DNN-Based Unit Selection Using Frame-Sized Speech Segments

Zhi-Ping Zhou, Zhen-Hua Ling

University of Science and Technology of China

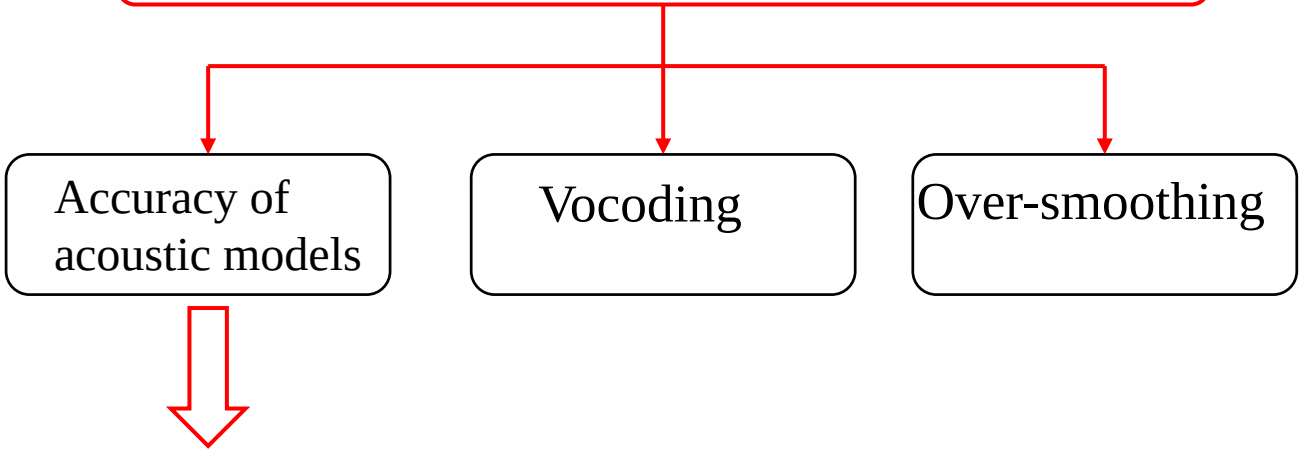
Oct. 20th, 2016

Outline

1. Motivation
2. Method
3. Experiments
4. Conclusions

Motivation

- HMM-based statistical parametric speech synthesis [Tokuda,2004]
  - Advantages: flexibility, small footprint, robustness；
  - Disadvantage: degraded speech quality.



