Speech Synthesis Workshop 2016
Sunnyvale, California, USA

Compare HMM, DNN and RNN on huge speaker dependent corpora for Japanese Text-to-Speech

Xin WANG, Shinji TAKAKI, Junichi YAMAGISHI
National Institute of Informatics, Japan
2016-09-14

contact: wangxin@nii.ac.jp
critical comments, suggestions, discussion are welcome

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Evaluate the performance of HMM, feedforward and recurrent neural network trained using different amount of data

Xin WANG, Shinji TAKAKI, Junichi YAMAGISHI
National Institute of Informatics, Japan
2016-09-14

contact: wangxin@nii.ac.jp
critical comments, suggestions, discussion are welcome

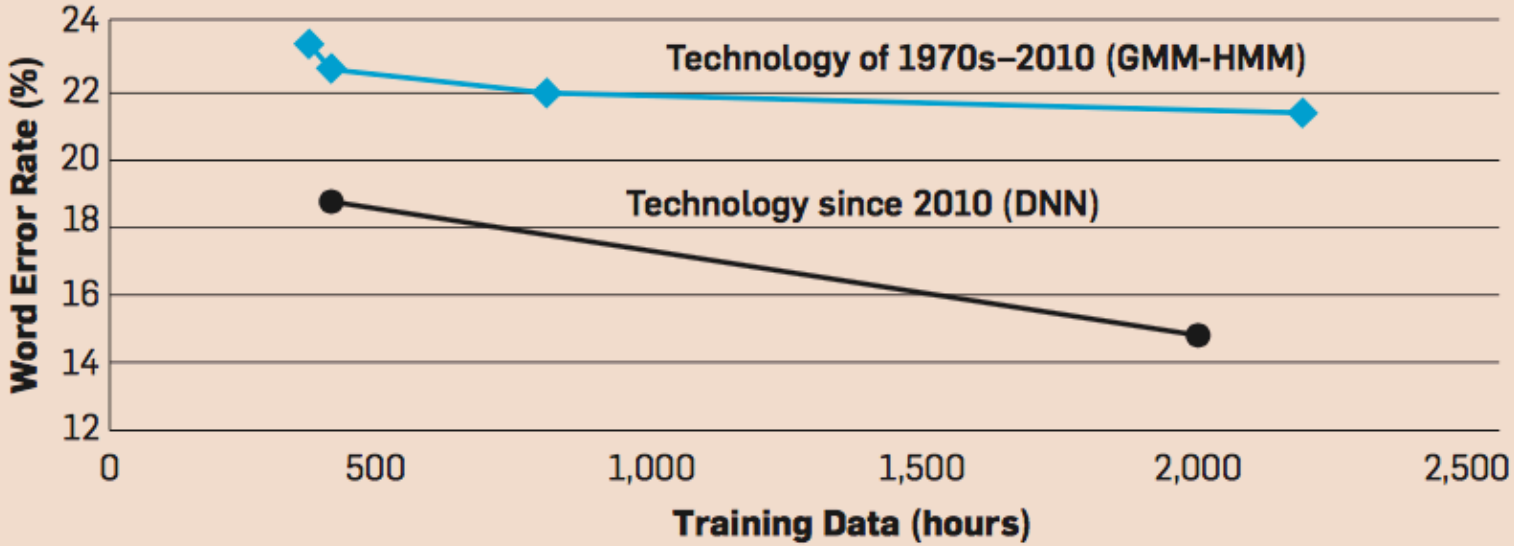
CONTENTS

- Motivation
- Corpora and system configuration
- Objective and subjective results
- Conclusion

MOTIVATION

Results of ASR [1]

Figure 3. There is no data like more data. Recognition word error rate vs. the amount of training hours for illustrative purposes only. This figure illustrates how modern speech recognition systems can benefit from increased training data.



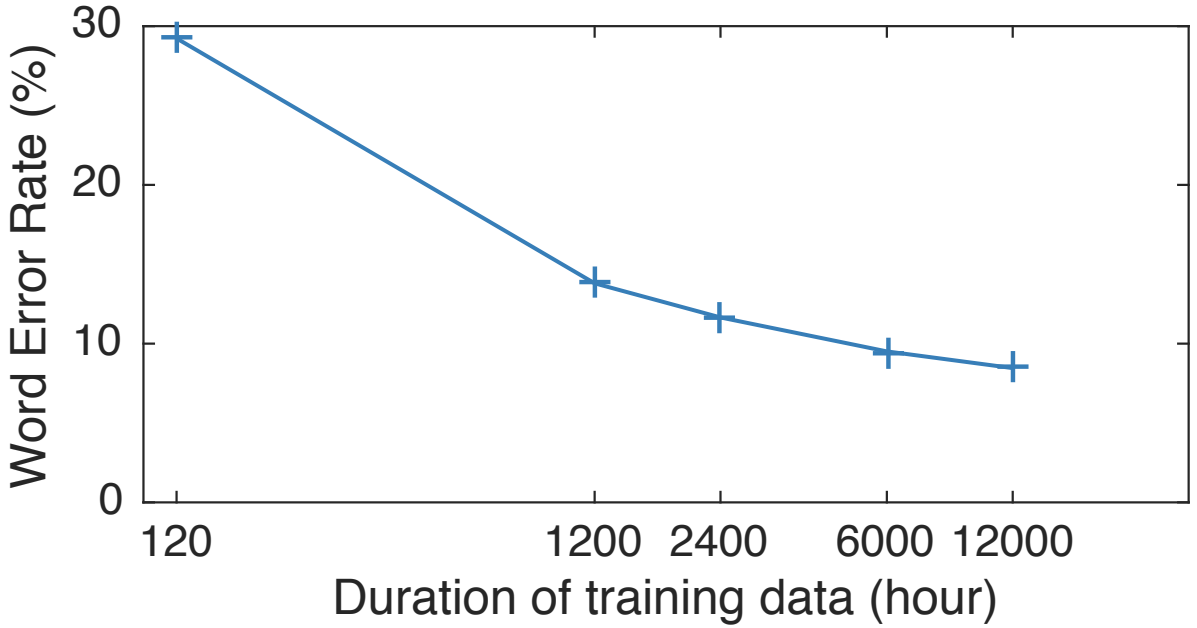
- Can we draw a similar figure for speech synthesis ?

[1] Huang, X., Baker, J. & Reddy, R., 2014. A Historical Perspective of Speech Recognition. *Commun. ACM*, 57(1), pp.94–103.

MOTIVATION

Corpora for ASR

- Large corpora (> 1000 hours)
  - speakers variation, noise types
  - …
- Benefits of using larger corpora [2]



[2] Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., … others. (2015). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. ICML, pp. 173–182, 2016.

MOTIVATION

For speech synthesis

- Large corpora ?
  - speaker-dependent
  - quality control
  - expert annotation

  ➢ corpora for speech synthesis are `expensive'

- Benefits of using larger corpora ?
  - quality VS quantity
    - accurate and fine annotation on a small corpus
    - large amount of data without expert annotation

MOTIVATION

For this work

- (Relatively) Large Japanese corpora are available

- What can we do with the data?

  It is better to do something simple than nothing at all. [3]

  - evaluate different kinds of acoustical model trained using different amount of data

[3] Church, K. (2004). Speech and language processing: Can we use the past to predict the future? In *International Conference on Text, Speech and Dialogue* (pp. 3–13).

CONTENTS

- Motivation
- Corpora and system configuration
- Objective and subjective results
- Conclusion

CORPORA

Overview of the corpora

- Originally designed for XIMERA, by ATR Japan[4,5]

Male voice

Genre	Size in hours
Novel	9.95
News	70.21
Sentence	2.1
Conversation	12.11
Word	4.93
Syllables	0.34
Voice check	9.41
Miscellaneous	1.63
Total	110.68

Female voice

Genre	Size in hours
Novel	17.27
News	20.46
Sentence	2.42
Conversation	3.86
Word	5.71
Voice check	9.83
Total	59.55

From Table 1, Table 2 of [4]

[4] Kawai, H., Toda, T., Ni, J., Tsuzaki, M., & Tokuda, K. (2004). XIMERA: A new TTS from ATR based on corpus-based technologies. In *SSW-5*.
[5] Kawai, H., Toda, T., Yamagishi, J., Hirai, T., Ni, J., Nishizawa, N., … Tokuda, K. (2006). XIMERA: A concatenative speech synthesis system with large scale corpora. *IEICE Trans. Inf. Syst.(Japanese Edition)*, 2688–2698.

EXPERIMENTS

Input/output features

- Input textual features
  - text analysis:
    - grapheme-to-phoneme: Open JTalk [6]
    - Part-of-Speech & morphological analysis: Mecab [7]
  - dimension: 389 for neural network
  - ➢ input features for the whole corpora are automatically analyzed, without expert annotation

[6] The HTS Working Group. (2015). The Japanese TTS System "Open JTalk." Retrieved from http://open-jtalk.sourceforge.net/
[7] Kudo, T. (n.d.). MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Retrieved from http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

EXPERIMENTS

Input/output features

- Output acoustic features

Name	Dimension	Note
Mel-generalized cepstrum coefficients (MGC)	60	Bark-scale, warping factor 0.77
F0 trajectory in Mel-scale	1	multiple F0 trackers + median filtering
Band aperiodicity	25	average value in critical band

  - based on STRAIGHT vocoder [8]
  - dimension: 259, with static, delta, delta-delta

[8] Kawahara, H., Masuda-Katsuse, I. & Cheveigne, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based {F}0 extraction: {P}ossible role of a \ repetitive structure in sounds. *Speech Communication*, 27, pp.187–207.

