# ADAPTING AND CONTROLLING DNN-BASED SPEECH SYNTHESIS USING INPUT CODES

Hieu-Thi Luong[*], Shinji Takaki,

Gustav Eje Henter, Junichi Yamagishi

National Institute of Informatics, Japan

[*] VNU-HCM University of Science, Vietnam

# Background

## Statistical parametric speech synthesis

- Remarkable progress thanks to DNN
- e.g. Wavenet, SampleRNN, GAN

## Flexibility of speech synthesizers

- HMM-based synthesis
  - Speaker/style adaptation, interpolation, multiple regression
- DNN-based synthesis
  - Black-box, not as flexible as HMM synthesizers yet

## Speaker, gender, and age codes: “**input codes**”

- Multi-speaker modelling
- Speaker adaptation

Objective and subjective results

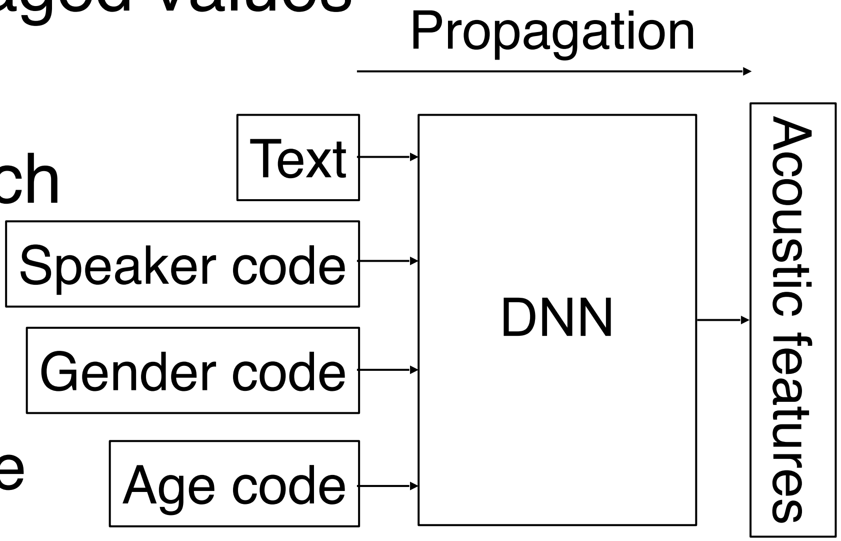
- Flexible manipulation

Demonstration

# Multi-speaker modelling using input codes

## Multi-speaker modelling using input codes

- Generate multiple speakers' voices from a single DNN
- Input codes: simple additional inputs that differentiate ID, gender and age of speakers
- Also good as an initial model for speaker adaptation
  - Input codes that use averaged values
    → Average voice
- Allow us to manipulate speech
  - e.g. flip the gender code
- Morphing
  - Change the code each frame



# Speaker codes

## One-hot vector codes

$\boldsymbol{s}_i = (s_1,\ s_2,\ ...,\ s_N)$ $s_i$ : $1$ other: $0$

$N$ : # speakers $i$ : index

Simple, Widely used

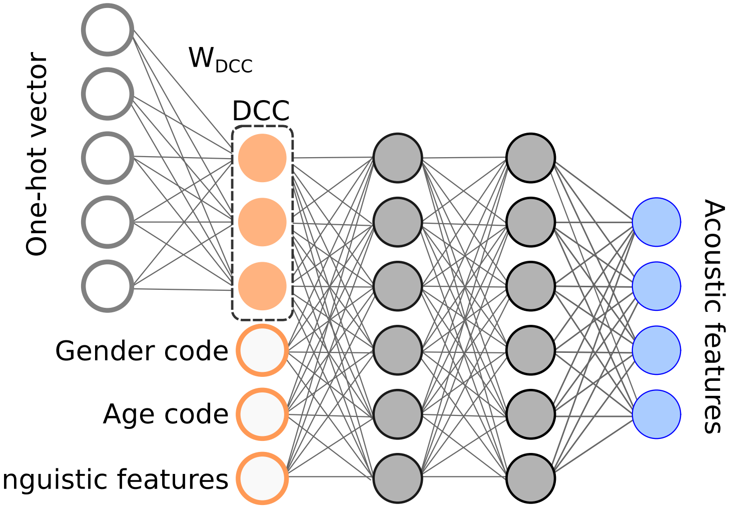
## Random-vector codes

$\boldsymbol{s}_i = (s_1,\ s_2,\ ...,\ s_K)$ $s_k$ : unique and random value

$K$ : Dimension

Easy to change the dimension

## Discriminant condition codes



– Projection of one-hot vector
– Project matrix: W_dcc
– Codes trained jointly with other parameters

Data-driven

# Gender and age codes

## One-hot vector

- – Gender: 2 dimensions (1st dim: female, 2nd dim: male)
- – Age: 7 dimensions (10’s, 20’s, …, 70’s)

## Binary/numerical representation (1 dimension)

- – Gender: binary (0: female, 1:male)
- – Age: use the age directly
- – Probably more intuitive easy-to-control representation

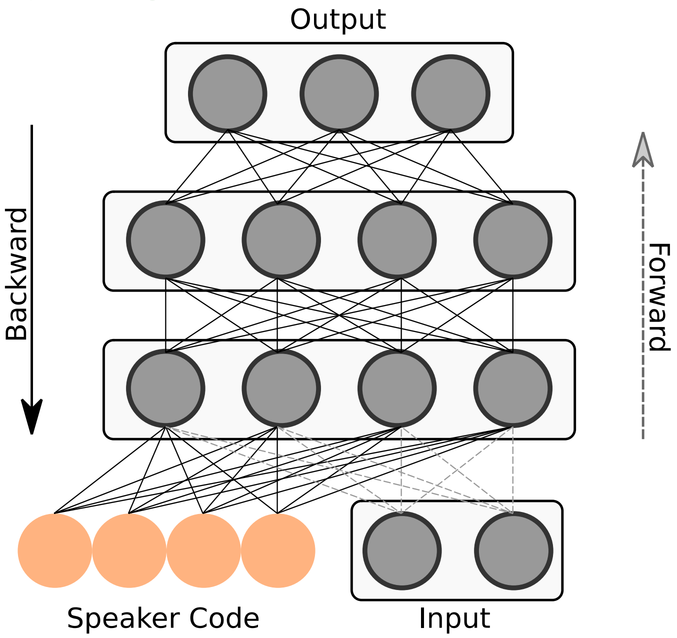
# Adaptation using input code: ‘***phantom code***’

## Estimate speaker code using adaptation data

- Estimation based on back-propagation [Bridle et al.; 90]
- Estimate the speaker code only, fix the other codes and other DNN parameters

## Update procedures of the speaker codes

- Initialize the codes with the average
- Update the codes
  - Fixed maximum number of epoch:
  - Fixed learning rate
- Choose codes that has minimum errors
- Simple!!
-



# Experimental conditions

| Multi-speaker | | | | Adaptation | | | |
|---|---|---|---|---|---|---|---|
| Age | Male | Female | Total | Age | Male | Female | Total |
| 10-20 | 8 | 8 | 16 | 10-20 | 0 | 2 | 2 |
| 21-30 | 8 | 8 | 16 | 21-30 | 2 | 2 | 4 |
| 31-40 | 8 | 8 | 16 | 31-40 | 2 | 2 | 4 |
| 41-50 | 8 | 8 | 16 | 41-50 | 1 | 2 | 3 |
| 51-60 | 8 | 8 | 16 | 51-60 | 2 | 2 | 4 |
| 61-70 | 8 | 8 | 16 | 61-70 | 2 | 2 | 4 |
| 71- | 8 | 8 | 16 | 71- | 0 | 2 | 2 |
| Total | 56 | 56 | **112** | Total | 9 | 14 | **23** |

– High-quality Japanese speech database

– Training: 112 speakers, 100 utterances per speaker, total of 11,170 utterances

– Adaptation: 23 speakers, 100 utterances per speaker

– Test: 10 different sentences per speaker

# Experimental conditions (contd.)

## Acoustic features (outputs)

- 60-dim STRAIGHT mel-cepstrum$+\Delta + \Delta^2$
- Voiced/unvoiced flag
- Log F0 $+\Delta + \Delta^2$
- 25-dim band-limited aperiodicities $+\Delta + \Delta^2$

## Text features (inputs)

- Open JTalk: 386-dim linguistic features
- Phone duration
  - Use oracle duration to compute objective measures
  - HMM-based forced-alignment
- Input codes

# Experimental conditions (contd.)

| | Speaker code<br>（S） | | Gender code<br>（G） | | Age code （A） | |
|---|---|---|---|---|---|---|
| Model label | Type | Size | Type | Size | Type | Size |
| ONE-S | One-hot | 112 | N/A | N/A | N/A | N/A |
| ONE-SGA’ | One-hot | 112 | One-hot | 2 | One-hot | 7 |
| ONE-SGA | One-hot | 112 | Binary | 1 | Numeric | 1 |
| RND112-SGA | Random | 112 | Binary | 1 | Numeric | 1 |
| RND008-SGA | Random | 8 | Binary | 1 | Numeric | 1 |
| DCC112-SGA | DCC | 112 | Binary | 1 | Numeric | 1 |
| DCC008-SGA | DCC | 8 | Binary | 1 | Numeric | 1 |

