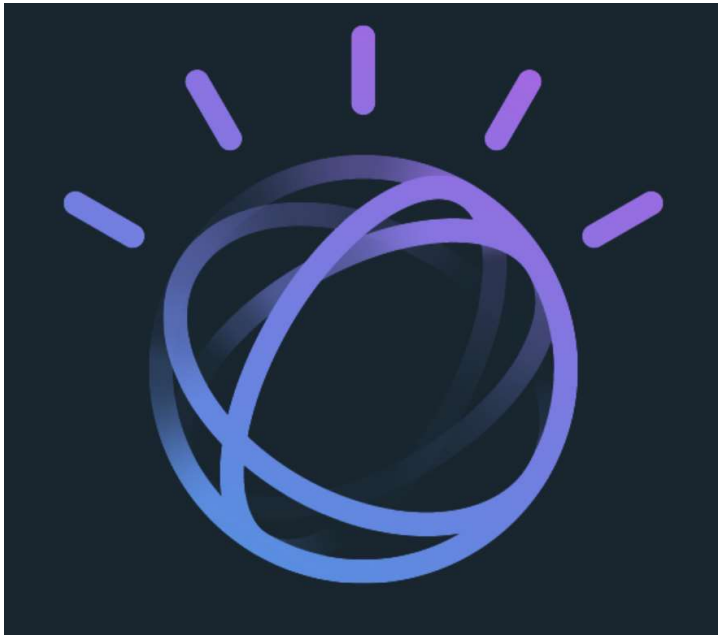
- Text-to-speech (TTS)
  - Text : linguistic & phonetic feature
  - Speech : phonetic & acoustic feature
  - Requires relatively complex model
  - Needs around 30 min of speech per voice [4]

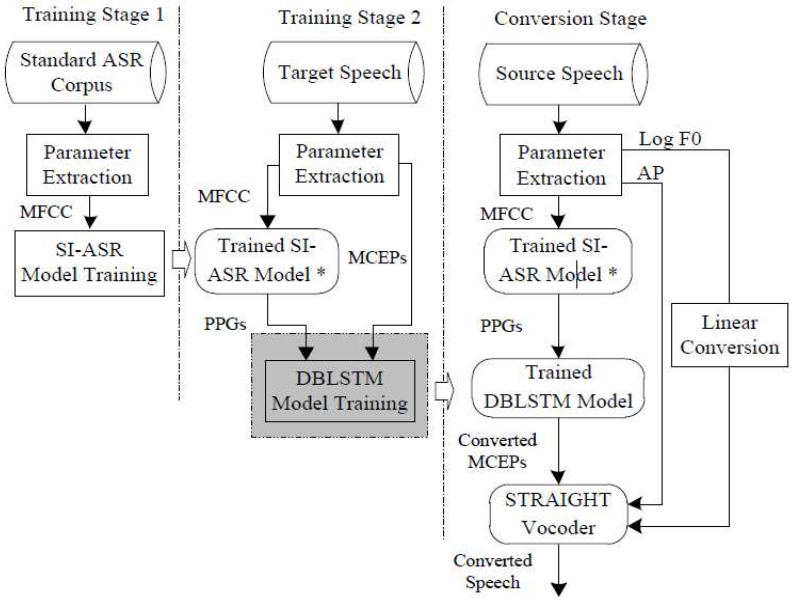


- Voice conversion
  - Inputs and outputs have same feature domain
  - Requires relatively simple model
  - Needs around 10 min of speech per voice [5]



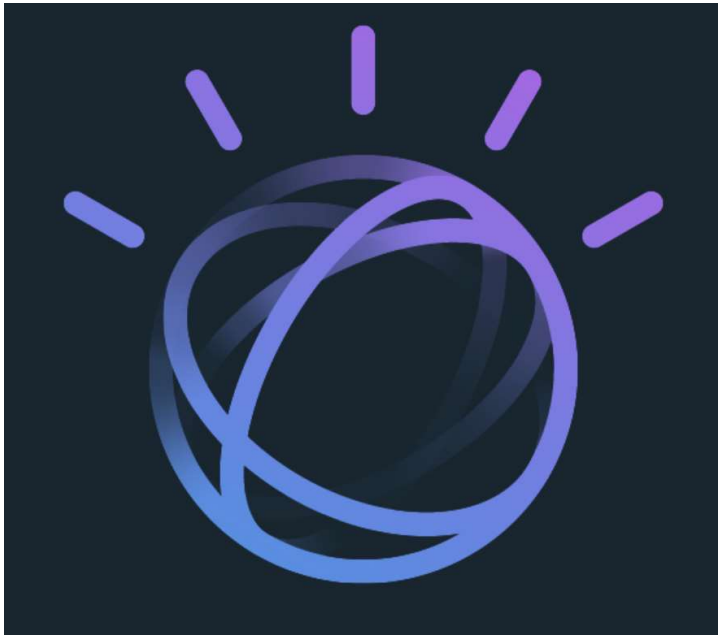
Introduction

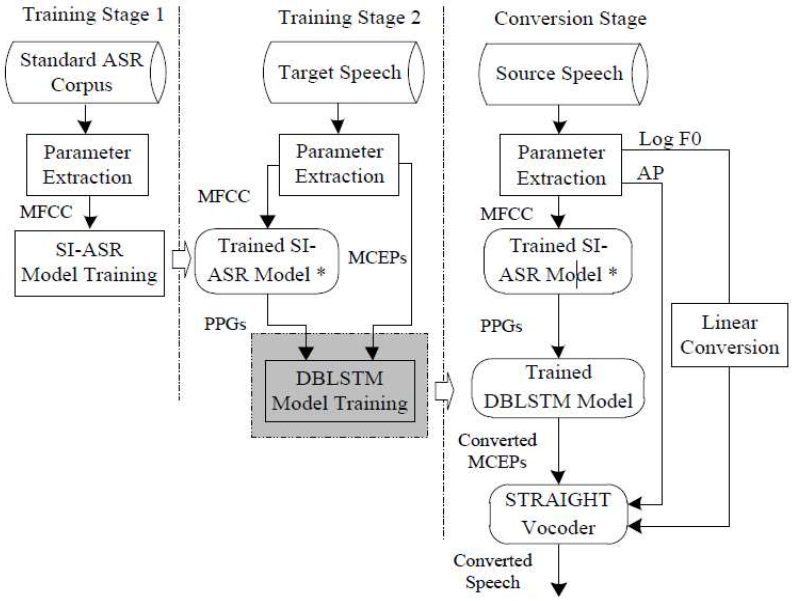




Project Intu [6] + Voice conversion [5]

