



<http://reverb2014.dereverberation.com/>

Summary of the REVERB challenge

Keisuke Kinoshita,
Marc Delcroix,
Takuya Yoshioka,
Tomohiro Nakatani
NTT Corporation

Emanuel Habëts
International AudioLabs Erlangen

Reinhold Haeb-Umbach,
Volker Leutnant
Paderborn Univ.

Armin Sehr
*Beuth Univ. of
Applied Sciences Berlin*

Walter Kellermann,
Roland Maas
Univ. of Erlangen-Nuremberg

Sharon Gannot
Bar-Ilan Univ.

Bhiksha Raj
Carnegie Mellon Univ.

Outline

- Motivation and design of the REVERB challenge
- Summary of the participants' systems
- Result summary
 - The ASR results
 - The SE (Speech Enhancement) results
- Concluding remarks

Motivation

- ☺ Recently, **substantial progress** made in the field of reverberant speech signal processing, including
 - Single- and multi-channel de-reverberation techniques
 - ASR techniques robust to reverberation
- ☹ **Lack of common evaluation framework**

➡ **REVERB challenge to provide a common evaluation framework for both ASR and SE studies**

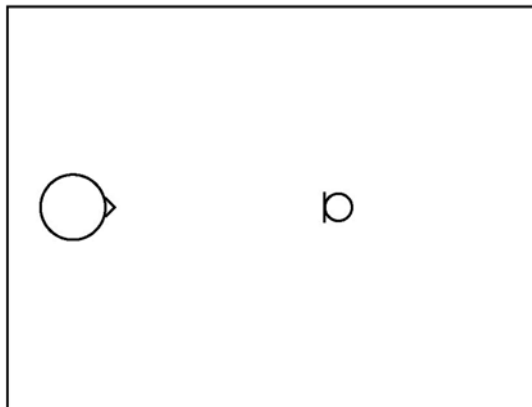
Target acoustic scenarios

- Reverberant
- Moderate stationary noise (\sim SNR* 20dB)
- 1ch, 2ch and 8ch scenarios

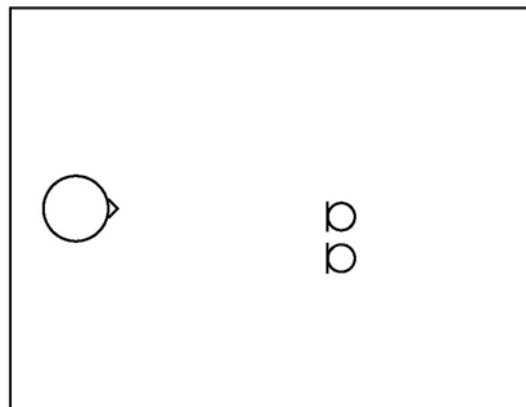


Fig: One of microphone arrays used

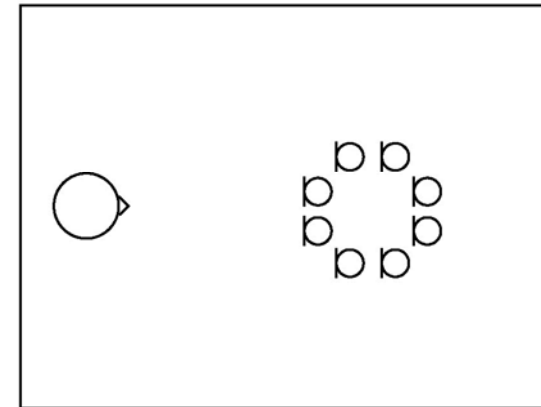
1ch scenario



2ch scenario



8ch circular-array scenario



* "S" includes direct signal and early reflections up to 50ms.

The challenge data (1/2)

- Based on Wall Street Journal Cambridge (WSJCAM0) 5K task
- **Real recordings (RealData)** ^{*1} & **simulated data (SimData)** ^{*2}
(Development and evaluation sets provided)
 - RealData for validity assessment in real reverb conditions
 - SimData for experiments in various reverb conditions
(A part of SimData simulates RealData in terms of the reverb time)
 - Text prompts used for both data were the same.
- Clean and multi-condition (simulated) training data provided

^{*1} RealData is available from the LDC catalog as a part of MC-WSJ-AV corpus (since April 2014).

^{*2} Materials required to generate SimData is available on our webpage. The data will soon be available through the LDC catalog. <http://catalog.ldc.upenn.edu/LDC2014S03>









The challenge data (2/2)

- Acoustic conditions for SimData and RealData

	Reverb time (T_{60})	Distance between speaker and mic
SimData	0.25s , 0.5s, 0.7s* (Room1, 2, 3)	near: 0.5m far: 2.0m
RealData	0.7s*	near: ~1.0m far: >2.5m

* *SimData room3 simulates RealData*

- Sound examples

	RealData (far)		SimData (Room2, far)	
	Male	Female	Male	Female
Clean/Headset				
Observed				

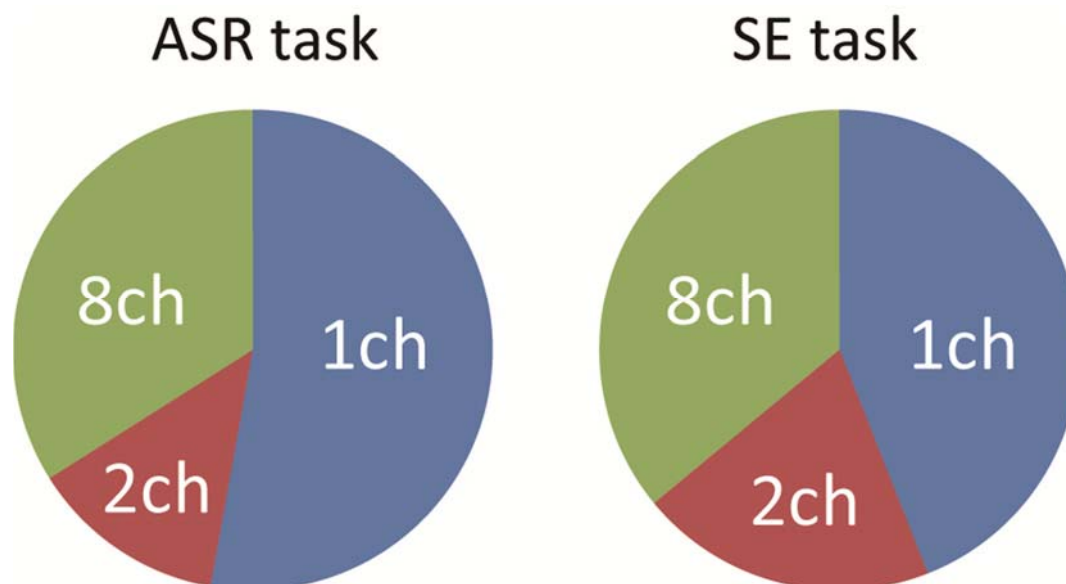
The challenge tasks: ASR and SE

- ASR task
 - Evaluation criterion: Word Error Rate (WER)
- SE task
 - Objective evaluation criteria
 - Intrusive measure (that requires reference clean speech)
 - Cepstrum distance (CD)
 - Freq-weighted segmental SNR (FWsegSNR)
 - Log likelihood ratio (LLR)
 - PESQ (optional)
 - Non-intrusive measure
 - Speech-to-reverb modulation ratio (SRMR)
 - Subjective evaluation criteria (web-based MUSHRA test)
 - Perceived amount of reverberation
 - Overall quality (i.e., artifacts, distortions, remaining reverb and etc)
- Same test & training data provided for both tasks

Number of submissions

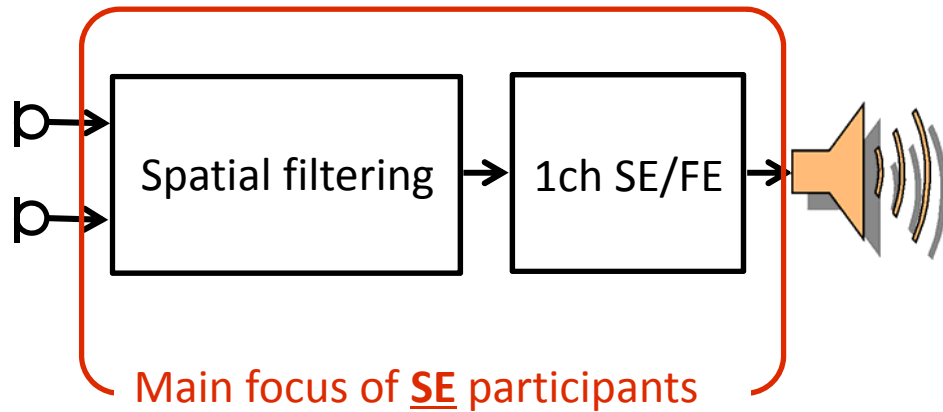
- 27 participants (i.e., # of papers)
 - 18 submissions (incl. 49 systems) to the ASR task
 - 14 submissions (incl. 25 systems) to the SE task

