Unsupervised speaker adaptation for DNN-based speech synthesis using input codes

Shinji Takaki[1], Yoshikazu Nishimura[2], Junichi Yamagishi[1]

[1] National Institute of Informatics
[2] alt Inc.

Background（1/2）

Statistical parametric speech synthesis

– Remarkable progress thanks to DNNs

Flexible and controllable speech synthesis

– Speaker, gender, and age codes: “input codes” [Luong+; 16]
  - Multi-speaker modeling
  - Flexible manipulation
  - Speaker adaptation

– Speaker adaptation using back-propagation [Luong+; 16]
  - Speech and text data of a target speaker are required

Speaker adaptation using only speech data

Background（2/2）

Speaker adaptation using a speaker-similarity vector

- Speaker-similarity vector : new speaker code
  - Text-independent ASV models are used
    - Posterior probabilities are concatenated to form the code
  - The code represents acoustic similarity to speakers
- Inputting the estimated code of a target speaker can generate the target speaker's voice

Speaker adaptation using low-quality speech data

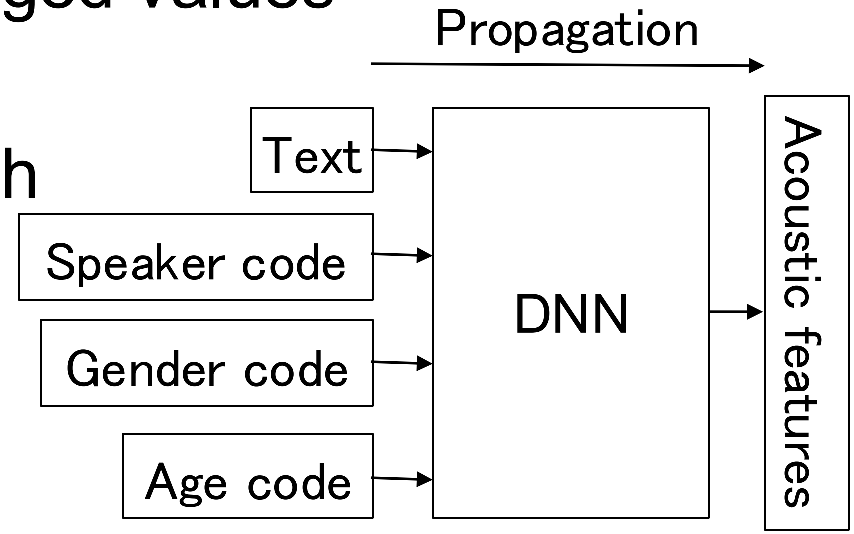
- A robust ASV model is required
- Model training using artificially created low-quality speech
  - Alleviating recording condition mismatch between training and adaptation data

Multi-speaker modelling using input codes

Multi-speaker modelling using input codes

- Input codes: simple additional inputs that differentiate ID, gender and age of speakers
- Generate multiple speakers' voices from a single DNN

- 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



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

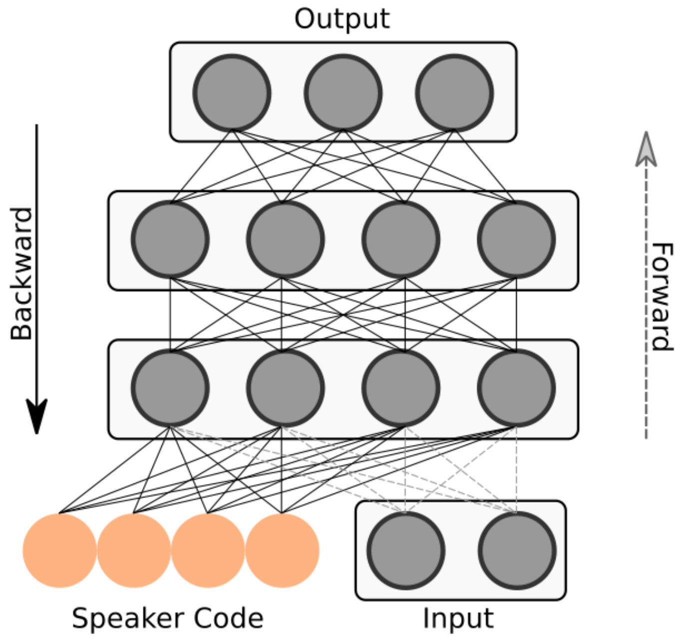
– Initialize the codes with the average

– Update the codes

- Fixed maximum number of epoch:
- Fixed learning rate

– Choose codes that has minimum errors

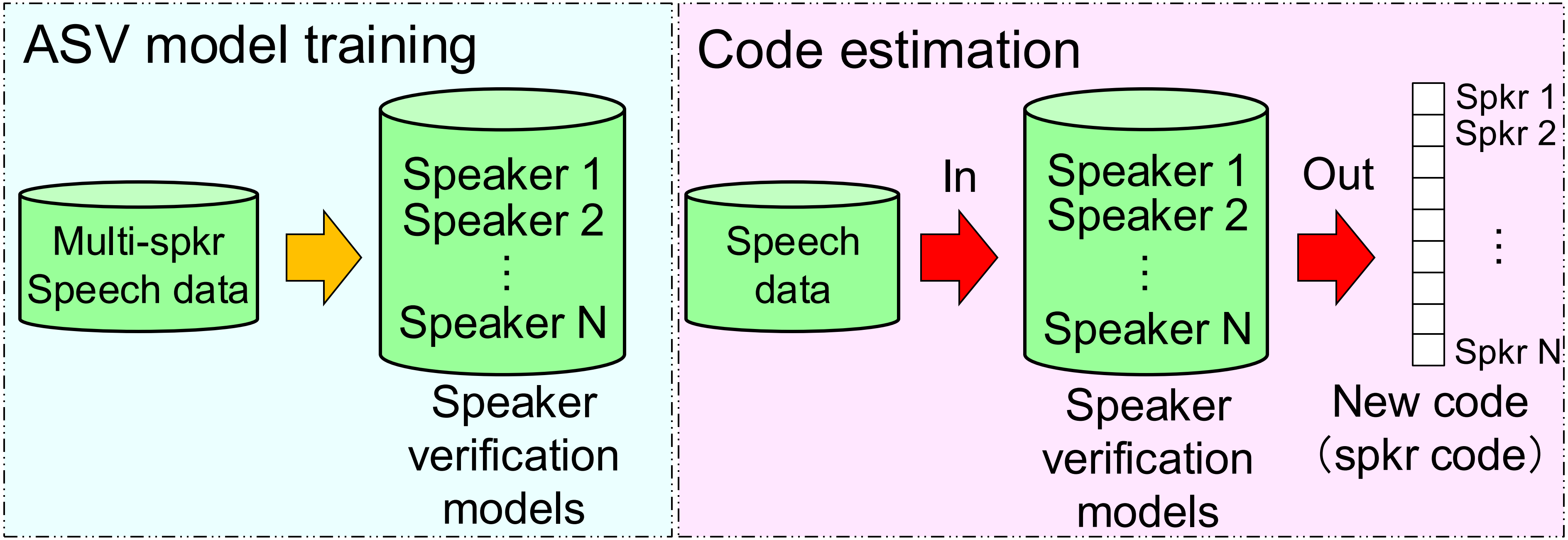
– Simple!!