EXPERIMENTS

Acoustic models

- System notation and configuration

Notation	System	Configuration
HMM	HMM-based system	the HTS 2.3 recipe
DBLSTM-RNN	Recurrent neural network with bi-directional long short term memory (LSTM) units	2 feedforward layers, 512 nodes/layer 2 bi-directional LSTM layers, 256 nodes/layer 1 linear projection output layer Number of parameter 1.59m
DNN	Deep feedforward neural network	1 feedforward layers, 1024 nodes/layer 3 feedforward layers, 512 nodes/layer 1 linear projection output layer Number of parameter 1.59m

EXPERIMENTS

Acoustic models

- Neural network toolkit: CURRENNT [8]
  - random initialization
  - stochastic gradient descent + early stopping
  - male voice: 500 utterances for validation, 500 for test
  - female voice: 260 utterances for validation, 260 for test

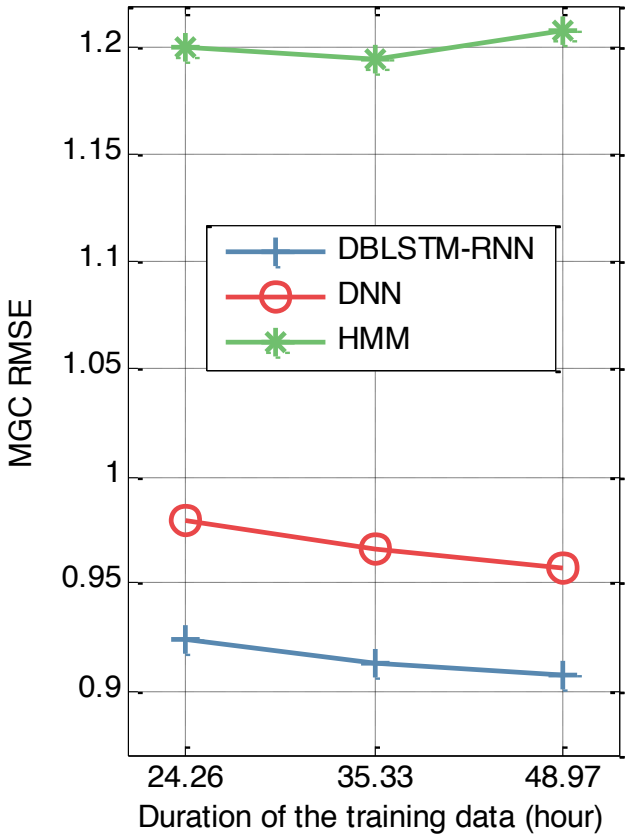
[8] Weninger, F., Bergmann, J., & Schuller, B. (2015). Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *The Journal of Machine Learning Research*, *16*(1), 547–551.

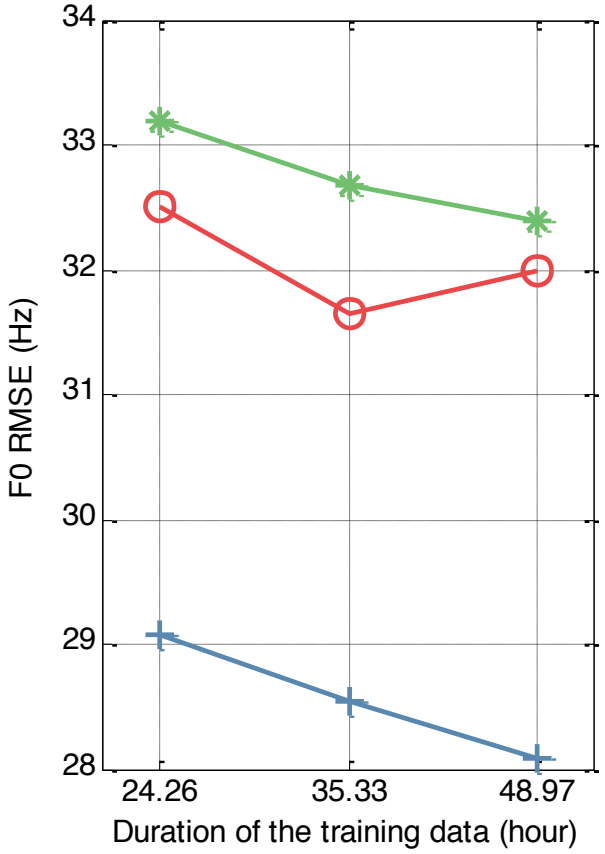
CONTENTS

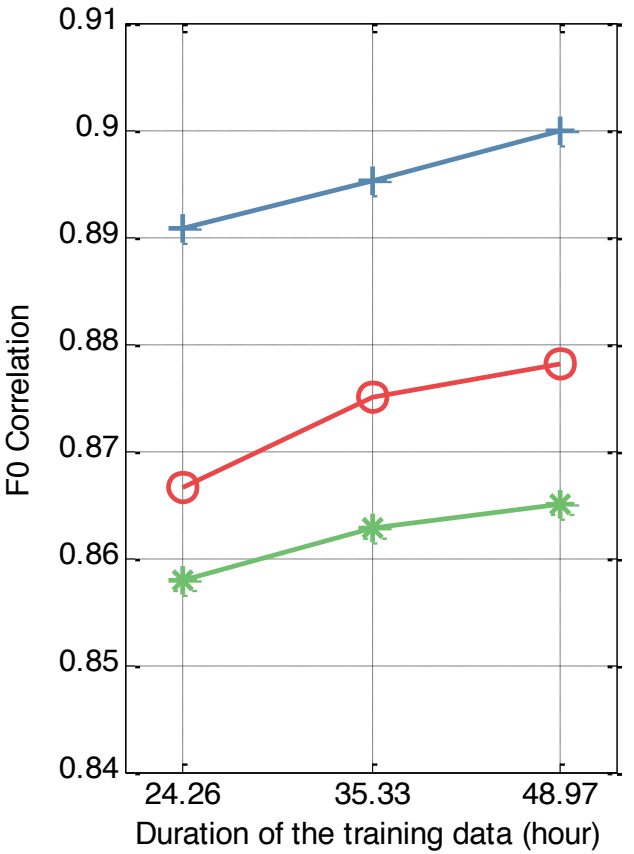
- Motivation
- Corpora and system configuration
- Objective and subjective results
- Conclusion