- Percentage of 1ch, 2ch and 8ch systems in each task -

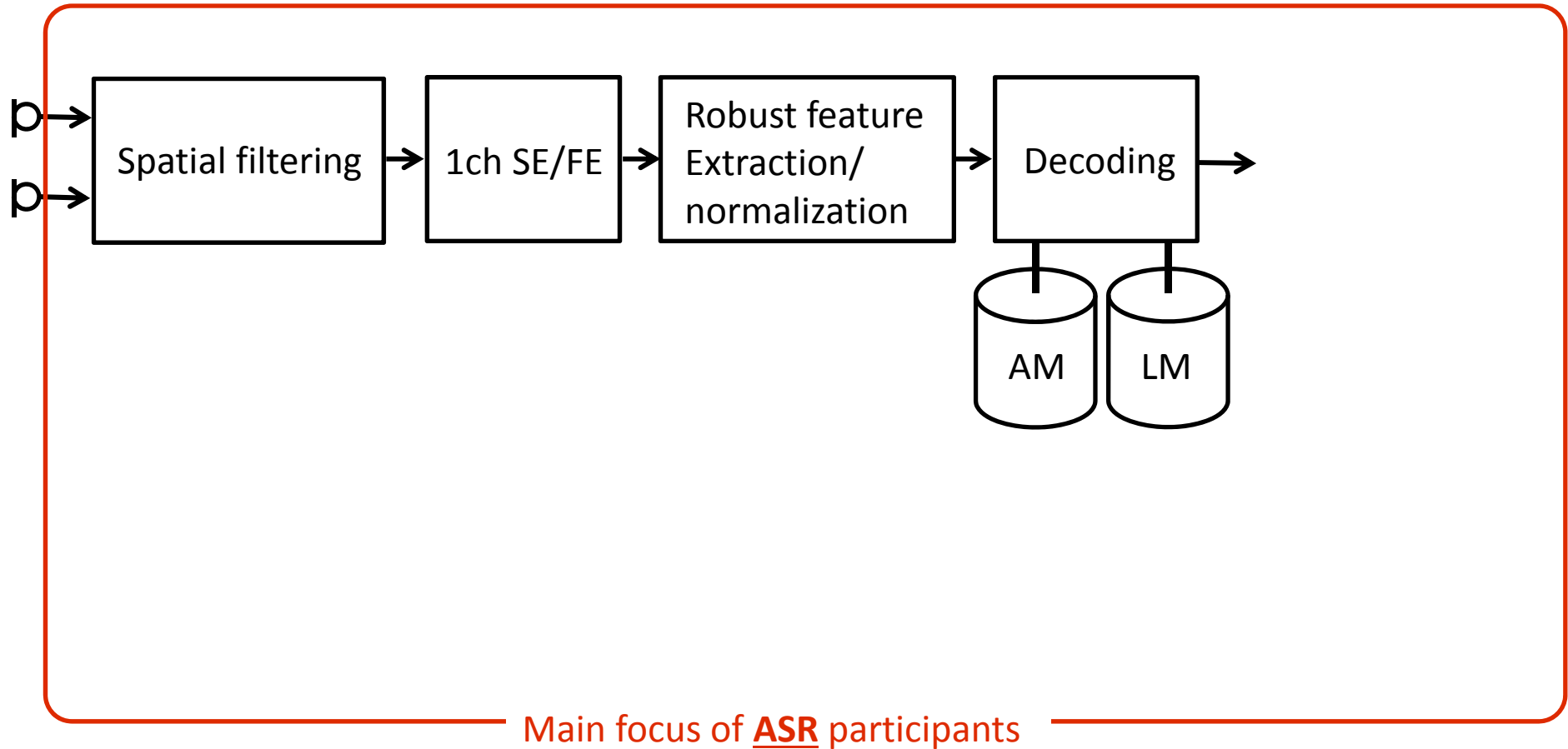


Quick introduction to the submitted participants' systems

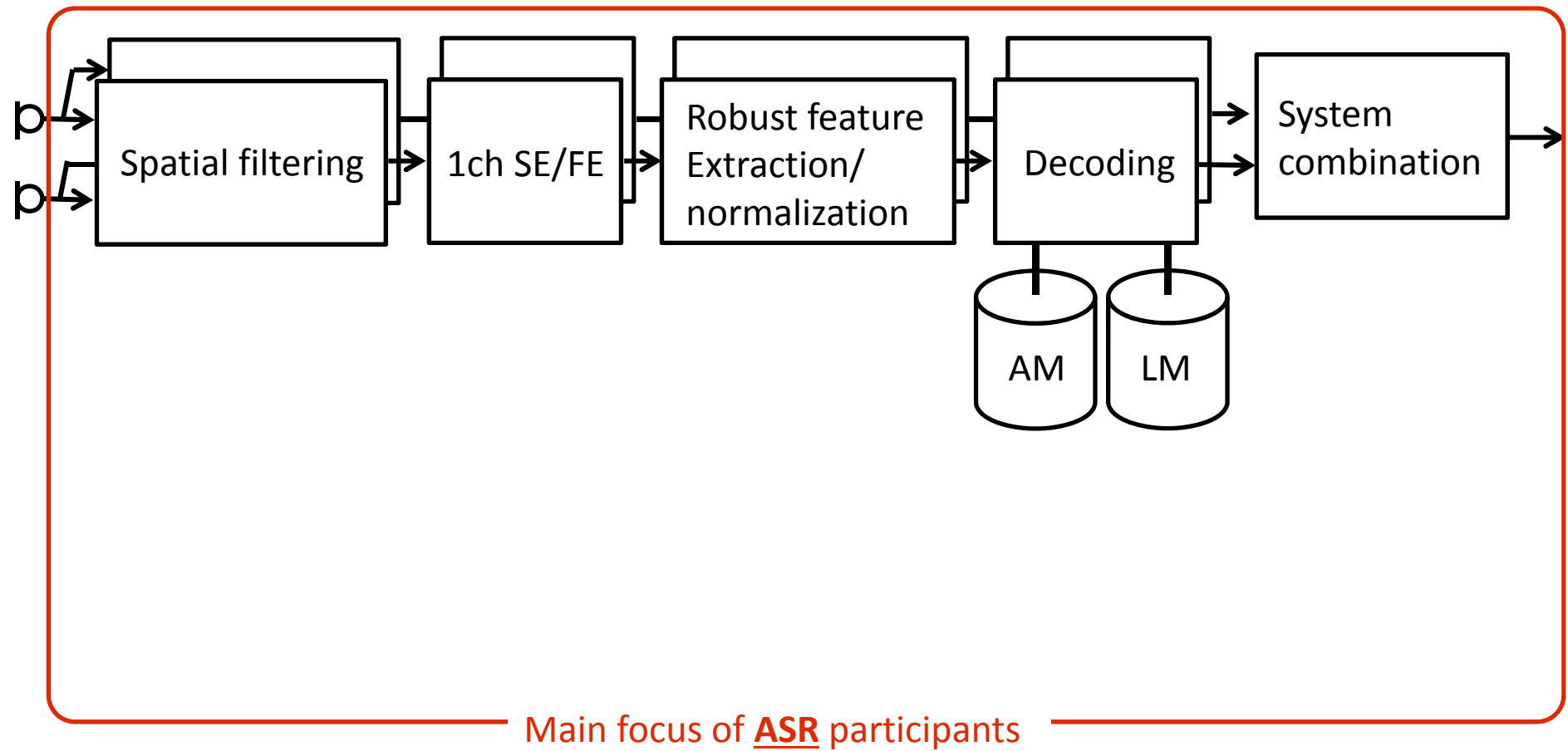
A wide variety of approaches submitted



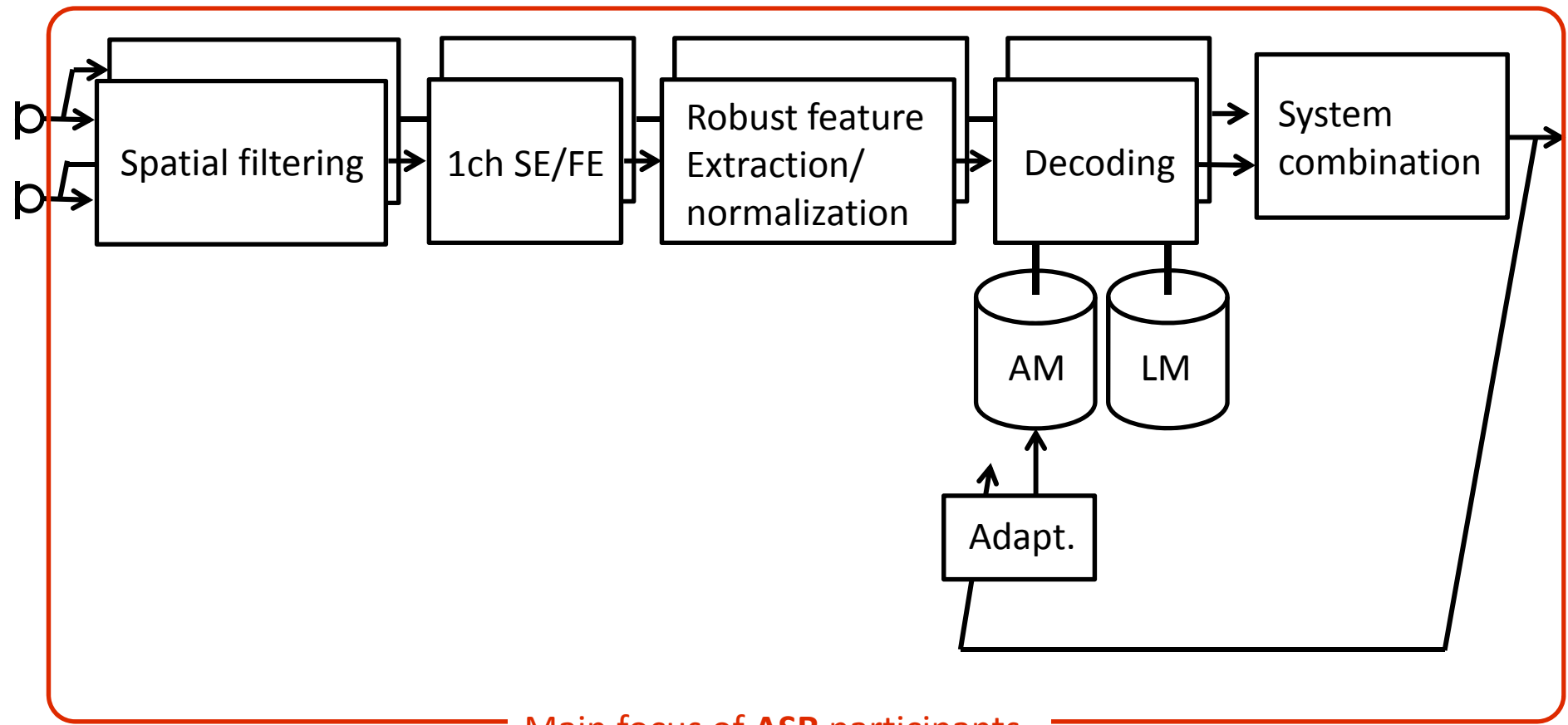
A wide variety of approaches submitted



A wide variety of approaches submitted



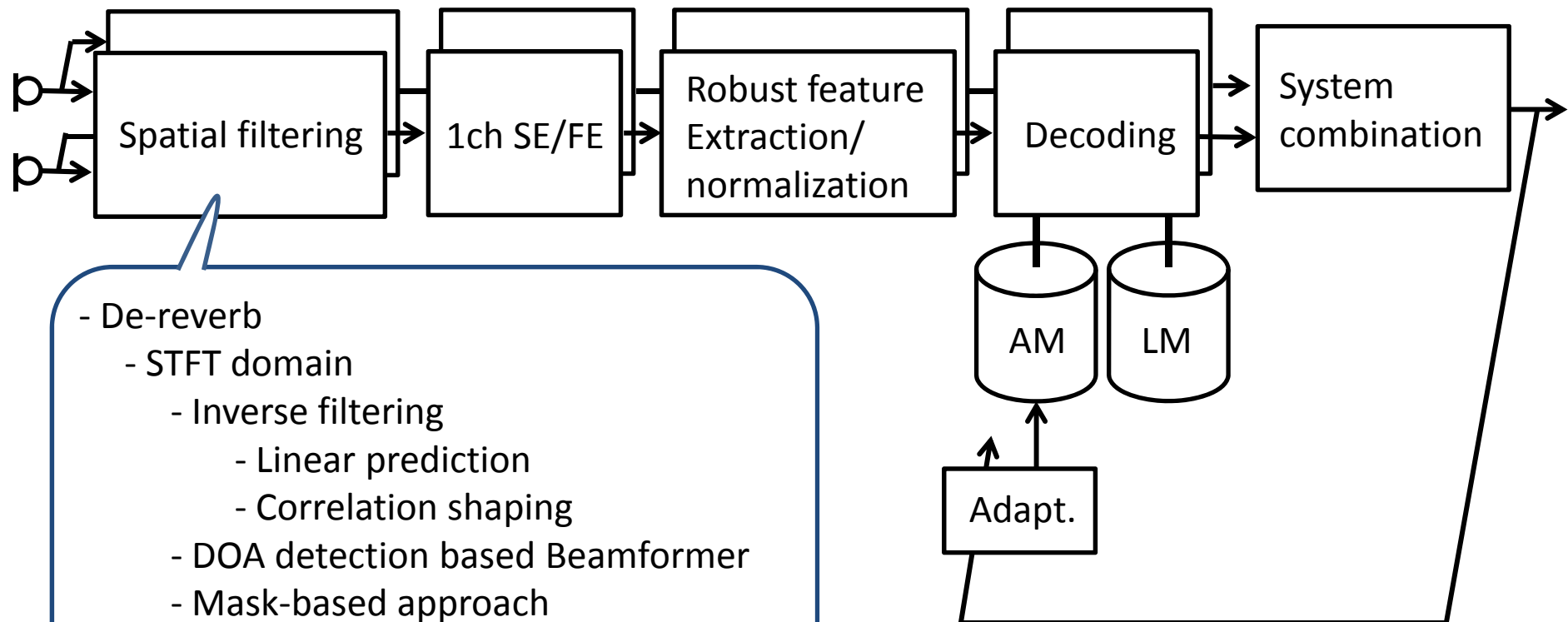
A wide variety of approaches submitted



Main focus of ASR participants

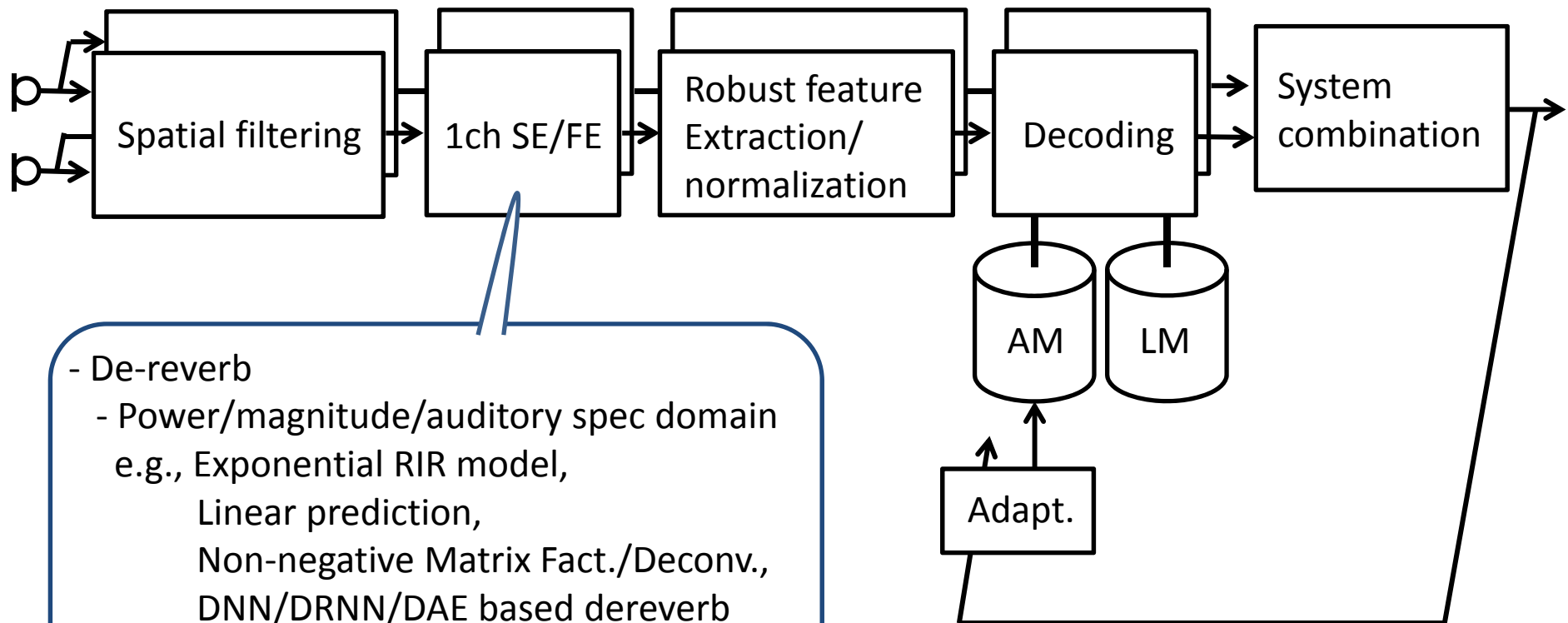
Submission ranges from 1ch/multi-channel SE algorithms to the ASR back-end algorithms.

Various approaches (1/4)



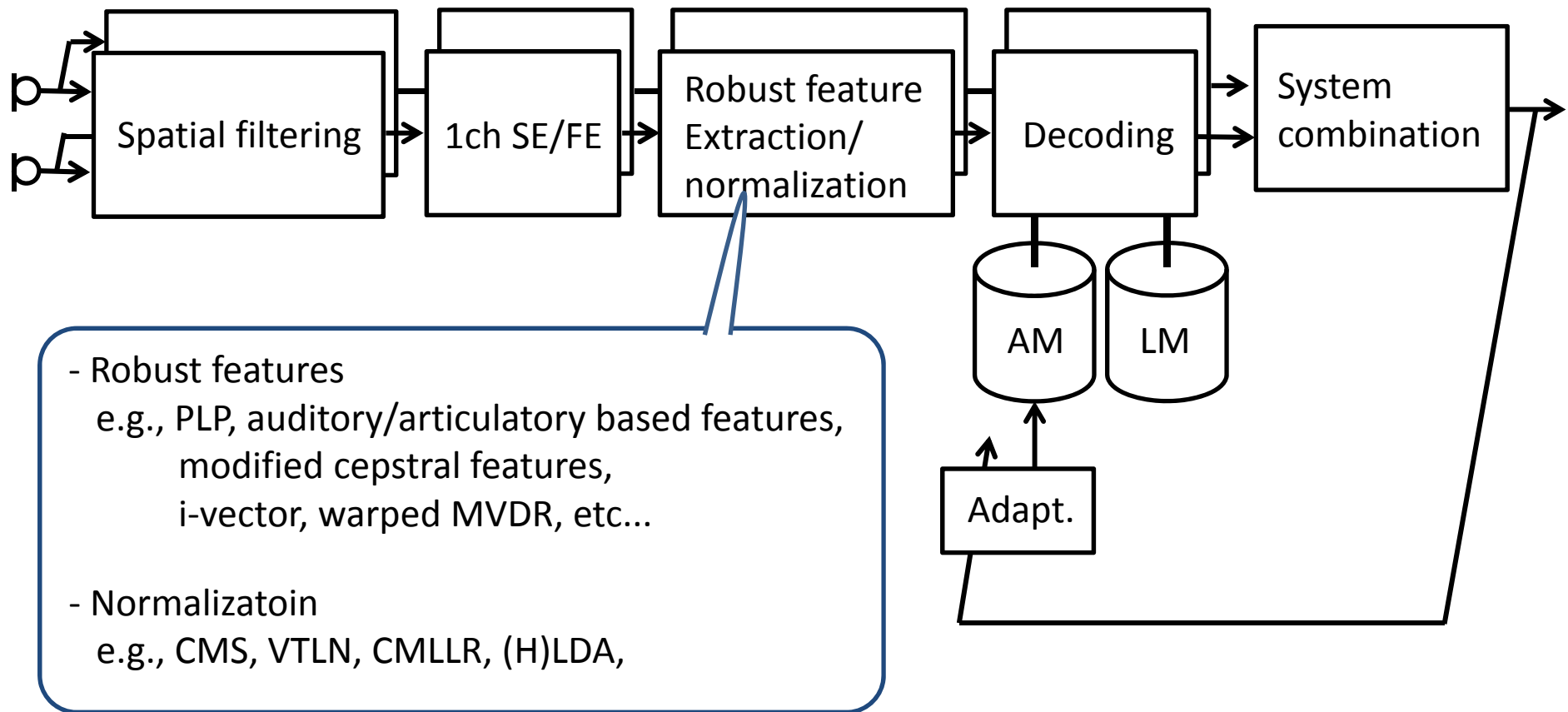
- De-reverb
 - STFT domain
 - Inverse filtering
 - Linear prediction
 - Correlation shaping
 - DOA detection based Beamformer
 - Mask-based approach
 - Phase-error filter
 - Magnitude spec domain
 - Estimation of nonnegative RIRs
- De-noising (STFT, auditory-feature domain)
e.g., MVDR, delay-sum, GSC, Mch-WF.

Various approaches (2/4)

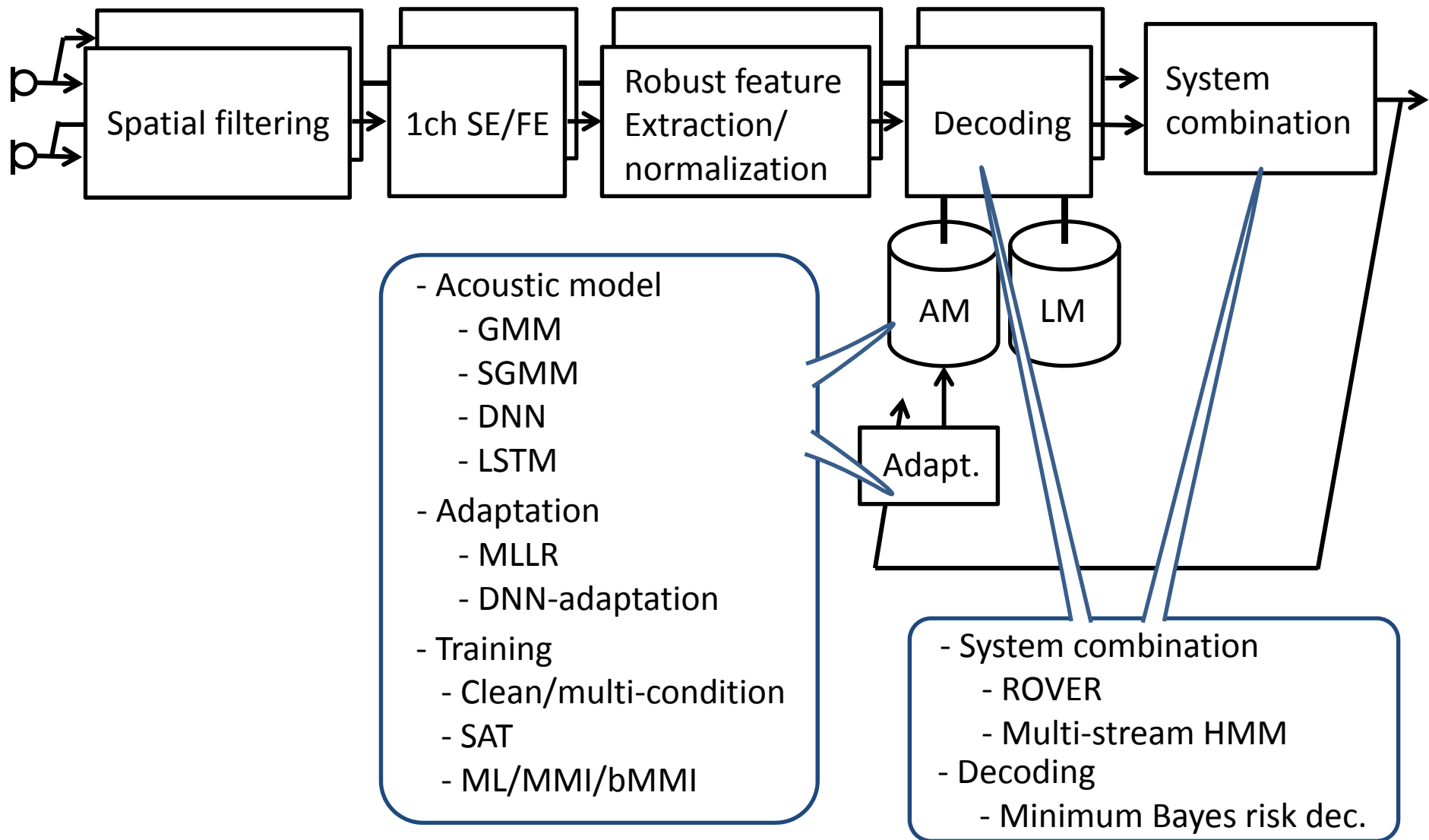


- De-reverb
 - Power/magnitude/auditory spec domain
e.g., Exponential RIR model,
Linear prediction,
Non-negative Matrix Fact./Deconv.,
DNN/DRNN/DAE based dereverb
 - Cepstral domain
e.g., Cepstral smoothing,
ML-based inverse filter estimation
- De-noising
e.g., SS, MMSE-STSA.