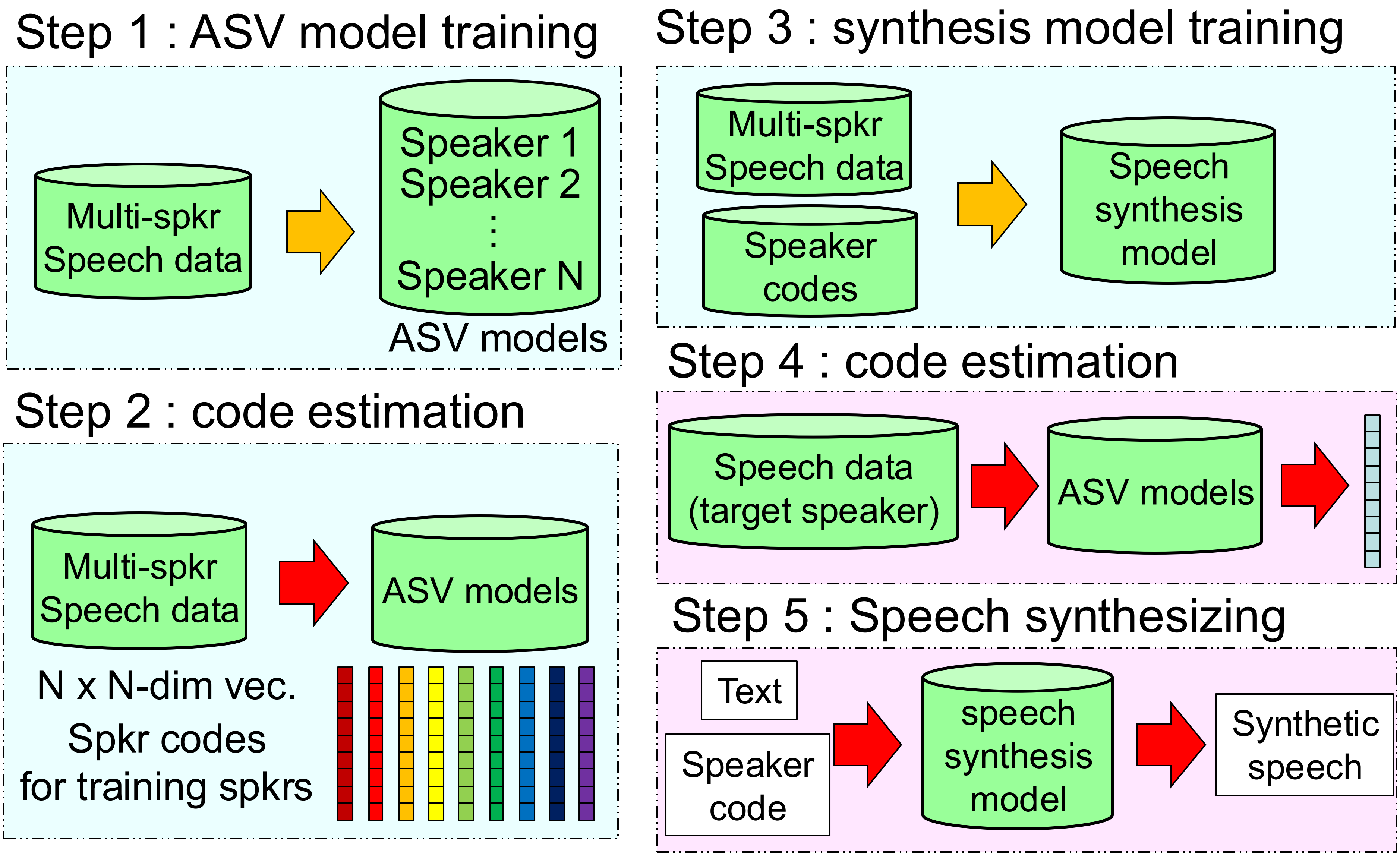
New speaker code : Speaker-similarity vector

Acoustic similarity to each of training speakers

- Replacing 1-hot vectors with speaker-similarity vectors
- Acoustic similarity is represented by posterior probability
  - Using text-independent ASV models
  - GMM-UBM or i-vector/PLDA is used



Flow of the proposed technique



Synthetic speech will vary if the speaker-similarity vector change

Training robust ASV model

Adaptation data is usually low quality

Training ASV models using low-quality data

- Alleviating recording condition mismatch
- Adding noise and reverberation to training speech data
  - An office room and a meeting room
  - Various SNRs
- Demand : Noise database [Thiemann+; 13]
- The Ace Challenge : Reverberation database [Hadad+; 14]
- Adaptation date is also artificially created
  - Office :
  - Meeting :

High SNR ⟷ Low SNR

Experimental conditions（1/3）

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（2/3）

Acoustic features（speaker verification）

- 19-dim MFCC+$\Delta + \Delta^2$ （MFCC）
- 19-dim WORLD mel-cepstrum+$\Delta + \Delta^2$ （MGC）
- 20-dim F0 features +$\Delta + \Delta^2$ （F0）
  - DCT is applied to F0 of prev., current and next 32 frames

Acoustic features（speech synthesis）

- 59-dim WORLD mel-cepstrum+$\Delta + \Delta^2$
- Voiced/Unvoiced parameter, Log F0+$\Delta + \Delta^2$
- 25-dim band apriodicity +$\Delta + \Delta^2$

Input（speech synthesis）

- 386-dim linguistic features, oracle phone duration
- Speaker, gender and age codes