RESULTS

Objective measure - female voice



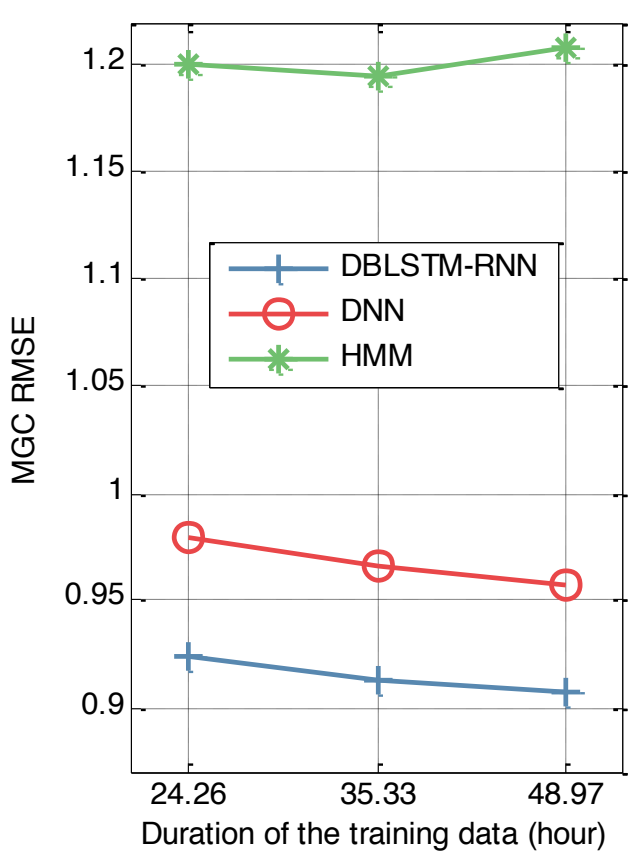


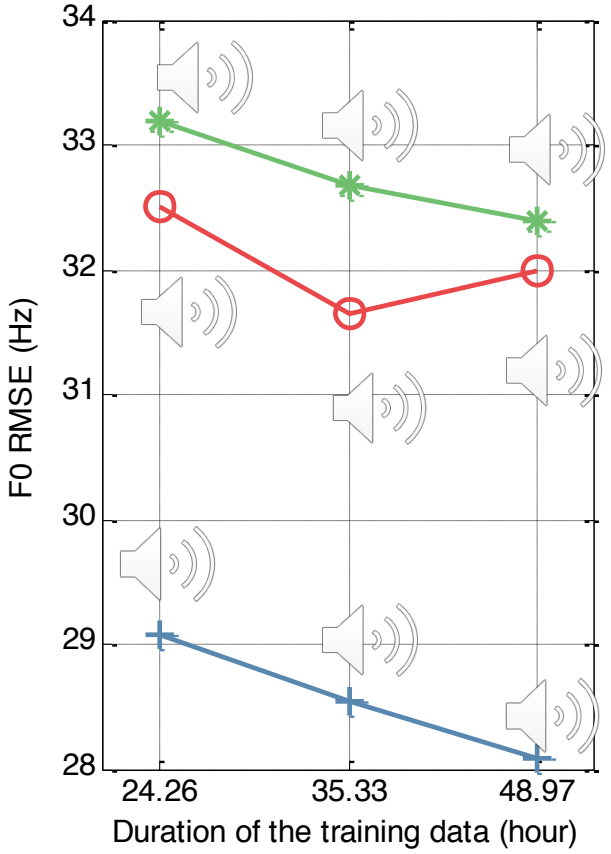


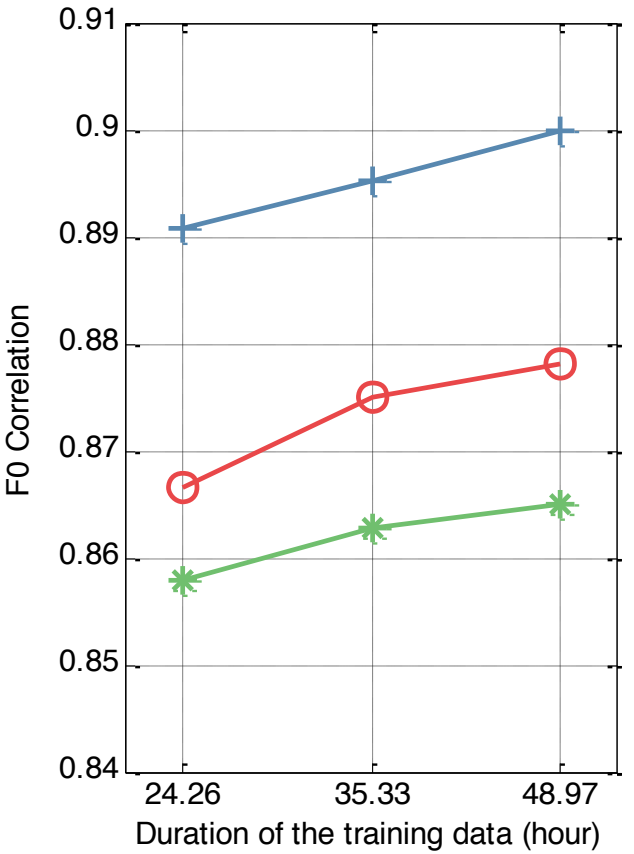
- DBLSTM-RNN will generate more accurate F0 with more data ?