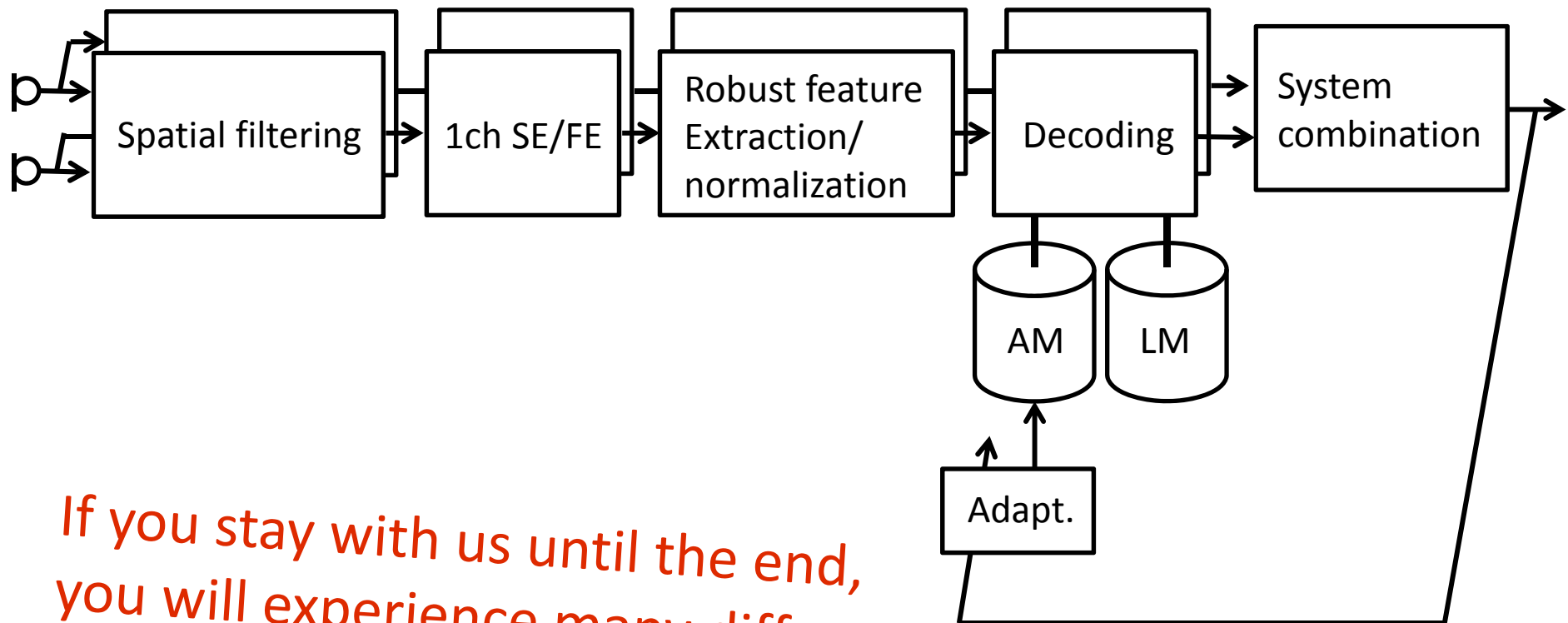
Various approaches (3/4)



Various approaches (4/4)



Various approaches (4/4)



*If you stay with us until the end,
you will experience many different
approaches tackling the same data*

Now, the results... 😊

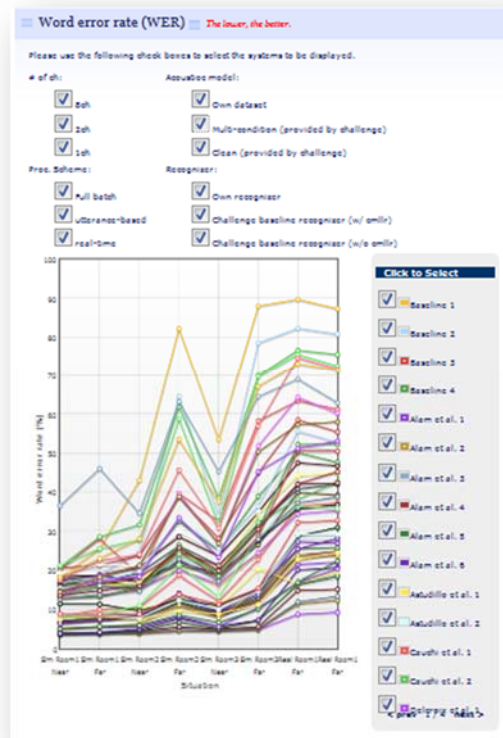
Results already publicly available

- Results for the ASR task

http://reverb2014.dereverberation.com/result_asr.html

- Results for the SE task

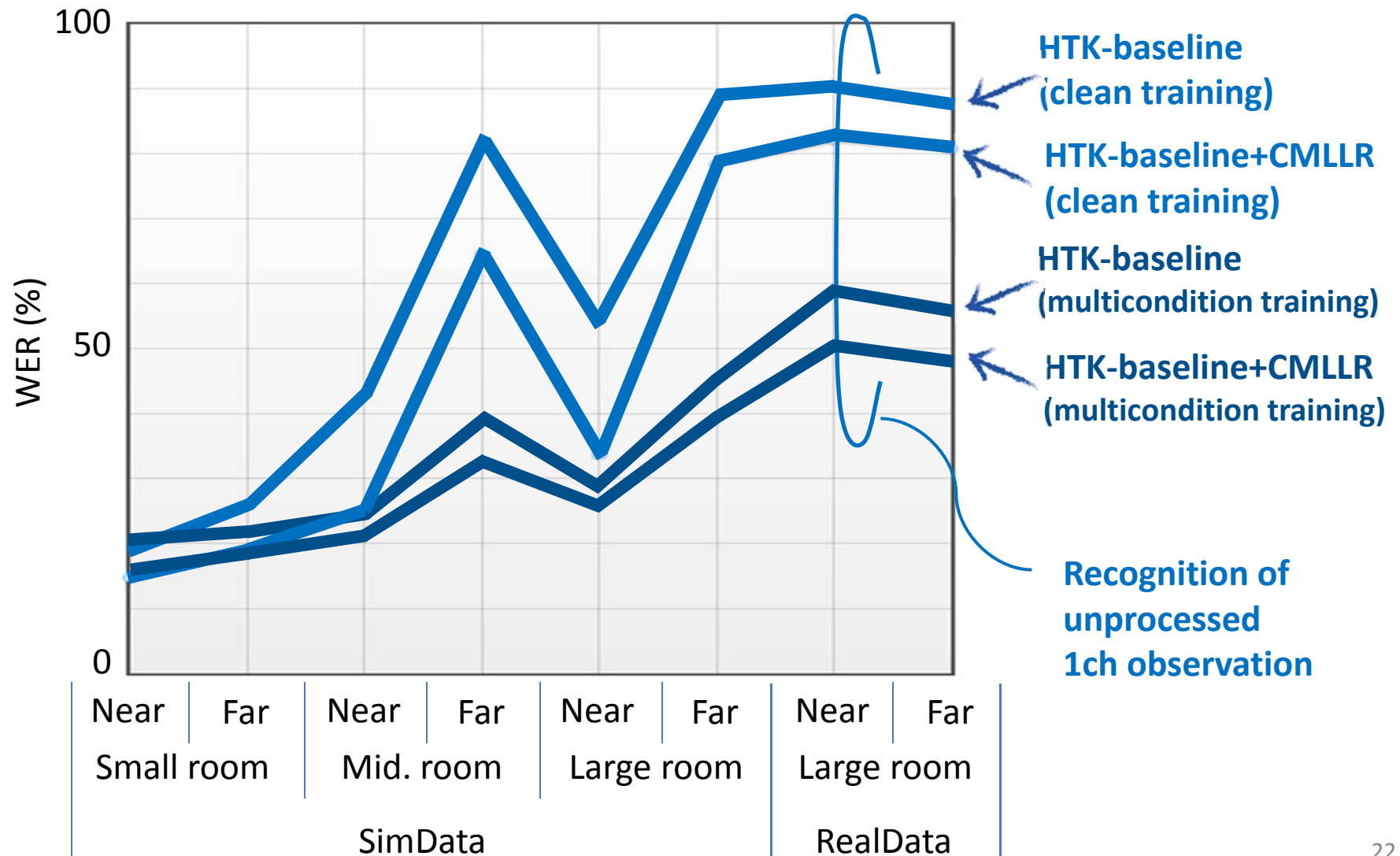
http://reverb2014.dereverberation.com/result_se.html



*Note:
More results (detailed/new/updated
results) are available in participants'
papers.*

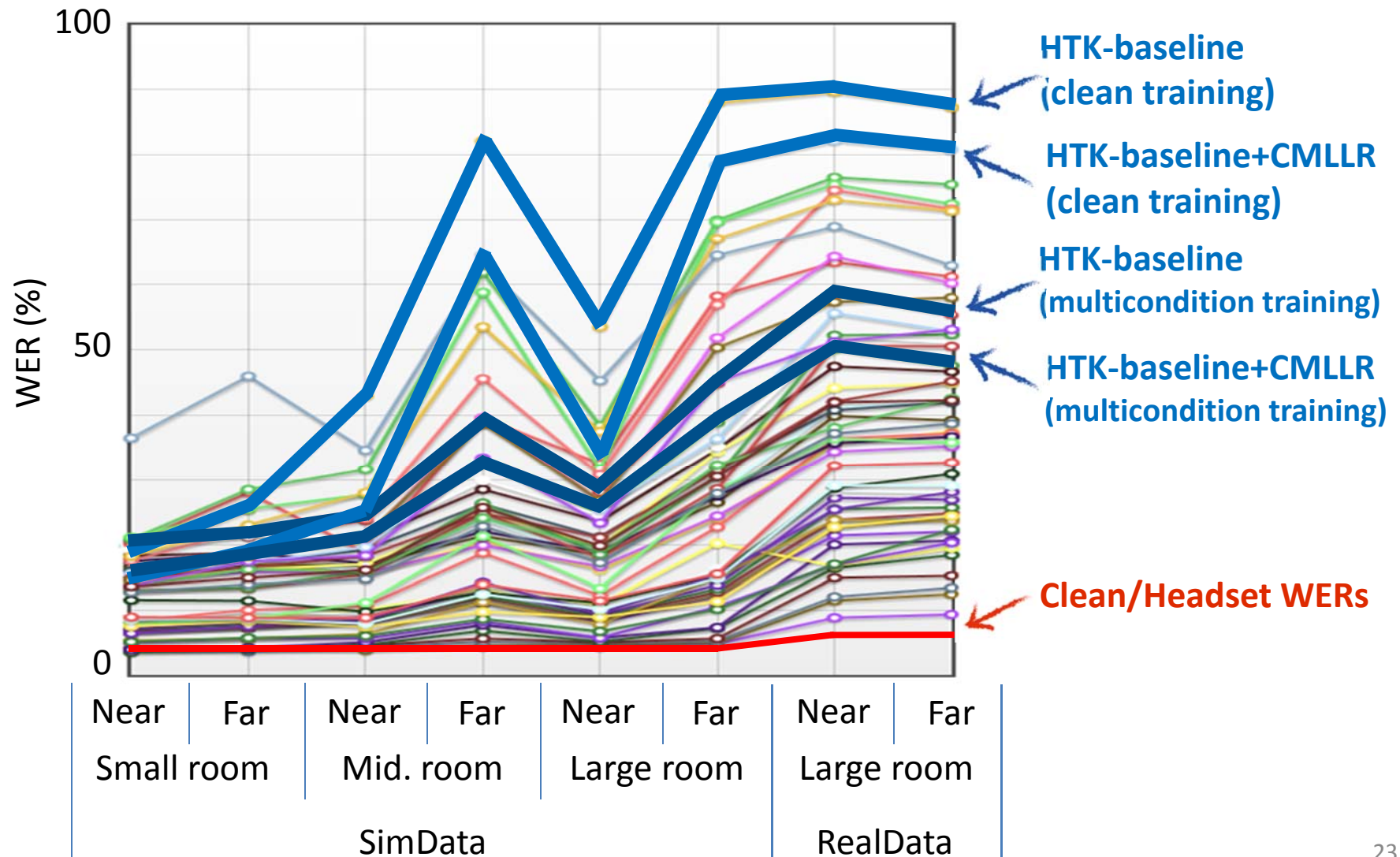
Let's start with the ASR results... 😊

ASR results: baselines



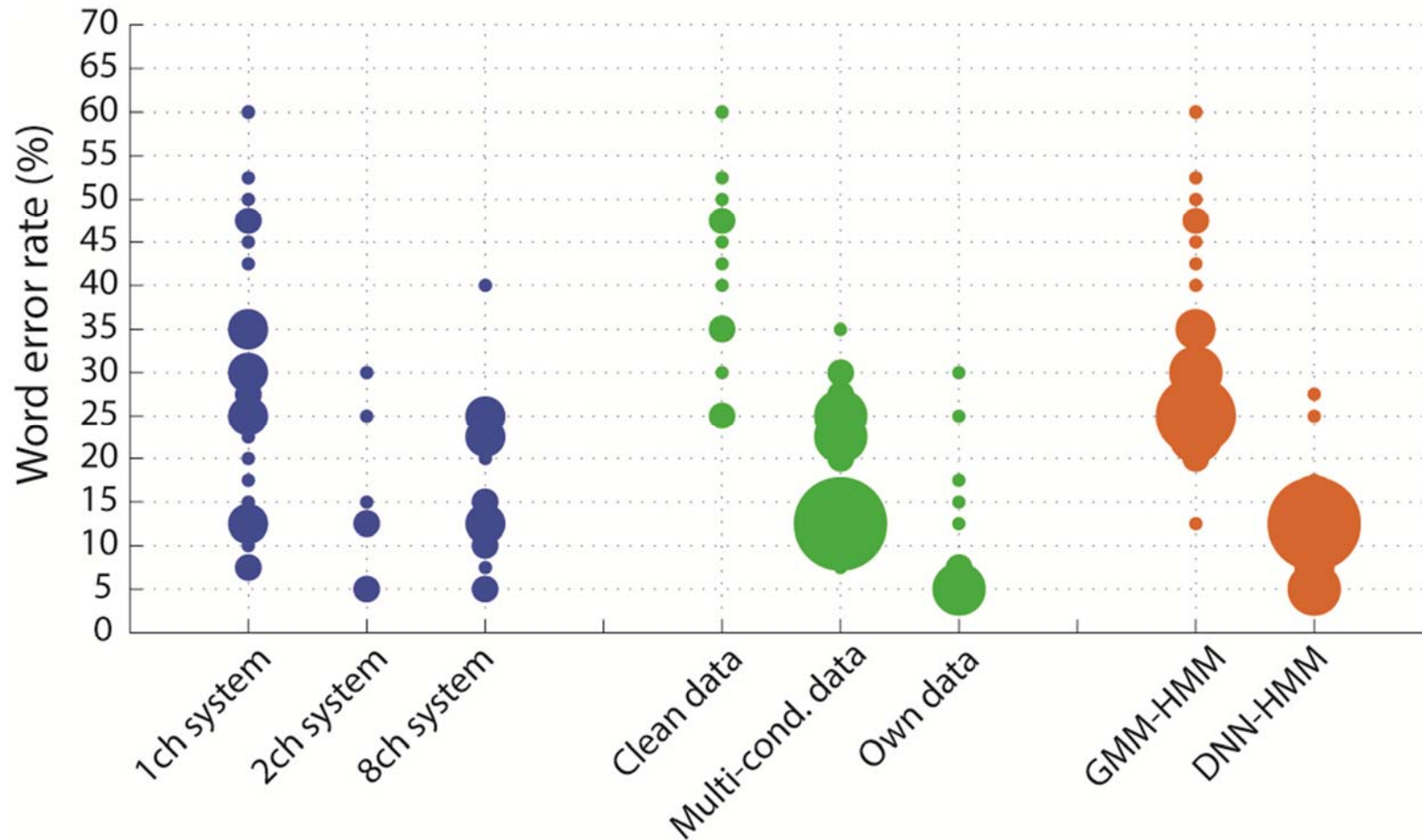
ASR results: at a glance

- All the submitted WERs (everything mixed, not a fair comparison)