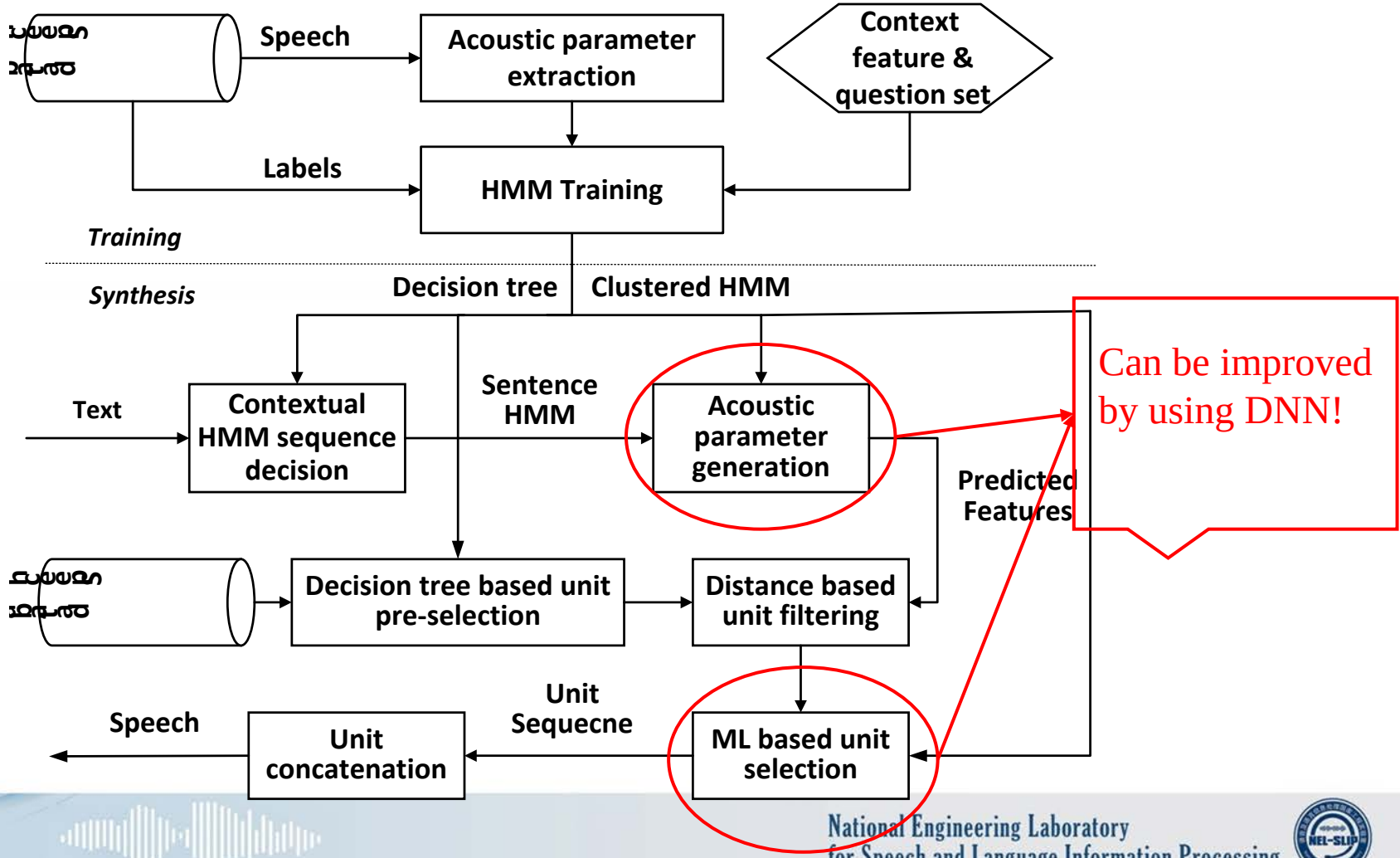
- DNN-based statistical parametric speech synthesis [Zen,2013]
  - lead to better performance than HMM-based acoustic model

Motivation

- Unit selection and waveform concatenation speech synthesis [Iwahashi,1992]
  - Advantage: better quality of synthesized speech;
  - Disadvantages: big corpus, unstable, discontinuity between two units.
- Frame-sized unit system [Hirai, 2004]
  - HMM-based unit selection using frame-sized speech segments was proposed. [Ling, 2006]

Motivation

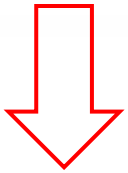
HMM-based unit selection using frame-sized speech segments



Motivation

HMM-based unit selection using frame-sized speech segments

Acoustic model: HMM vs DNN



DNN-based unit selection using frame-sized speech segments

Outline

1. Motivation
2. Method
3. Experiments
4. Conclusions

Method

Framework

➢Unit size

- Frames of 5ms length.

➢Two cost functions

- Target cost
  - Calculate the distances of acoustic features between a candidate unit and a target unit predicted by DNN.
- Concatenation cost
  - measures the discontinuity between two consecutive candidate units using our DNN model.

➢Unit selection procedure

- dynamic programming (DP) search using target costs and concatenation costs. [Sakoe,1978]

Method

Assuming that the sentence to be synthesized has N frames, $u = \{u_1, u_2, \cdots u_N\}$ is a candidate sequence, and $\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots \boldsymbol{w}_N\}$ are the context information of all frames, the optimal sequence u^* is determined as follows:

$$u^* = \arg\min_u C(u, w)$$

where

$$C(u, w) = \sum_{n=1}^{N} C_{t\arg}(u_n, w_n) + W_{con} \sum_{n=T+1}^{N} C_{con}(u_{n-T}, \cdots u_n, w_n)$$