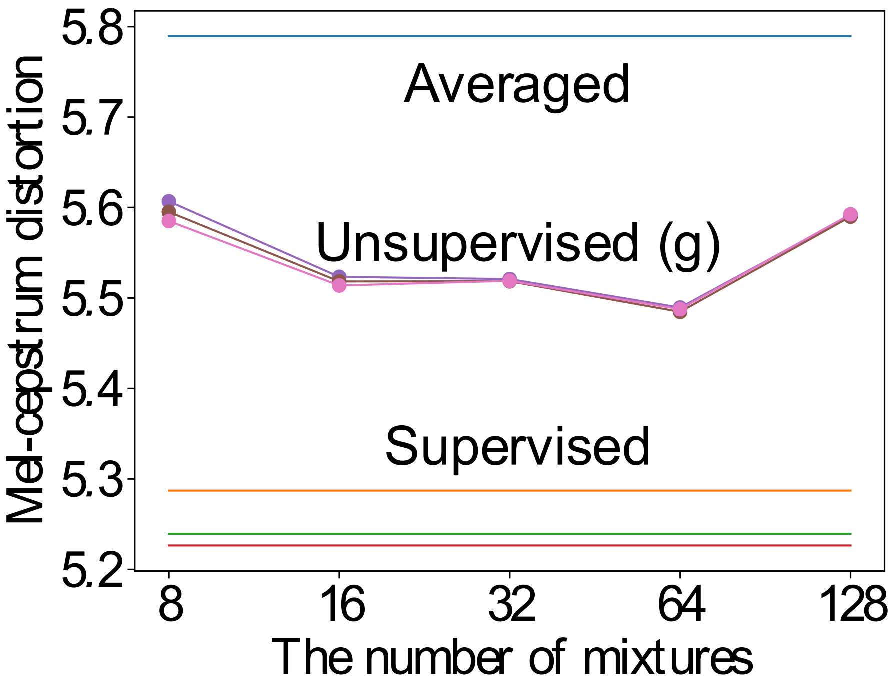
Experimental conditions（3/3）

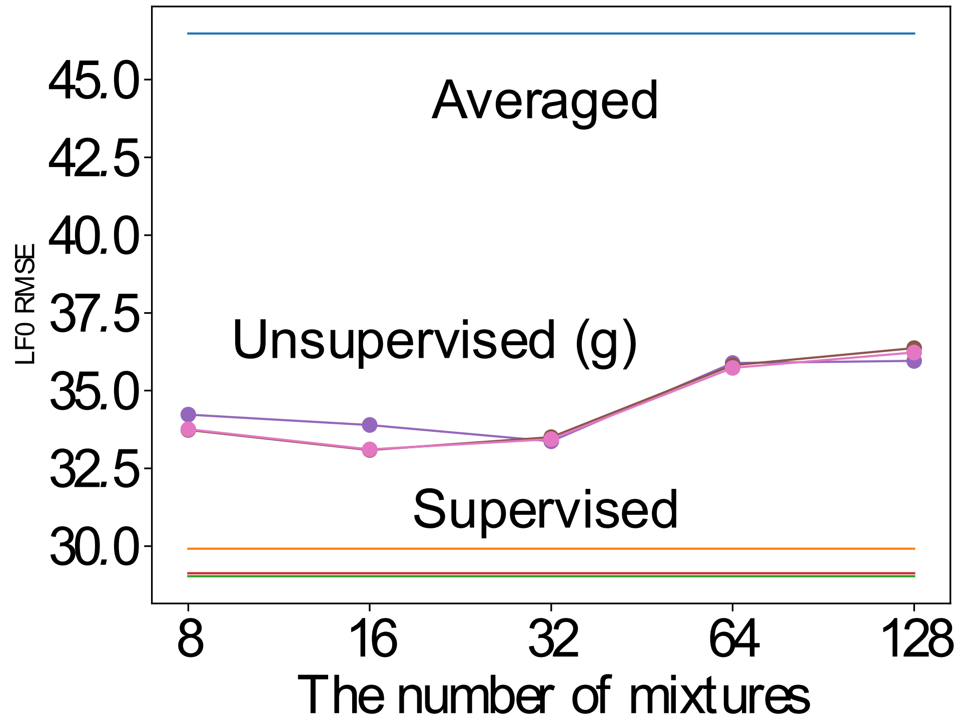
Systems	Multi-speaker model	Adaptation
Averaged	One-hot vector	-
Supervised	One-hot vector	Vector estimated by BP
Unsupervised (g)	Speaker-similarity vec. estimated from GMM-UBM	
Unsupervised (i)	Speaker-similarity vec. estimated form i-vector/PLDA	

– SIDEKIT was used to train ASV models
  - GMM-UBM（#mixtures：8, 16, 32, 64, 128）
  - i-vector/PLDA
    - –The number of mixtures for extracting i-vector：64
    - –i-vector dimension：400