ASR results analysis with bubble chart

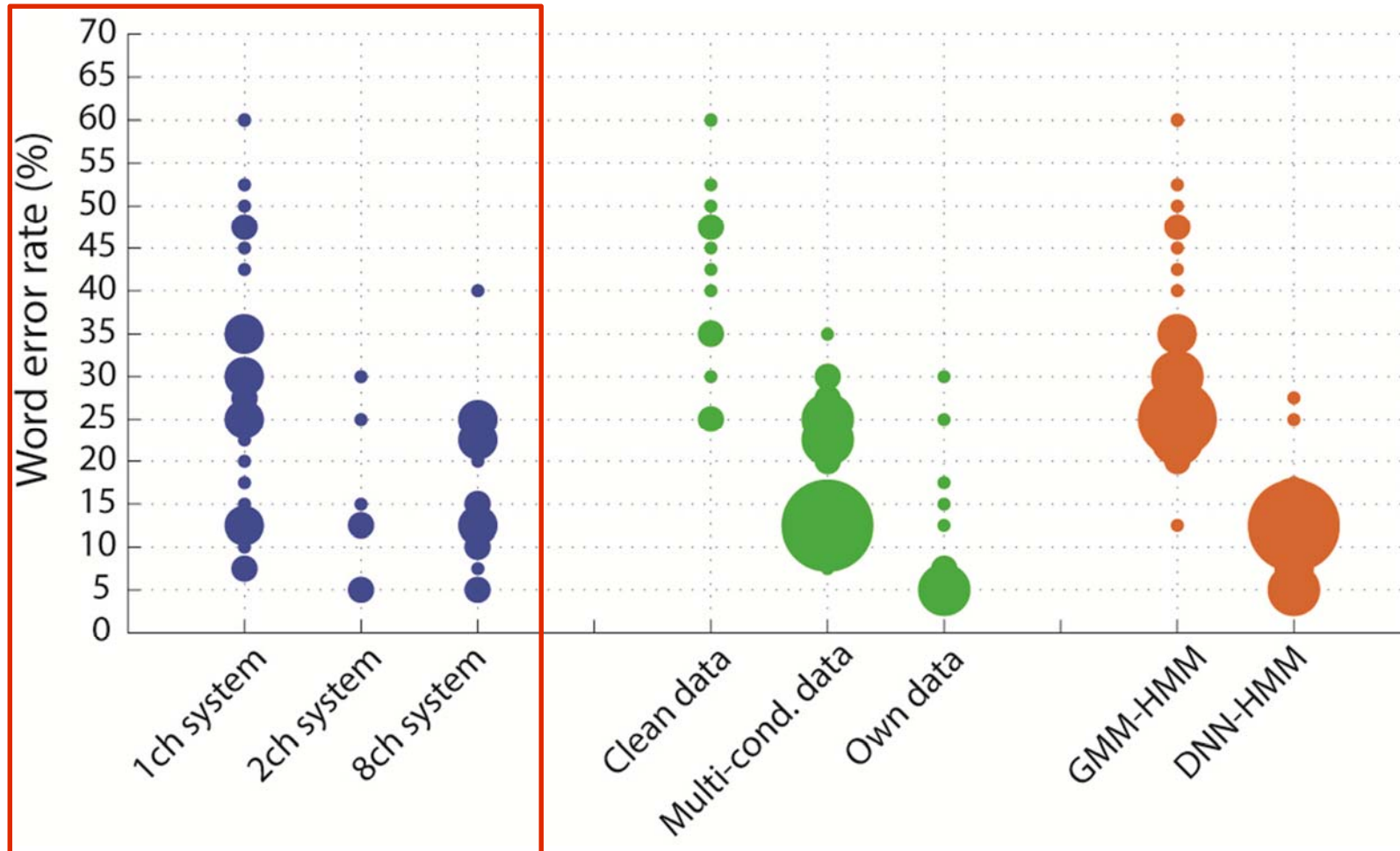
- Relationship between (averaged) WER and # of mic., data and acoust. model



The size of a circle indicates the # of systems in the corresponding category

ASR results analysis with bubble chart

Results per 1ch, 2ch and 8ch systems

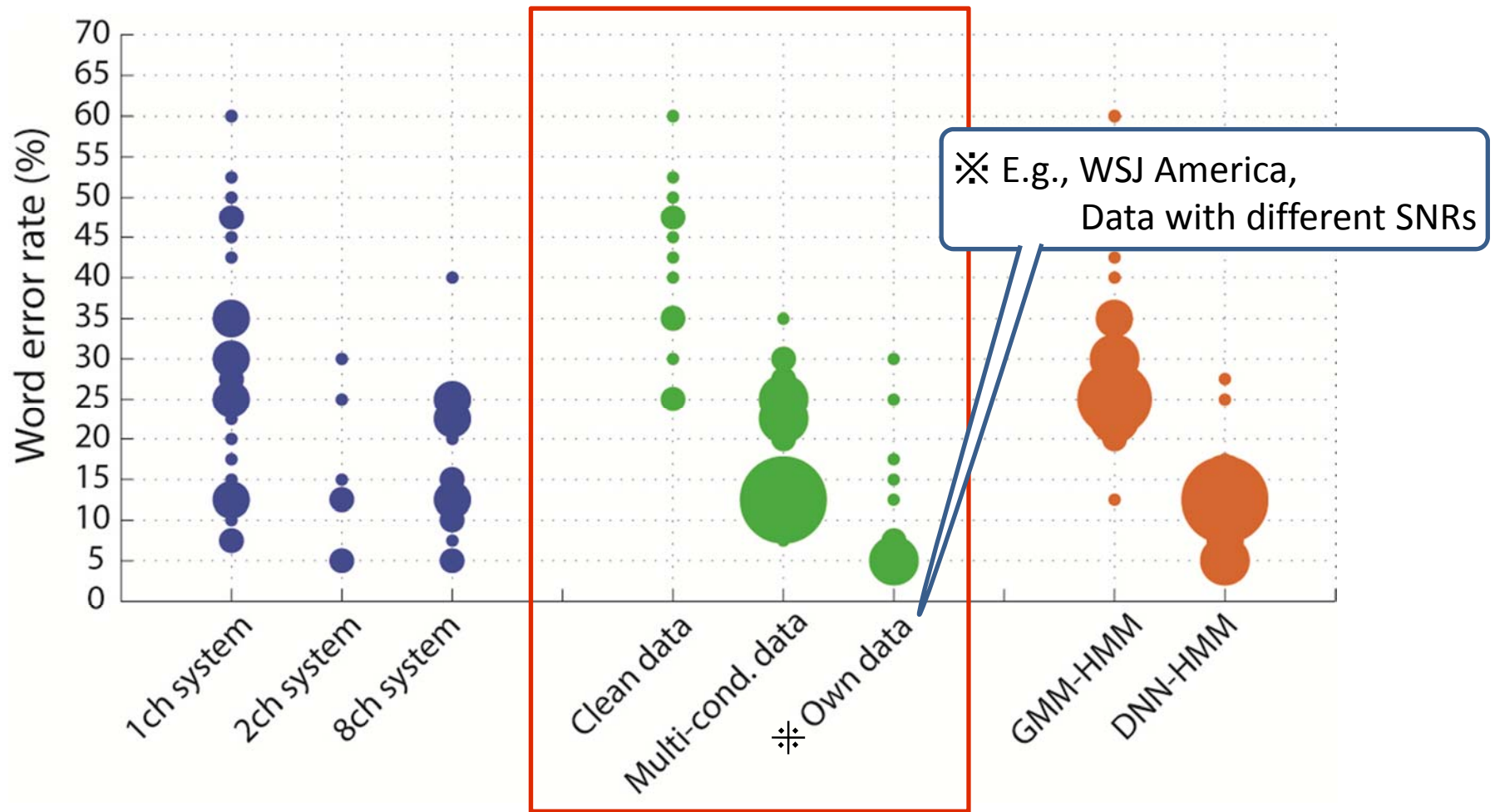


More microphones lead to better performance

ASR results analysis with bubble chart

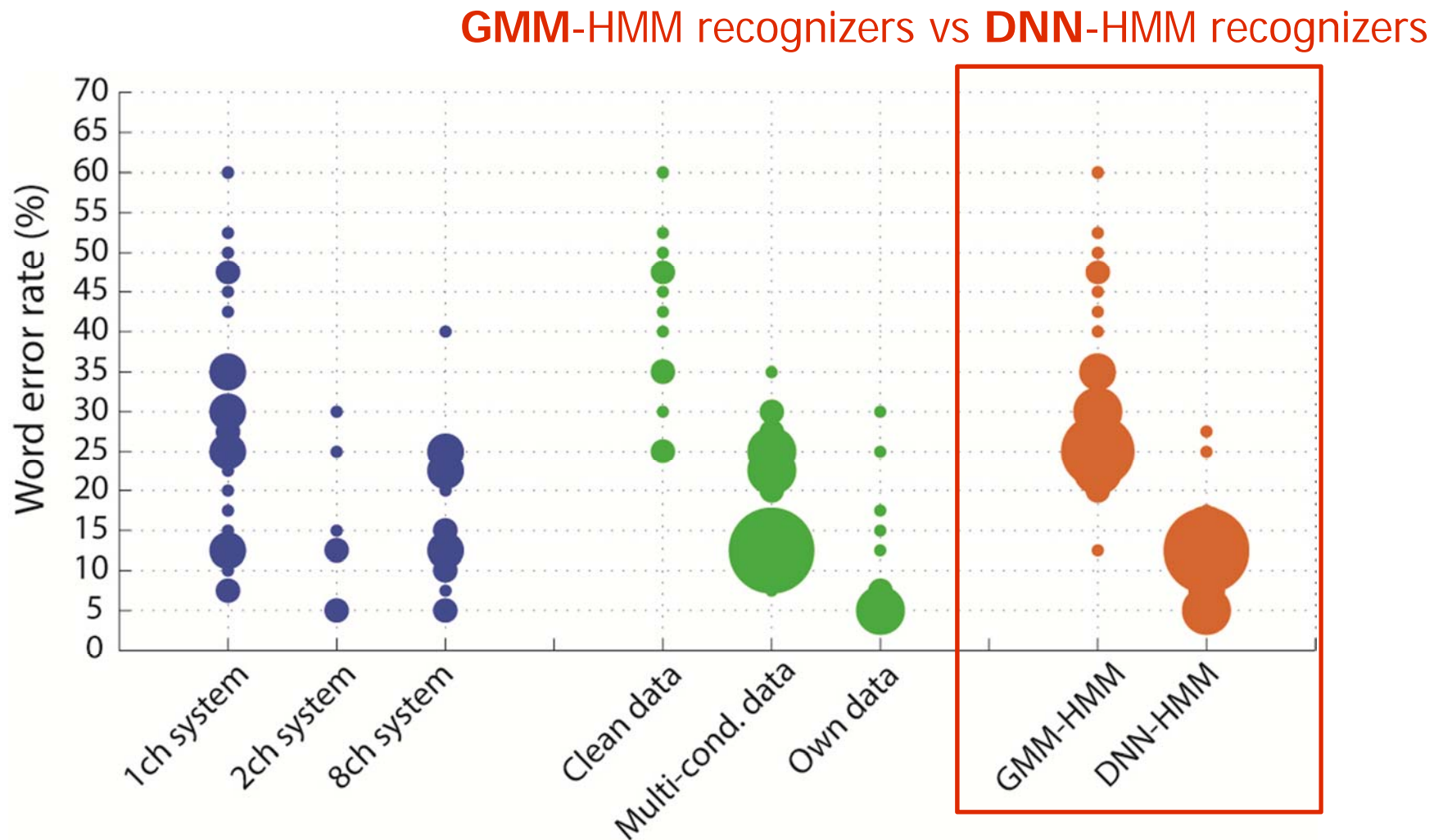


Training data: "Clean" vs "multi-condition" vs "own data"



More training data (acoustic variety) lead to better performance

ASR results analysis with bubble chart

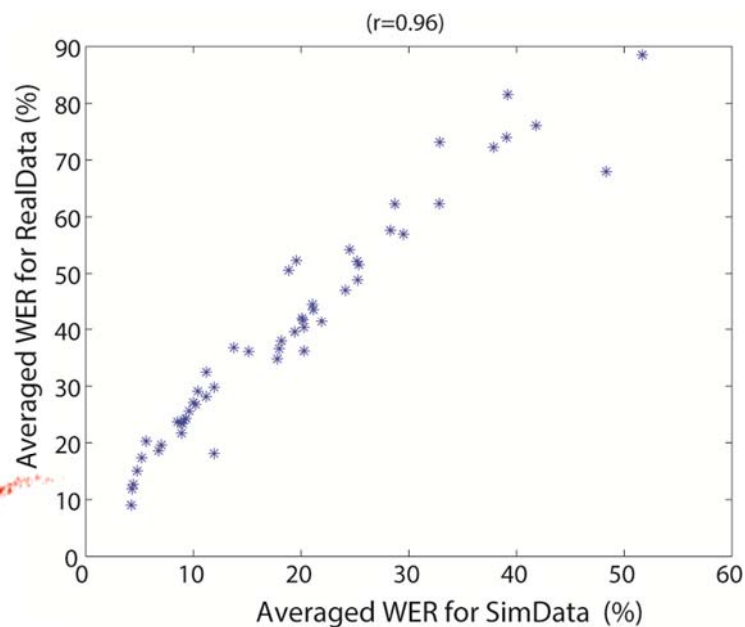


- The top-performing systems often employ DNN-HMM
- Resultant performance may differ due to the front-end proc. and the DNN config. etc

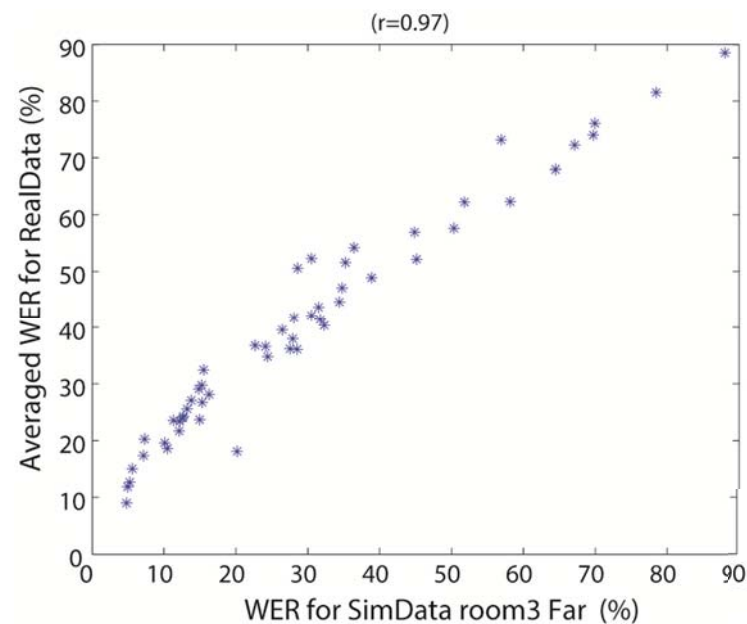
ASR results analysis: SimData vs RealData

- Relationship between SimData scores and RealData scores

SimData vs RealData




SimData Room3 Far vs RealData



Very strong correlation between SimData and RealData scores
(Even stronger between SimData Room3 Far and RealData)

ASR results: Some remarks...