RESULTS

Objective measure - female voice

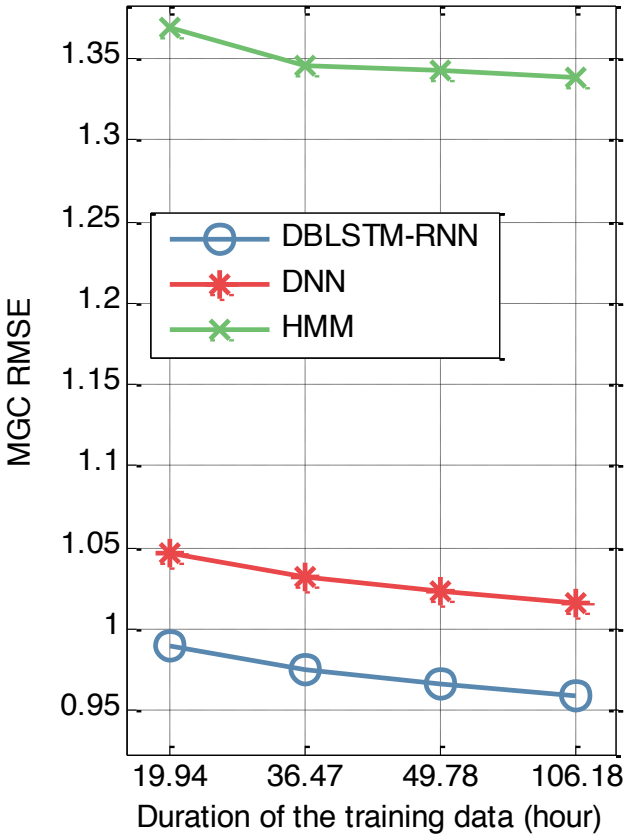


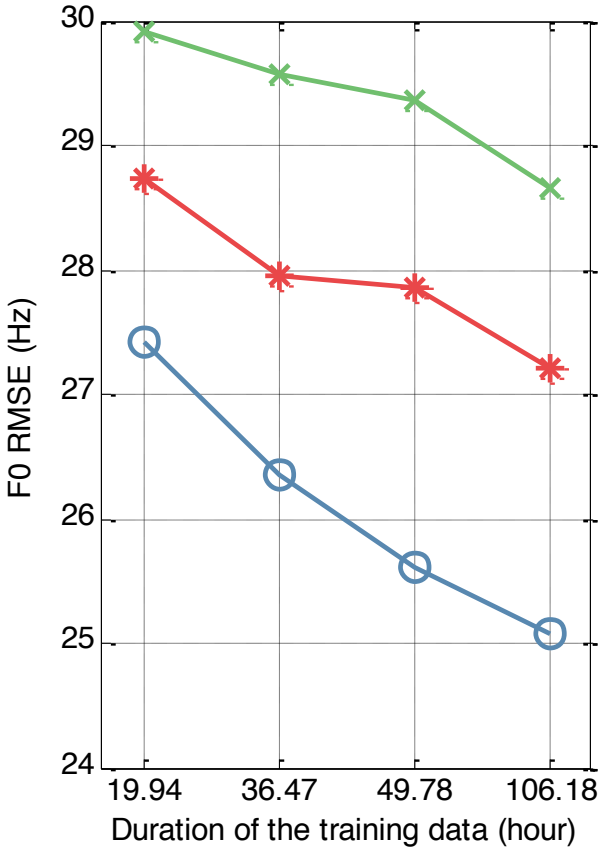


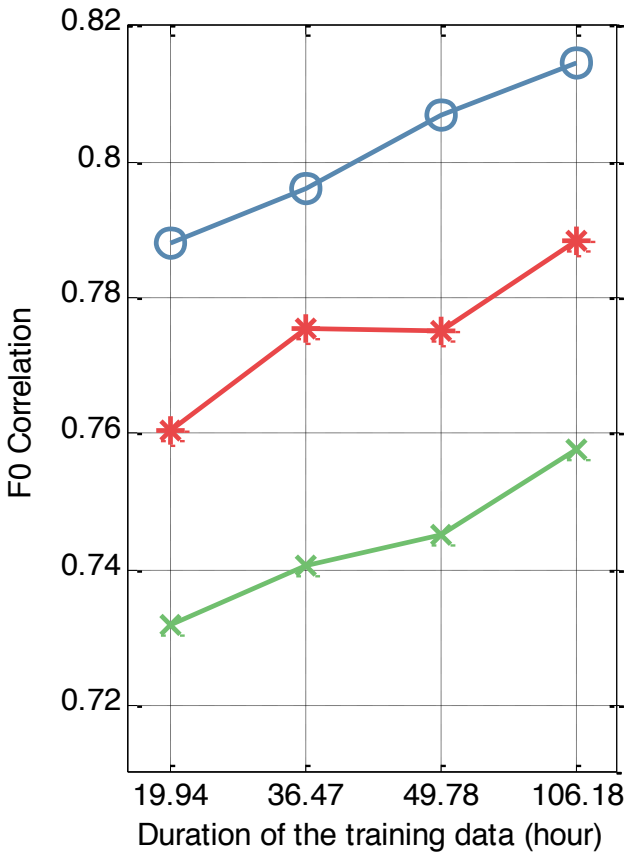


RESULTS

Objective measure - male voice



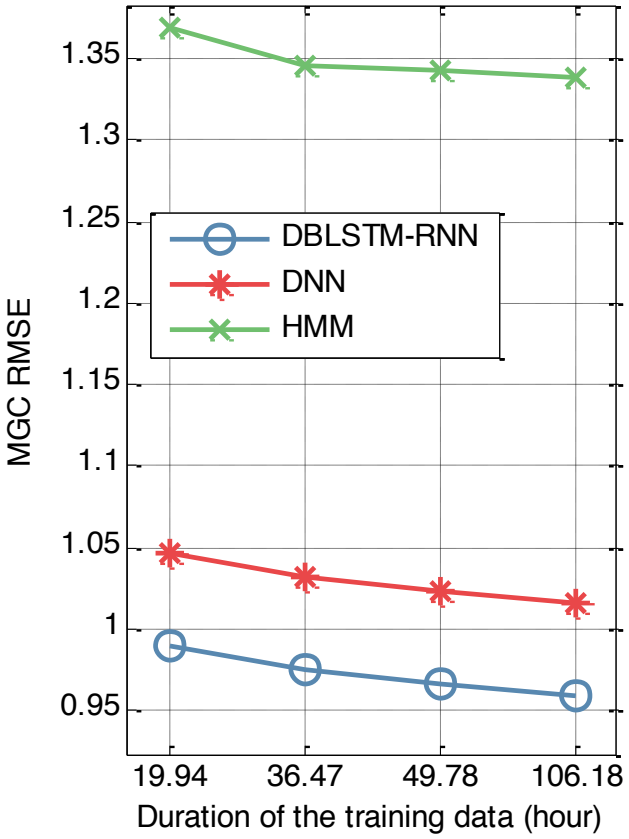


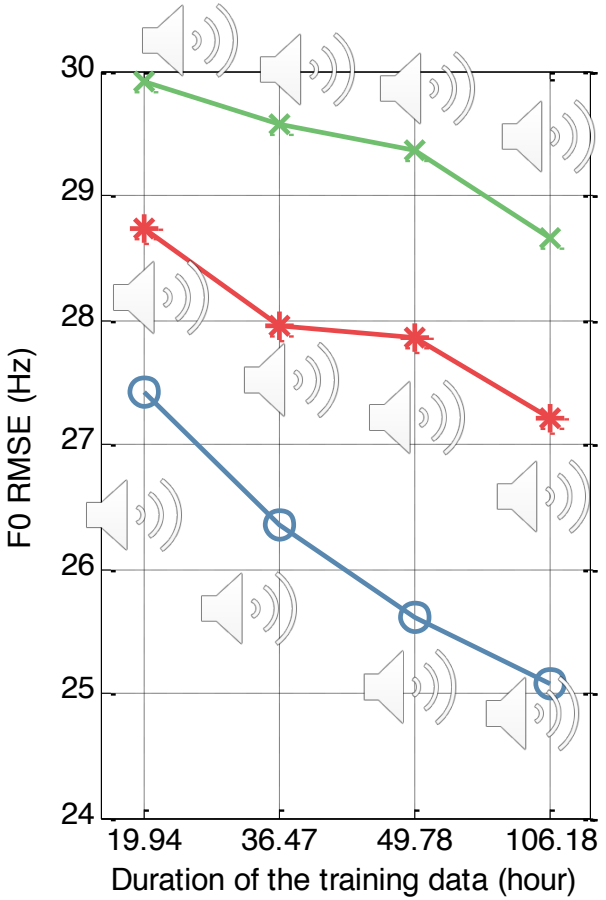


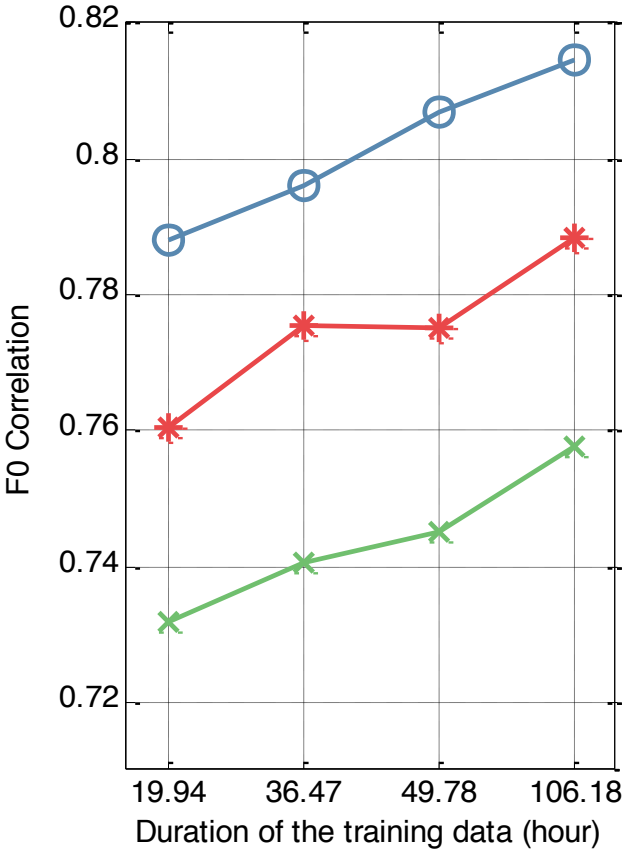
- DBLSTM-RNN will generate more accurate F0 with more data ?