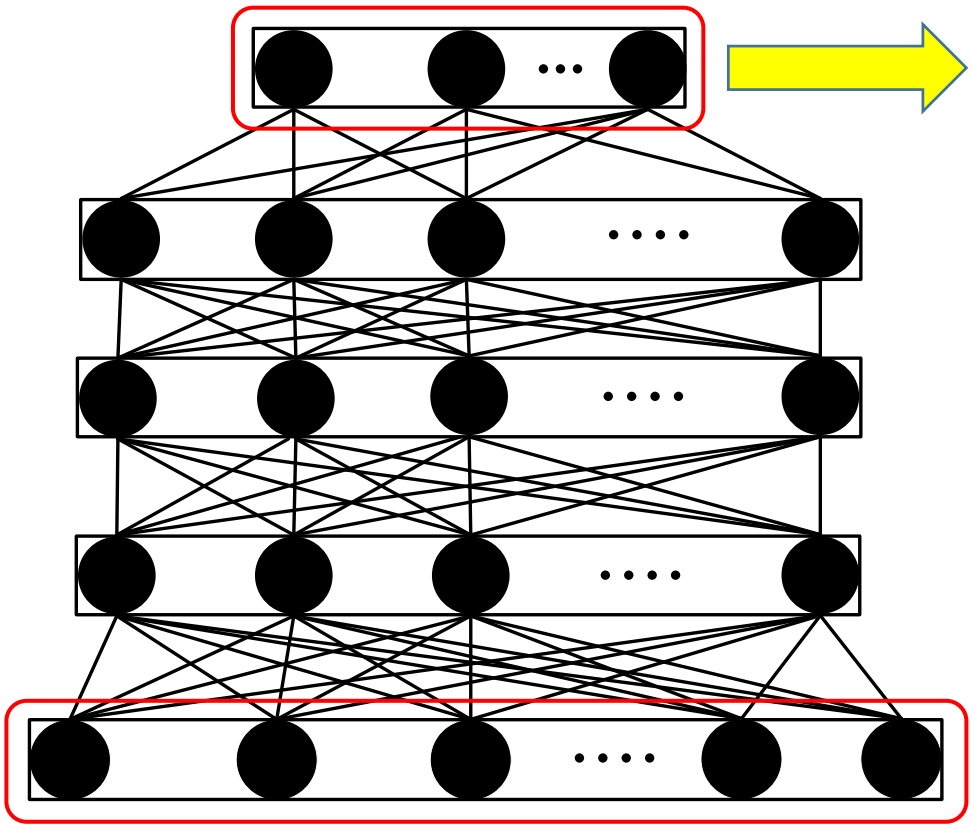
$$C_{t\arg}(u_n, w_n) = \left\| f(u_n) - f_p(w_n) \right\|^2$$

$$C_{con}(u_{n-T}, \cdots u_n, w_n) = \left\| f(u_n) - f_c(w_n, u_{n-T}, \cdots u_{n-1}) \right\|^2$$

Method

DNN-based target cost calculation

$$C_{t\arg}(u_n, w_n) = \left\| f(u_n) - f_p(w_n) \right\|^2$$



DNN target

Acoustic features: mel-cepstra, logF0 and u/v flag;