– Speech synthesis model
  - Feed forward DNN（Hidden layers : 5, units : 1024）

Objective result



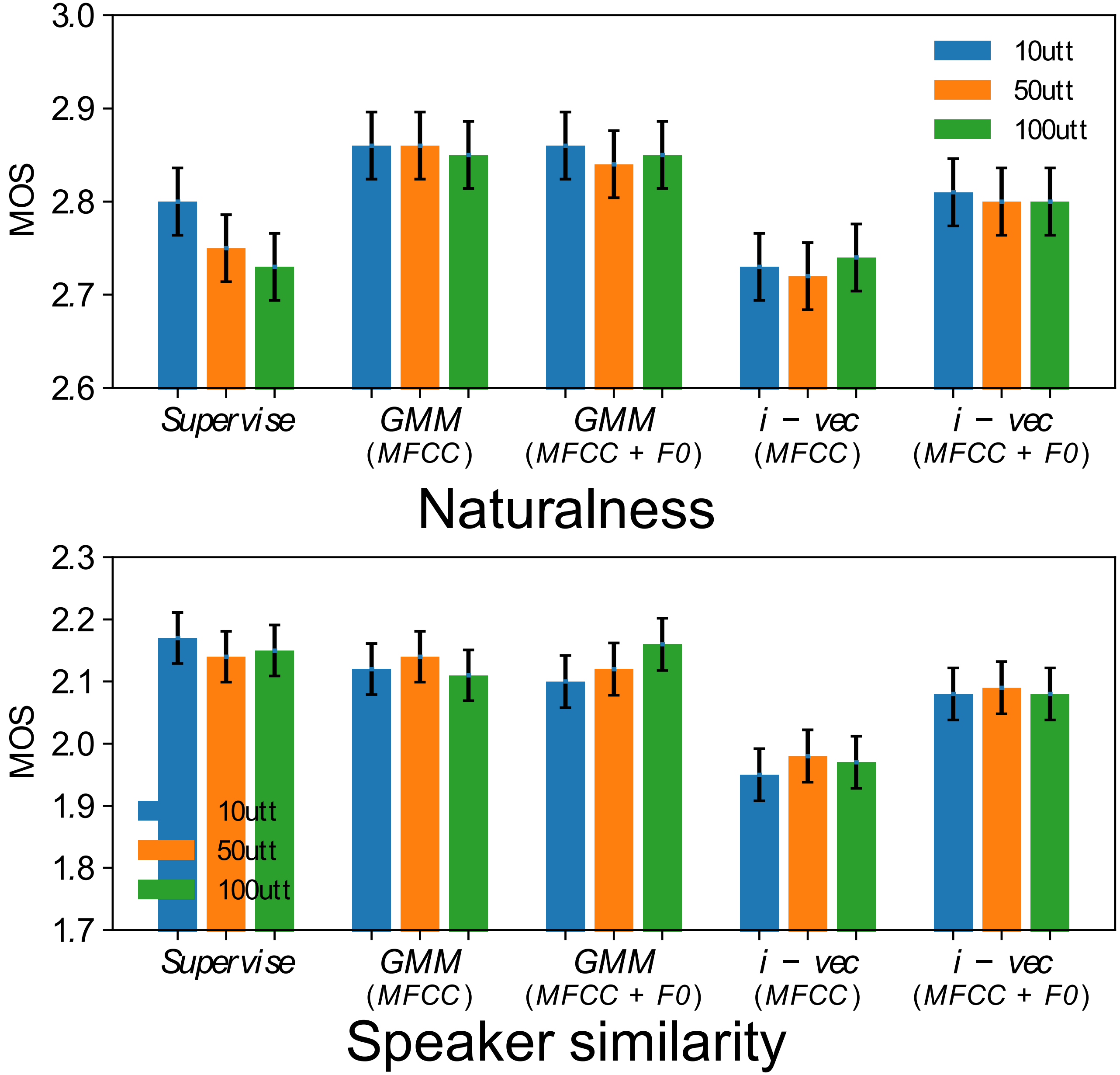


Supervised < Unsupervised (g) < Averaged

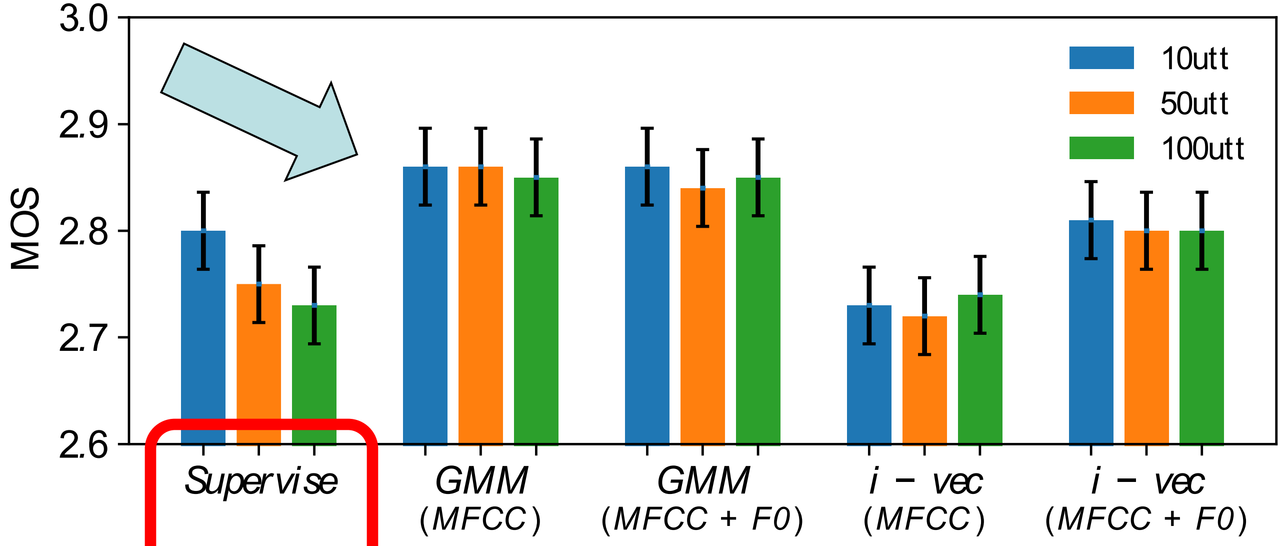
– The proposed technique successfully performed speaker adaptation