- 
- Strategies often present in the top-performing systems include:
 - Some kind of dereverberation (STFT/Amp spec/feature domain)
 - Linear Multi-ch filtering (MVDR, DS, etc) often for denoising
 - Strong backend (e.g., DNN-HMM recognizer, sophisticated adaptation, robust feature extraction, multi-condition training)
 - System combination
 - However, it's hard to tell the exact impact of each SE/ASR technique.
(It's something we should discover at this workshop!)
 - Some more works required to achieve the clean/headset performance.
(E.g., for RealData, the headset WER is roughly 60% of the best performing system.)

Now, the SE part... 😊

- An important question in the SE task-

Most submissions managed to improve the objective measures (cf. webpage, presentations), but **how about their subjective qualities?**

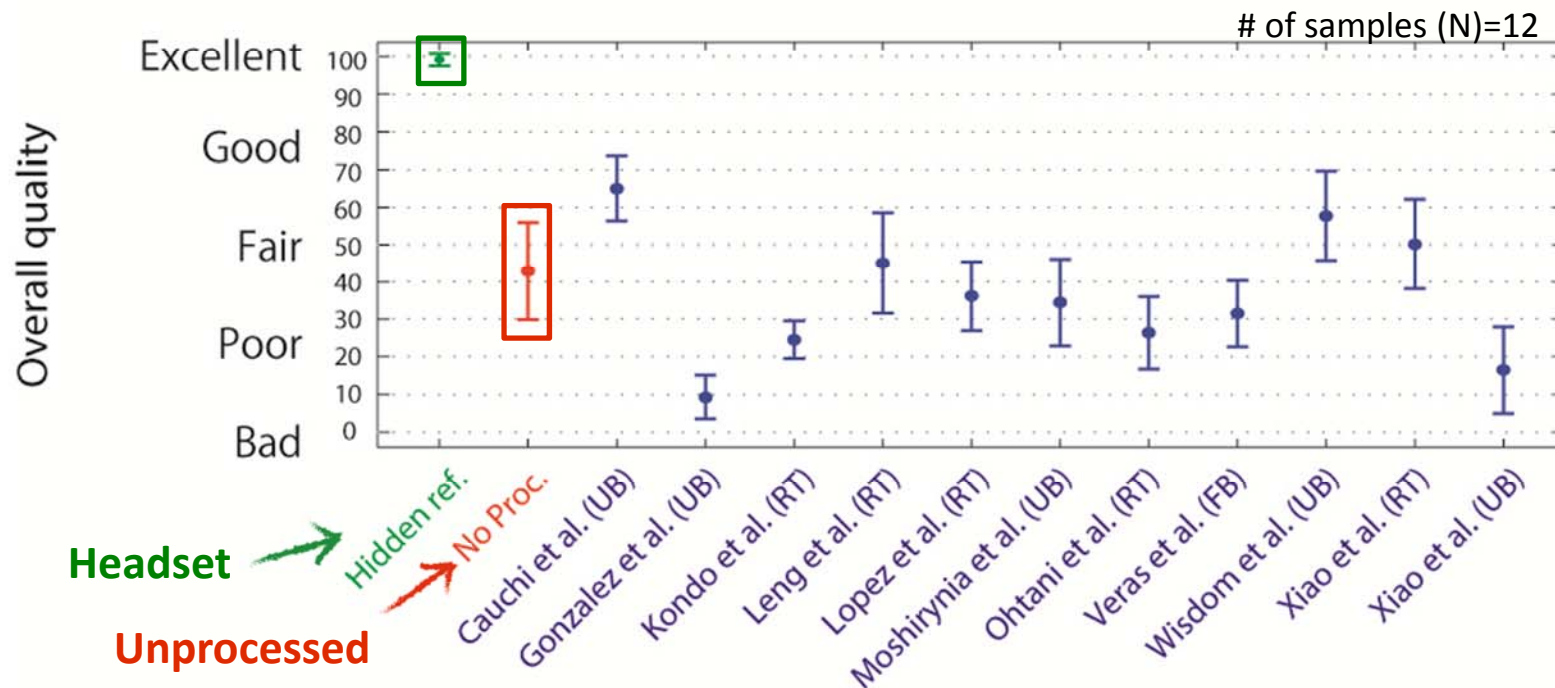
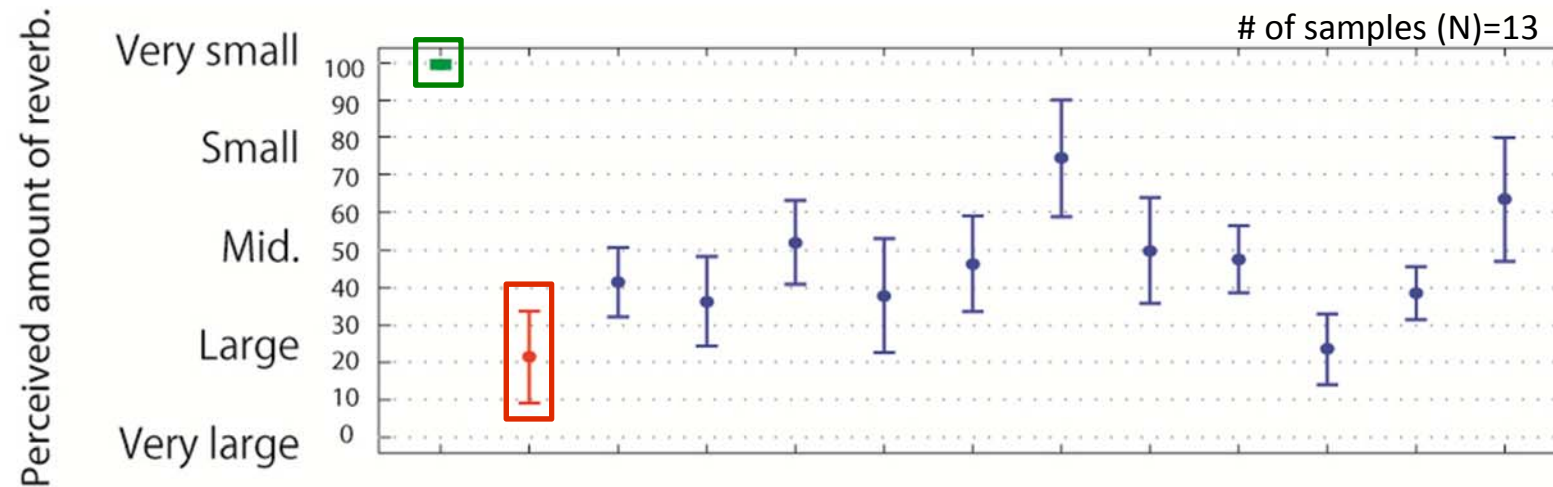


Subjective evaluation: test outline

- MUSHRA (MULTiple Stimuli with Hidden Reference and Anchor) test
- **Web-based listening test** (not well controlled)
- Test carried out separately for 1ch, 2ch and 8ch systems
- Evaluation conditions (4 conditions): SimData room2 near & far
RealData near & far
- **2 evaluation metrics**
 - Perceived amount of reverberation
 - *Very large/Large/Mid./Small/Very small*
 - Test materials: clean (hidden ref.) + No proc. + test systems
 - Overall quality (i.e., artifacts, distortions, remaining reverb and etc)
 - *Bad/Poor/Fair/Good/Excellent*
 - Test materials: clean (hidden ref) + no proc.
+ a 3.5kHz lowpass of the reverberant speech,
+ test systems

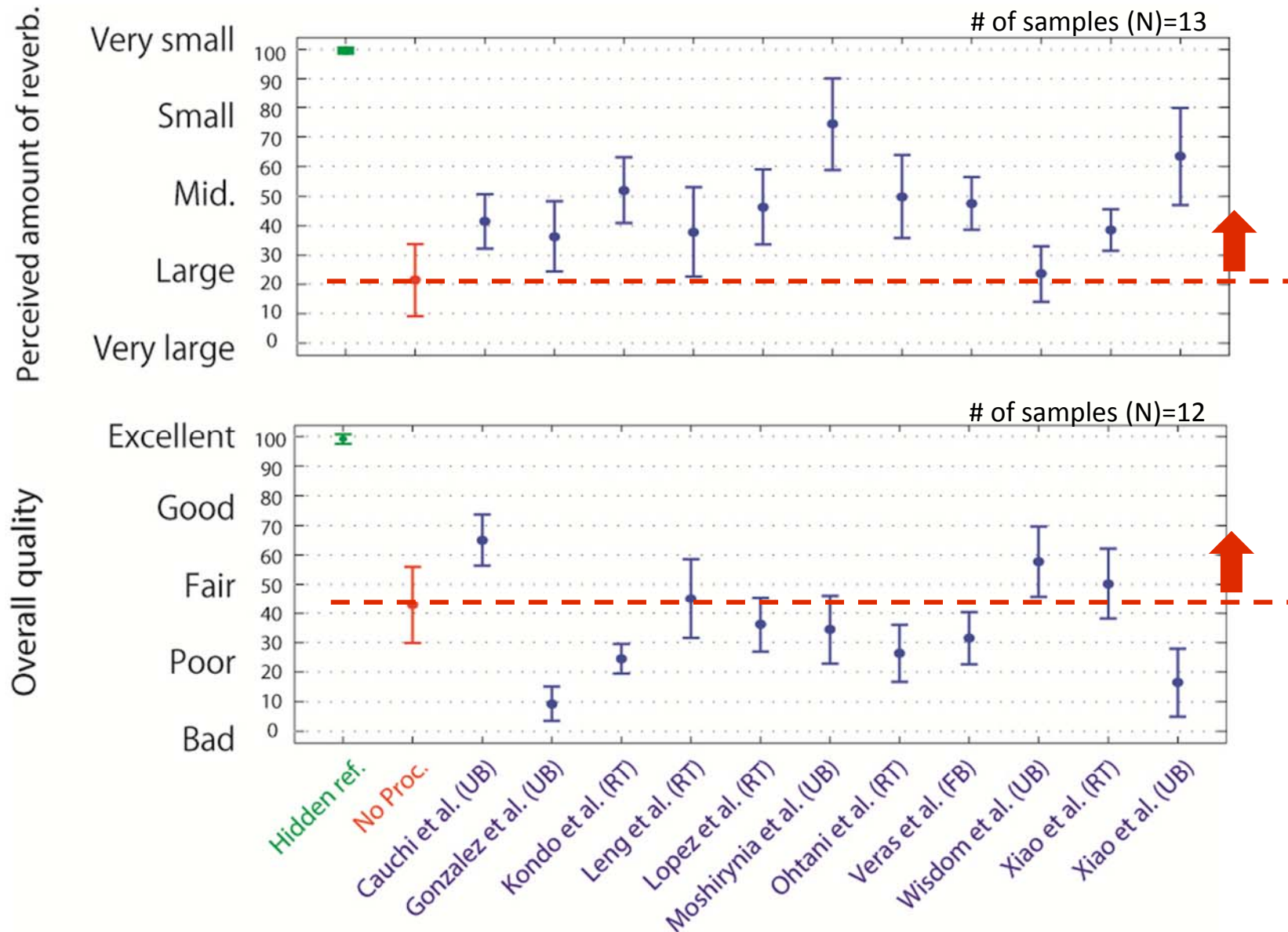
Subjective eval. result : 1ch

(Result at *RealData far* condition)



Subjective eval. result : 1ch

(Result at RealData far condition)

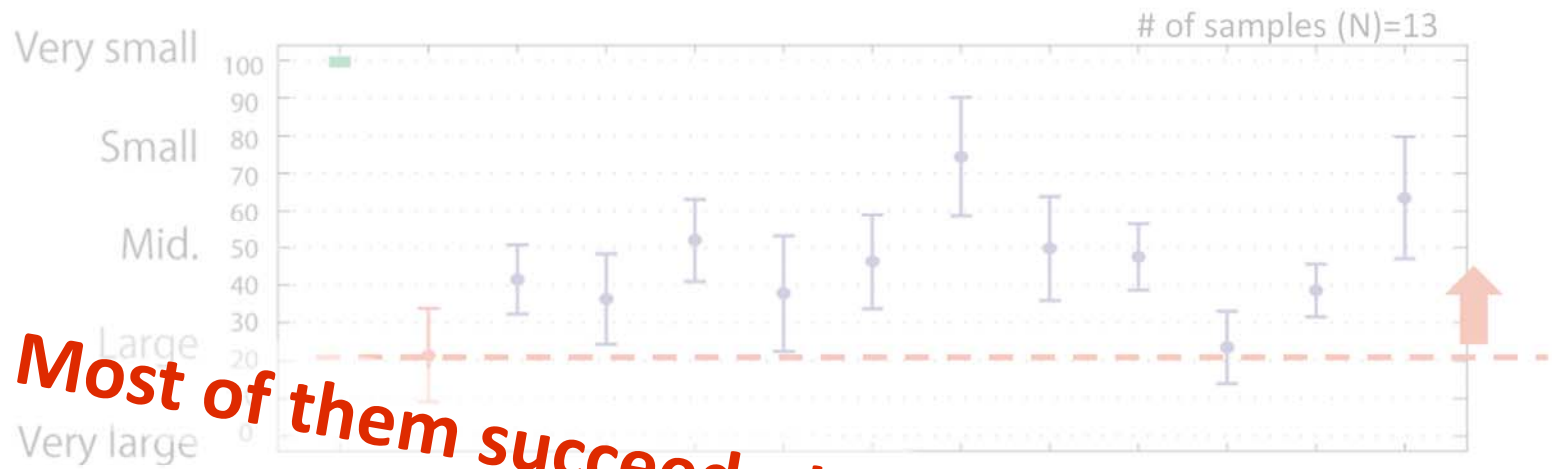


Subjective eval. result : 1ch

(Result at RealData far condition)

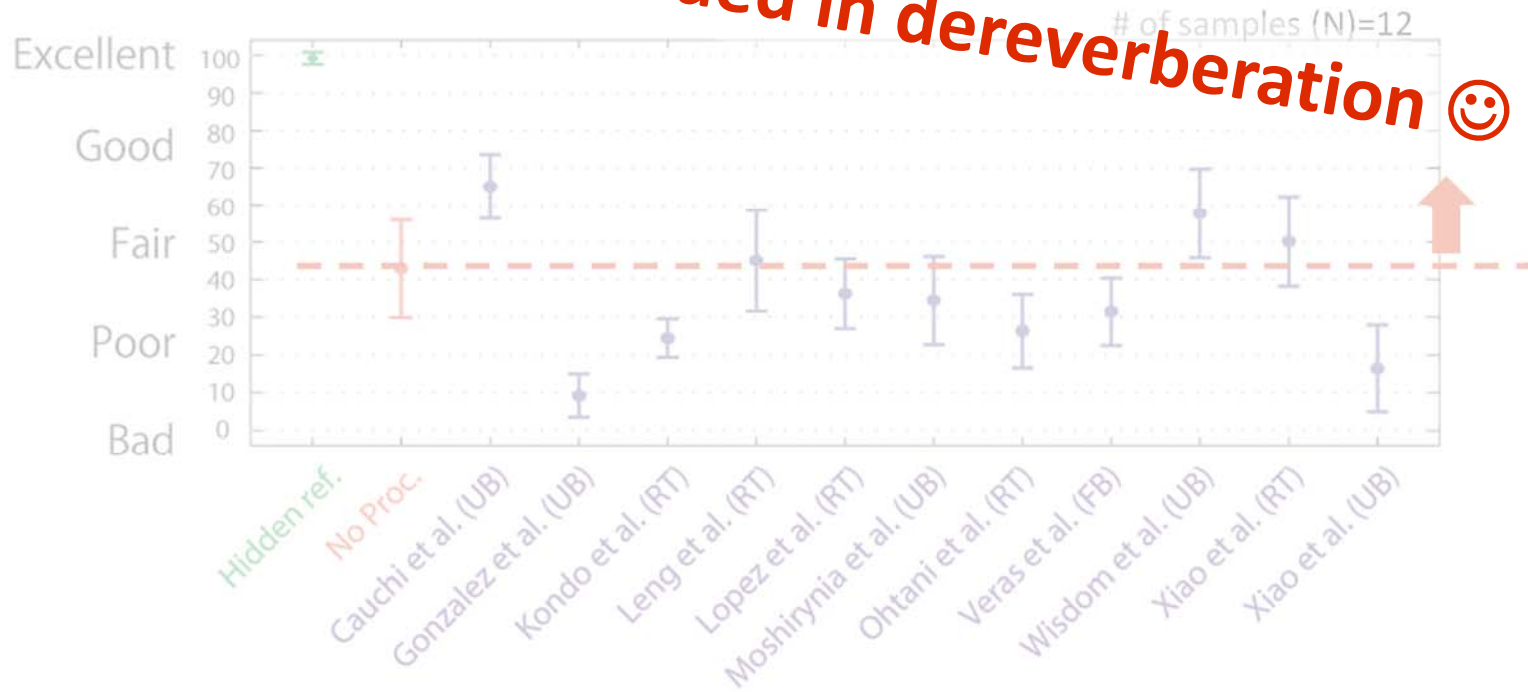


Perceived amount of reverb.