Introduction



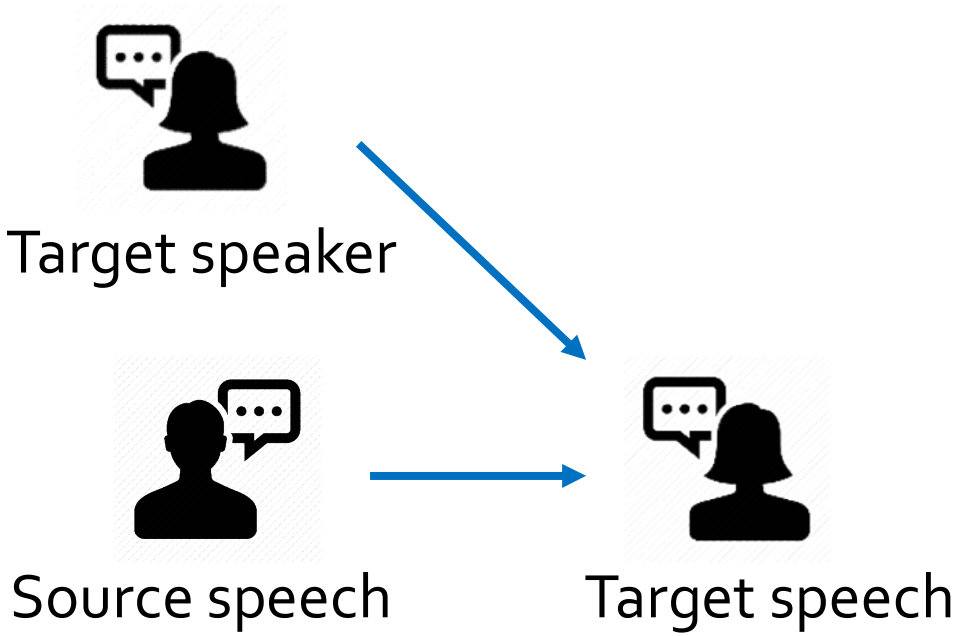


Pre-trained using public speech data

+

Trained using 10~ min of target speech

Introduction



- 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

Contribution

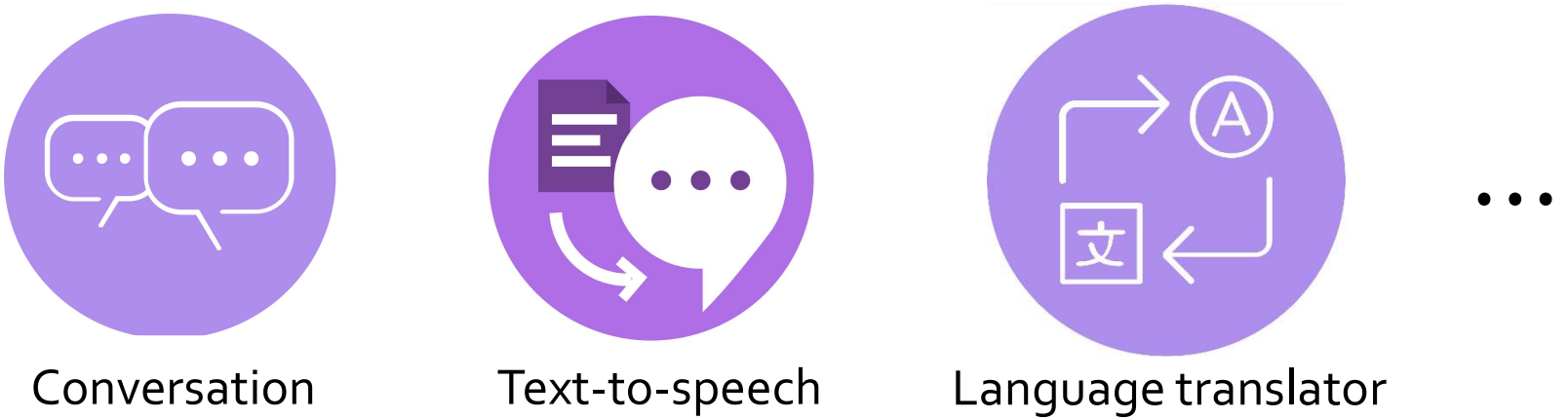
- Voice customization for ML-as-a-Service design
- Methods for inference time optimization
- Analysis for proper amount of target speech

Outline

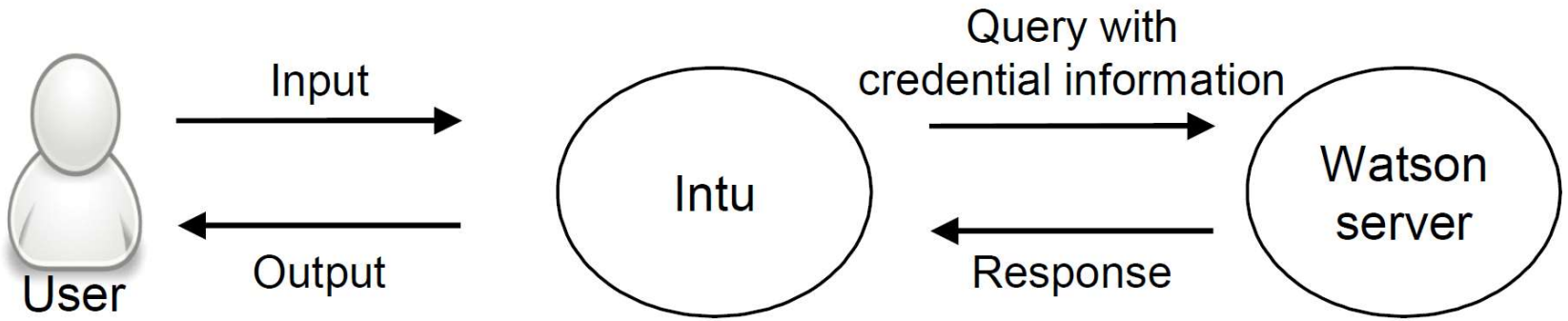
- Introduction
- Related Work
  - IBM Watson and Project Intu
  - Text-to-speech (TTS)
  - Voice conversion
- Model Description
  - Voice conversion network (VCN)
  - Intu
- Experiments and Discussion
- Conclusion

Related Work – IBM Watson and Project Intu

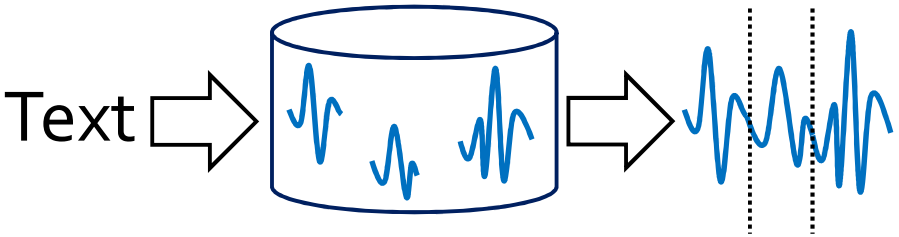
- IBM Watson : API service for cognitive task



- Project Intu : A platform for intelligent personal assistant service

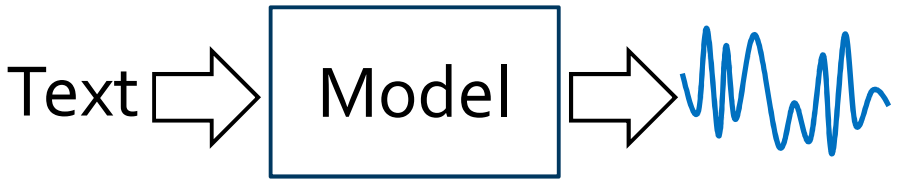


Related Work – Text-to-speech



Unit selection

- Decision tree [7]
- Recurrent Neural Network (RNN)
  - IBM Watson's text-to-speech [8]



Parametric generation

- Hidden Markov model [9]
- Dilated convolution neural network
  - WaveNet [4]

  Recurrent Neural Network (RNN)
  - Deep Voice [10]