– As expected, the results of supervised systems are better

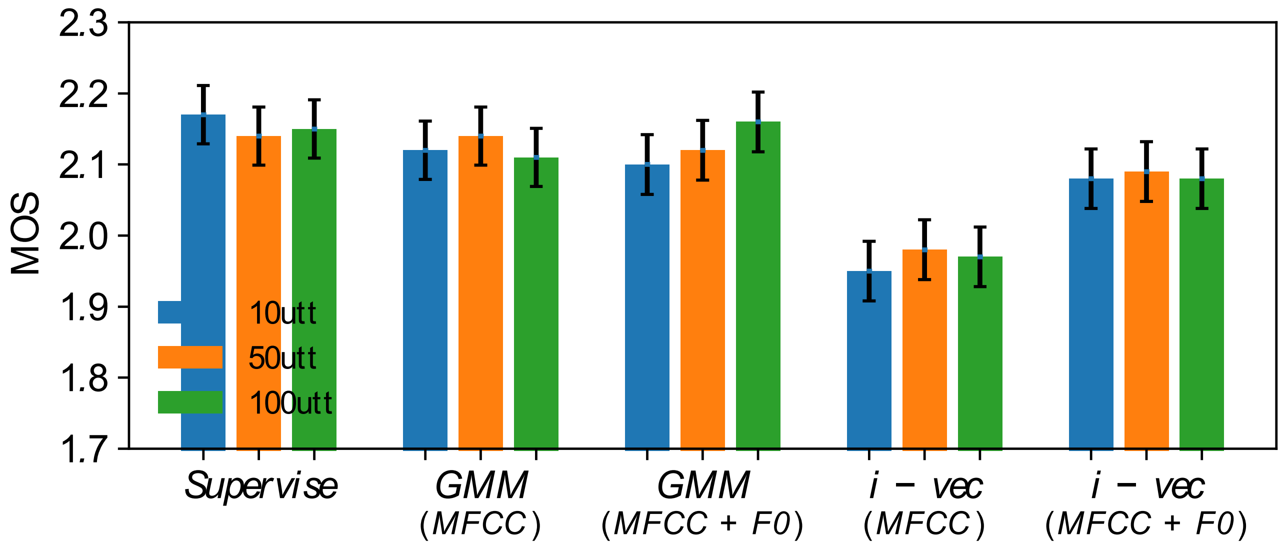
Subjective evaluation results



Subjective evaluation results

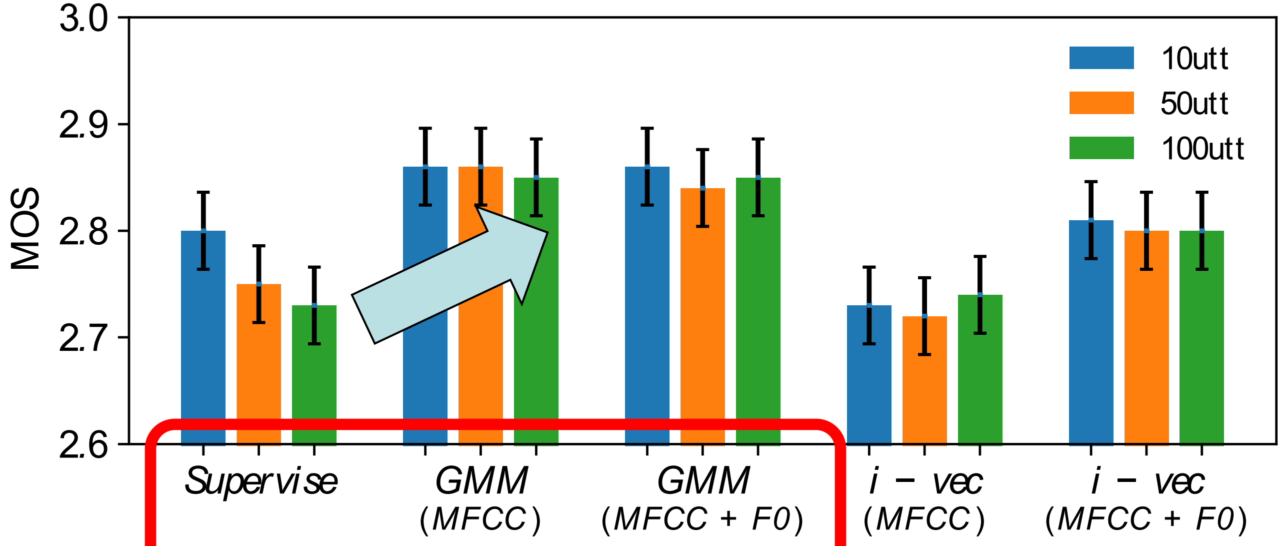


Naturalness

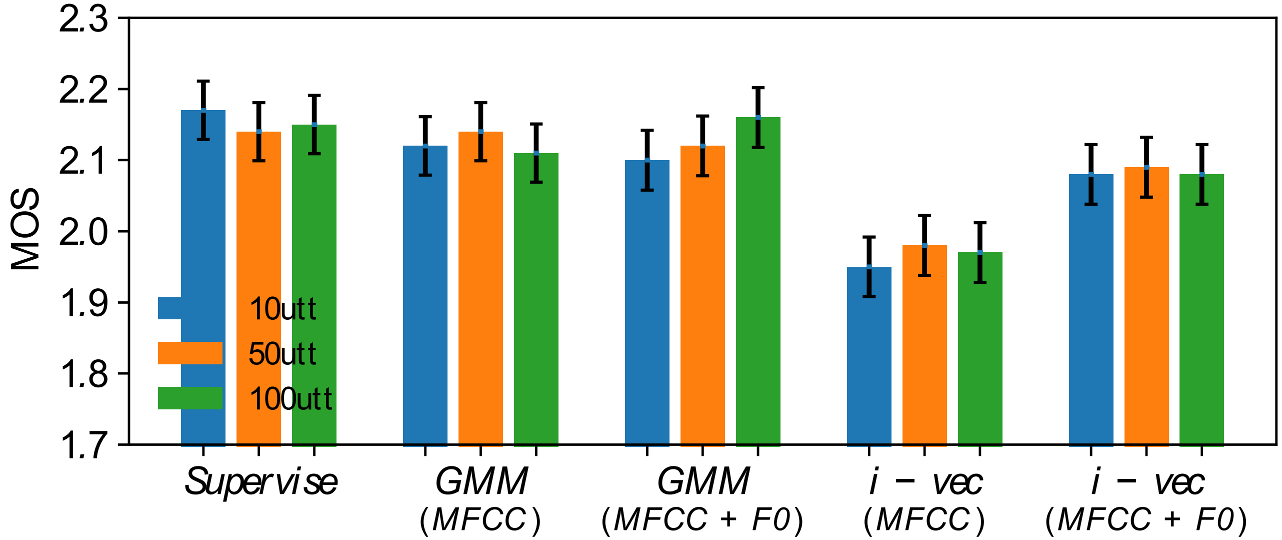


Speaker similarity

Subjective evaluation results

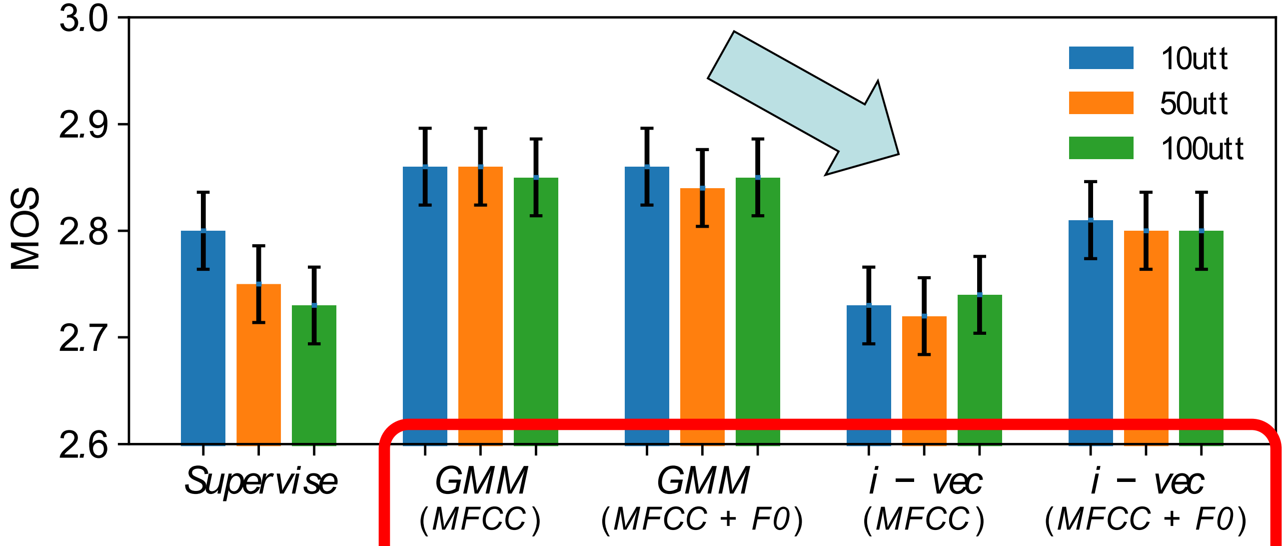


Naturalness

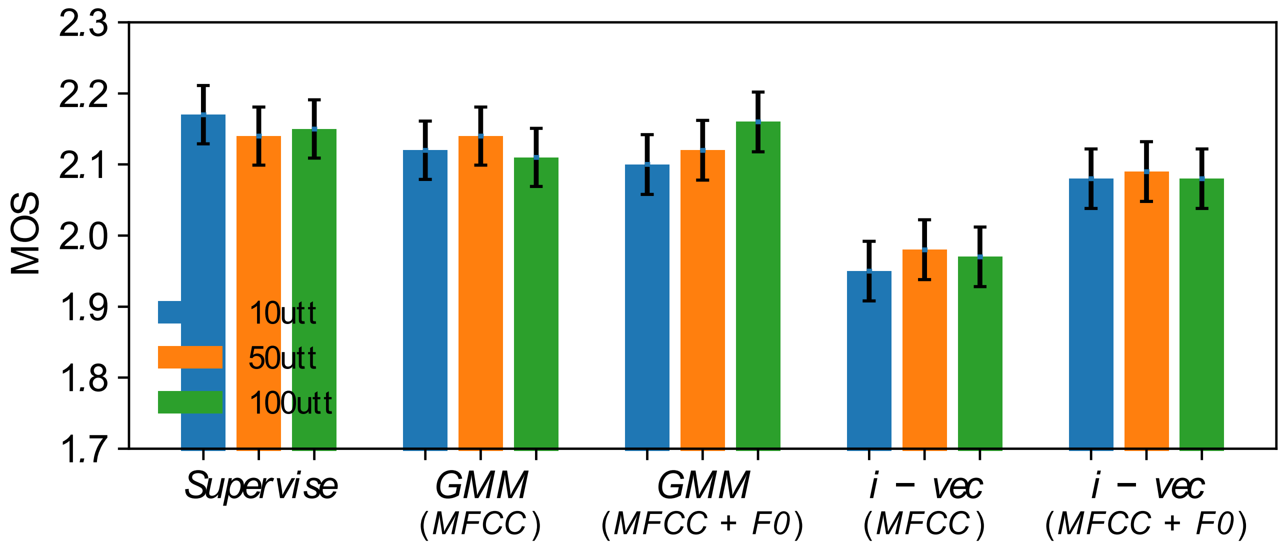


Speaker similarity

Subjective evaluation results

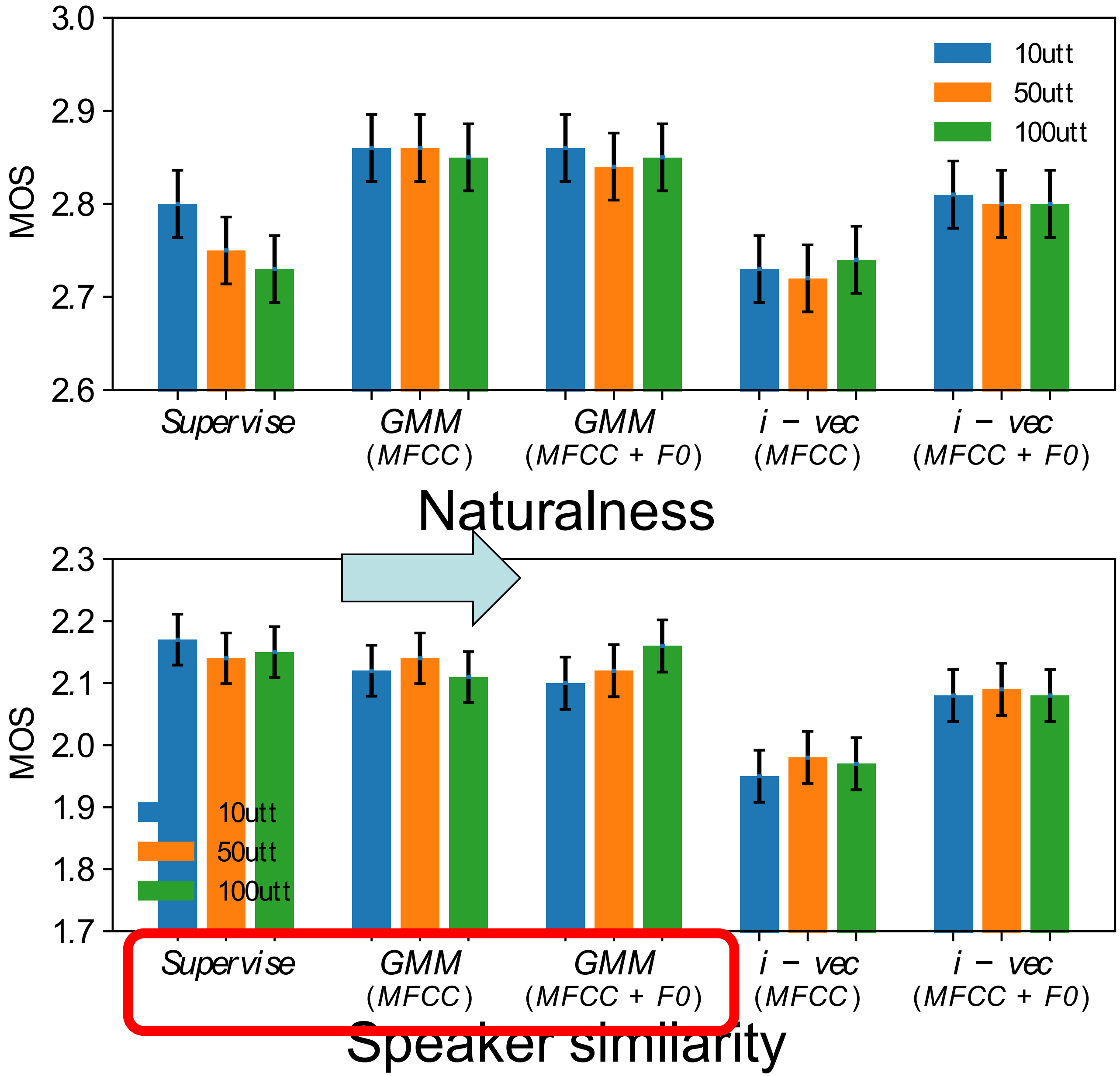


Naturalness



Speaker similarity

Subjective evaluation results