– Simple feed forward DNN

– Hidden layers: 5, units: 1024

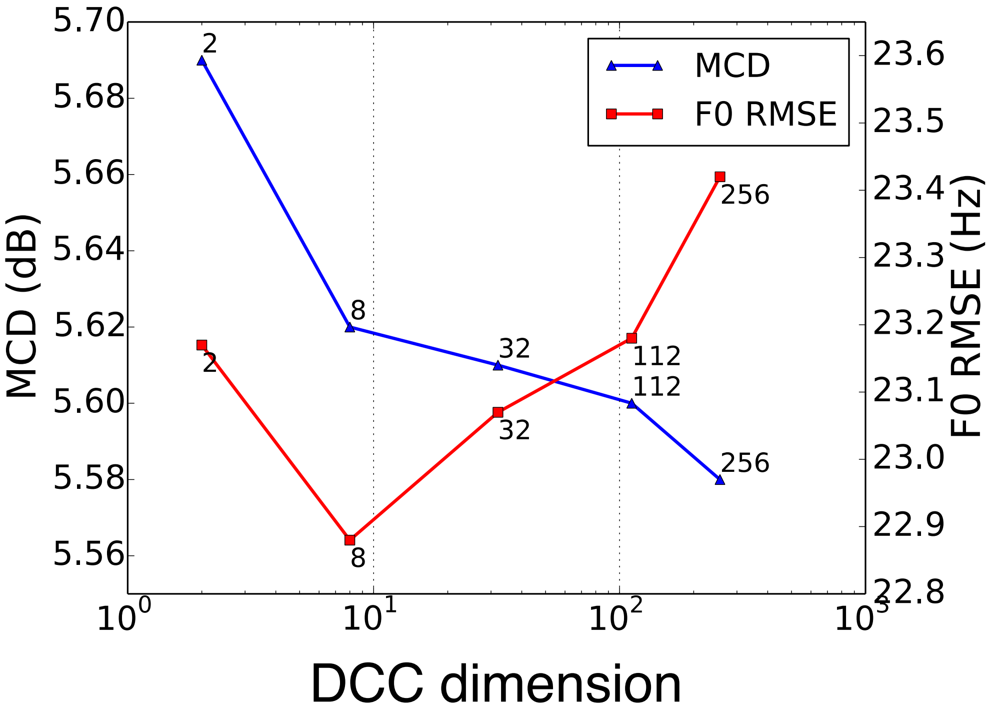
– Activation function: sigmoid

– Learning rate: 0.05, 10 epochs

– Objective measures: Mel-cepstral distortion, F0 RMSE

# Objective evaluation: Multi-speaker modelling

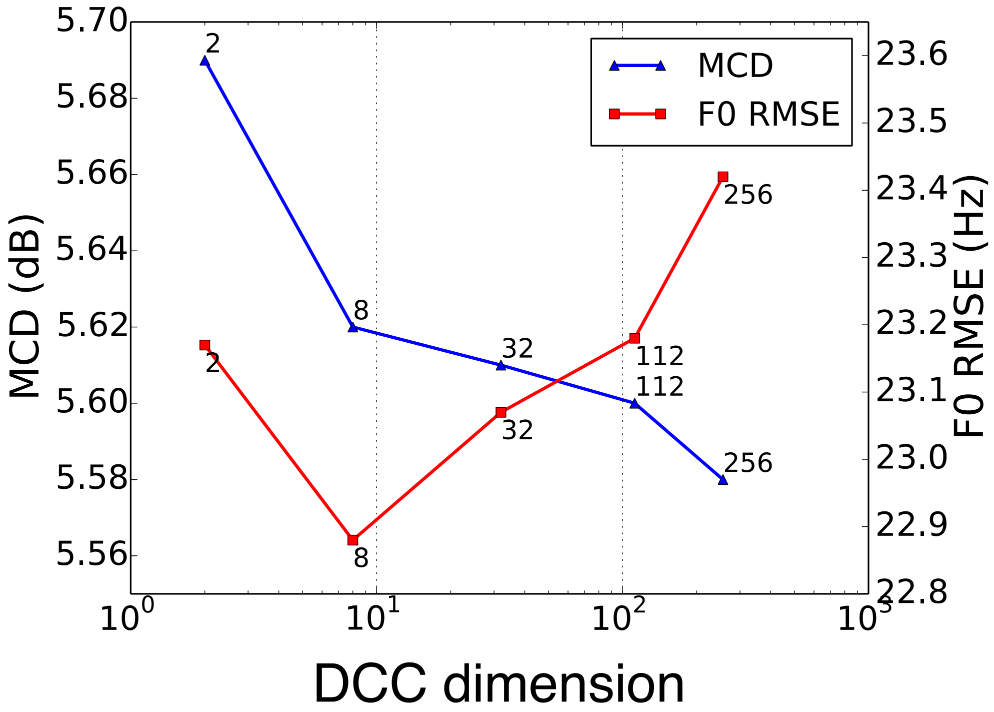
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc' | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |



- Evaluation using training speakers' codes
- ONE-c: correct code
- All systems are better than 'ONE-a (average voice)'
  - Possible to model multiple speakers simultaneously
- No significant differences between code representations
- Preferable DCC dimension depends on acoustic features

# Objective evaluation: Multi-speaker modelling

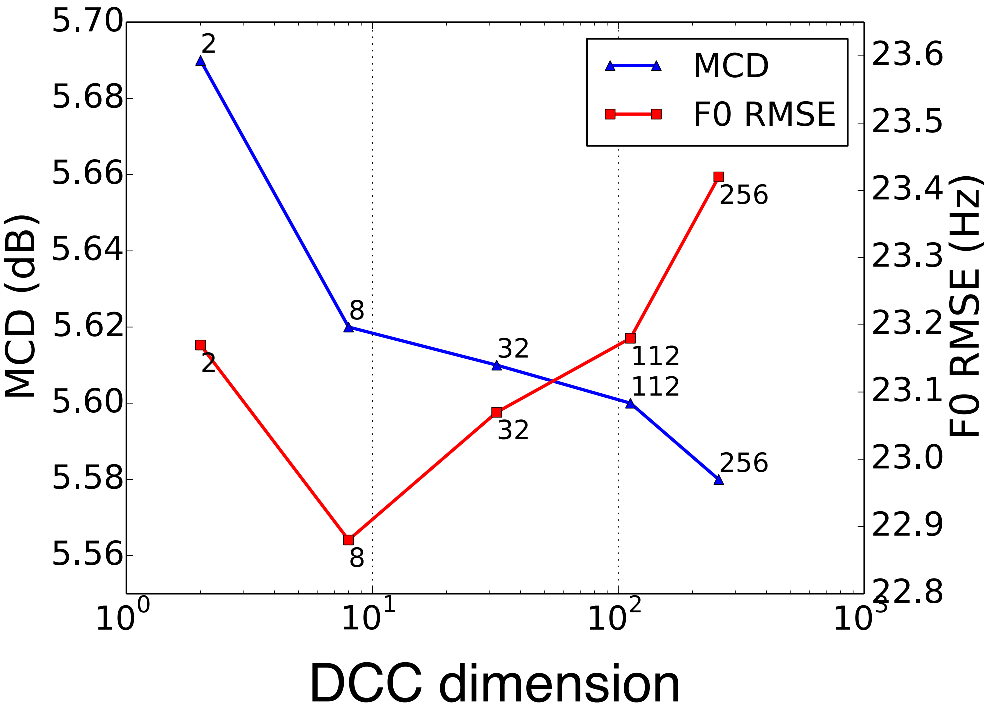
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |



- Evaluation using training speakers’ codes
- ONE-c: correct code
- All systems are better than ‘ONE-a (average voice)’
  - <u>Possible to model multiple speakers simultaneously</u>
- No significant differences between code representations
- Preferable DCC dimension depends on acoustic features

# Objective evaluation: Multi-speaker modelling

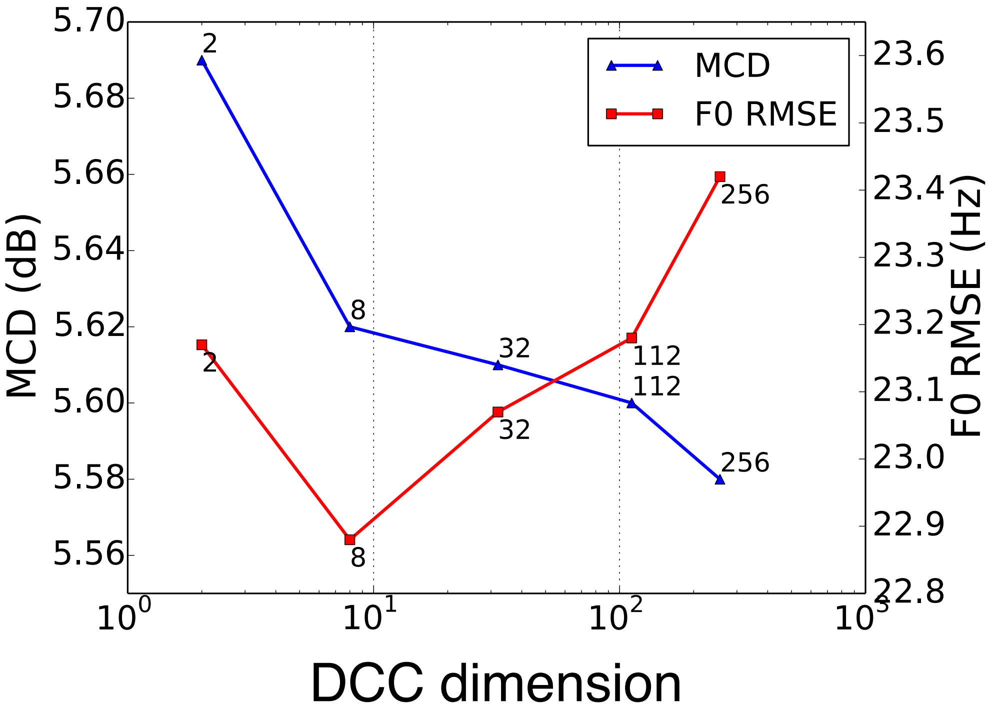
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |



