Multiple Feed-forward Deep Neural Networks for Statistical Parametric Speech Synthesis

Shinji Takaki[1], SangJin Kim[2], Junichi Yamagishi[1,3], JongJin Kim[2]

[1] National Institute of Informatics, [2] Naver Labs, [3] The Centre for Speech Technology Research, University of Edinburgh

1. introduction

Statistical parametric speech synthesis (SPSS)
- HMM-based speech synthesis [Tokuda et al.; 00]
- DNN-based speech synthesis [Zen et al.; 12]

DNNs have high potential in SPSS
- Further Investigation of DNNs for other tasks in SPSS is needed

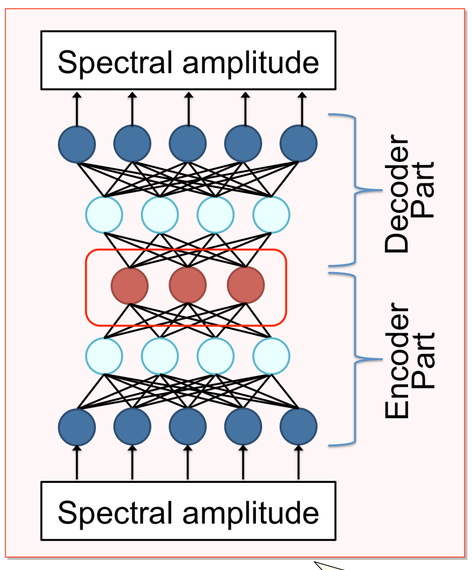
Standard components for SPSS
- Acoustic feature extraction (Mel-cep, LSP)
- Acoustic modeling (HMM, DNN)
- Smoothing (MLPG with delta, Recurrent NN)
- Enhancement (GV, Post-filter)

Feed-forward DNNs are used in this work

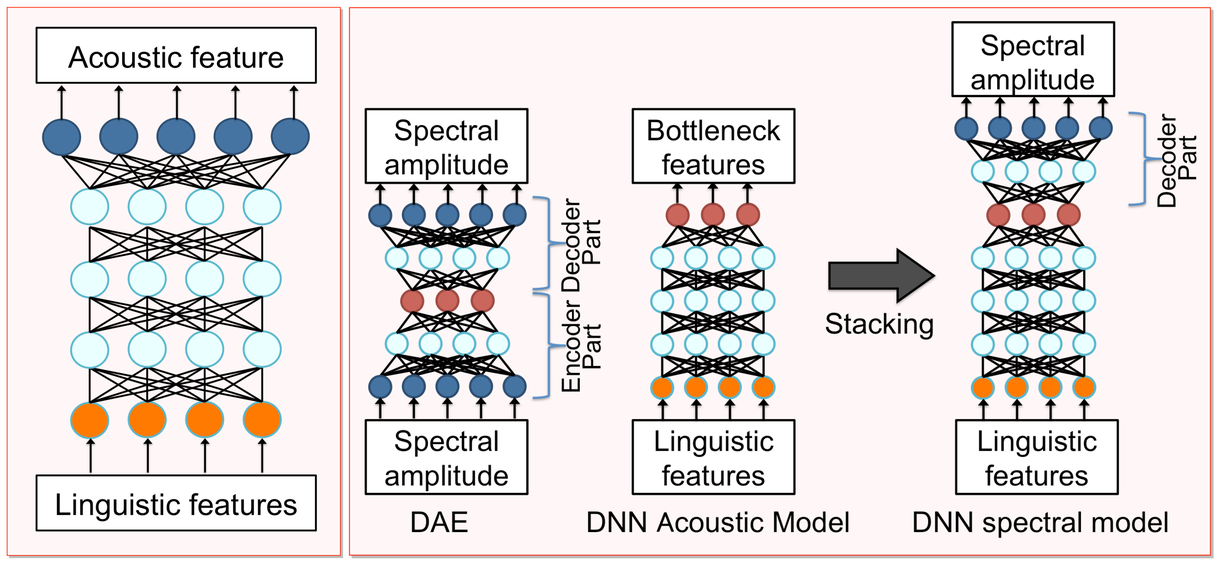
All standard steps of SPSS are performed using DNNs