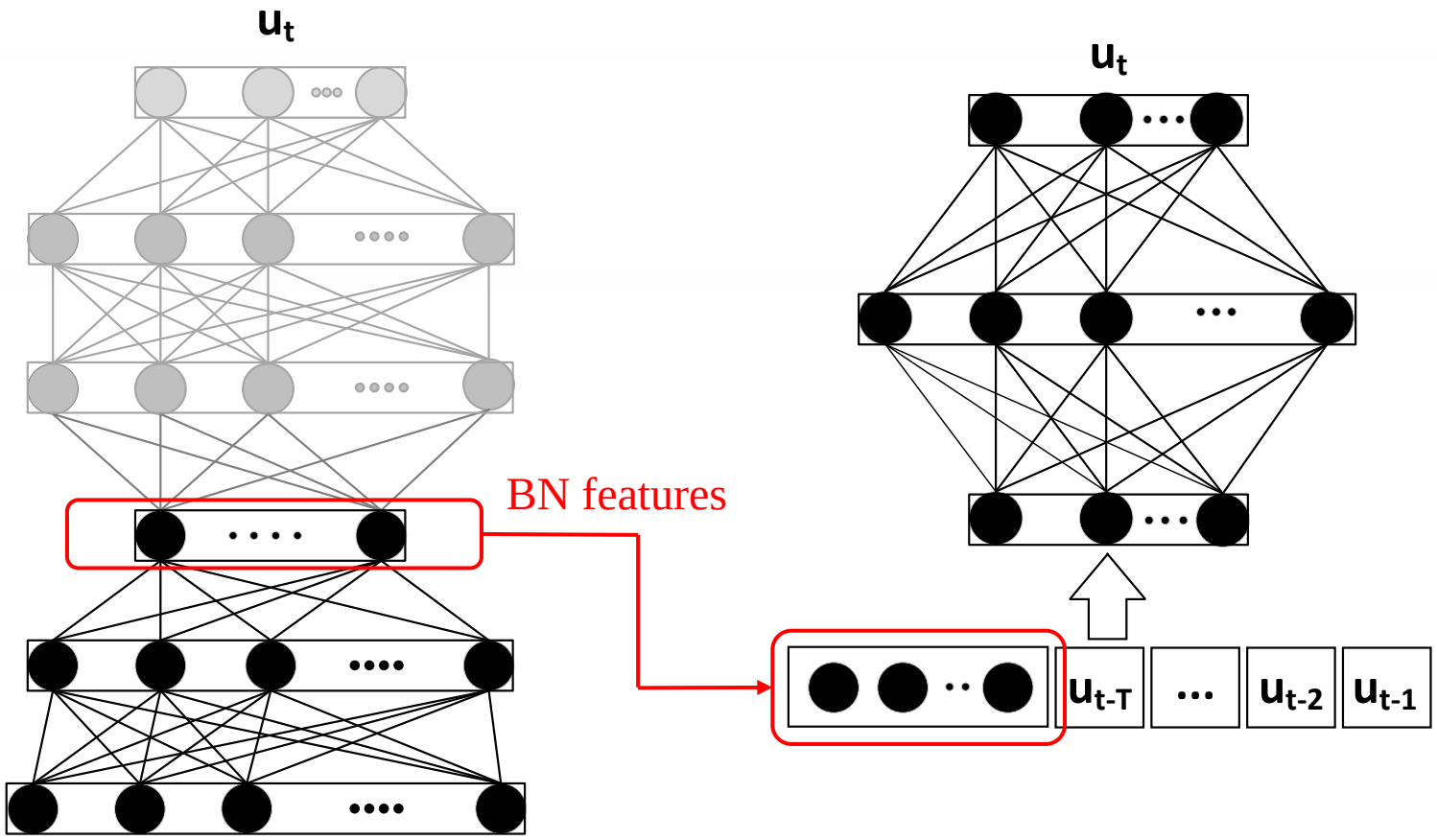
DNN input

Context information: including binary answers to question set, length of segment and position of frame.

Method

DNN-based concatenation cost calculation

$$C_{con}(u_{n-T},\cdots u_n, w_n)=\left\| f(u_n) - f_c(w_n, u_{n-T},\cdots u_{n-1})\right\|^2$$



Outline

1. Motivation
2. Method
3. Experiments
4. Conclusions

Experiments

Experimental condition

- CMU Arctic database, female *slt*, 1132sentences;
- 1000 sentences for training, 66 sentences for validation, 66 sentences for test;
- Acoustic features: 13 order mel-cepstra, F0 and u/v;
- 1534 context-dependent information;
- DNN training
  - Learning rate: 0.0001;
- Subjective listening test
  - Randomly selected 20 sentences, evaluated by 20 listeners on Amazon Mechanical Turk.