- Evaluation using training speakers’ codes
- ONE-c: correct code
- All systems are better than ‘ONE-a (average voice)’
  - Possible to model multiple speakers simultaneously
- No significant differences between code representations
- Preferable DCC dimension depends on acoustic features

# Objective evaluation: Multi-speaker modelling

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |



– Evaluation using training speakers’ codes

– ONE-c: correct code

– All systems are better than ‘ONE-a (average voice)’

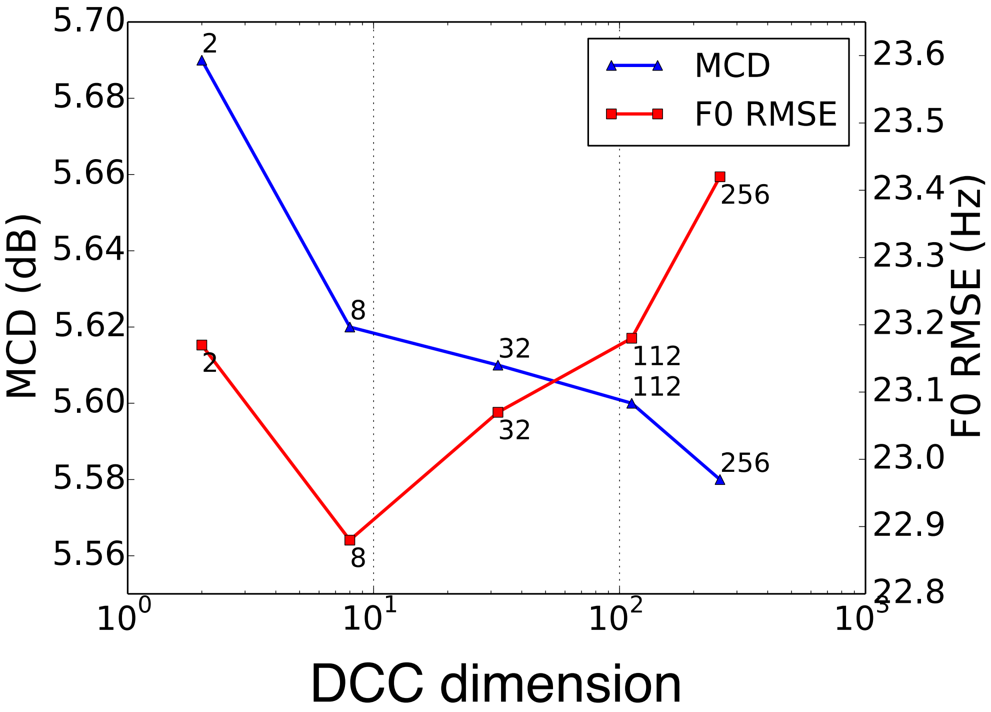
  – Possible to model multiple speakers simultaneously

– No significant differences between code representations

– Preferable DCC dimension depends on acoustic features

# Objective evaluation: Multi-speaker modelling

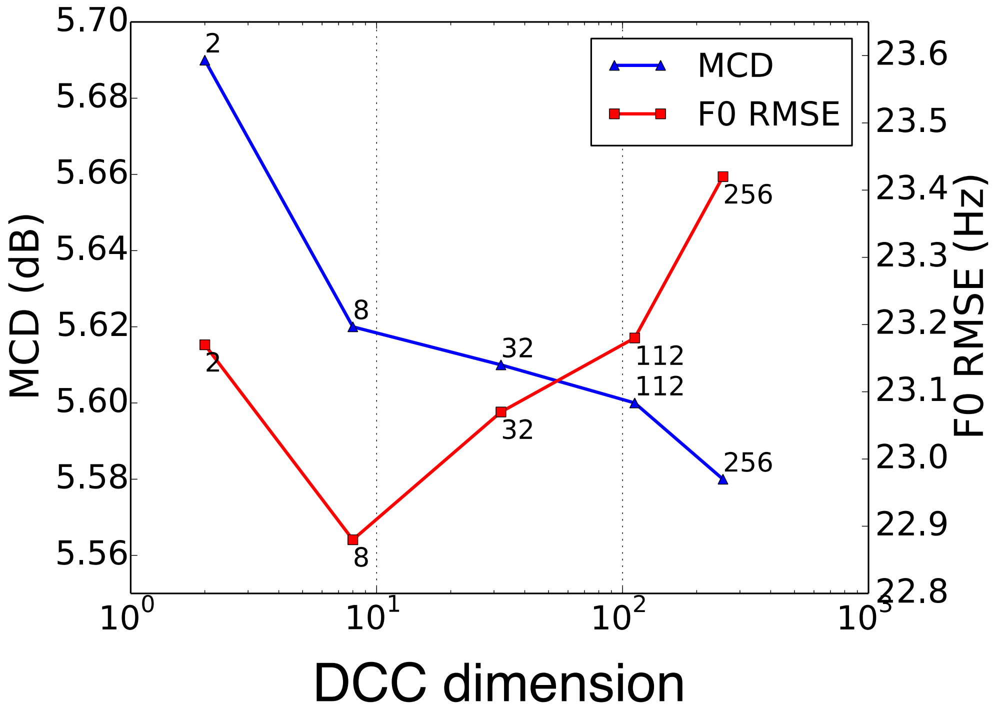
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |



- Evaluation using training speakers’ codes
- ONE-c: correct code
- All systems are better than ‘ONE-a (average voice)’
  - <u>Possible to model multiple speakers simultaneously</u>
- No significant differences between code representations
- Preferable DCC dimension depends on acoustic features

# Objective evaluation: Multi-speaker modelling

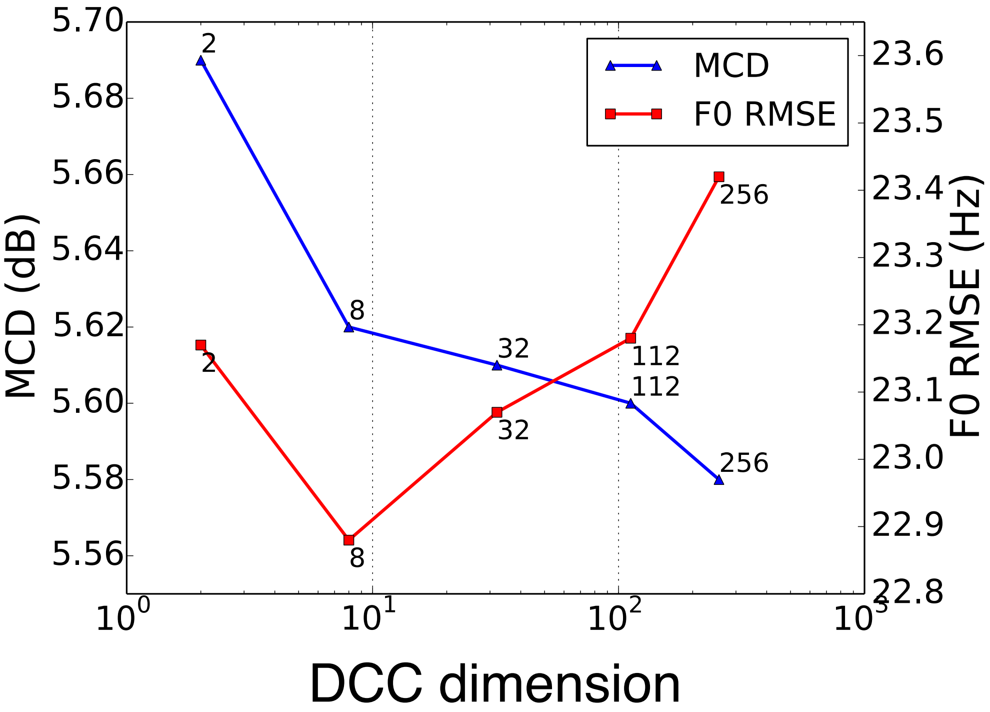
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |



- – Evaluation using training speakers’ codes
- – ONE-c: correct code
- – All systems are better than ‘ONE-a (average voice)’
    - – <u>Possible to model multiple speakers simultaneously</u>
- – No significant differences between code representations
- – Preferable DCC dimension depends on acoustic features