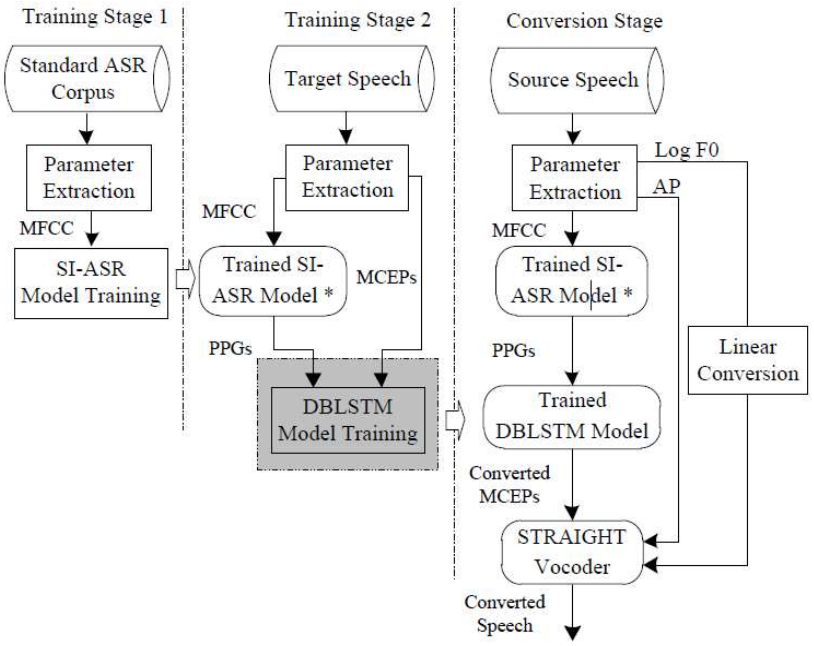
Related Work – Voice conversion

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

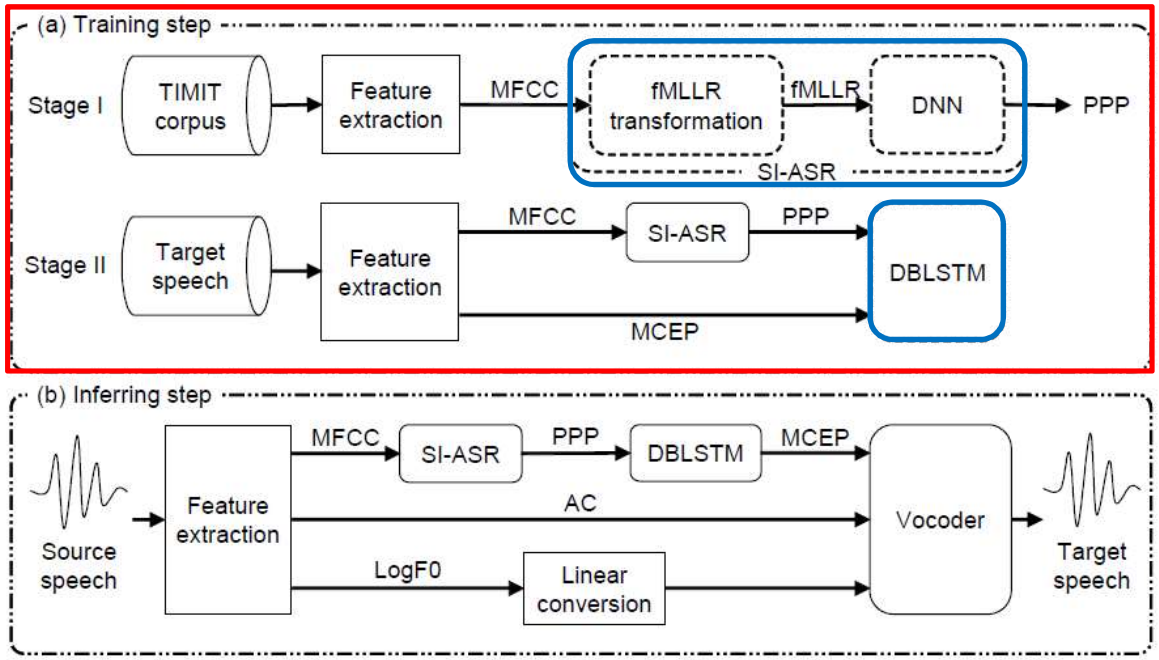


Outline

- 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)

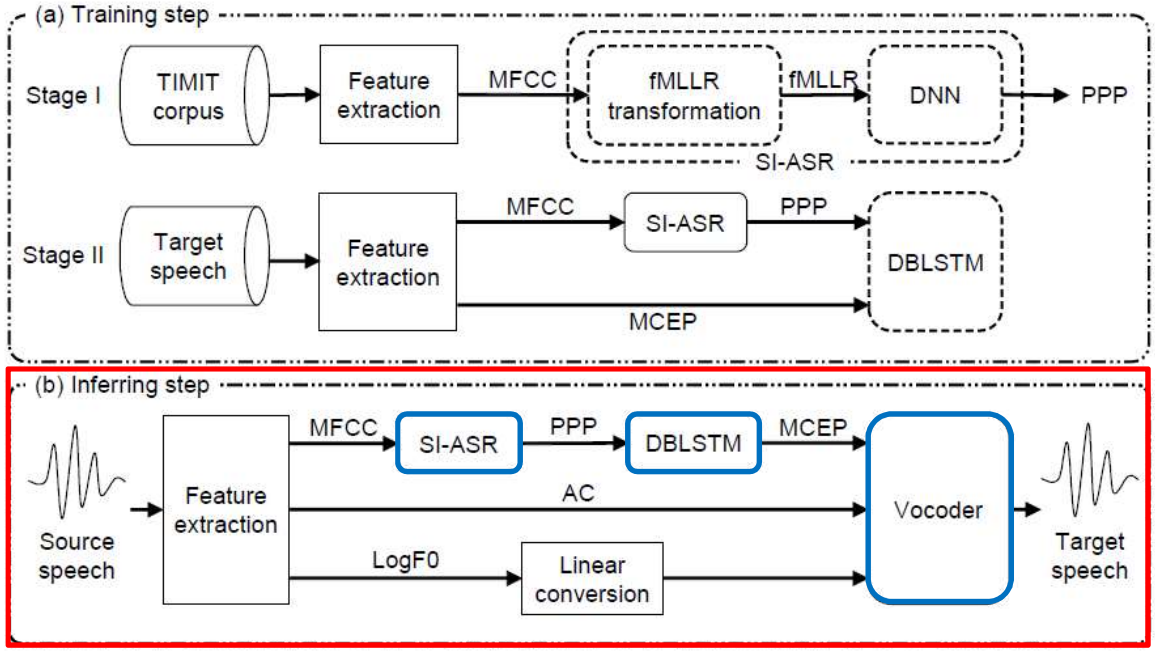
- Overall structure of VCN



- Training step
  - Stage I : raw wave → linguistic feature
  - Stage II : linguistic feature → target speaker's acoustic feature

Model Description – Voice conversion network (VCN)

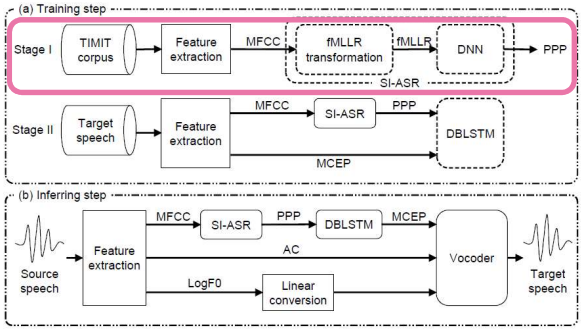
- Overall structure of VCN

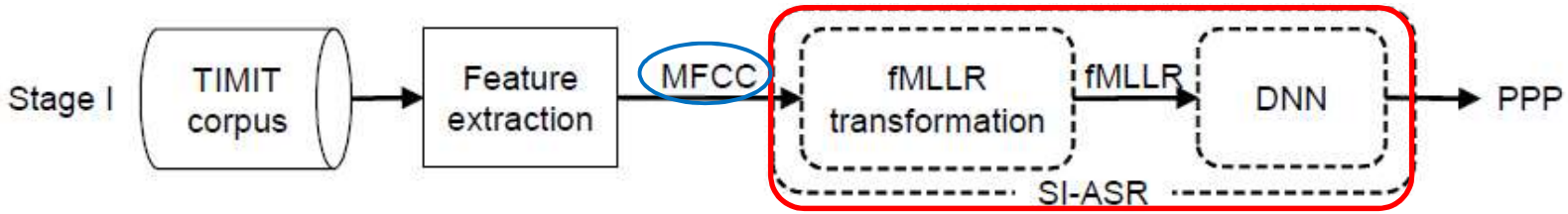


- Inferring step
  - Stage III : source speech → target speech

Model Description – Voice conversion network (VCN)