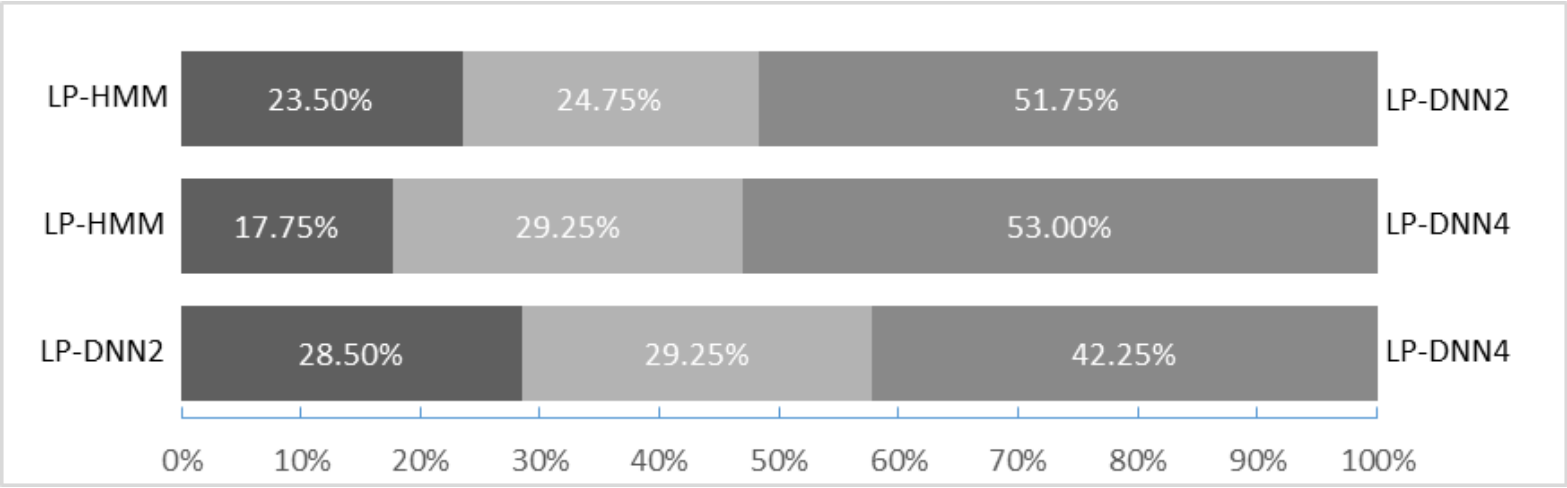
Experiments

Systems for comparing

System ID	Target cost	Concatenation cost
ML-HMM	HMM / log prob.	HMM / log prob.
ML-DNN2	HMM / log prob.	DNN (T=2)
ML-DNN4	HMM / log prob.	DNN (T=4)
HMM-DNN4	HMM / distance	DNN (T=4)
DNN-DNN4	DNN / distance	DNN (T=4)
HMM-GV	HMM-based SPSS with a GV model	

Experiments

Subjective preference test results



The naturalness of synthesized speech got improved after replacing the concatenation costs calculation with our proposed DNN-based approach.