RESULTS

Objective measure - male voice

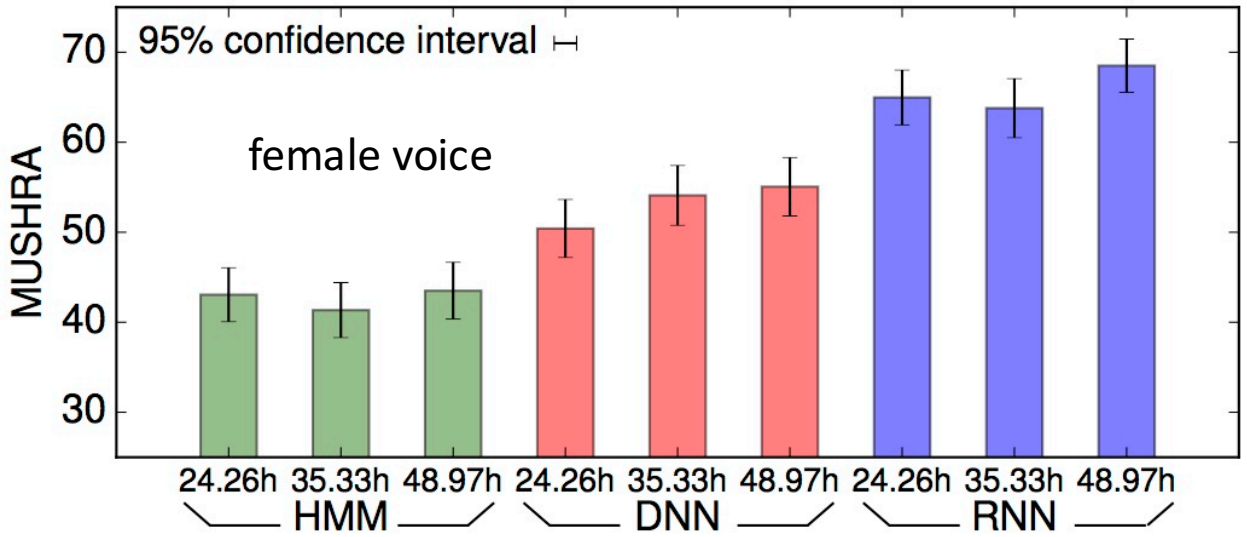


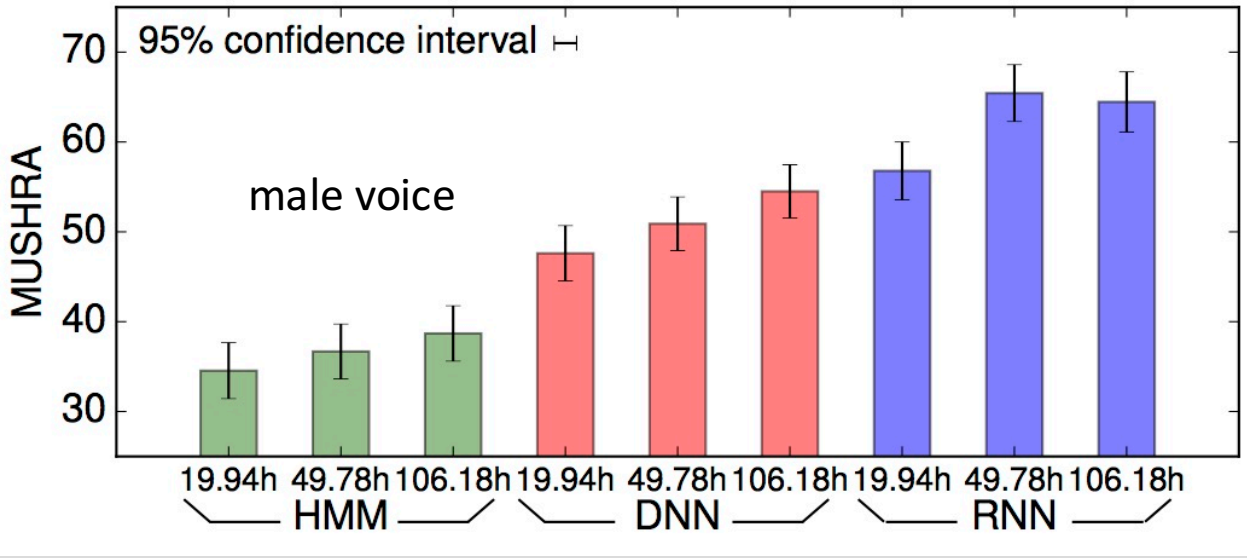




RESULTS

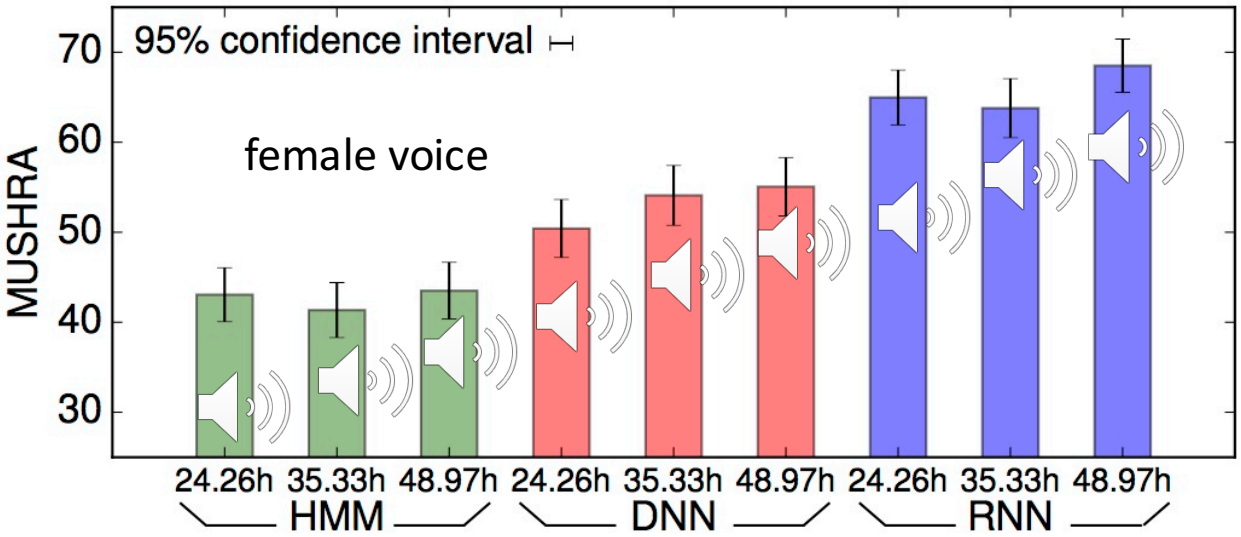
Subjective measure - MUSHRA

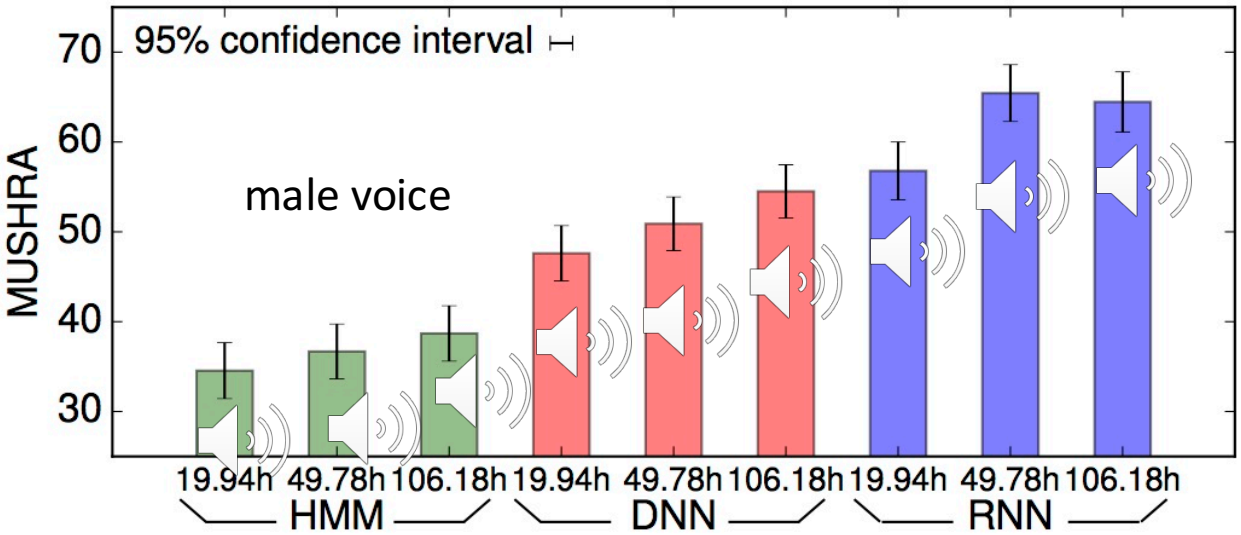




RESULTS

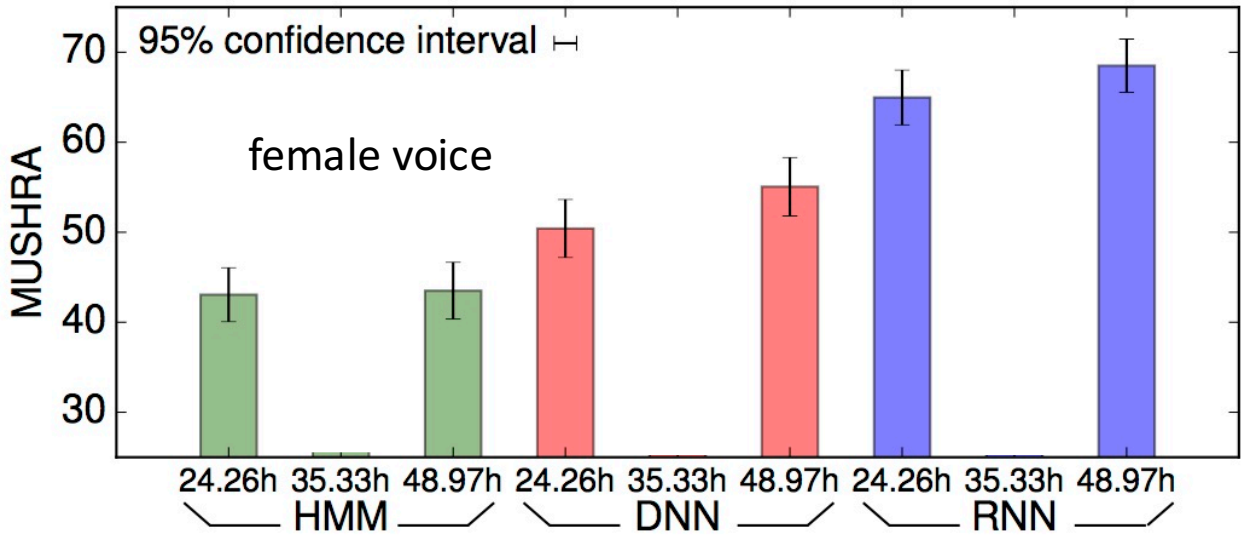
Subjective measure - MUSHRA

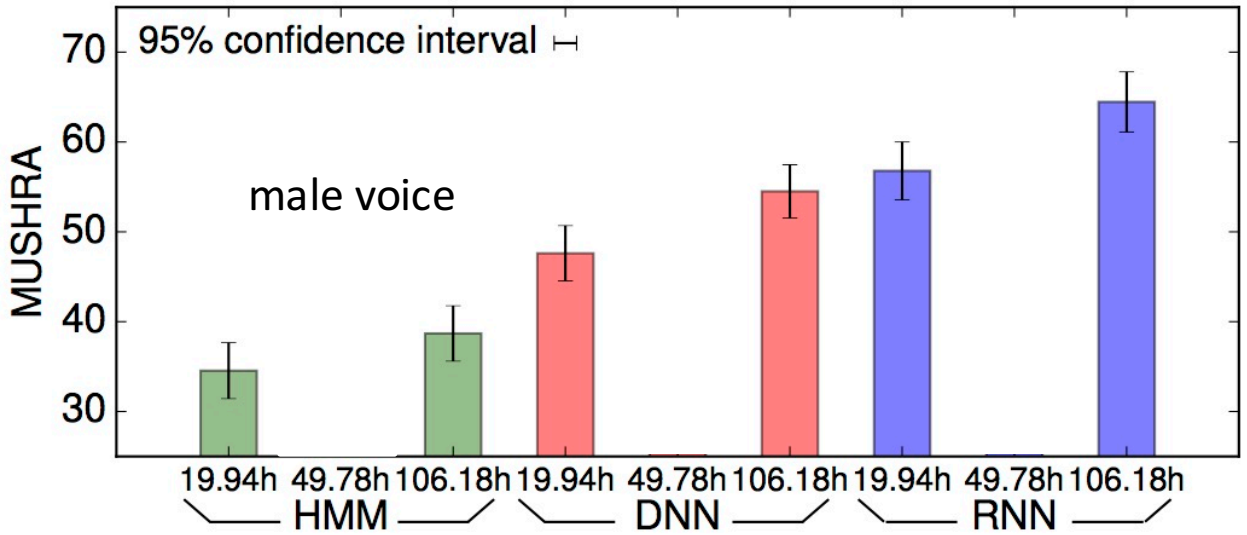




RESULTS

Subjective measure - MUSHRA





CONTENTS

- Motivation
- Corpora and system configuration
- Objective and subjective results
- Conclusion

CONCLUSION

Answer to the previous question

- Benefits of using more data ?
  - objective result:
    - F0 => more accurate synthetic F0 (DBLSTM-RNN)
    - MGC => the improvement gradually decreases
  - subjective result:
    - improvement can be observed
- Using more data without fine annotation is less effective than using a more effective model ?

CONCLUSION

Future questions

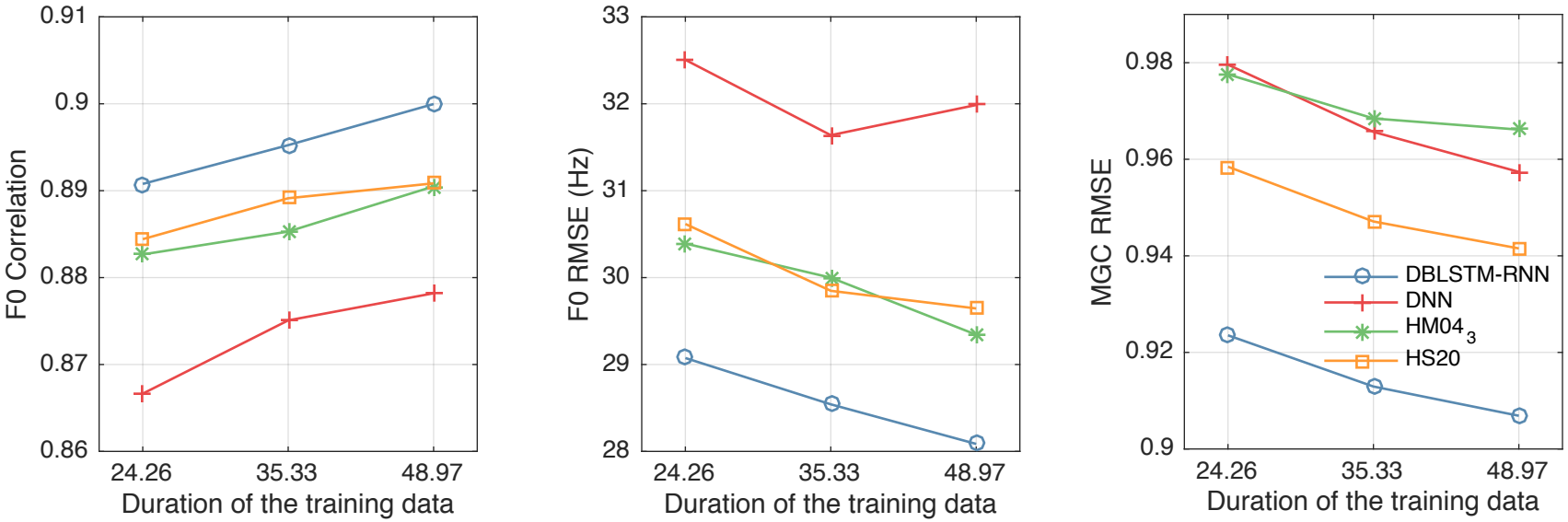
- The gap between DBLSTM-RNN and DNN on F0 is due to the recurrent connection ?
- Is there any pitch-related feature better than F0 trajectory ?
- …