Experiments using low-quality speech data

SNR

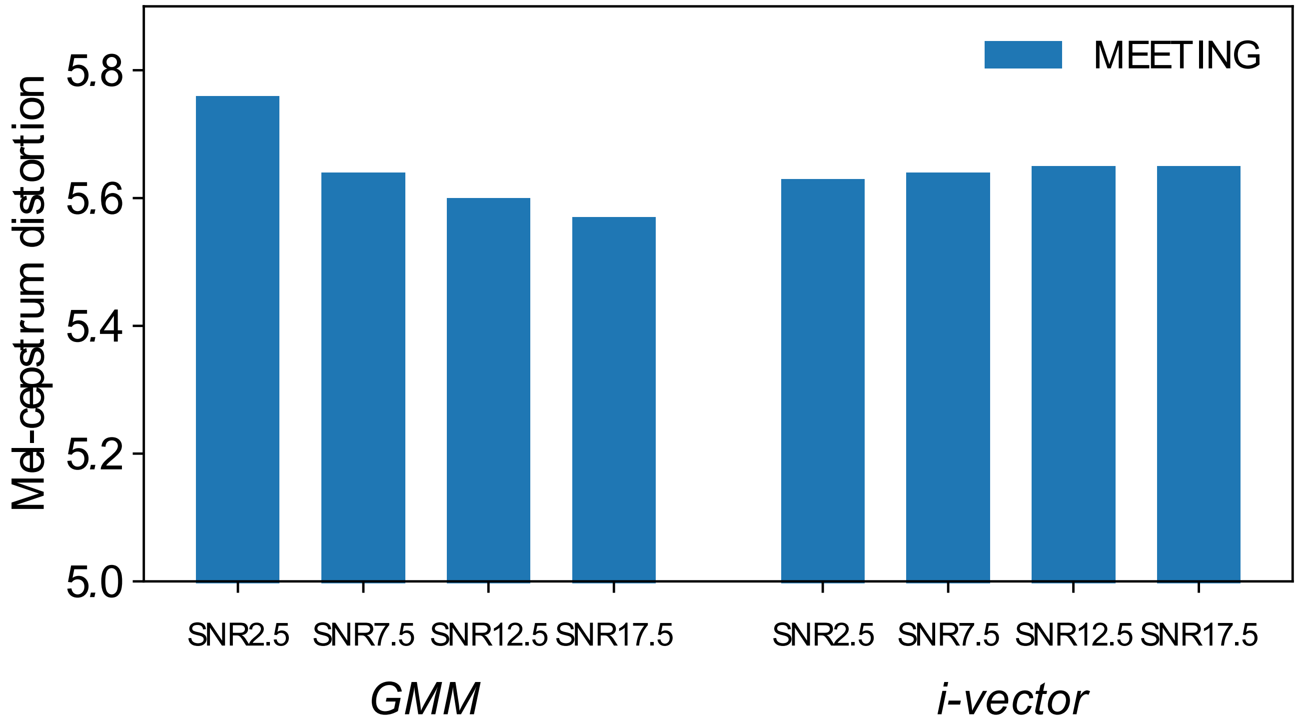
- Training data : 2.5-, 7.5-, 12.5-, or 17.5-dB
- Adaptation data : 0.0-, 5.0-, 10.0-, or 15.0-dB

Quality types of training and adaptation data

Training data	Adaptation data	Quality condition
CLEAN	CLEAN	ideal
CLEAN	OFFICE	mismatched
CLEAN	MEETING	mismatched
OFFICE	OFFICE	matched
MEETING	MEETING	matched

Objective evaluation（1/2）

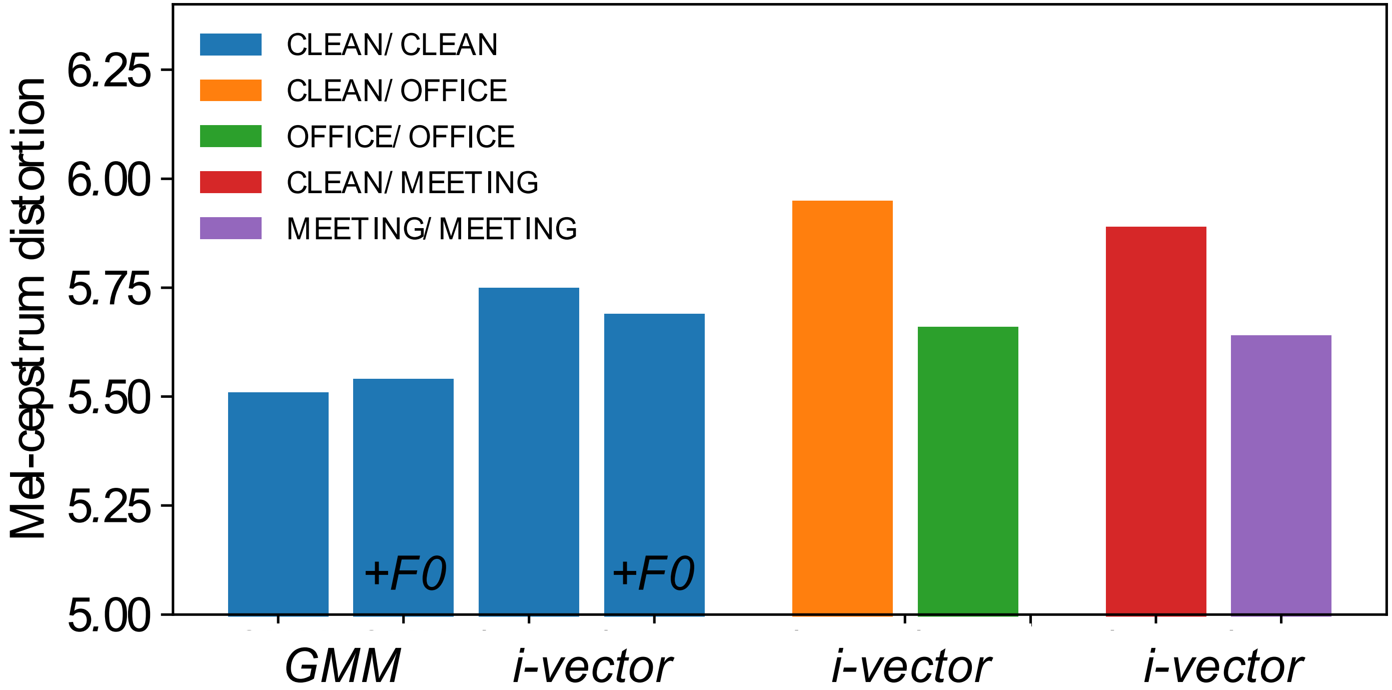
Matched condition



– The performance of i-vector/PLDA was almost the same in all SNR cases
  - i-vector/PLDA was robust for the proposed adaptation