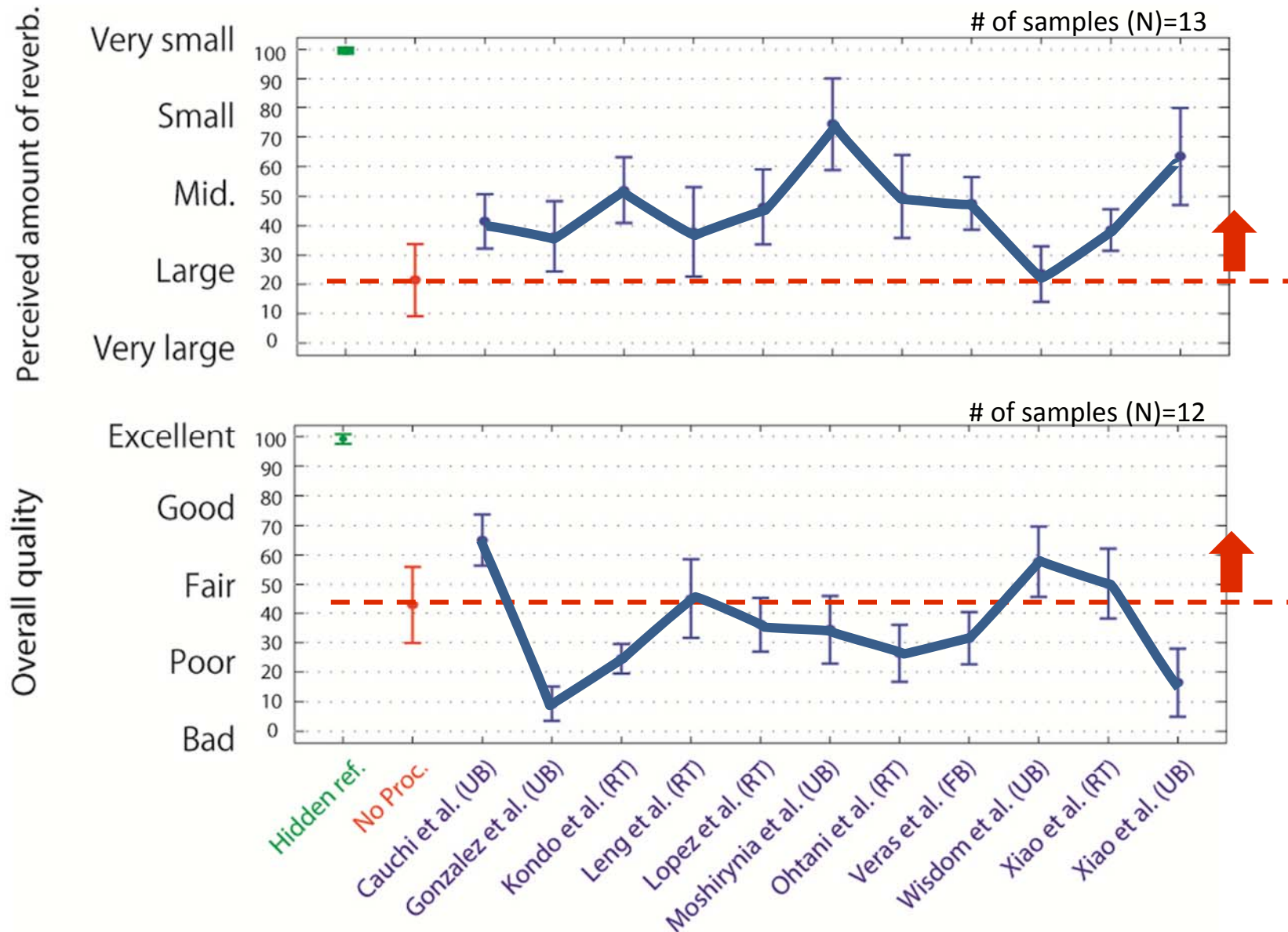
Most of them succeeded in dereverberation 😊

Overall quality



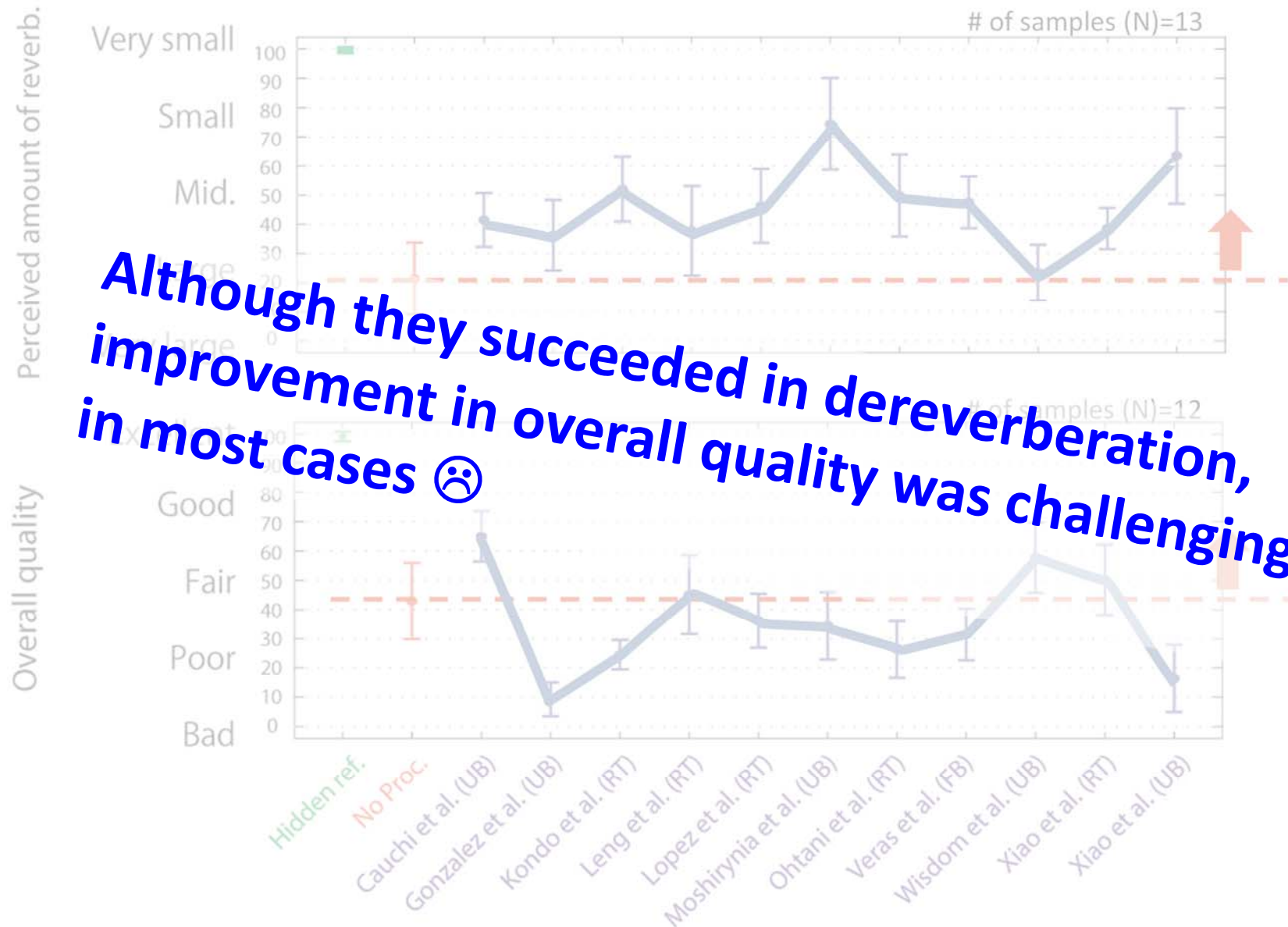
Subjective eval. result : 1ch

(Result at RealData far condition)



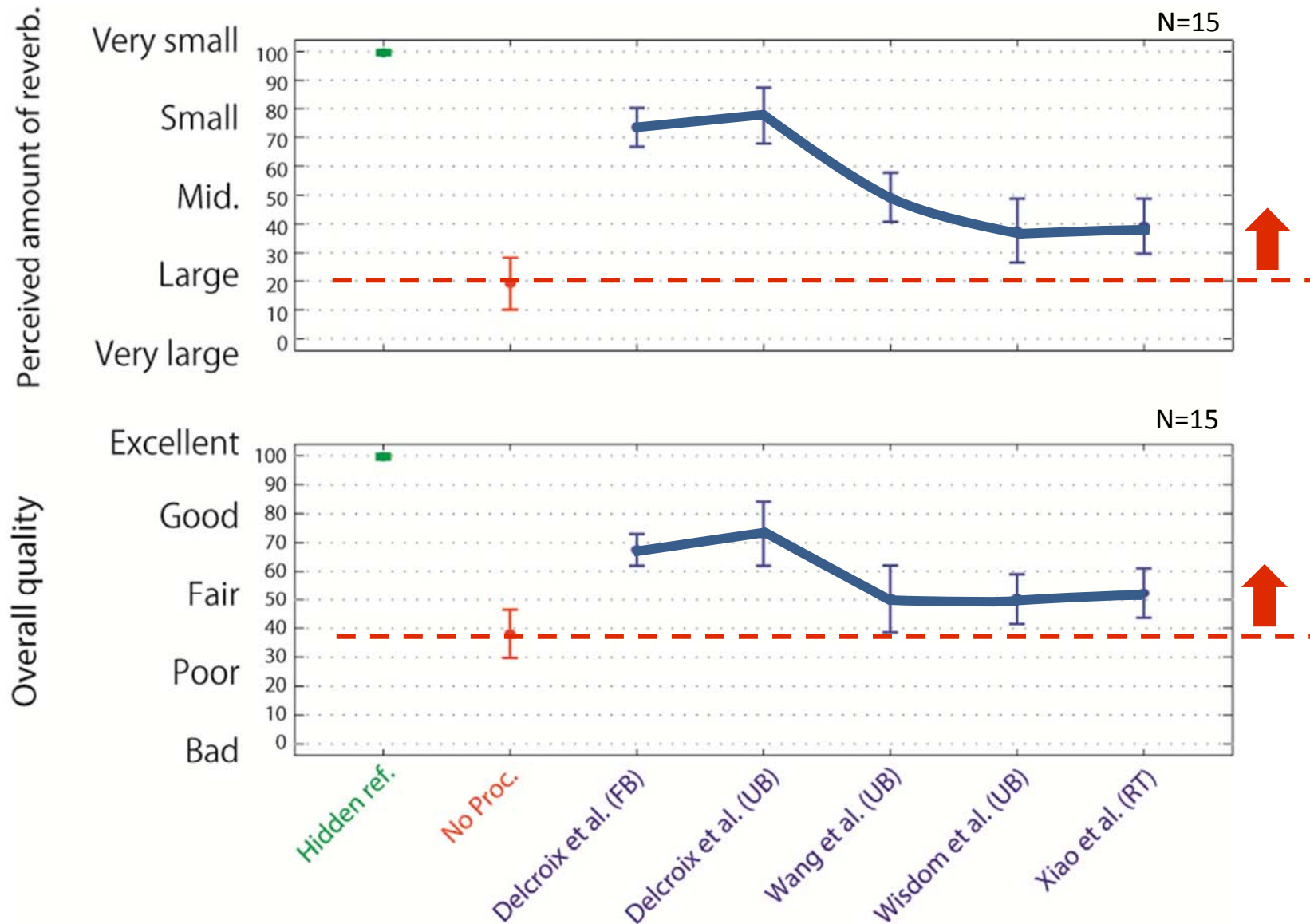
Subjective eval. result : 1ch

(Result at RealData far condition)



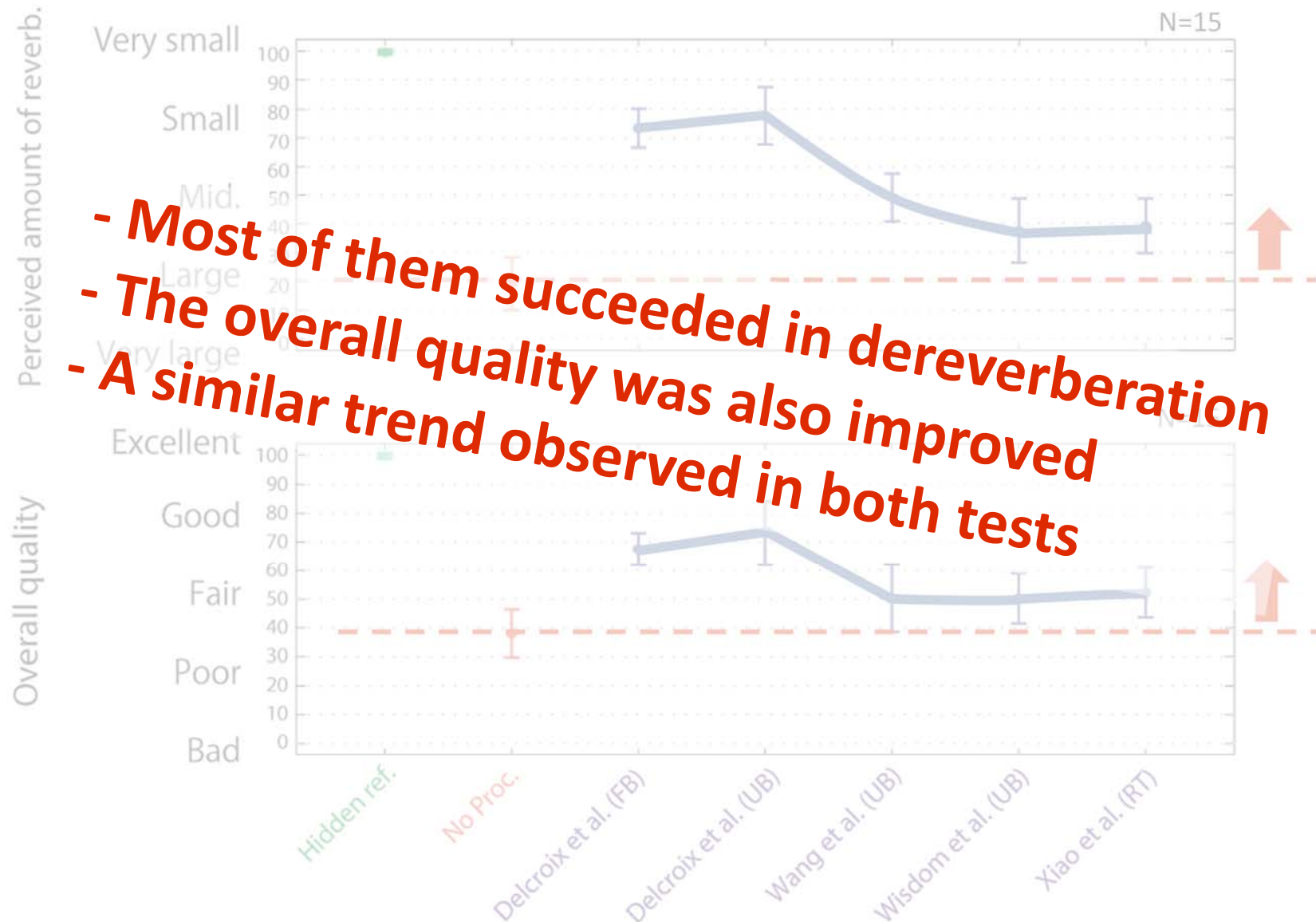
Subjective eval. result : 2ch

(Result at RealData far condition)