CONCLUSION

One question being explored

- The gap between DBLSTM-RNN and DNN on F0 is due to the recurrent connection ?

on the female voice



- a highway network is a feedforward neural network
- **HM04$_2$**: a multi-stream highway network, 4 tanh hidden layers
- **HS20**: a single-stream highway network, 20 tanh hidden layers

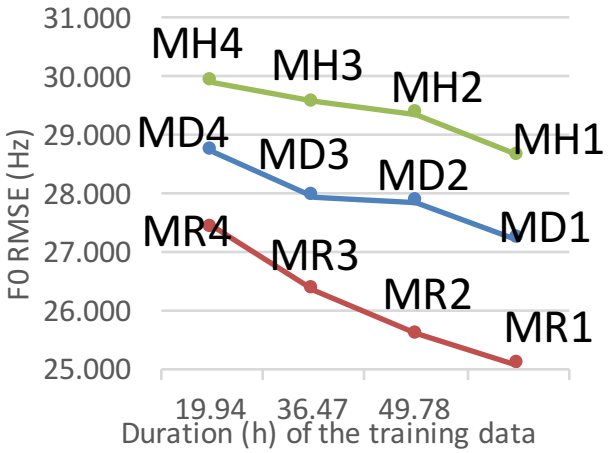
CONCLUSION

Not available due to Japanese law on personal information protection

~~http://tonywangx.github.io~~

- ~~Female voice (60M) with 15 samples/system~~
  ~~https://www.dropbox.com/s/dhh26clsjxaiwy7/JVOICE_F009_15_0502_x04.tar.gz?dl=0~~
- ~~Male voice (90M) with 15 samples/system~~
  ~~https://www.dropbox.com/s/p49zfunh27enk3x/JVOICE_M007_15_0502_x05.tar.gz?dl=0~~

- Naming of the folder in the above package
  - FN and MN: natural voice
  - $#&:
    - $: F female, M male
    - #: H HMM, D DNN, R RNN
    - &: 1 full training set, 2 less data …
  - see example on the right



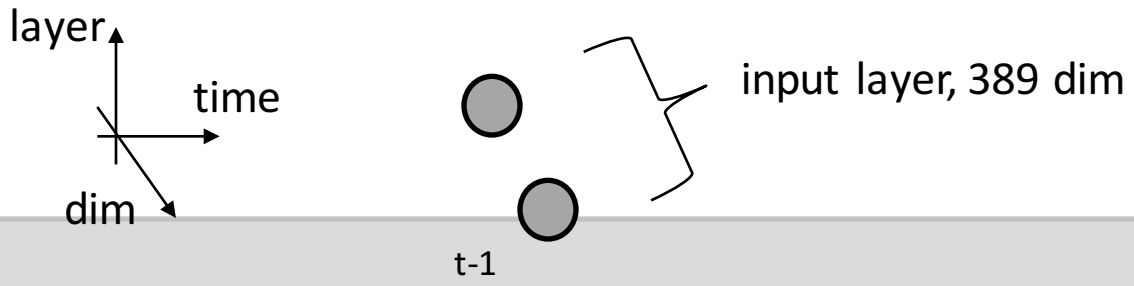
Thanks for your attention

Q & A

EXPERIMENTS

Acoustic models

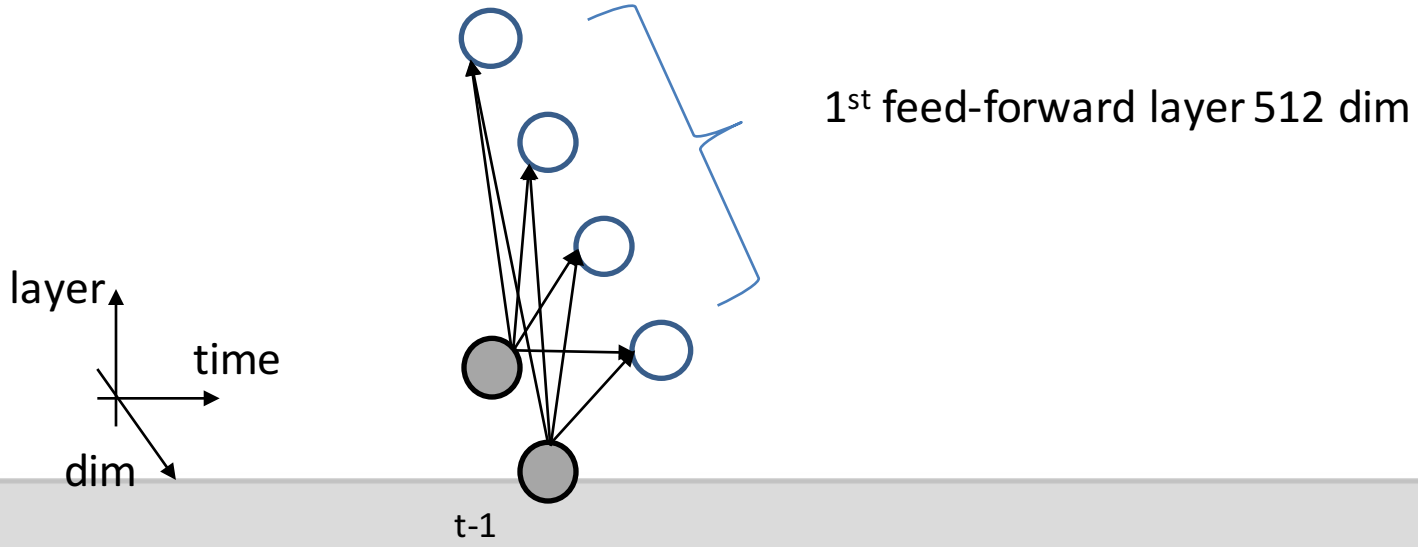
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

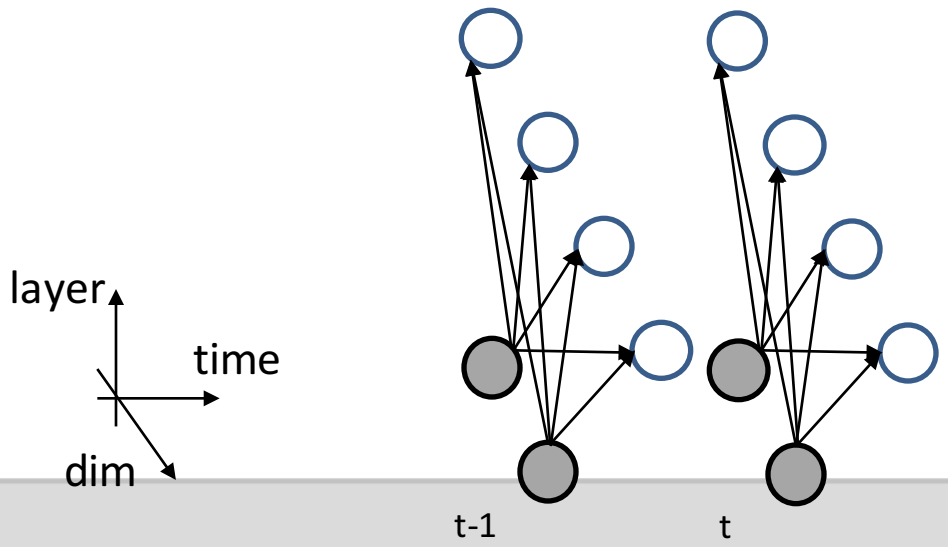
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