Objective evaluation（2/2）

Labels show conditions（Training/Adaptation）

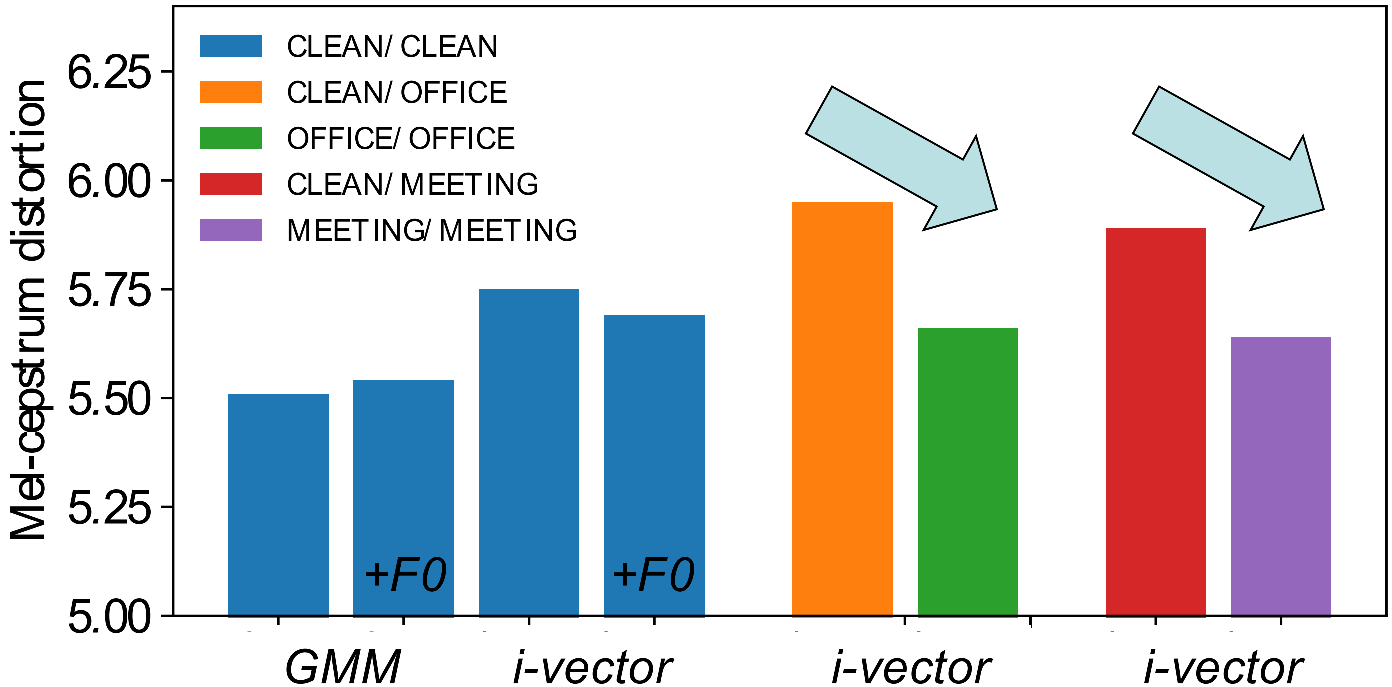


– Ideal < matched < mismatched

- Using speech data whose quality is matched to adaptation data for ASV model training improves the performance

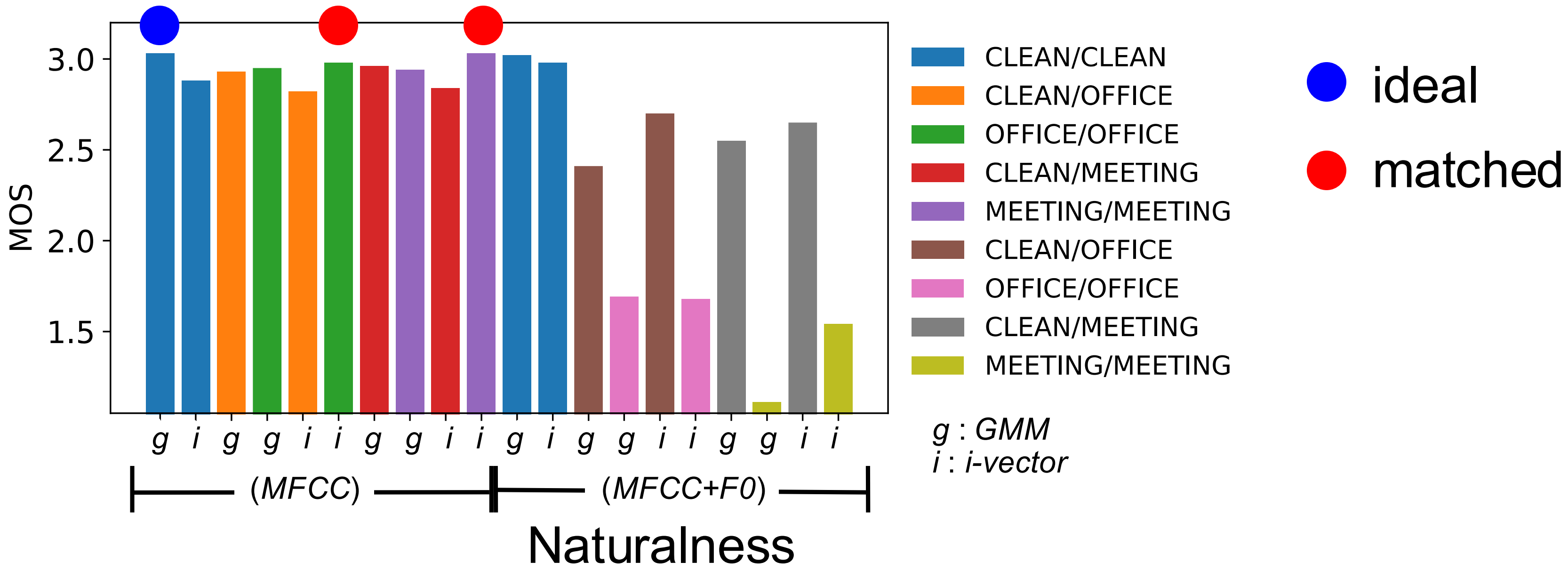
Objective evaluation (2/2)