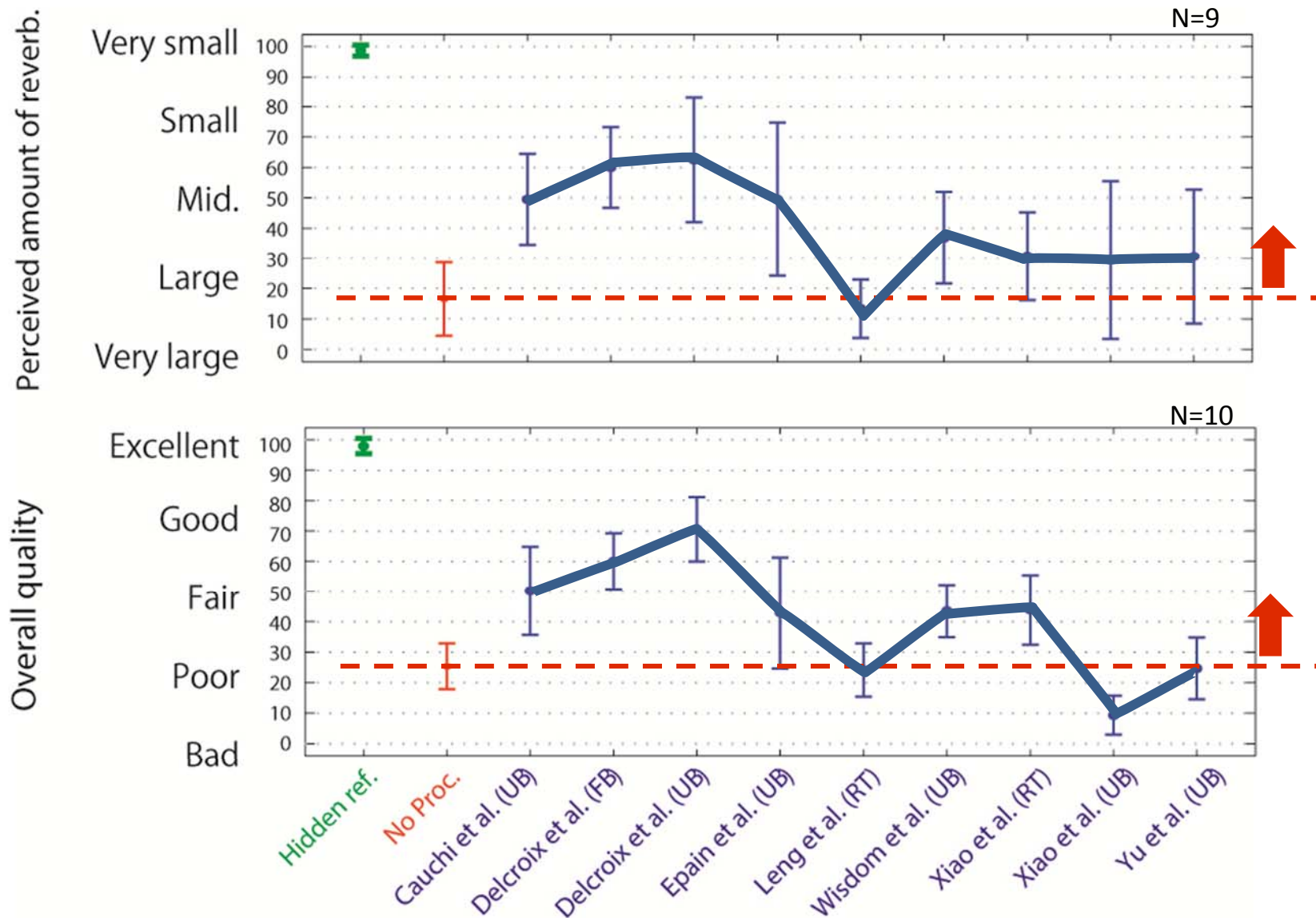
Subjective eval. result : 2ch

(Result at RealData far condition)



Subjective eval. result : 8ch

(Result at RealData far condition)



Subjective eval. result : 8ch

(Result at RealData far condition)



- Another important question-

How does the subjective score correlate with the objective measures?



SE results: subjective vs objective

- Relationship between the subjective scores and each objective score

Correlation with the scores of the “perceived amount of reverberation” test

	CD	FWSegSNR	LLR	SRMR	PESQ
Averaged correlation coeff.	-0.70	0.71	-0.43	0.62	0.77

Correlation with the scores of the “overall quality” test

	CD	FWSegSNR	LLR	SRMR	PESQ
Averaged correlation coeff.	-0.35	0.39	-0.21	0.12	0.28

- Amount of dereverberation can be roughly measured with the objective measures such as CD, FWSegSNR, PESQ.
- The overall quality is not well captured with the objective measures used.

There may be more appropriate objective measures that correlate well with the subjective scores.

SE results: Some remarks...

- 1ch dereverberation is still a challenge task
(Much room to be improved!)
- Some multi-channel dereverberation methods are found to be effective in various conditions.
- More appropriate objective quality measure should be considered, which well coincides with subjective scores.

Conclusions...

- A wide variety of approaches submitted to both the ASR and the SE tasks
- ASR task
 - Most submissions managed to bring improvement over the baseline systems
 - The top-performing systems tend to be quite sophisticated both in the front-end and the back-end
- SE task
 - Most submissions succeeded in dereverberation
 - Improvement in the overall quality was not always easy
 - Better objective scores maybe necessary

Important questions to be discussed...

- How was the challenge framework? How can we do better?
- Is this challenge already overcome?
- Which directions/methodologies are essential to pursue?
 - For improving ASR performance
 - For improving SE performance
- Collaboration between SE and ASR necessary?

Let's discover our own answers during the workshop and discuss at the panel session 😊

**Thank you... and now
let's start the workshop!**

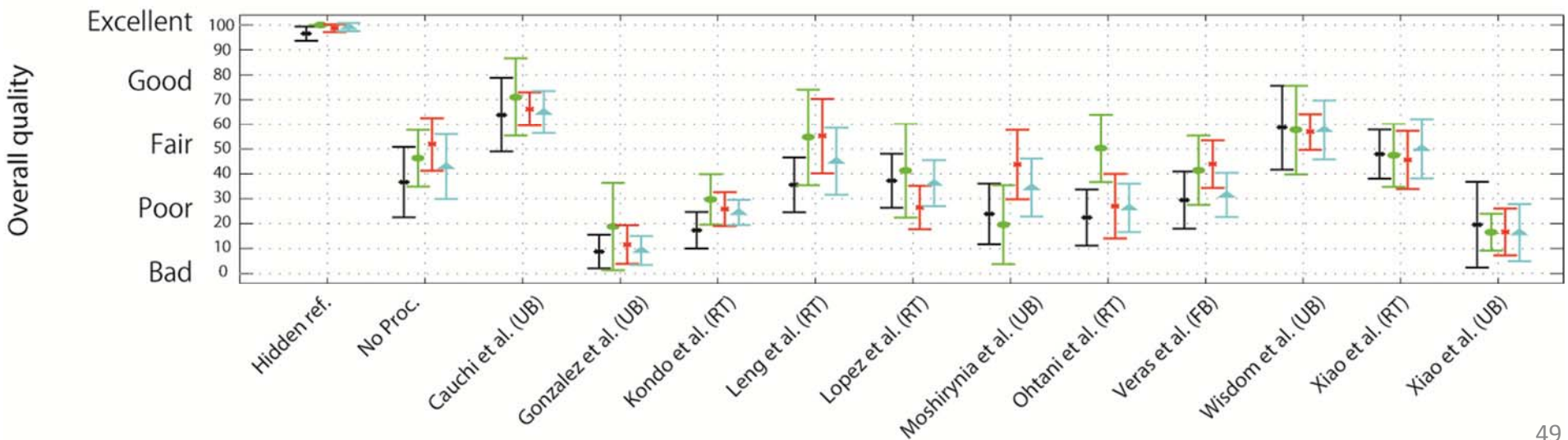
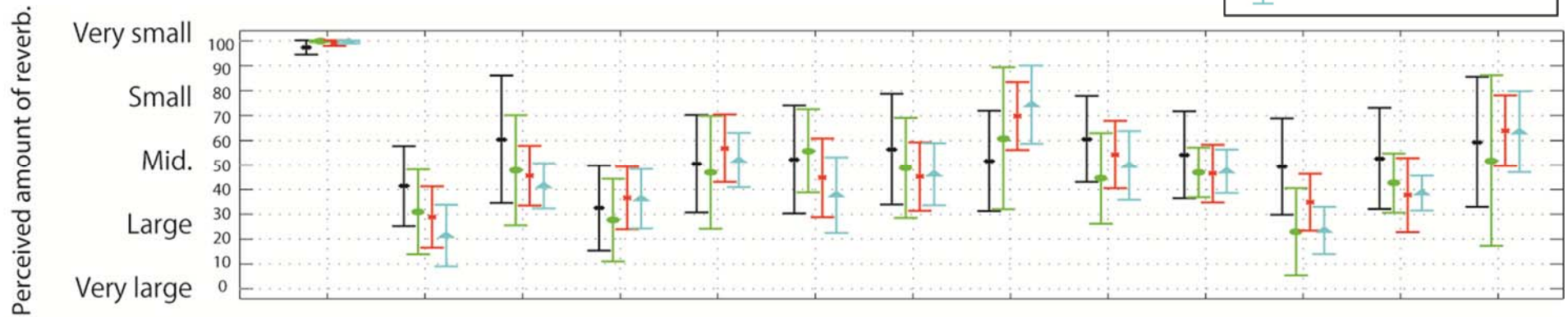
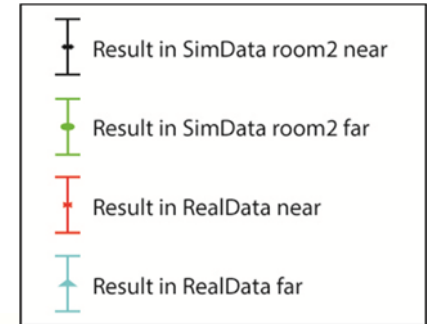
Appendix

Intermediate result of the subjective quality test for 1ch systems

Notes:

- It is not recommended to directly compare the numbers obtained with the different reverberation conditions.
- All mean scores are plotted with their associate 95% confidence intervals.
- About notations

- RT: real-time processing - UB: utterance-batch processing - FB: full-batch processing



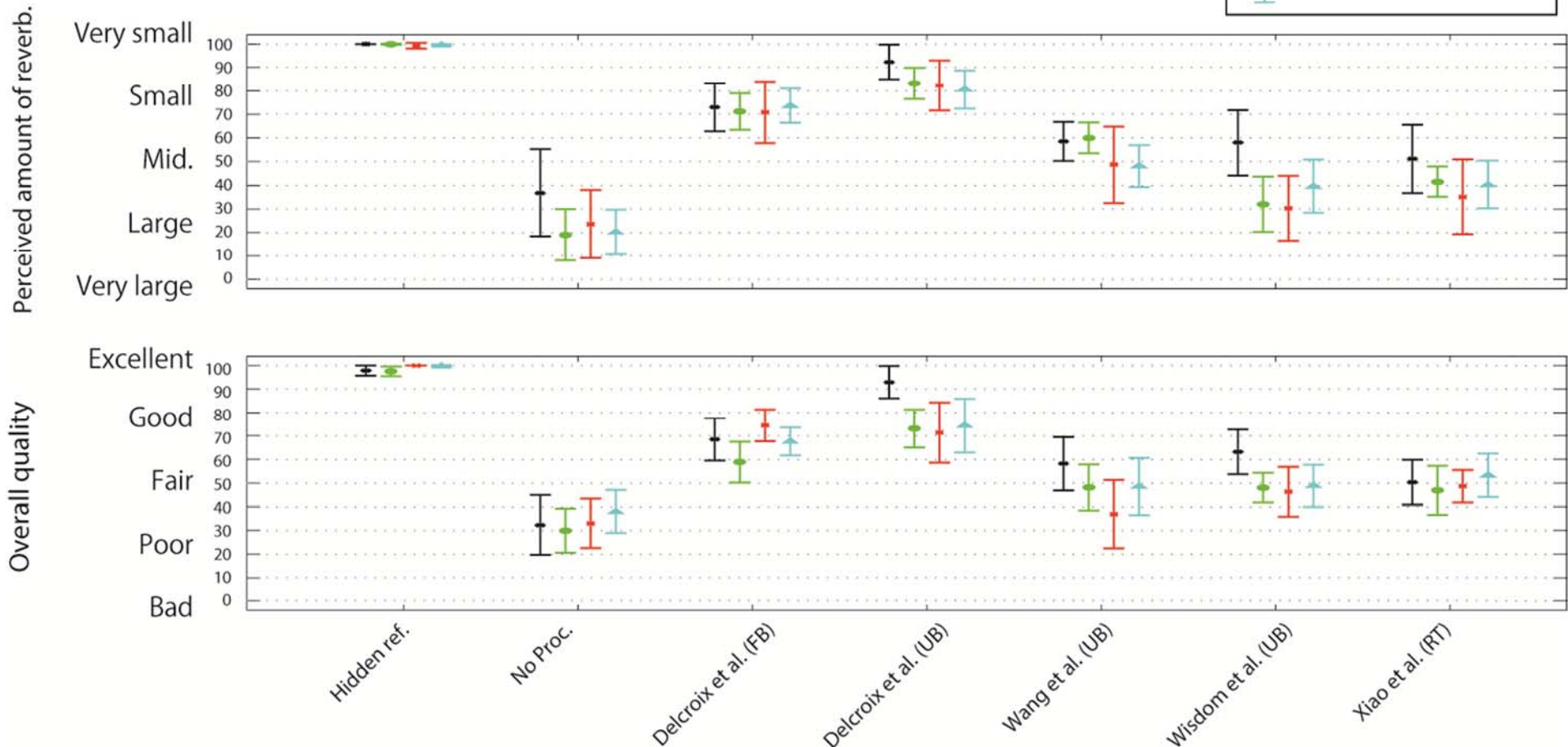
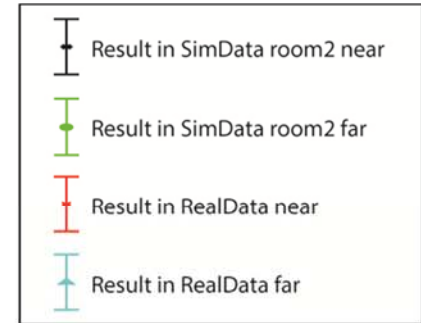
Intermediate result of the subjective quality test for 2ch systems

Notes:

- It is not recommended to directly compare the numbers obtained with the different reverberation conditions.
- All mean scores are plotted with their associate 95% confidence intervals.