- Stage I



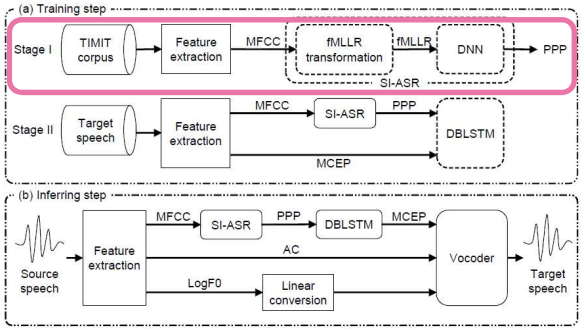


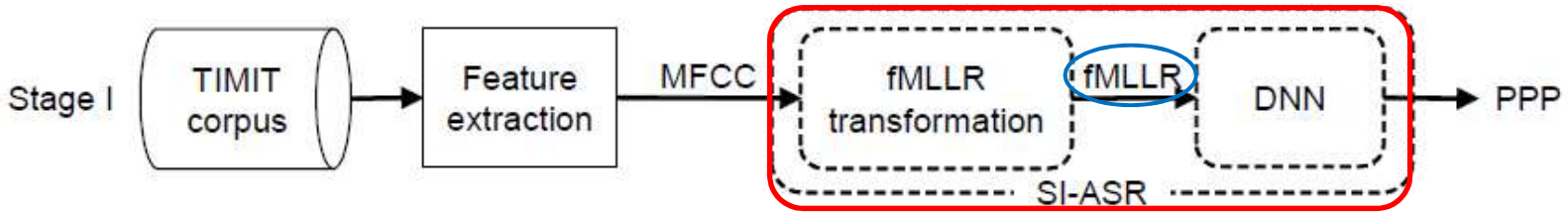
Raw wave → speaker-independent linguistic feature

- 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



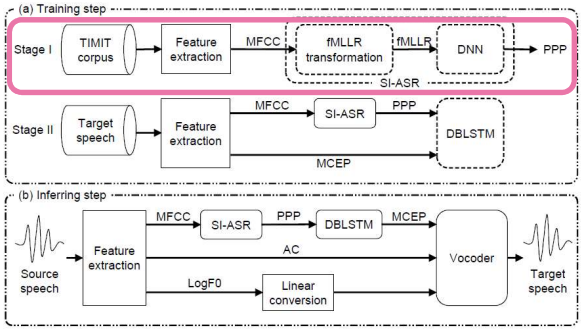


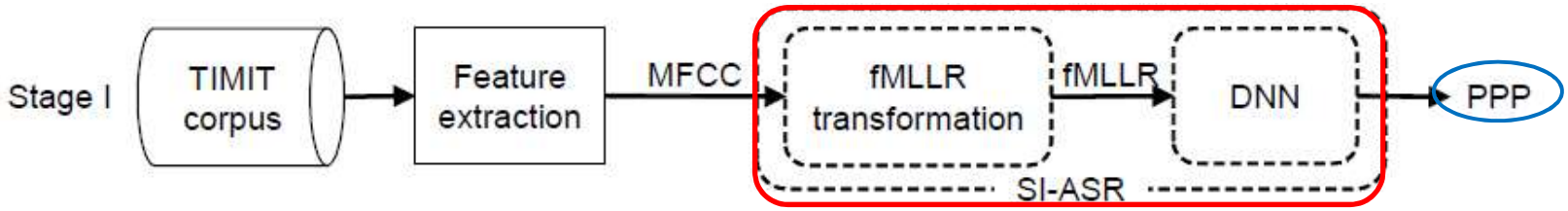
Raw wave → speaker-independent linguistic feature

- 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





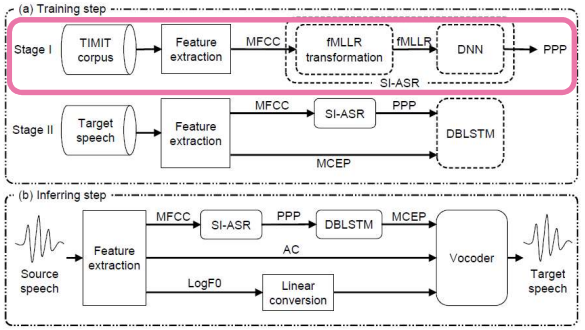
Raw wave → speaker-independent linguistic feature

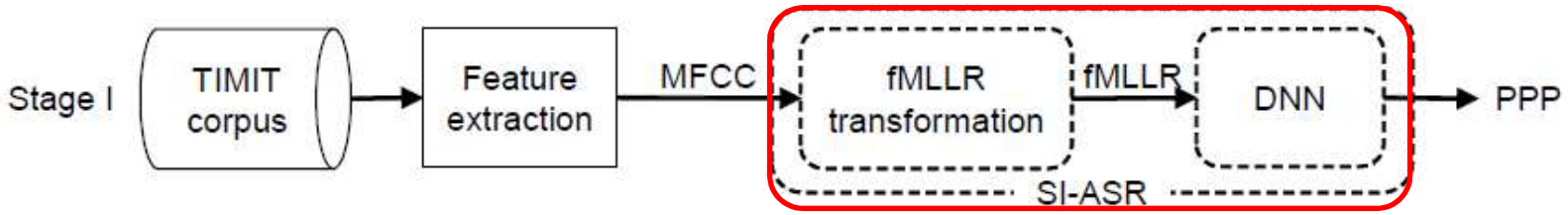
- Phonetic Class Posterior Probabilities (PPPs)
  - Probabilities of phonetic class for each piece of speech
  - Phoneme's representation is limited
  - The number of class of triphone is too large
  - Senone is cluster of triphones which are similar

Cat
Phoneme : k, æ, t
Triphone : /kæt/

Model Description – Voice conversion network (VCN)

- Stage I



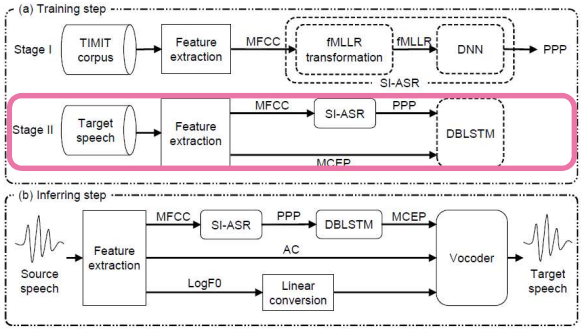


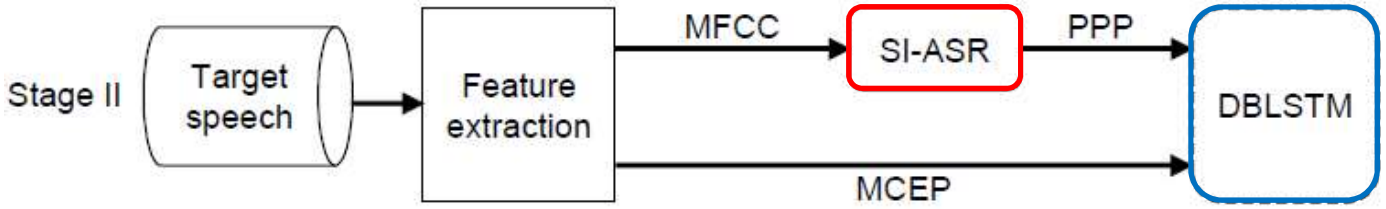
Raw wave → speaker-independent linguistic feature

- 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



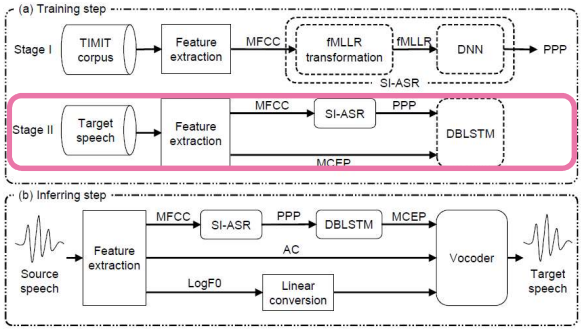


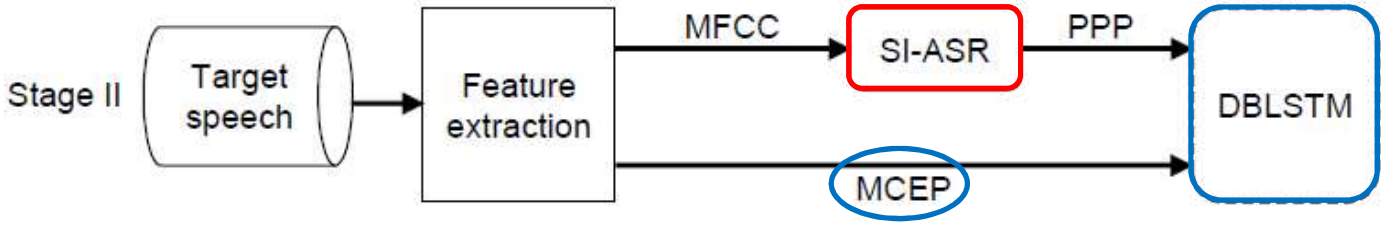
SI-phonetic feature → acoustic feature of target speaker