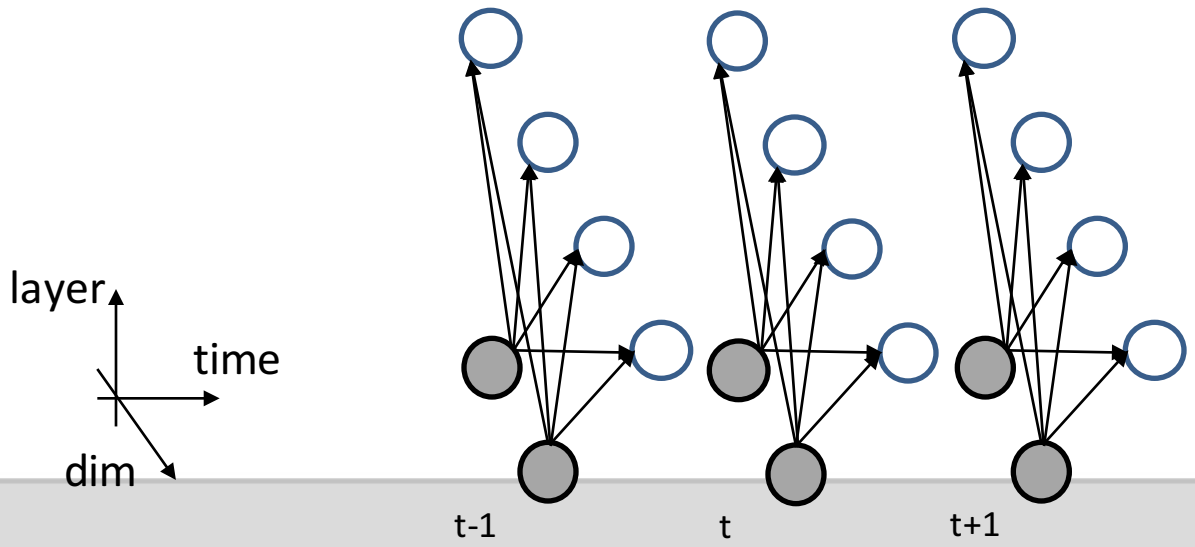
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

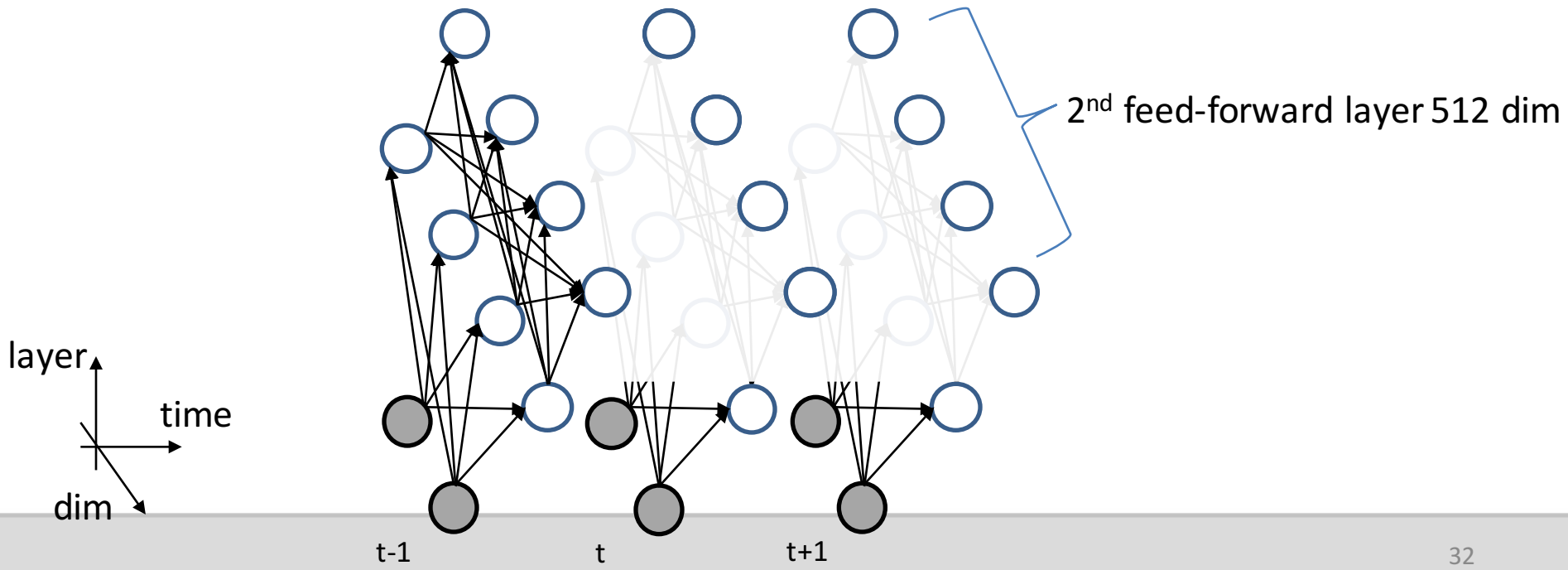
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

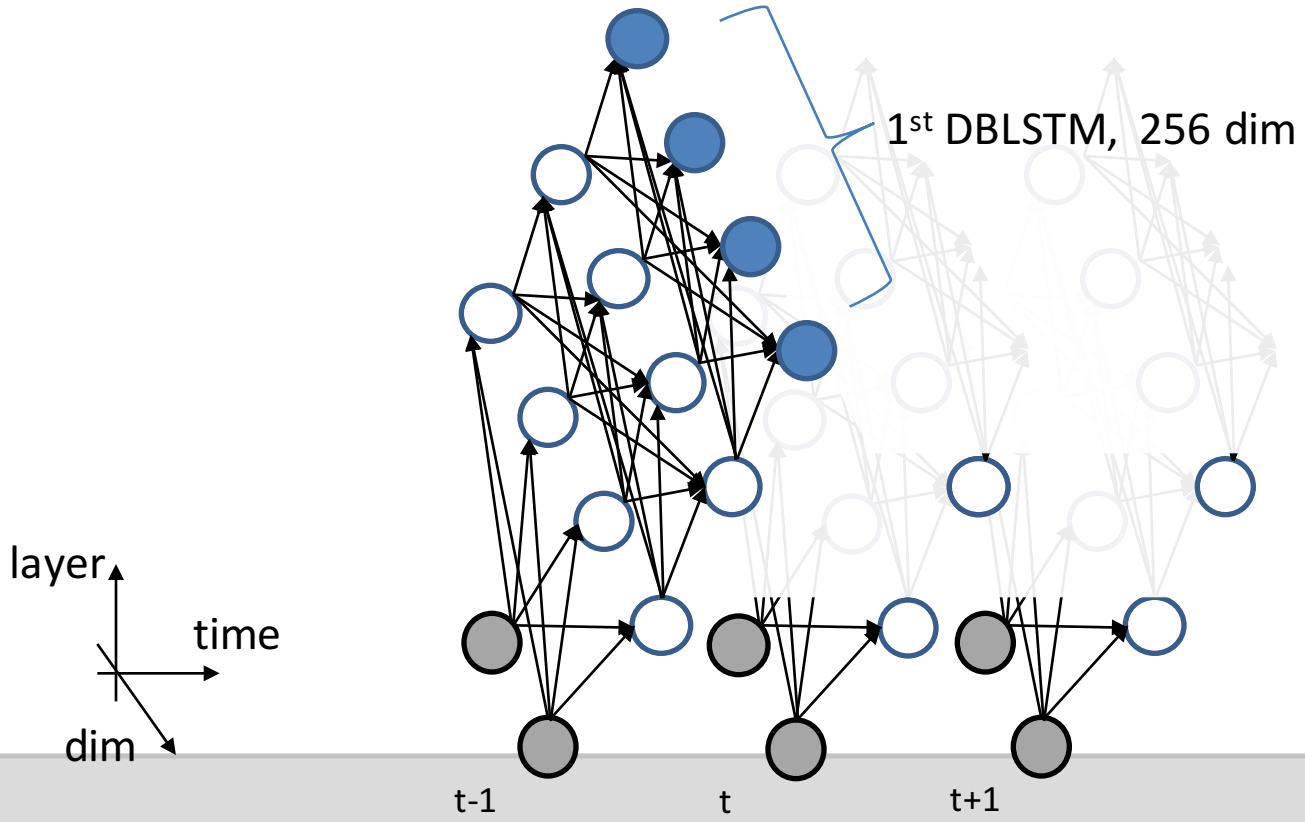
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

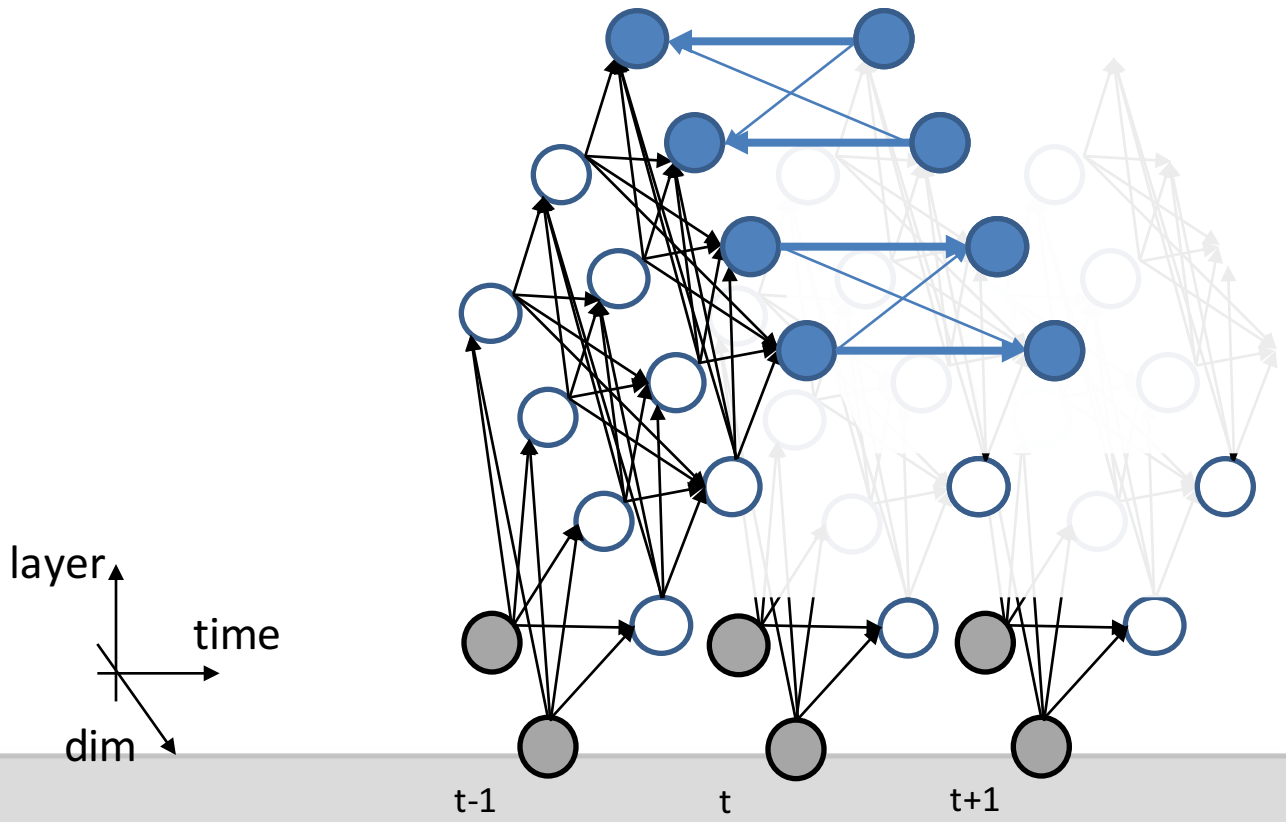
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