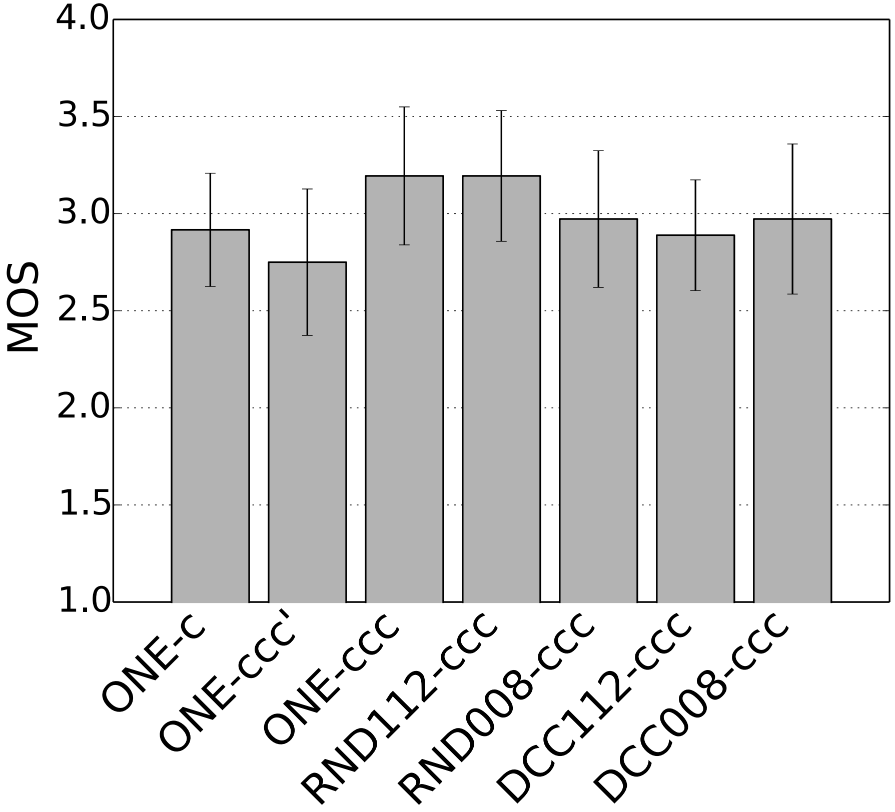
# Objective evaluation: Multi-speaker modelling

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |

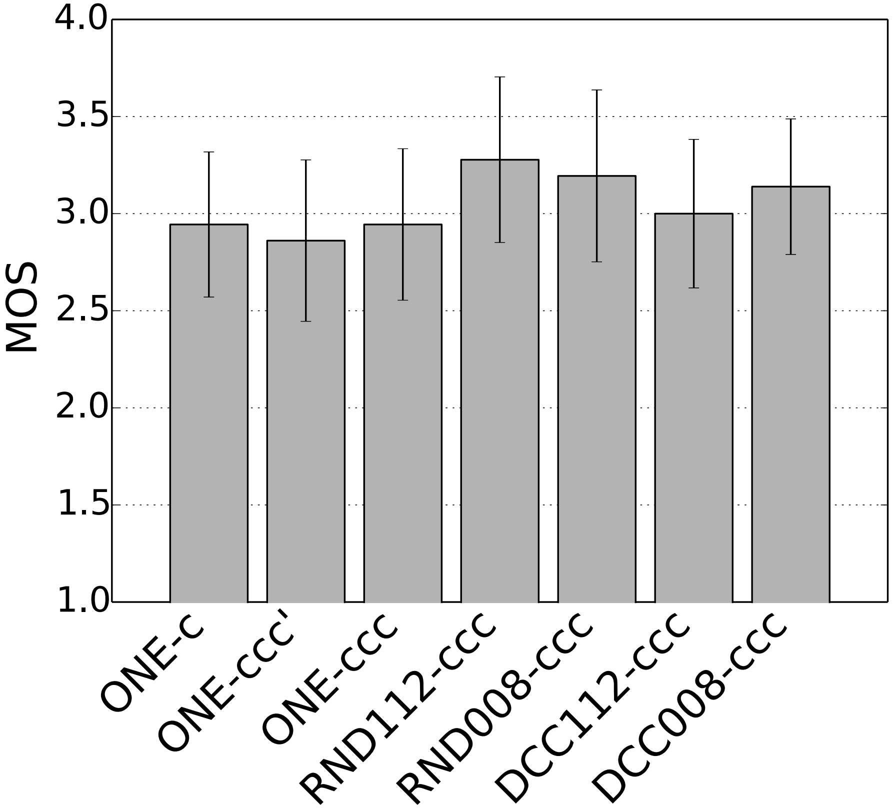


- – Evaluation using training speakers’ codes
- – ONE-c: correct code
- – All systems are better than ‘ONE-a (average voice)’
  - – Possible to model multiple speakers simultaneously
- – No significant differences between code representations
- – Preferable DCC dimension depends on acoustic features

# Subjective evaluation: Multi-speaker modelling



Quality



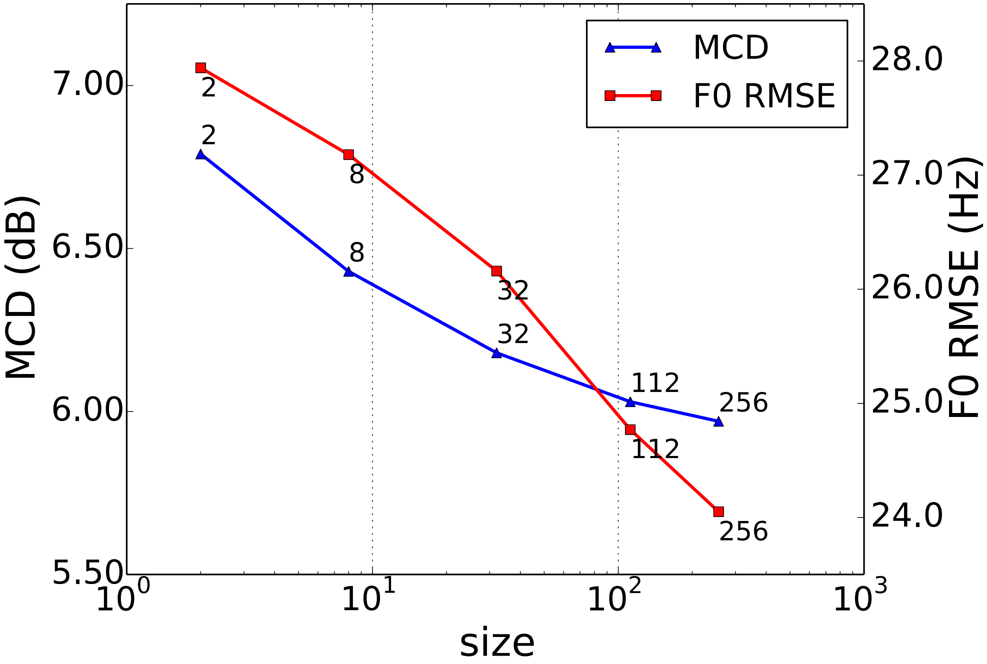
Speaker similarity

– No significant differences

– Tested code representations have similar performance

# Objective evaluation: Adaptation

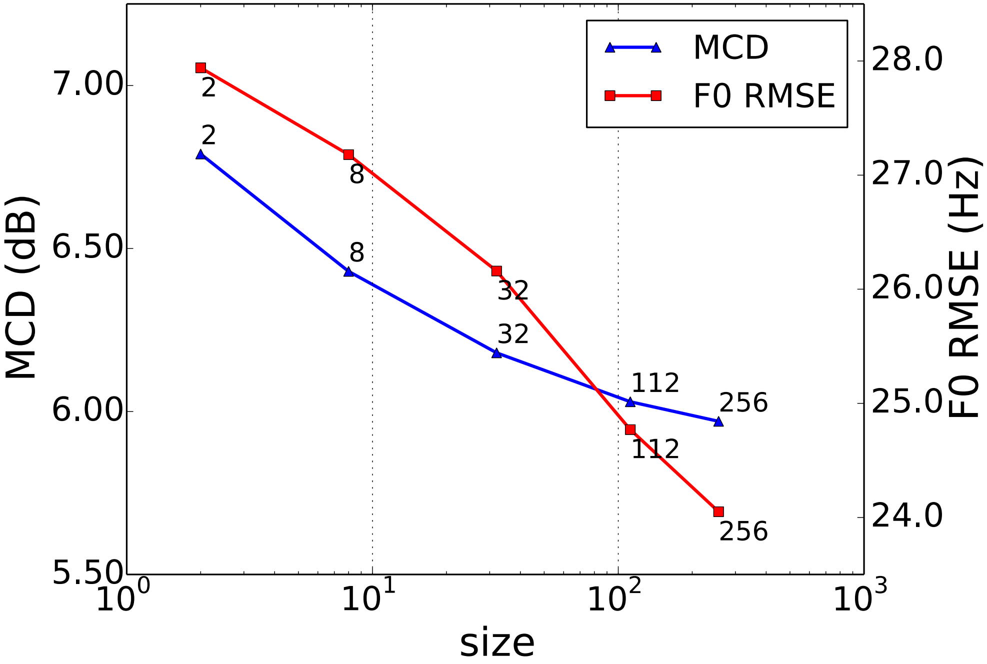
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |