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

Model Description – Voice conversion network (VCN)

- Stage II



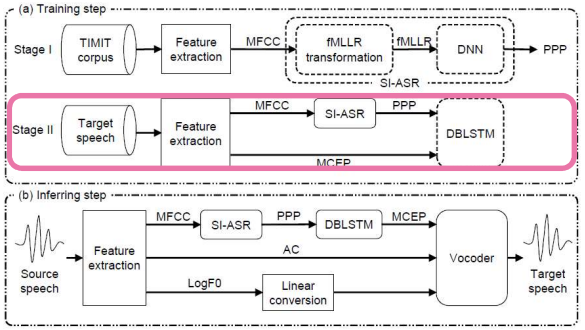


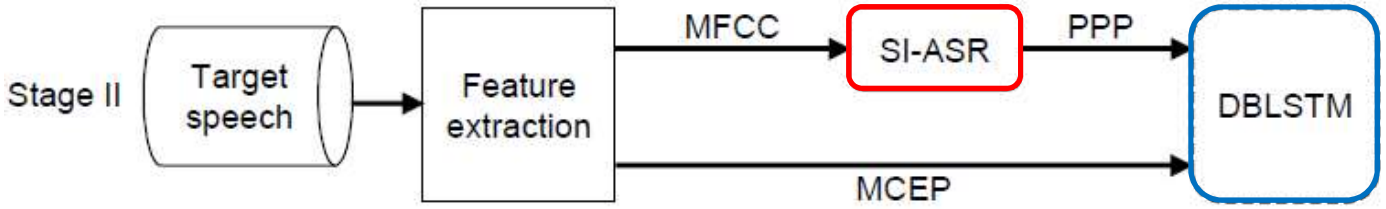
SI-phonetic feature → acoustic feature of target speaker

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

Model Description – Voice conversion network (VCN)

- Stage II



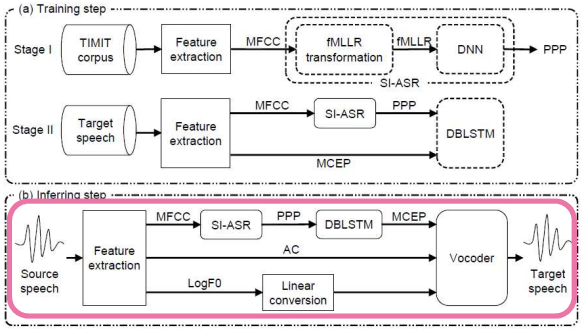


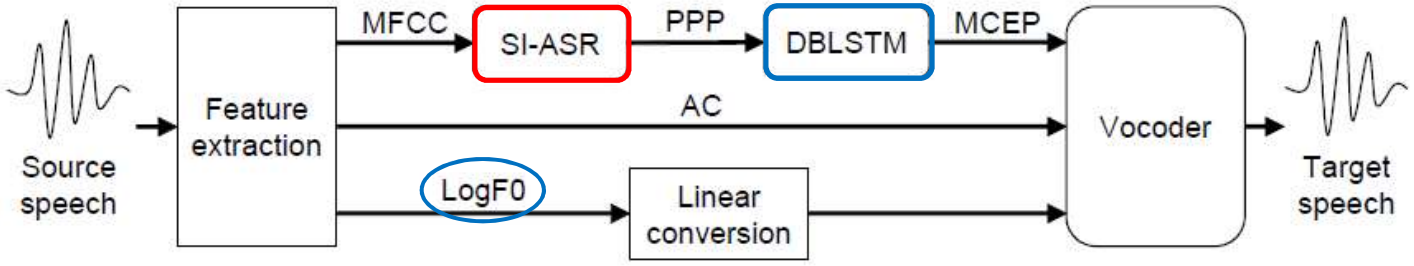
SI-phonetic feature → acoustic feature of target speaker

- 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

Model Description – Voice conversion network (VCN)

- Stage III



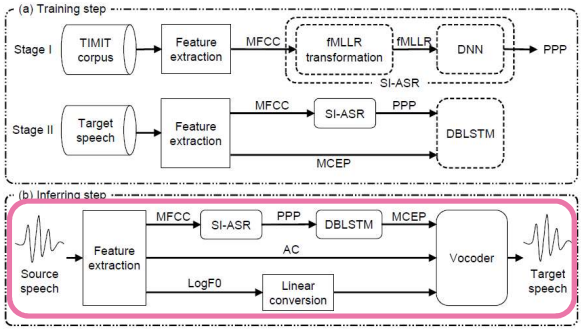


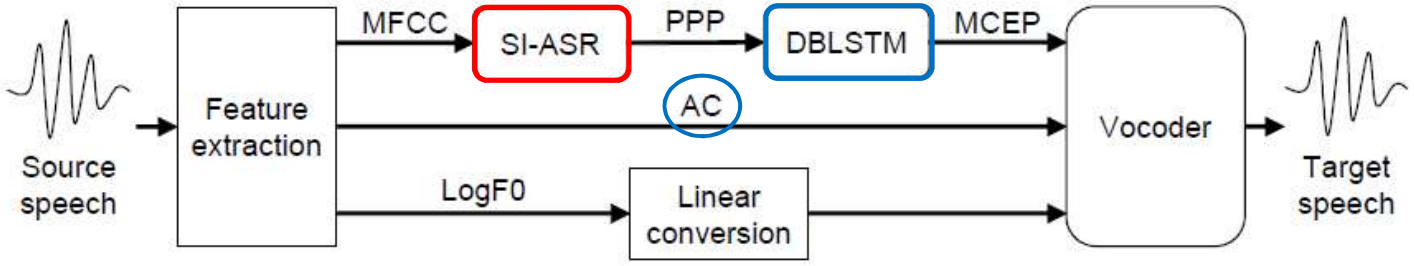
Acoustic feature → raw wave of target speech

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

Model Description – Voice conversion network (VCN)

- Stage III



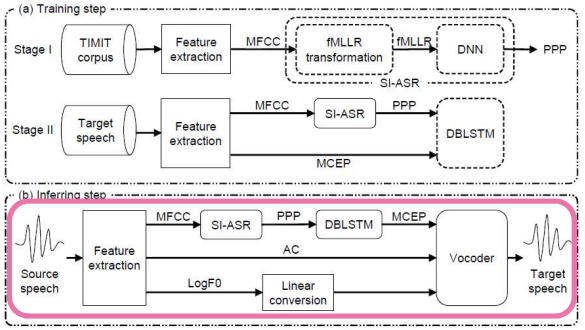


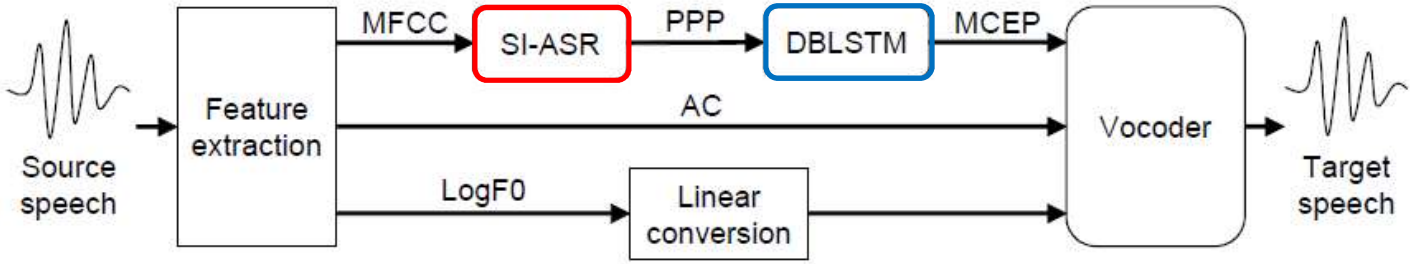
Acoustic feature → raw wave of target speech