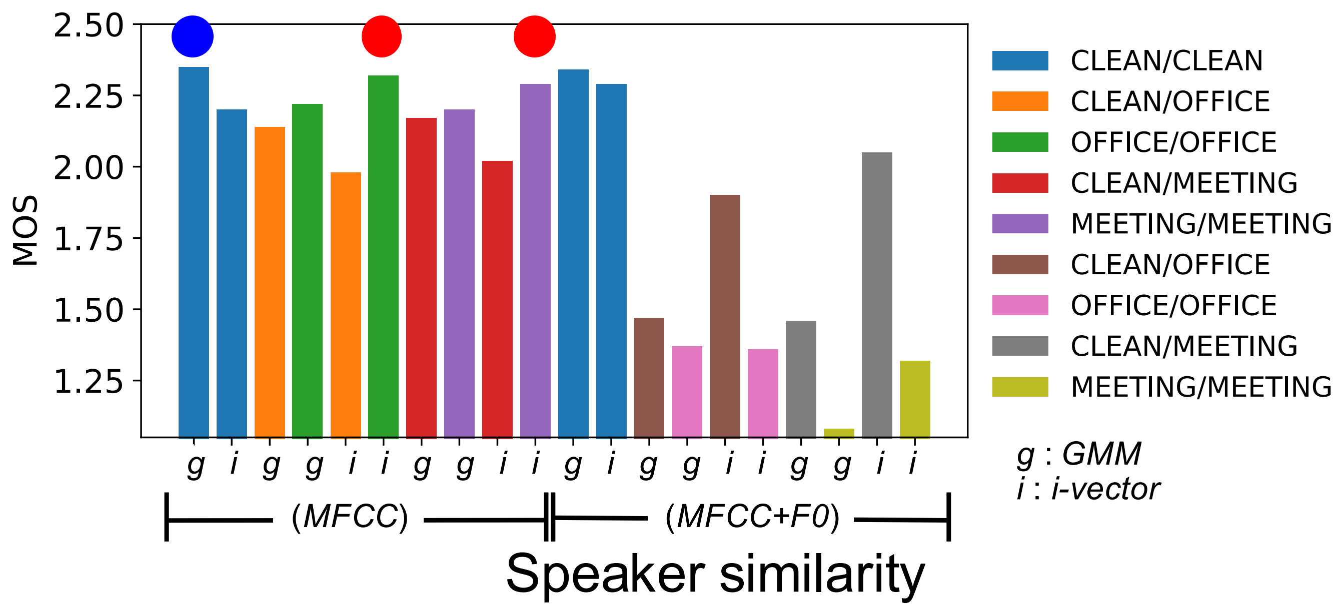
Labels show conditions (Training/Adaptation)



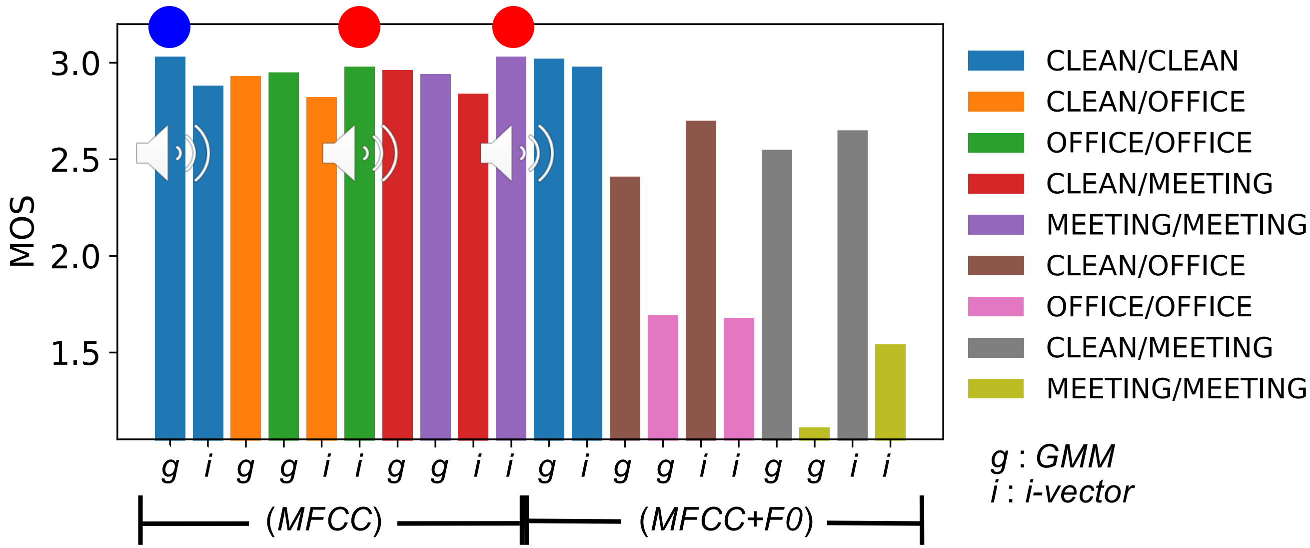
- Ideal < matched < mismatched
  - Using speech data whose quality is matched to adaptation data for ASV model training improves the performance

Subjective evaluation

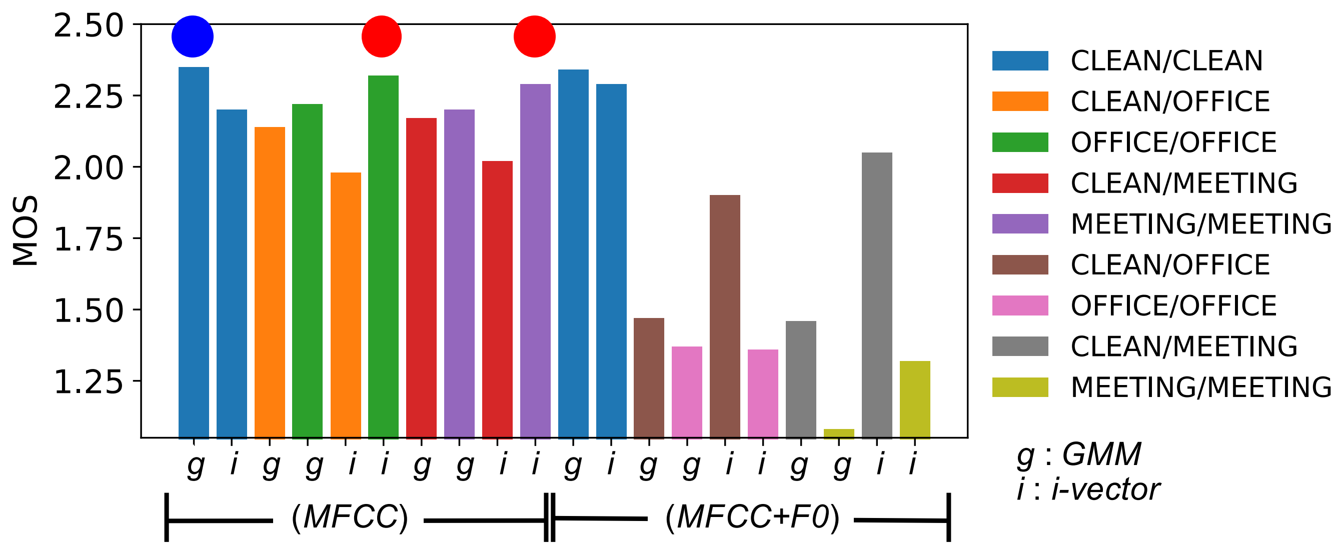




Subjective evaluation



Naturalness



Speaker similarity

ideal

matched

Target

Conclusion

Unsupervised adaptation for speech synthesis

- Adaptation using a speaker-similarity vector
  - The proposed technique change the speaker characteristics
- Robustness against low-quality adaptation data
  - Using speech data whose quality is matched to adaptation data for model training improved the performance

Future work

- MP3 or AMR codec speech
- Speech recorded under real conditions