- – Evaluation using codes estimated from adaptation data
- – ONE-e: estimated code
- – All estimated systems are better than ‘ONE-a (average voice)’
  - – Possible to adapt DNN to a new target speaker
- – Larger dimensions are better for RND and DCC based codes

# Objective evaluation: Adaptation

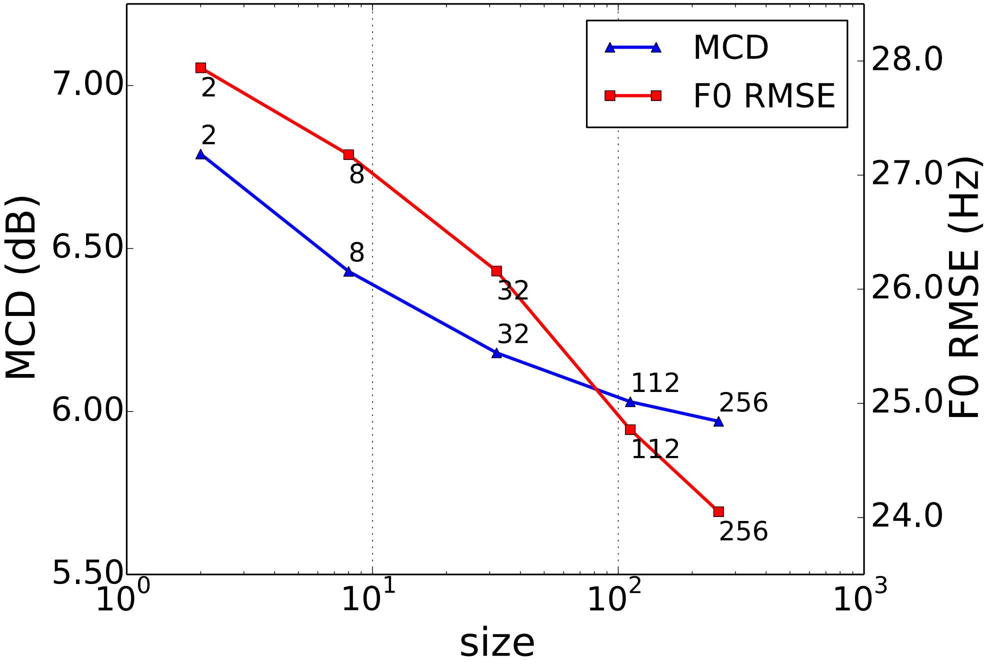
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |



- Evaluation using codes estimated from adaptation data
- ONE-e: estimated code
- All estimated systems are better than ‘ONE-a (average voice)’
    - Possible to adapt DNN to a new target speaker
- Larger dimensions are better for RND and DCC based codes

# Objective evaluation: Adaptation

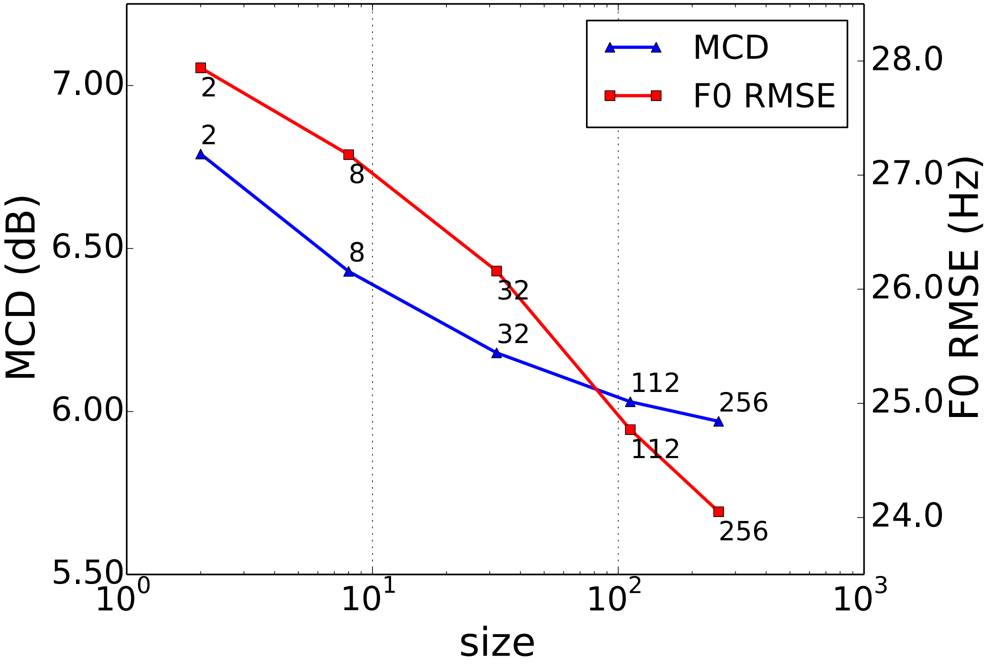
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |



- Evaluation using codes estimated from adaptation data
- ONE-e: estimated code
- All estimated systems are better than ‘ONE-a (average voice)’
  - Possible to adapt DNN to a new target speaker
- Larger dimensions are better for RND and DCC based codes

# Objective evaluation: Adaptation

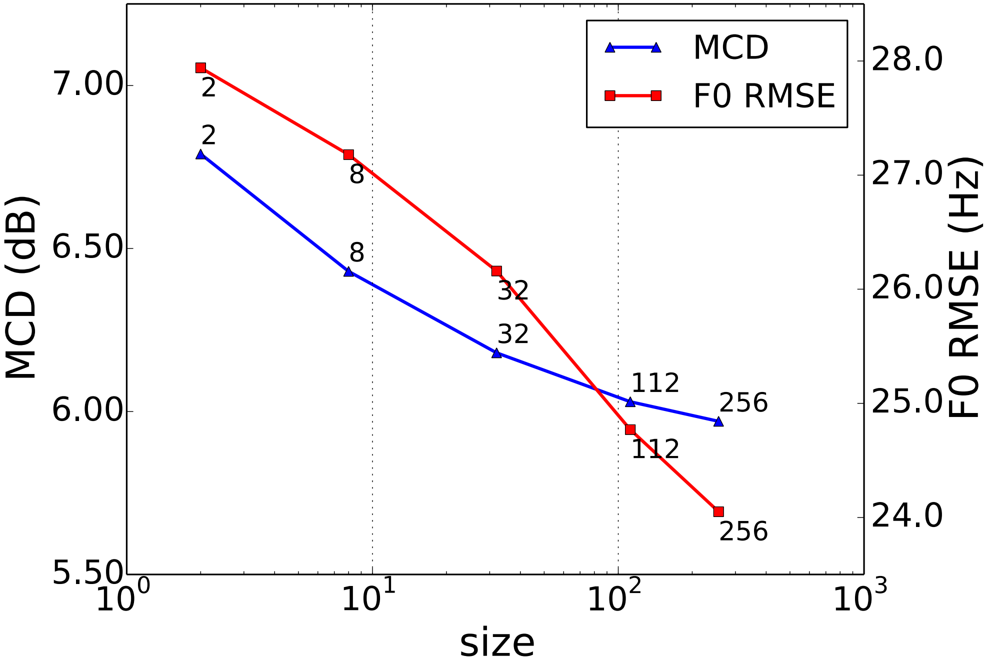
| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |



- Evaluation using codes estimated from adaptation data
- ONE-e: estimated code
- All estimated systems are better than ‘ONE-a (average voice)’
  - <u>Possible to adapt DNN to a new target speaker</u>
- Larger dimensions are better for RND and DCC based codes

# Objective evaluation: Adaptation

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |



- Evaluation using codes estimated from adaptation data
- ONE-e: estimated code
- All estimated systems are better than ‘ONE-a (average voice)’
    - <u>Possible to adapt DNN to a new target speaker</u>
- Larger dimensions are better for RND and DCC based codes

# Known speakers vs. unknown speakers

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |

Training speakers (known)

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |

Adaptation speakers (unknown)

– Known speakers: **5.6 dB**, unknown speakers: **6 dB**

– Unknown speakers have worse errors expectedly

– More improved adaptation methods required

– F0 adaptation performance seems to be comparable

# Known speakers vs. unknown speakers

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |

Training speakers (known)

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |

Adaptation speakers (unknown)

– Known speakers: **5.6 dB**, unknown speakers: **6 dB**

– Unknown speakers have worse errors expectedly

– More improved adaptation methods required

– F0 adaptation performance seems to be comparable

# Known speakers vs. unknown speakers

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |

Training speakers (known)

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |

Adaptation speakers (unknown)

- Known speakers: **5.6 dB**,  unknown speakers: **6 dB**
- Unknown speakers have worse errors expectedly
- More improved adaptation methods required
- F0 adaptation performance seems to be comparable

# Known speakers vs. unknown speakers

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |

Training speakers (known)

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |

Adaptation speakers (unknown)