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

Model Description – Voice conversion network (VCN)

- Stage III



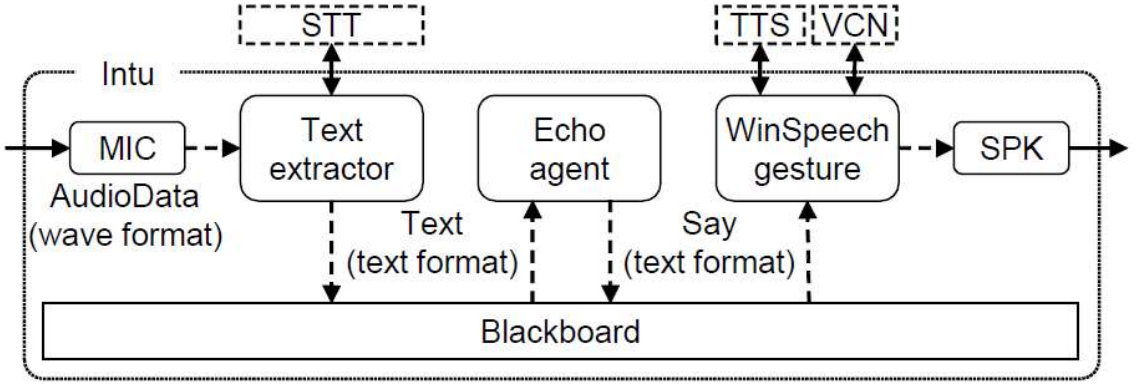


Acoustic feature → raw wave of target speech

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

Model Description – Intu

- Intu structure : echoing model



(MIC) input speech
→ (Text extractor) speech to text → (Echo agent) change the type
→ (WinSpeech gesture) text to speech → voice conversion (VCN)
→ (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

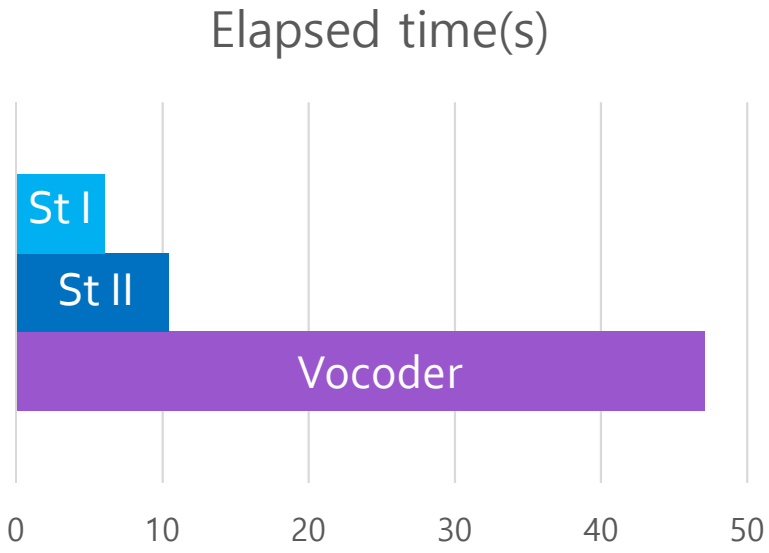
Two experiments

1. Additional **time** measurement
2. Varying **size** of the target speech samples

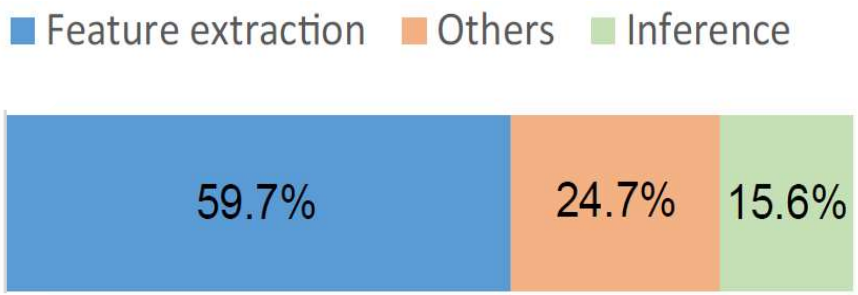
Experiments and Discussion

- Additional time measurement

Results - w.r.t step



Results - w.r.t task



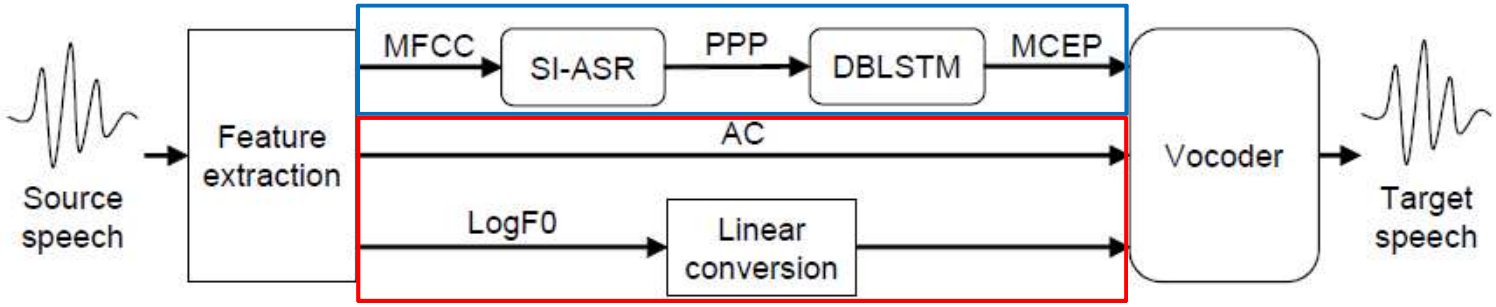
Feature extraction is a major factor of time delay

Experiments and Discussion

- Additional time measurement

Main proposal 1 – Parallel processing

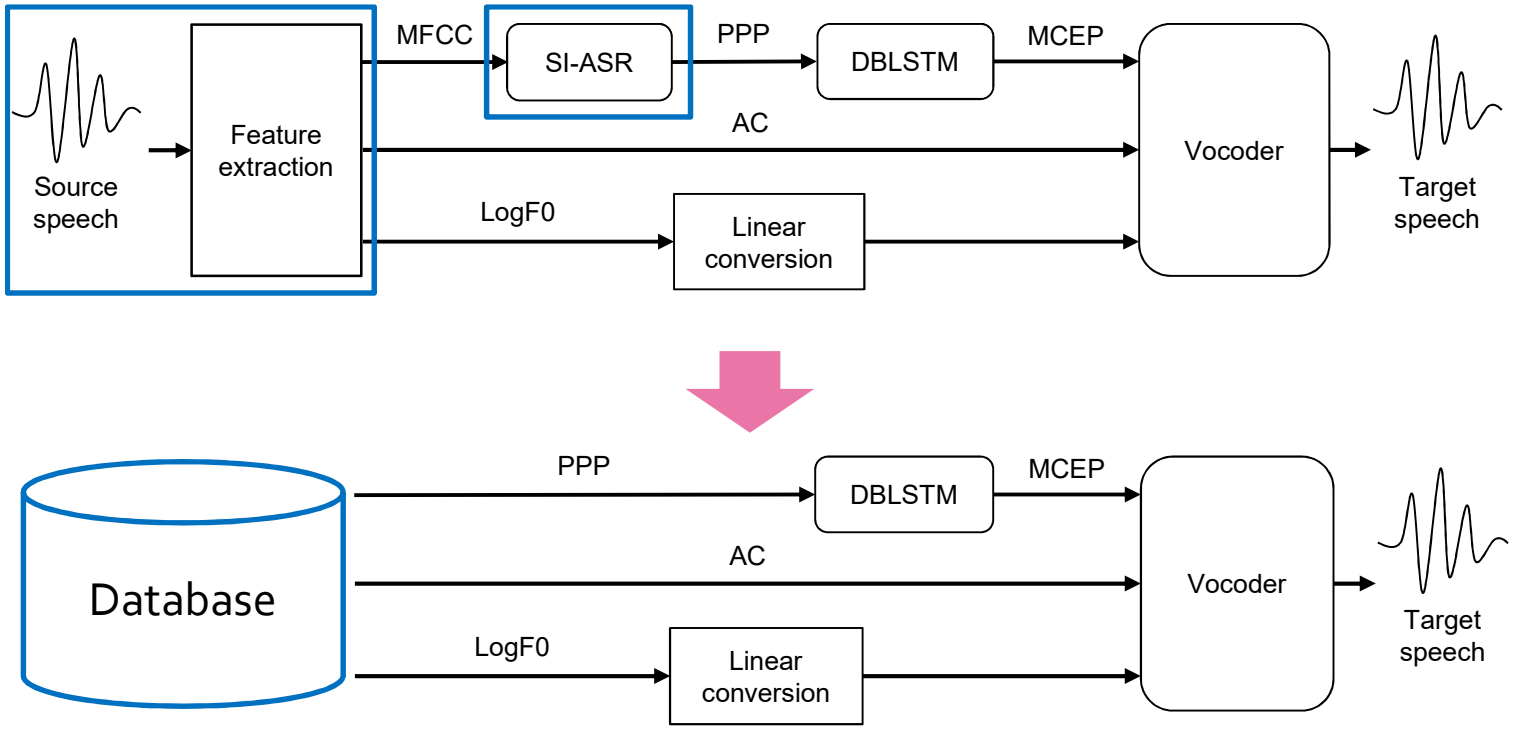
← SI-ASR and DBLSTM processes are independent of early process of vocoder