2. Feed-forward DNNs for SPSS

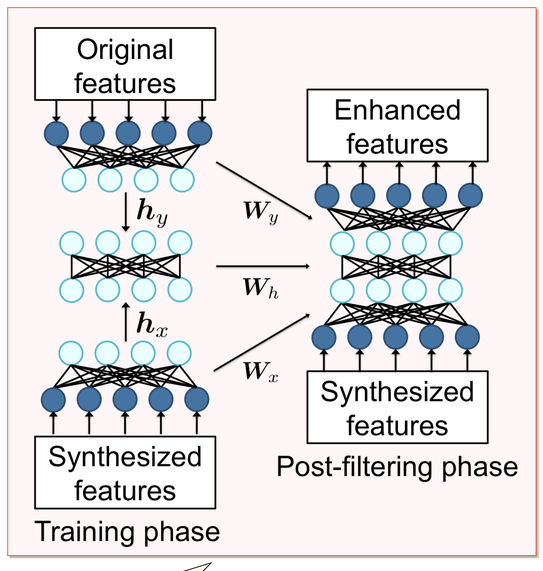
Feature Extraction



Acoustic modeling [Takaki et al.; 15]



Post-filtering [Chen et al.; 14]



Deep Auto-encoder (DAE)
- Typical purpose is dimensionality reduction
- Same features are used as input and output (Spectra obtained from STRAIGHT)
- Outputs of a encoder part can be used as dimensionality reduced features

Non-linear
- Vocal tract has a non-linear

Statistical and unsupervised approach
- Data driven, speaker dependent
- Automatically extract appropriate feature

Extracting low-dimensional spectral parameter

Low-dimentional spectral parameter
→ Quality loss

Direct synthesis of spectral amplitudes
- Catch the spectral fine structure

Difficulty of DNN training
- Local maxima, Vanishing gradient
- High Dimensionality
  - Mel-cepstrum : 60 dims.
  - Spectral amplitude : 2049 dims.

→ Efficient training technique would be needed

Pre-training with a DAE and a DNN AM
- The general flow for constructing SPSS

Function-wise pre-training for DNN-based speech synthesis

Model of the difference between synthesized and natural spectra

Consecutive inputs & outpus
- Considering the differences in the time-frequency domain
- Spectral peak enhancement
- Spectral smoothing

Smoothing and enhancement steps are simultaneously performed

Combination of DNNs are used for constructing the proposed system

3. Experiments

・Methods

US	Unit selection based speech synthesis
HMM	HMM-based speech synthesis with GV
MDNNs1	Proposed technique F0 and aperiodicity measures : *HMM* synthesis
DNN	Conventional single DNN speech synthesis with a signal processing-based post-filter
MDNNs1	Proposed technique F0 and aperiodicity measures : *DNN* synthesis

Acoustic features for HMM / DNN : Spectral parameter, log F0, 25-dim band aperiodicity and their $\Delta + \Delta^2$