– Known speakers: **5.6 dB**, unknown speakers: **6 dB**

– Unknown speakers have worse errors expectedly

– More improved adaptation methods required

– F0 adaptation performance seems to be comparable

# Known speakers vs. unknown speakers

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.64 | 52.06 |
| ONE-c | 5.62 | 23.79 |
| ONE-ccc’ | 5.61 | 23.80 |
| ONE-ccc | 5.58 | 23.43 |
| RND112-ccc | 5.60 | 23.48 |
| RND008-ccc | 5.60 | 23.06 |
| DCC112-ccc | 5.60 | 23.18 |
| DCC008-ccc | 5.62 | 22.88 |

Training speakers (known)

| Strategy | MCD | F0 RMSE |
|---|---|---|
| ONE-a | 7.49 | 53.90 |
| ONE-e | 6.02 | 23.71 |
| ONE-ecc’ | 6.06 | 23.75 |
| ONE-ecc | 6.01 | 23.54 |
| RND112-ecc | 6.46 | 24.98 |
| RND008-ecc | 6.49 | 29.44 |
| DCC112-ecc | 6.03 | 24.77 |
| DCC008-ecc | 6.43 | 27.18 |

Adaptation speakers (unknown)

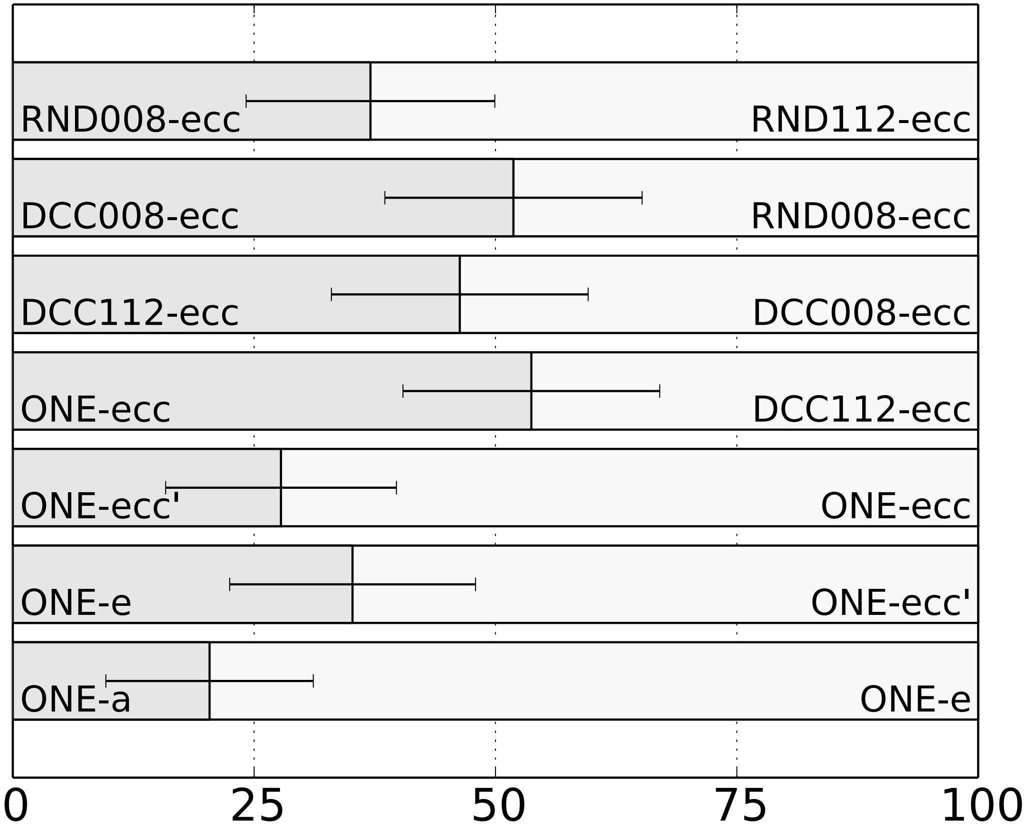
– Known speakers: **5.6 dB**, unknown speakers: **6 dB**

– Unknown speakers have worse errors expectedly

– More improved adaptation methods required

– F0 adaptation performance seems to be comparable

# Subjective evaluation: AB test, adaptation



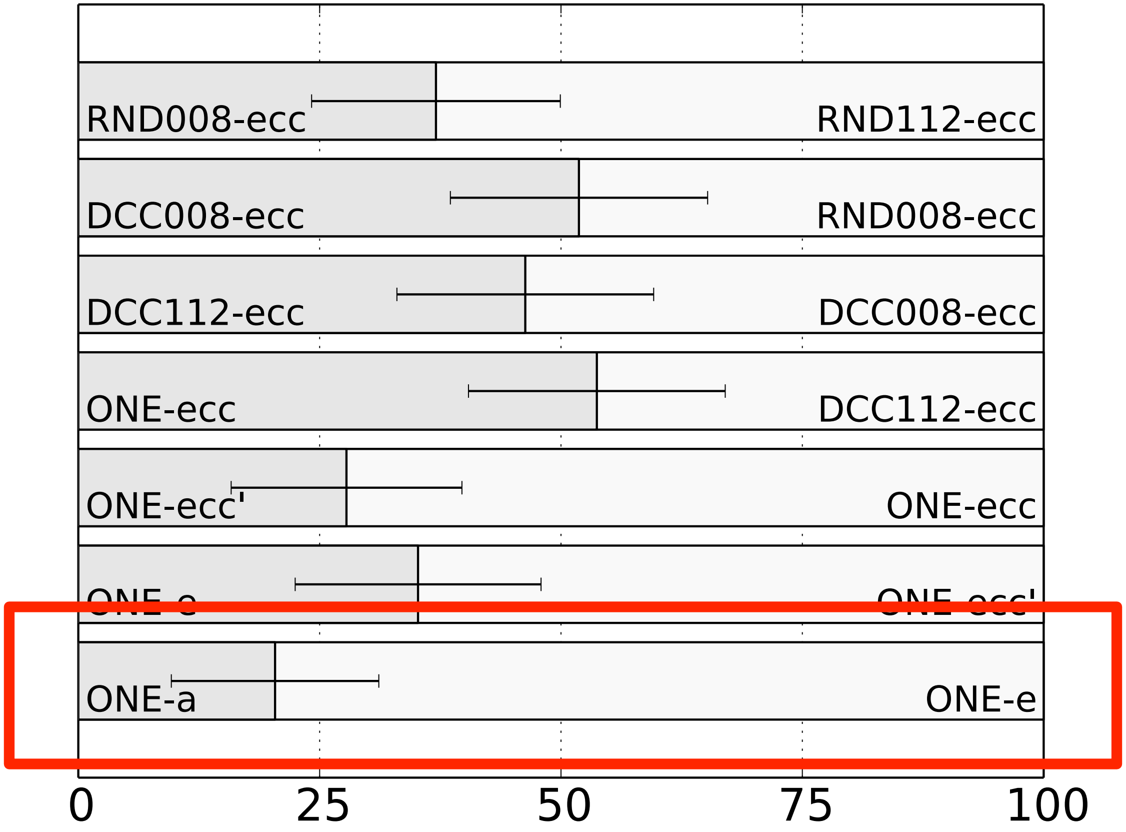
– Speaker similarly judgement via AB test

– Estimated codes have more similarity than average code

– Using gender and age codes improves speaker similarity

– Numeric gender/age representation is better

# Subjective evaluation: AB test, adaptation



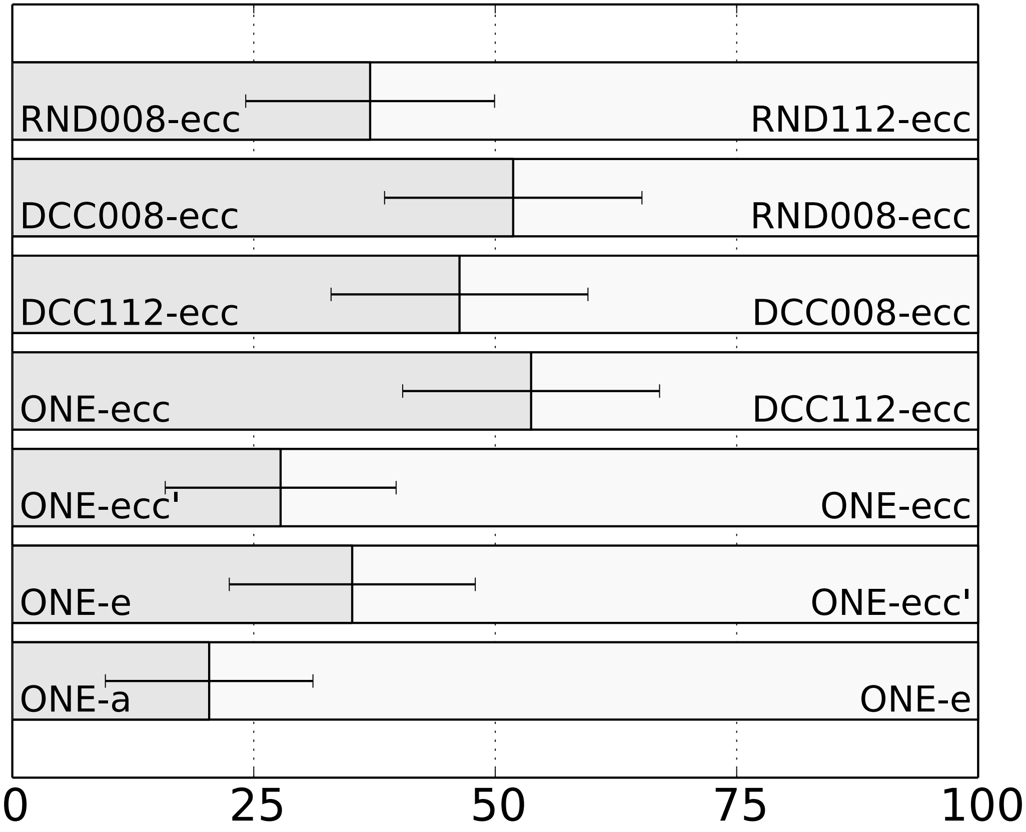
– Speaker similarly judgement via AB test