Experiments and Discussion

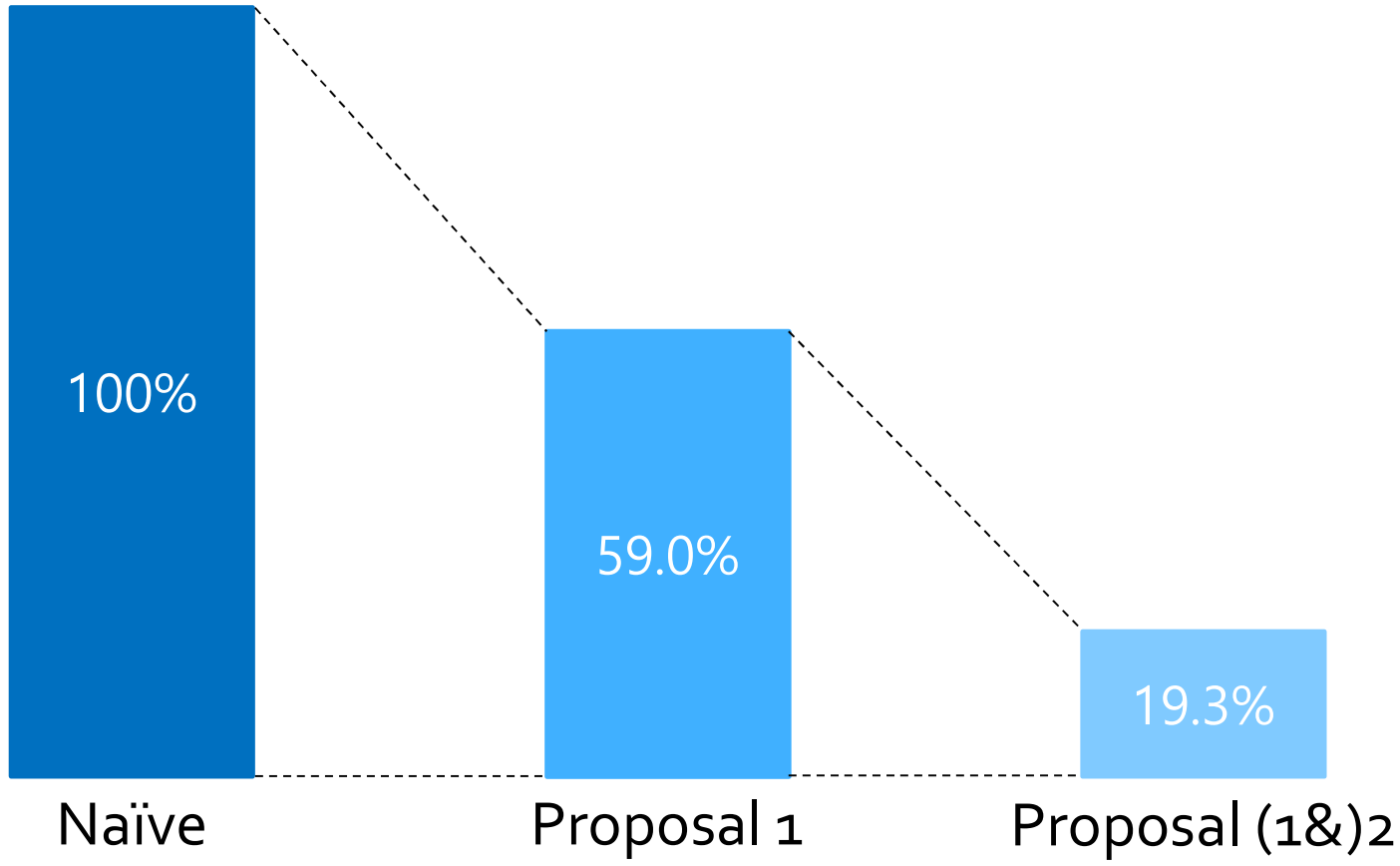
- Additional time measurement

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



Experiments and Discussion

- Additional time measurement



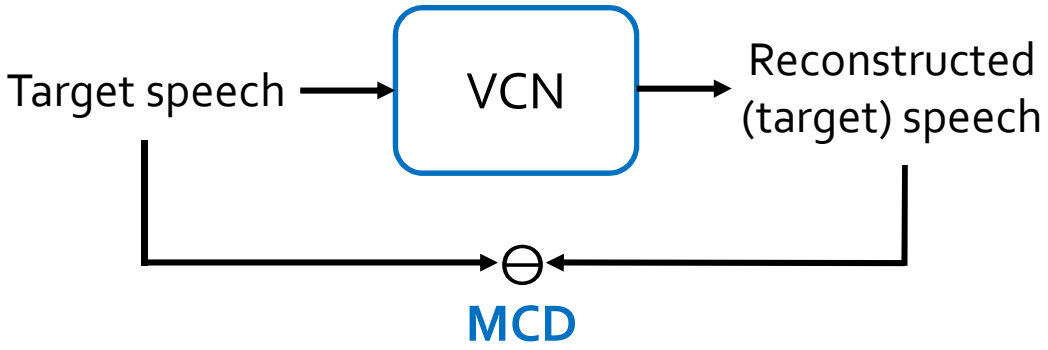
80.7% time reduction

Varying the size of target speech samples

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

  MCEP distance between original & reconstructed target speech [5]

$$MCD(dB) = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{D}(c_d - c_d^{converted})^2}$$



Varying the size of target speech samples

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

  MCEP distance between original & reconstructed target speech [5]

$$MCD(dB) = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{D}(c_d - c_d^{converted})^2}$$