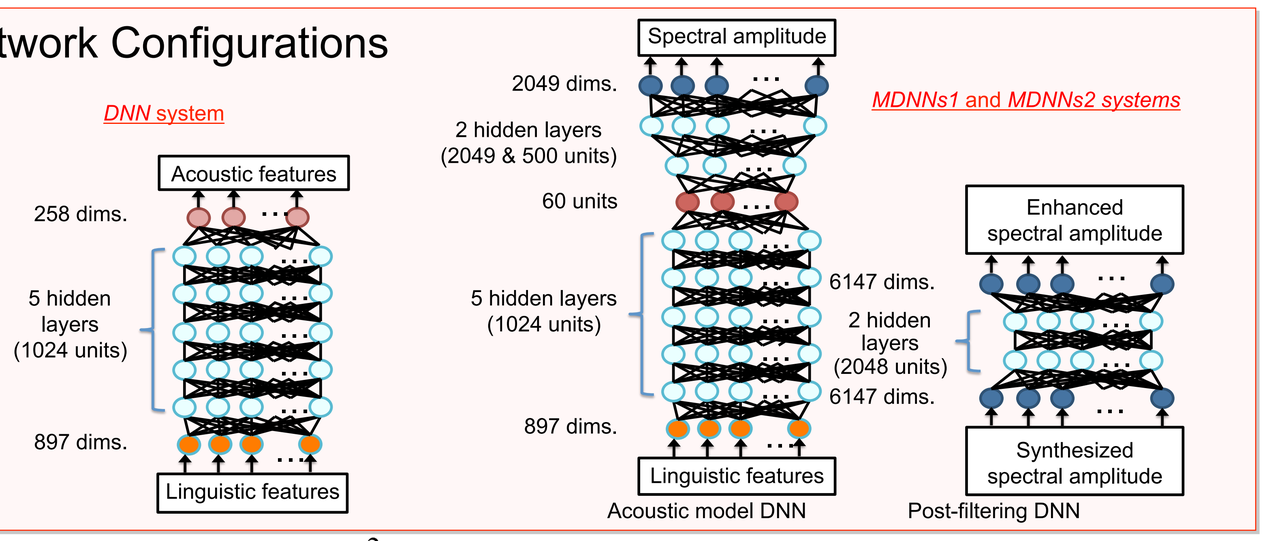
・English (Test : 15 samples * 33 subjects)

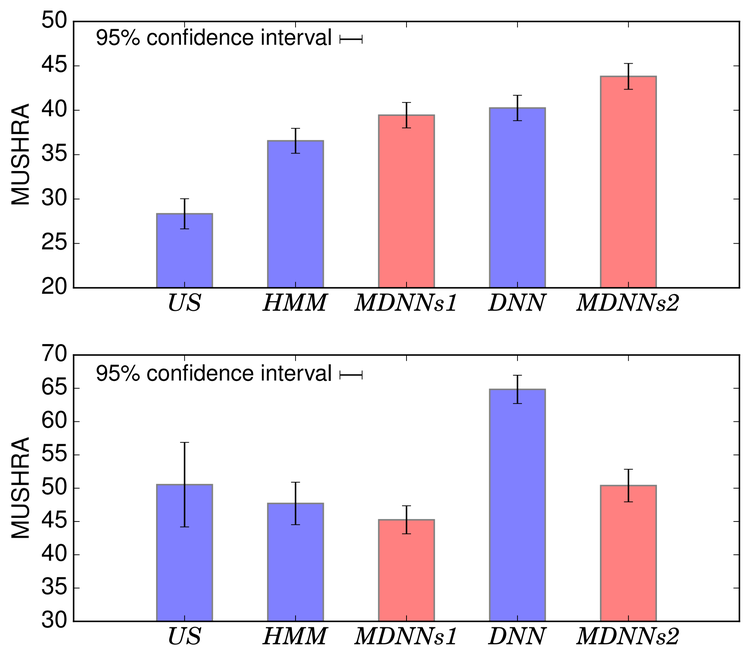
Database	Professional female 12,085 utts. (17 hours)
Test set	200 sentences
Sampling rate	48 kHz
FFT points / Cepstrum dims	4096 (2049-dim) / 59

・Korean (Test : 15 samples * 23 subjects)

Database (Manual alignments)	Professional female 11,937 utts. (38 hours)
Test set	200 sentences
Sampling rate	16 kHz
FFT points / Cepstrum dims	2048 (1025-dim) / 39

Network Configurations





US is rated higher than *HMM* in Korean
- Manual alignments & Large corpus size

DNN outperformed *US* in both
- Although there were not many artifacts in *US*, subjects did not prefer *US* samples in Korean

HMM v.s. *MDNNs1* and *DNN* v.s. *MDNNs2*
- Proposed combination systems produce more natural sounds in English
- In Korean, completely opposite outcome to the English findings

→ Investigation into 16 kHz sample is needed

Future work
- Proposed framework for F0 and aperiodicity
- Recurent and convolution networks