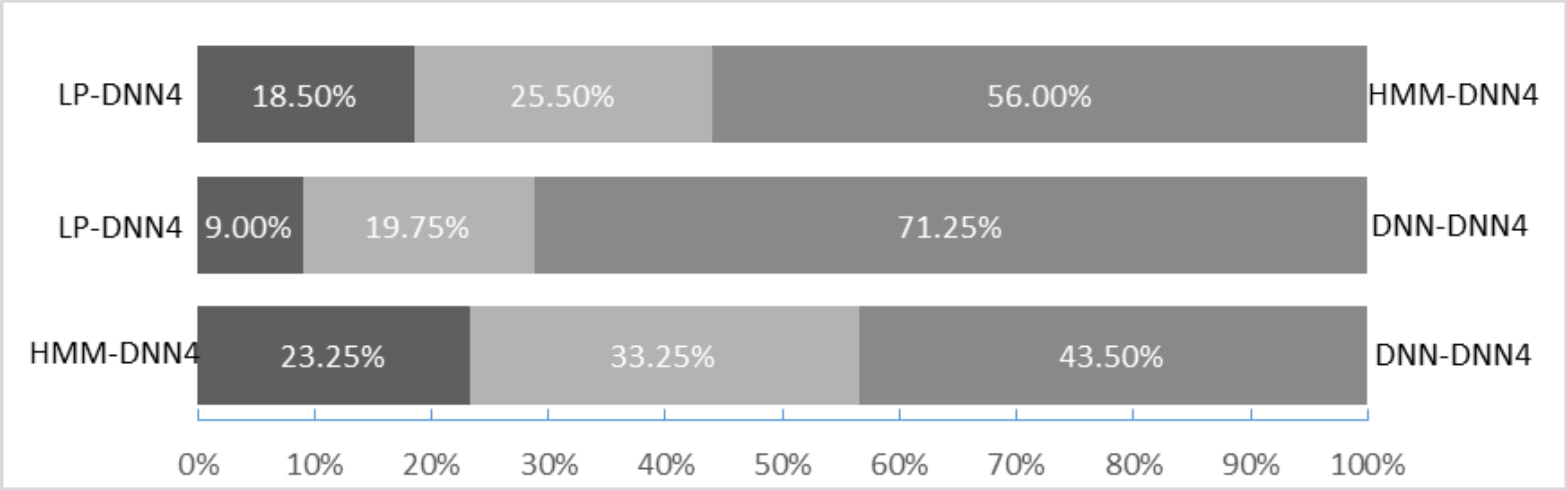
Experiments

Systems for comparing

System ID	Target cost	Concatenation cost
ML-HMM	HMM / log prob.	HMM / log prob.
ML-DNN2	HMM / log prob.	DNN (T=2)
ML-DNN4	HMM / log prob.	DNN (T=4)
HMM-DNN4	HMM / distance	DNN (T=4)
DNN-DNN4	DNN / distance	DNN (T=4)
HMM-GV	HMM-based SPSS with a GV model	

Experiments

Subjective preference test results

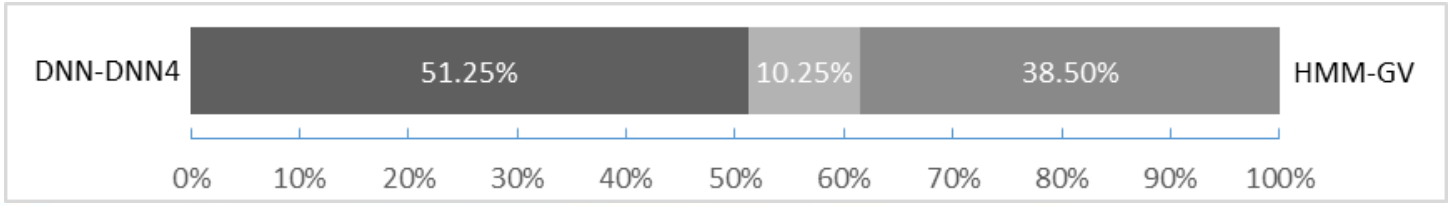


- The quality was improved when changing the target costs calculation from probabilities of HMMs to acoustic distances.
- After replacing HMMs with DNN to predict the target acoustic features for distances calculation, the quality improved further.

Experiments

Systems for comparing

System ID	Target cost	Concatenation cost
ML-HMM	HMM / log prob.	HMM / log prob.
ML-DNN2	HMM / log prob.	DNN (T=2)
ML-DNN4	HMM / log prob.	DNN (T=4)
HMM-DNN4	HMM / distance	DNN (T=4)
DNN-DNN4	DNN / distance	DNN (T=4)
HMM-GV	HMM-based SPSS with a GV model	



Outline

1. Motivation
2. Method
3. Experiments
4. Conclusions

Conclusions

- The DNN-based target prediction model can improve the accuracy of the predicted acoustic features compared with the HMM.
- Both the DNN-based target cost calculation and the DNN-based concatenation cost calculation can lead to better naturalness of synthetic speech in our listening tests.
- But the computation complexity is very high, to reduce the computation complexity will be the tasks of our future work.

References

[Tokuda,2004] Tokuda, K., H. Zen, and A. W. Black. "HMM-based approach to multilingual speech synthesis." *Text to speech synthesis: New paradigms and advances*(2004): 135-153.

[Zen,2013] Zen, Heiga, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.

[Iwahashi,1992] Iwahashi, Naoto, Nobuyoshi Kaiki, and Yoshinori Sagisaka. "Concatenative speech synthesis by minimum distortion criteria." *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. Vol. 2. IEEE, 1992.

[Hirai, 2004] Hirai, Toshio, and Seiichi Tenpaku. "Using 5 ms segments in concatenative speech synthesis." Fifth ISCA Workshop on Speech Synthesis. 2004.

[Ling, 2006] Ling, Zhen-Hua, and Ren-Hua Wang. "HMM-based unit selection using frame sized speech segments." Ninth International Conference on Spoken Language Processing. 2006.

[Sakoe,1978] Sakoe, Hiroaki, and Seibi Chiba. "Dynamic programming algorithm optimization for spoken word recognition." IEEE transactions on acoustics, speech, and signal processing 26.1 (1978): 43-49.

Demo

System ID			
ML-HMM	1.wav	2.wav	3.wav
ML-DNN2	1.wav	2.wav	3.wav
ML-DNN4	1.wav	2.wav	3.wav
HMM-DNN4	1.wav	2.wav	3.wav
DNN-DNN4	1.wav	2.wav	3.wav
HMM-GV	1.wav	2.wav	3.wav

Q&A

Thank You!