– Estimated codes have more similarity than average code

– Using gender and age codes improves speaker similarity

– Numeric gender/age representation is better

# Subjective evaluation: AB test, adaptation



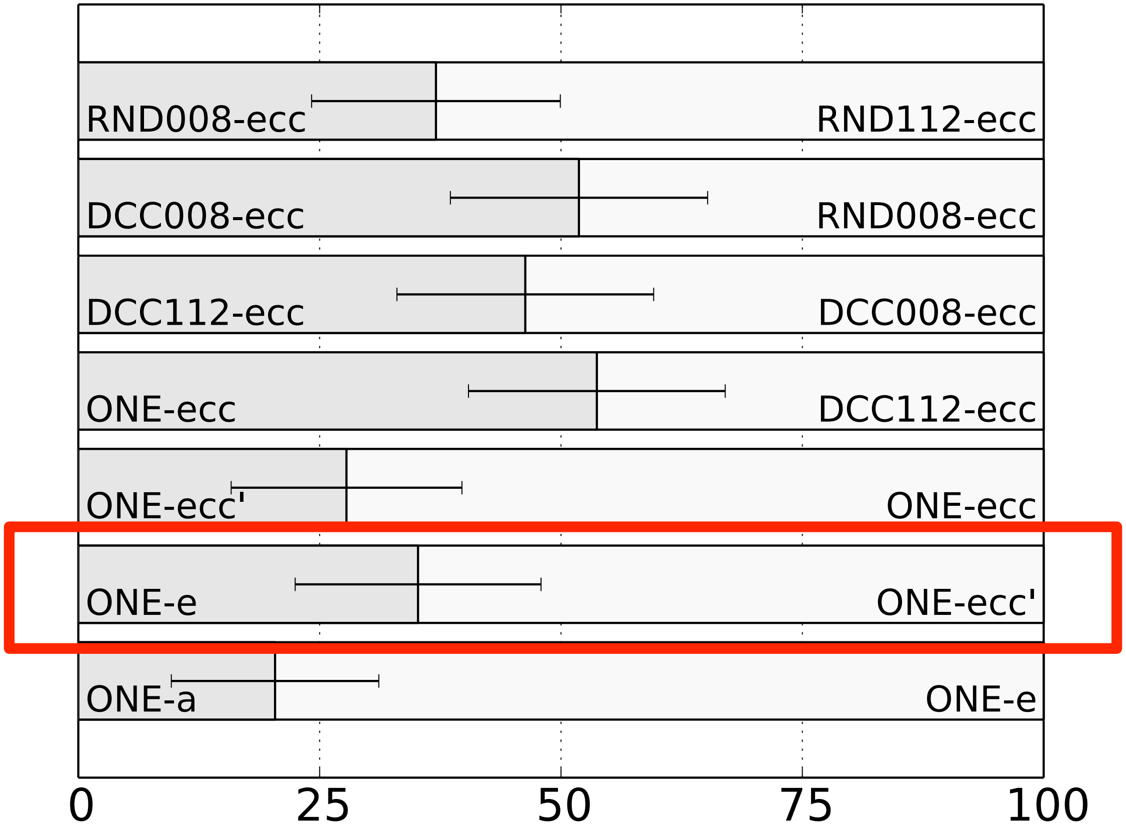
– Speaker similarly judgement via AB test

– Estimated codes have more similarity than average code

– Using gender and age codes improves speaker similarity

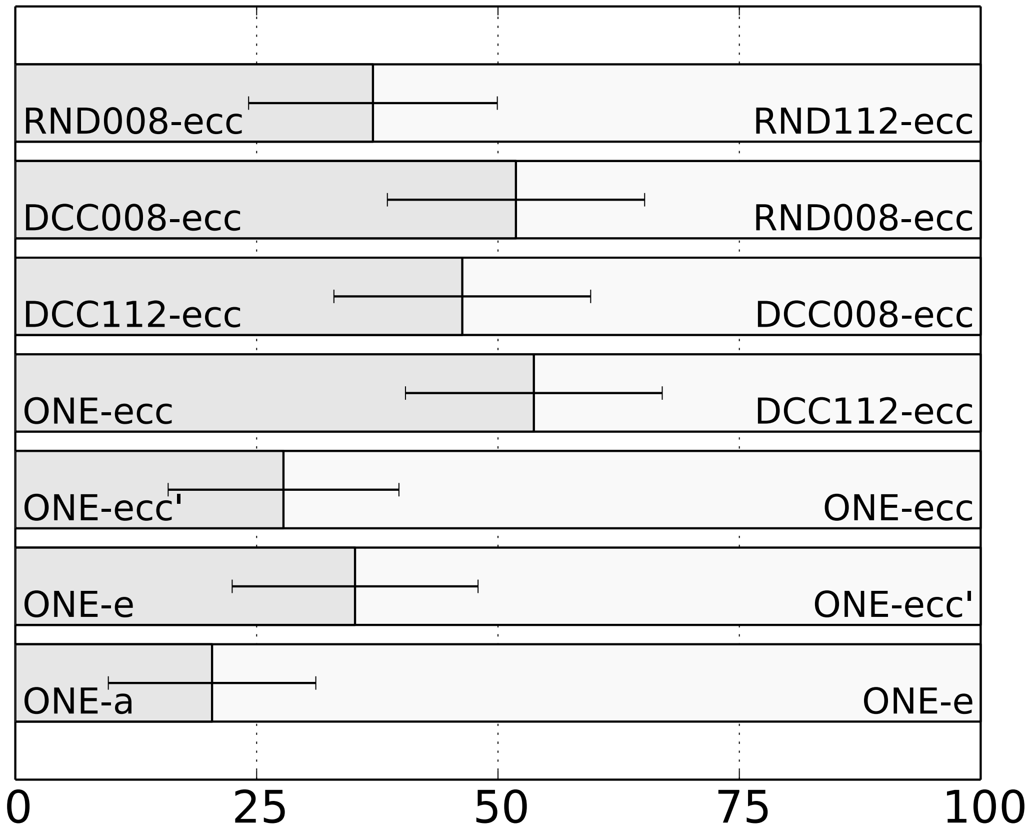
– Numeric gender/age representation is better

# Subjective evaluation: AB test, adaptation



- – Speaker similarly judgement via AB test
- – Estimated codes have more similarity than average code
- – Using gender and age codes improves speaker similarity
- – Numeric gender/age representation is better

# Subjective evaluation: AB test, adaptation



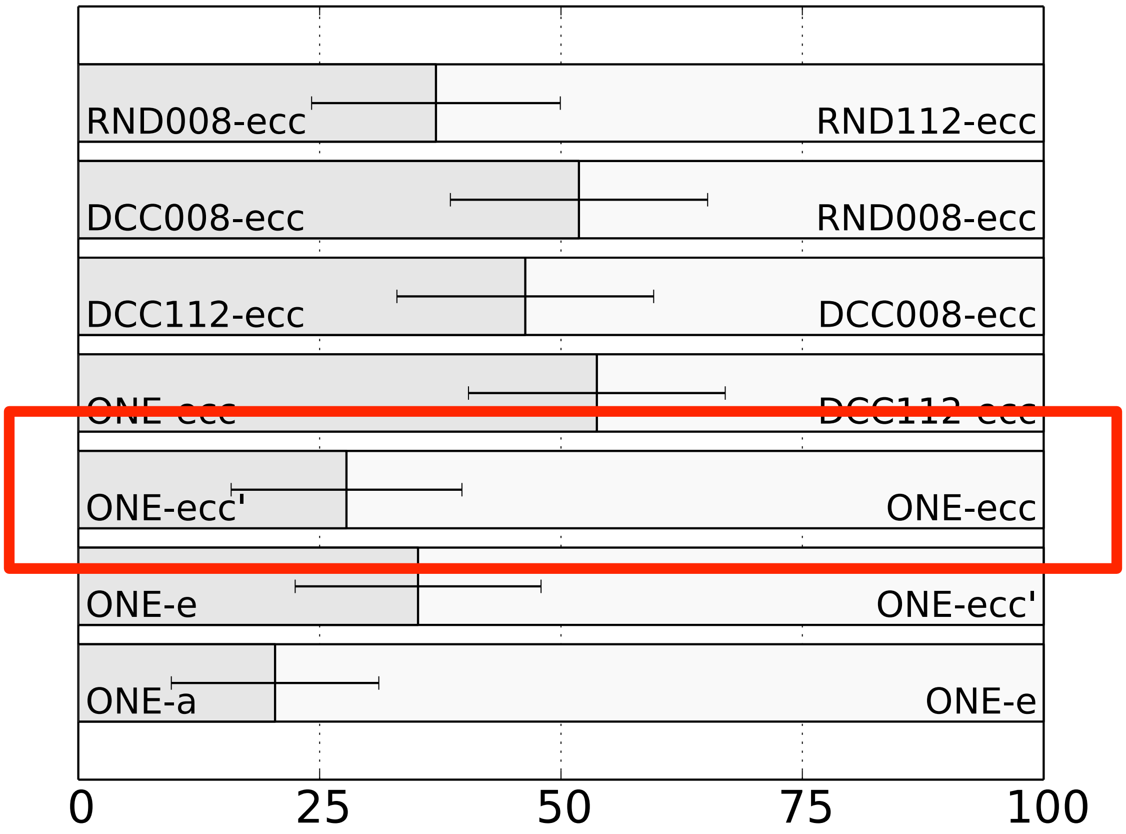
– Speaker similarly judgement via AB test