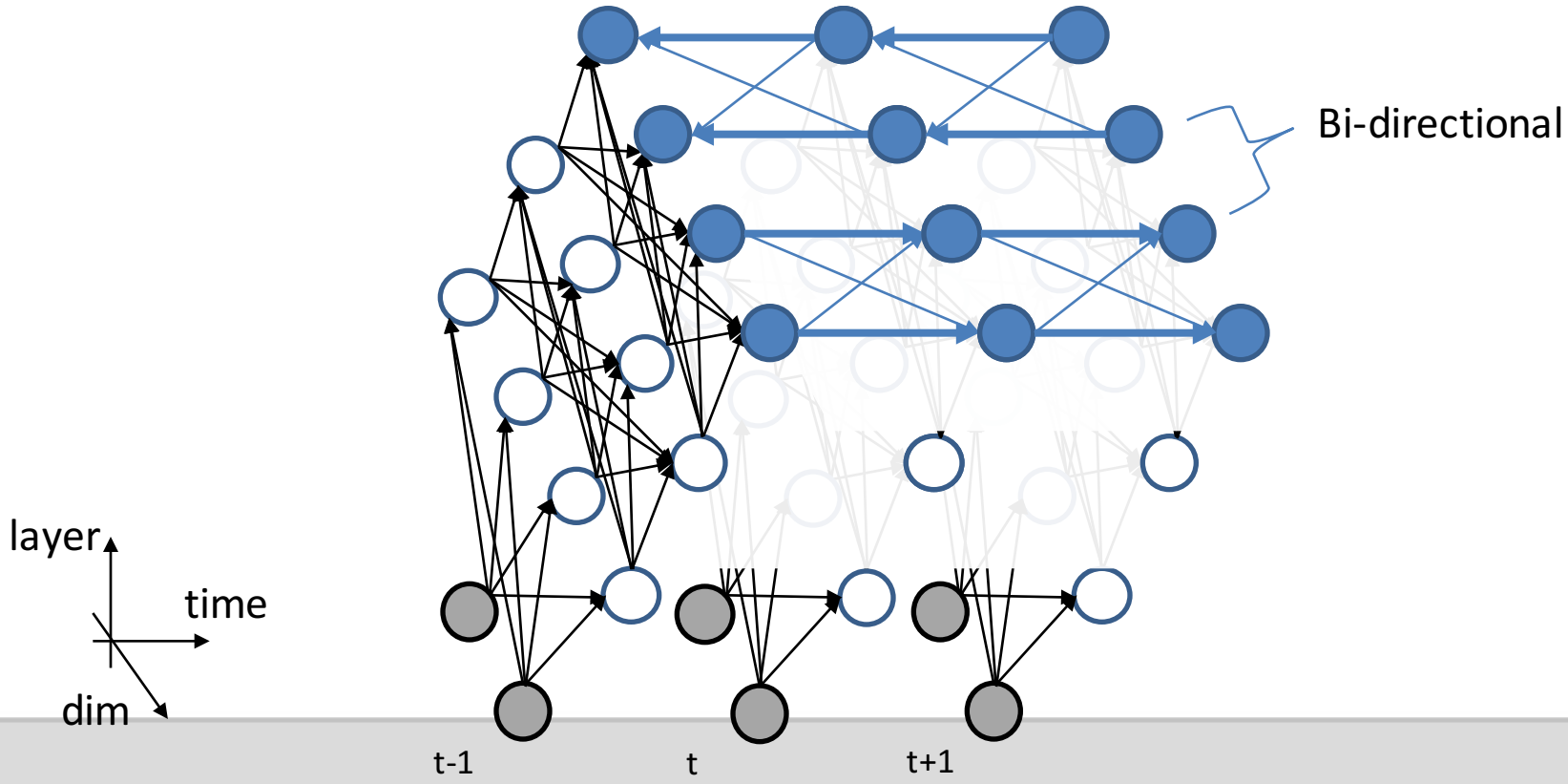
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

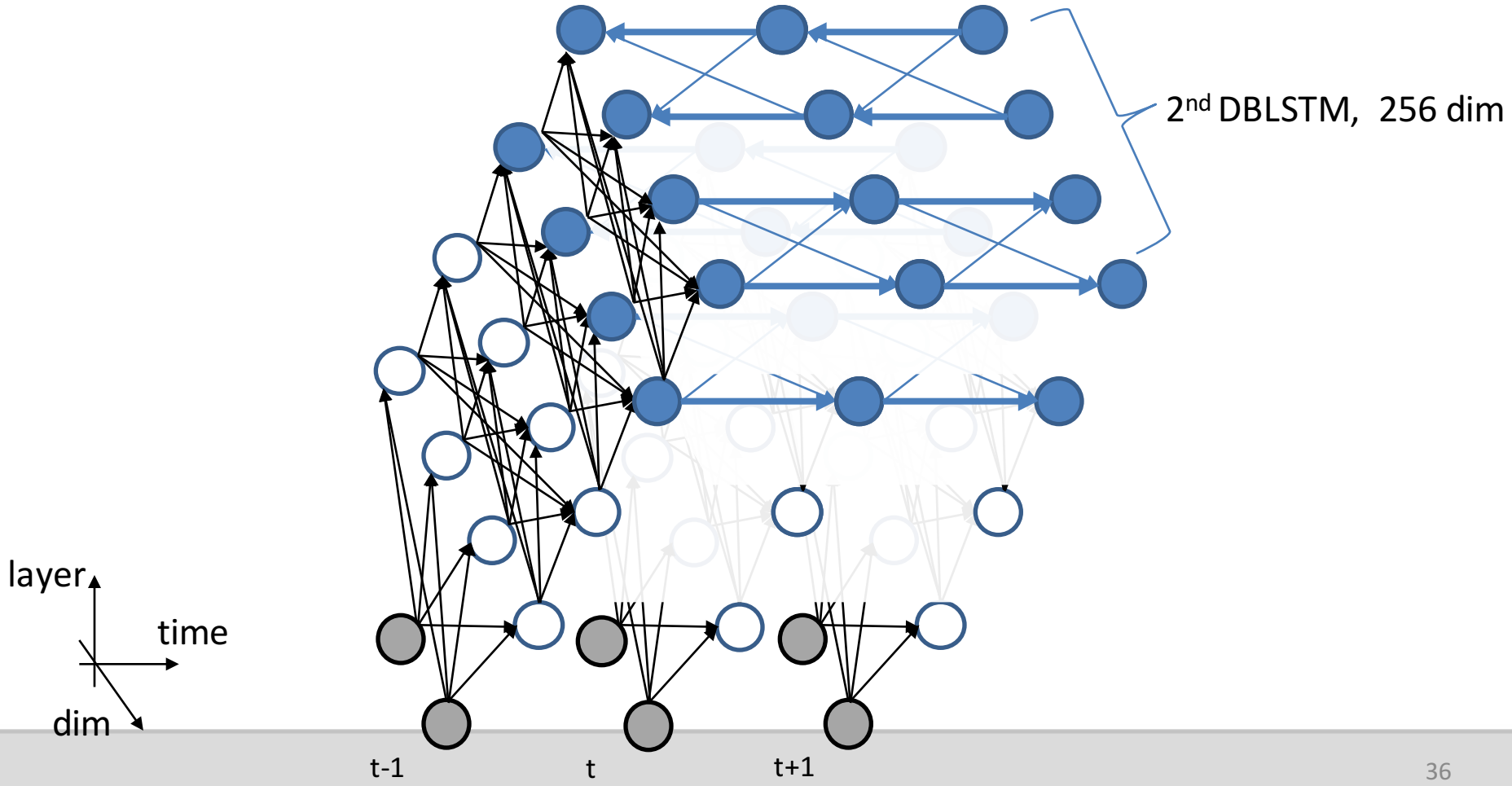
- DBLSTM-RNN



EXPERIMENTS

Acoustic models

- DBLSTM-RNN



EXPERIMENTS

Acoustic models

- DBLSTM-RNN