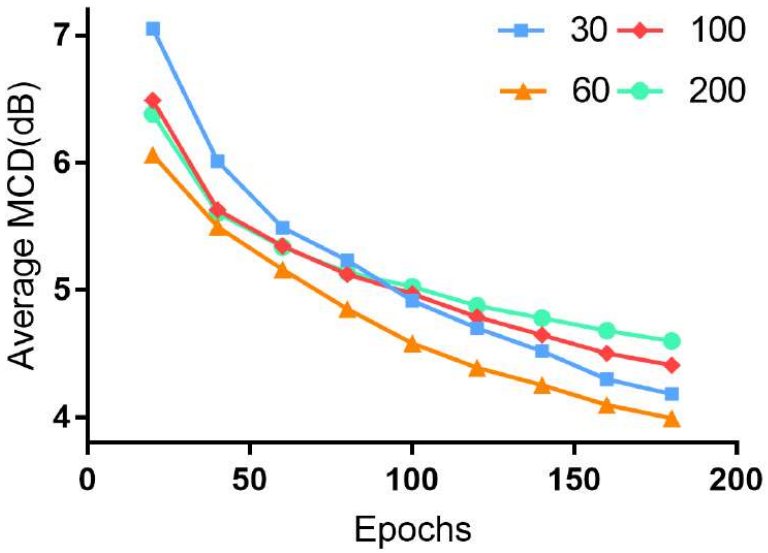
High MCD

Low MCD

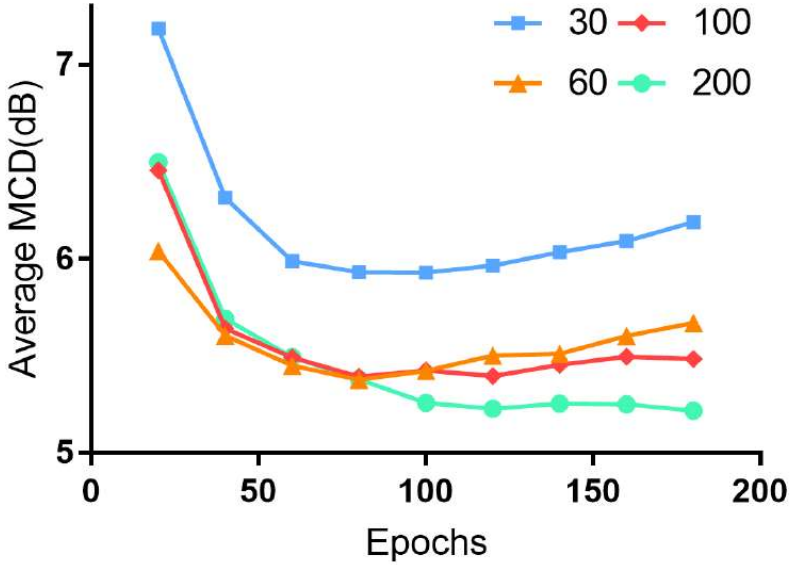
Experiments and Discussion

- Varying the size of target speech samples

MCD for training set



MCD for validation set



∴ 100+ of target speech samples avoid overfitting

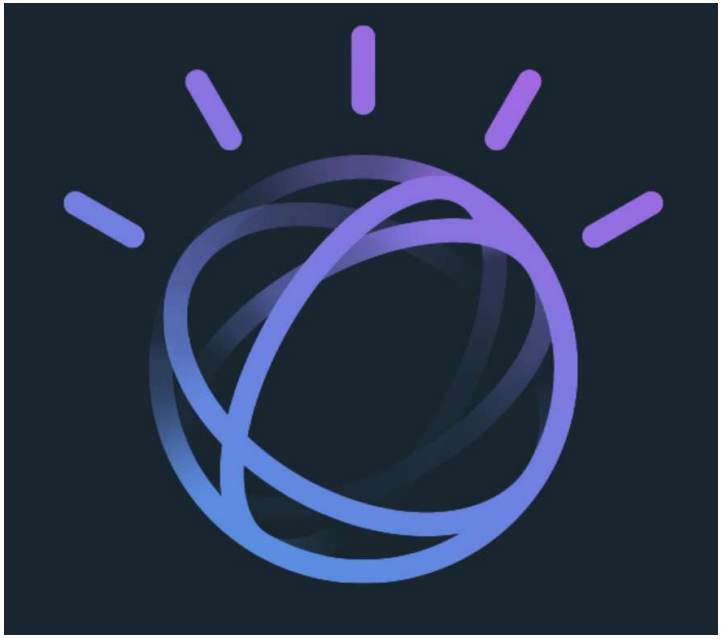
= 10~ min of target speech

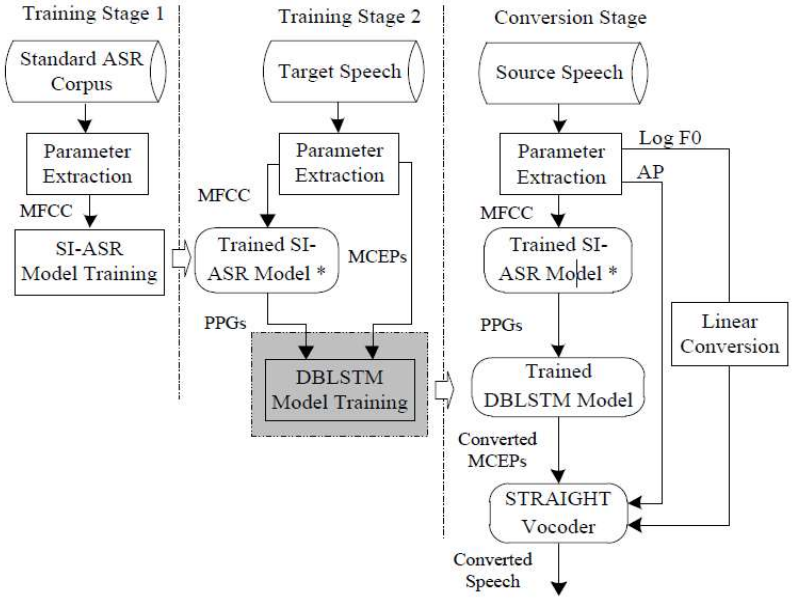
Outline

- 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

[1] "Amazon Echo." https://images-na.ssl-images-amazon.com/images/I/41-v1fozy0L.jpg

[2] "Google Home." https://i5.walmartimages.ca/images/Large/309/511/6000197309511.jpg?odnBound=460

[3] "Apple Siri." https://images.techhive.com/images/article/2016/11/siri-mac-icon-100694914-large.jpg

[4] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio.," *arXiv preprint arXiv:1609.03499*, 2016.

[5] L. Sun, K. Li, H. Wang, *et al.*, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training.," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pp. 1–6, IEEE, 2016.

[6] "IBM Project Intu." https://www.ibm.com/watson/developercloud/project-intu

[7] "IBM Watson: Text-to-Speech." https: //www.ibm.com/watson/developercloud/ doc/text-to-speech/index.html.

[8] R. Fernandez, A. Rendel, B. Ramabhadran, *et al.*, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks.," pp. 2268–2272, Interspeech, 2014.

[9] T. Yoshimura, K. Tokuda, T. Masuko, *et al.*, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis.," in *Sixth European Conference on Speech Communication and Technology*, pp. 2347–2350, 1999.

[10] S. Arik, M. Chrzanowski, A. Coates, *et al.*, "Deep voice: Real-time neural text-to-speech.," *arXiv preprint arXiv:1702.07825*, 2017.

[11] M. Abe, S. Nakamura, K. Shikano, *et al.*, "Voice conversion through vector quantization.," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[12] Y.Stylianou,O.Cappe ́ ,andE.Moulines,"Continuous probabilistic transform for voice conversion.," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[13] Z. Wu, E. Chng, H. Li, *et al.*, "Conditional restricted Boltzmann machine for voice conversion.," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*, pp. 104–108, IEEE, 2013.

[14] T. Nakashika, R. Takashima, T. Takiguchi, *et al.*, "Voice conversion in high-order eigen space using deep belief nets.," pp. 369–372, Interspeech, 2013.

[15] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians.," pp. 1145–1148, Interspeech, 2006.

[16] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The Kaldi speech recognition toolkit.," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584, IEEE Signal Processing Society, 2011.

[17] "Reference manual for speech signal processing toolkit ver. 3.10." http://sp-tk.sourceforge.net/.

[18] "Wikipedia: Fundamental frequency." https://en.wikipedia.org/wiki/Fundamental_frequency

[19] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based fo extraction: Possible role of a repetitive structure in sounds.," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[20] "The DARPA TIMIT acoustic-phonetic continuous speech corpus (TIMIT)." https://catalog.ldc. upenn.edu/docs/LDC93S1/timit.readme. html.