– Estimated codes have more similarity than average code

– Using gender and age codes improves speaker similarity

– Numeric gender/age representation is better

# Subjective evaluation: AB test, adaptation



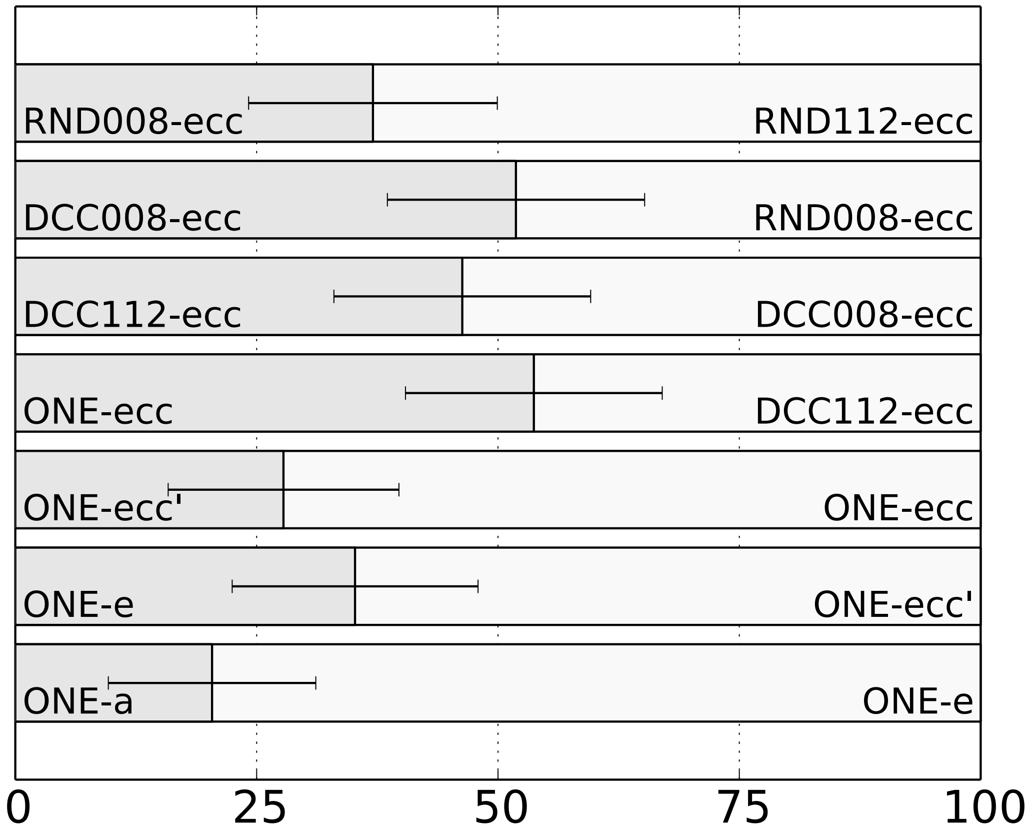
– Speaker similarly judgement via AB test

– Estimated codes have more similarity than average code

– Using gender and age codes improves speaker similarity

– Numeric gender/age representation is better

# Subjective evaluation: AB test, adaptation



– Speaker similarly judgement via AB test

– Estimated codes have more similarity than average code

– Using gender and age codes improves speaker similarity

– Numeric gender/age representation is better

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

## Speaker adaptation

| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

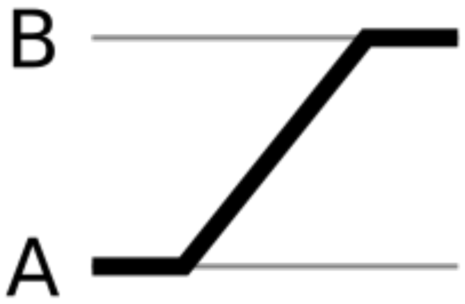
# Demonstration

## Speaker adaptation

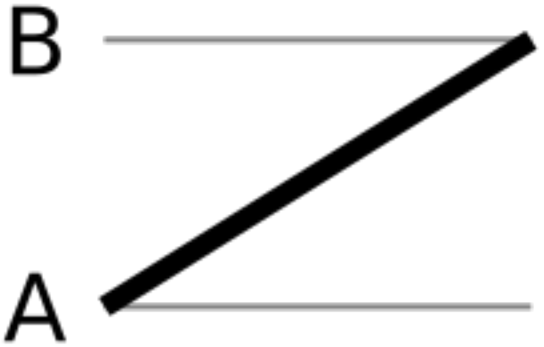
| Target | One-a | One-e | One-ecc |
|---|---|---|---|
| | | | |
| | | | |

# Demonstration

Speaker interpolation

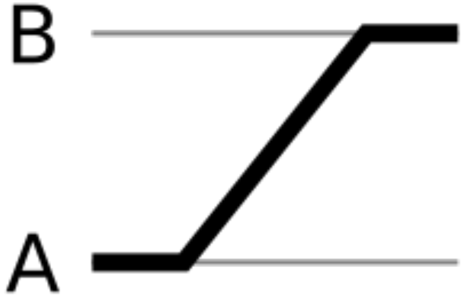


Gender morphing

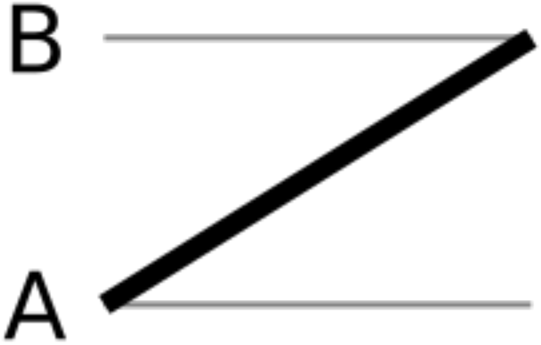


# Demonstration

Speaker interpolation

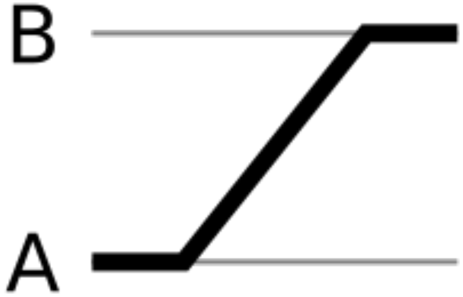


Gender morphing

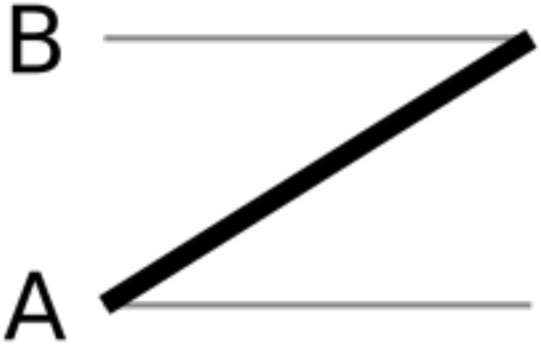


# Demonstration

Speaker interpolation



Gender morphing



# Conclusions

## DNN speech synthesis systems using input codes

- Multi-speaker modelling
- Speaker adaptation
- Manipulation and control

Large-scale speech database
Objective and subjective tests

- **Flexible DNN speech synthesizers**
  - Input codes seem to be effective
- Gender and age codes
  - Improve speaker-adaptation performance
  - Similar finding to HMM-based speaker adaptation

## Future work

- Evaluation using LSTM-RNN and waveform models
- Improved adaptation methods
- Investigate different input codes

# Other inputs codes that we're investigating

## Emotions

RNN-based audio examples where emotions are manipulated using emotional codes

## Speaking skills

Voice talent, semi voice talent, amateur

Annotations of hundreds of speakers completed

# Other inputs codes that we're investigating

## Emotions

RNN-based audio examples where emotions are manipulated using emotional codes

## Speaking skills

Voice talent, semi voice talent, amateur

Annotations of hundreds of speakers completed

Thank you very much

Q&A

contact: jyamagis@nii.ac.jp