- About notations

- RT: real-time processing
- UB: utterance-batch processing
- FB: full-batch processing

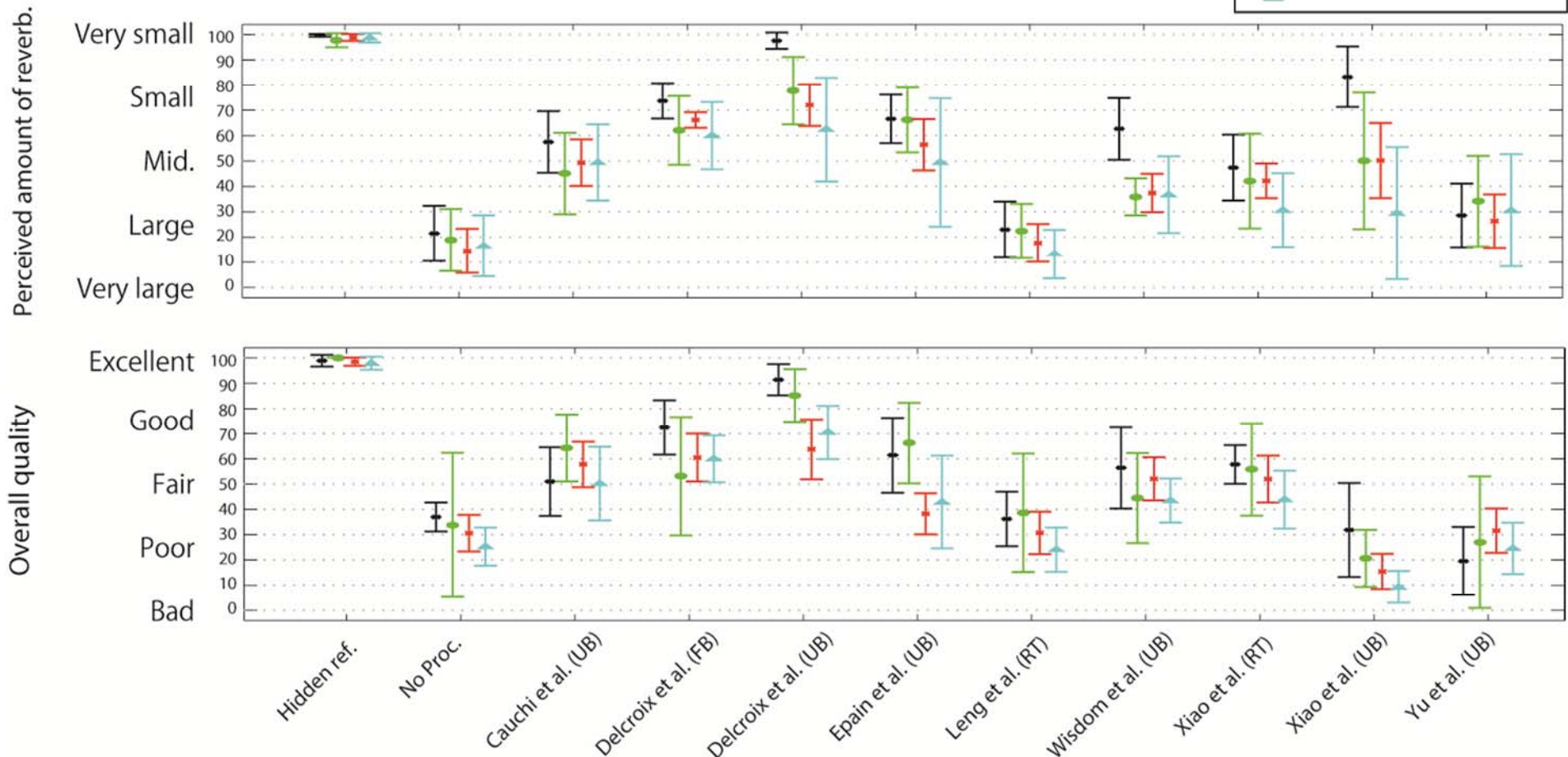


Intermediate result of the subjective quality test for 8ch systems

Notes:

- It is not recommended to directly compare the numbers obtained with the different reverberation conditions.
- All mean scores are plotted with their associate 95% confidence intervals.
- About notations

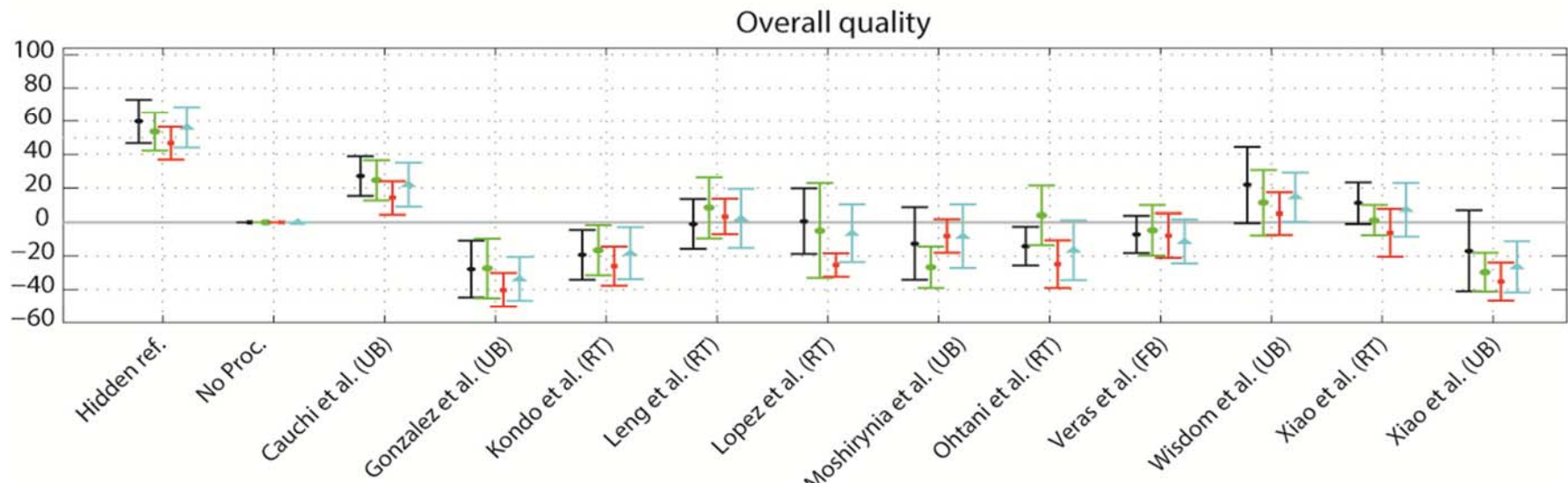
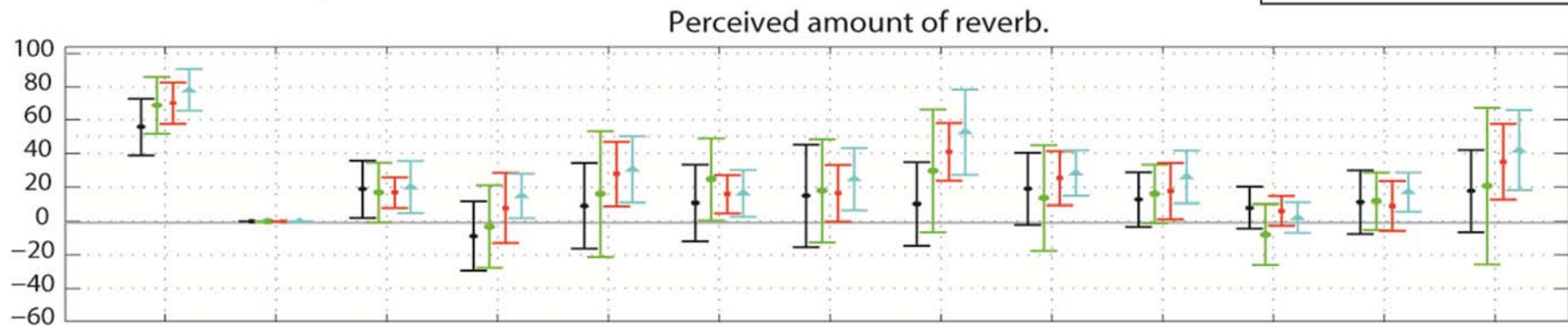
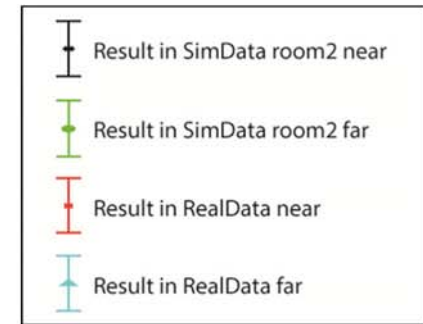
- RT: real-time processing - UB: utterance-batch processing - FB: full-batch processing



Differential score based on the MUSHRA score: 1ch systems

Notes:

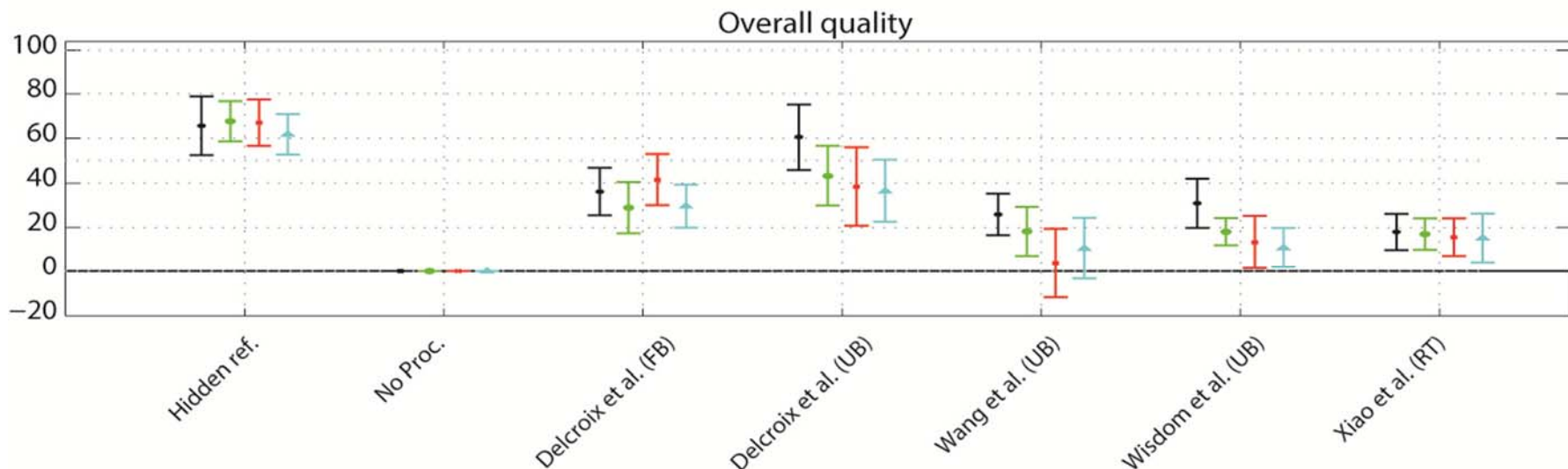
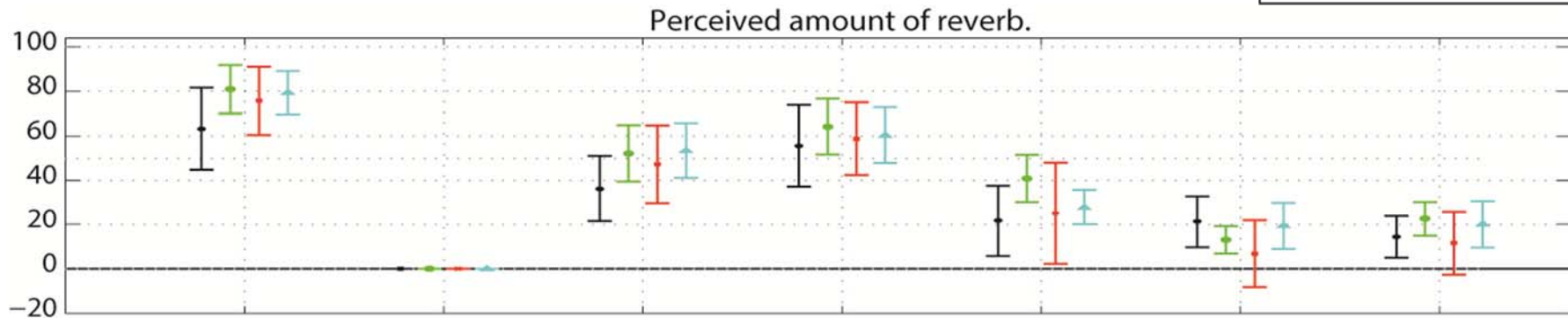
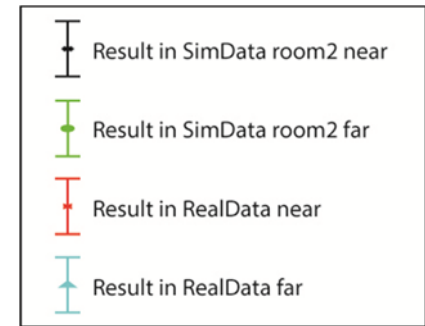
- The differential scores were calculated by subtracting the scores for the unprocessed signal from all the scores to remove potential biases [1].
- It is not recommended to directly compare the numbers obtained with the different reverberation conditions.
- All mean scores are plotted with their associate 95% confidence intervals.



Differential score based on the MUSHRA score: 2ch systems

Notes:

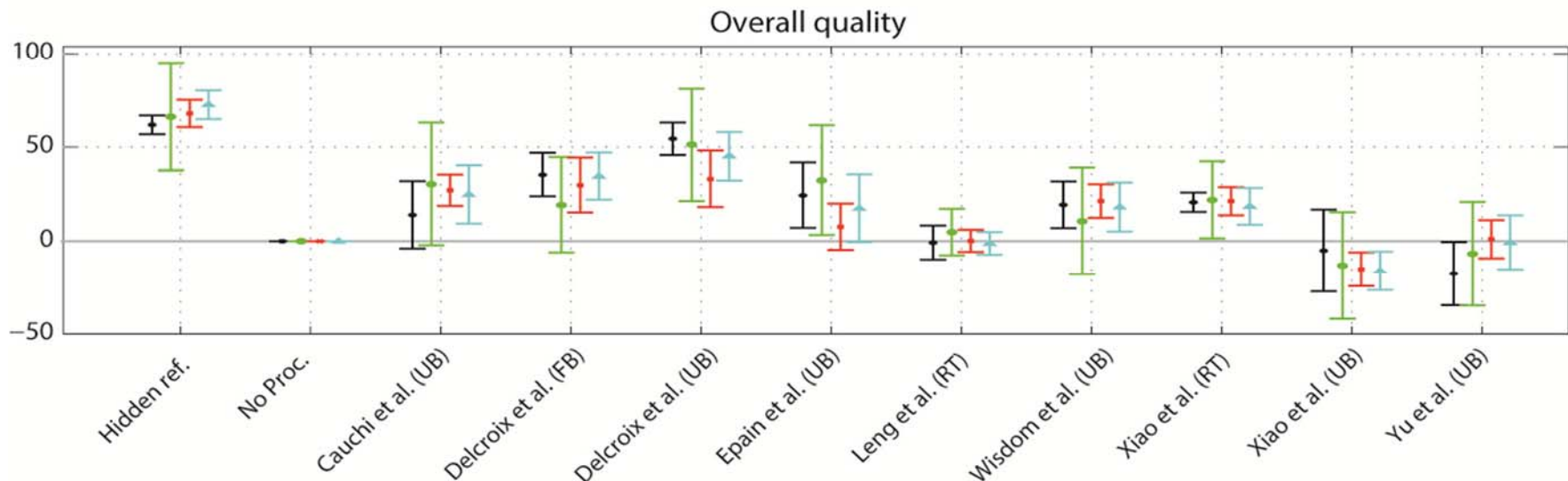
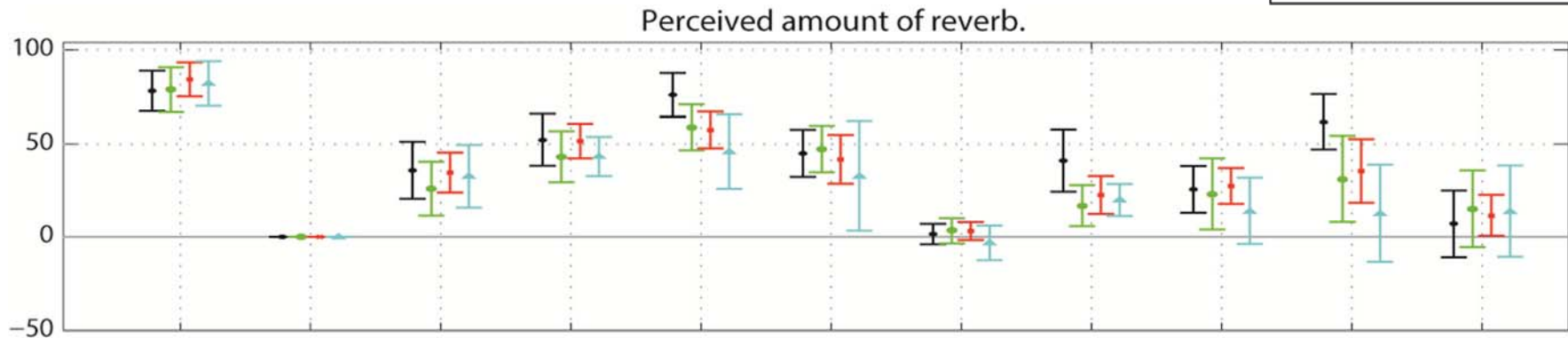
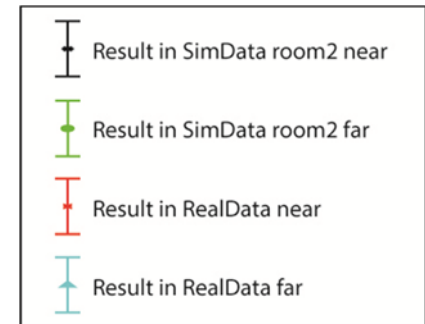
- The differential score was calculated by subtracting the score for the unprocessed signal from all the scores to remove potential biases [1].
- It is not recommended to directly compare the numbers obtained with the different reverberation conditions.
- All mean scores are plotted with their associate 95% confidence intervals.



Differential score based on the MUSHRA score: 8ch systems

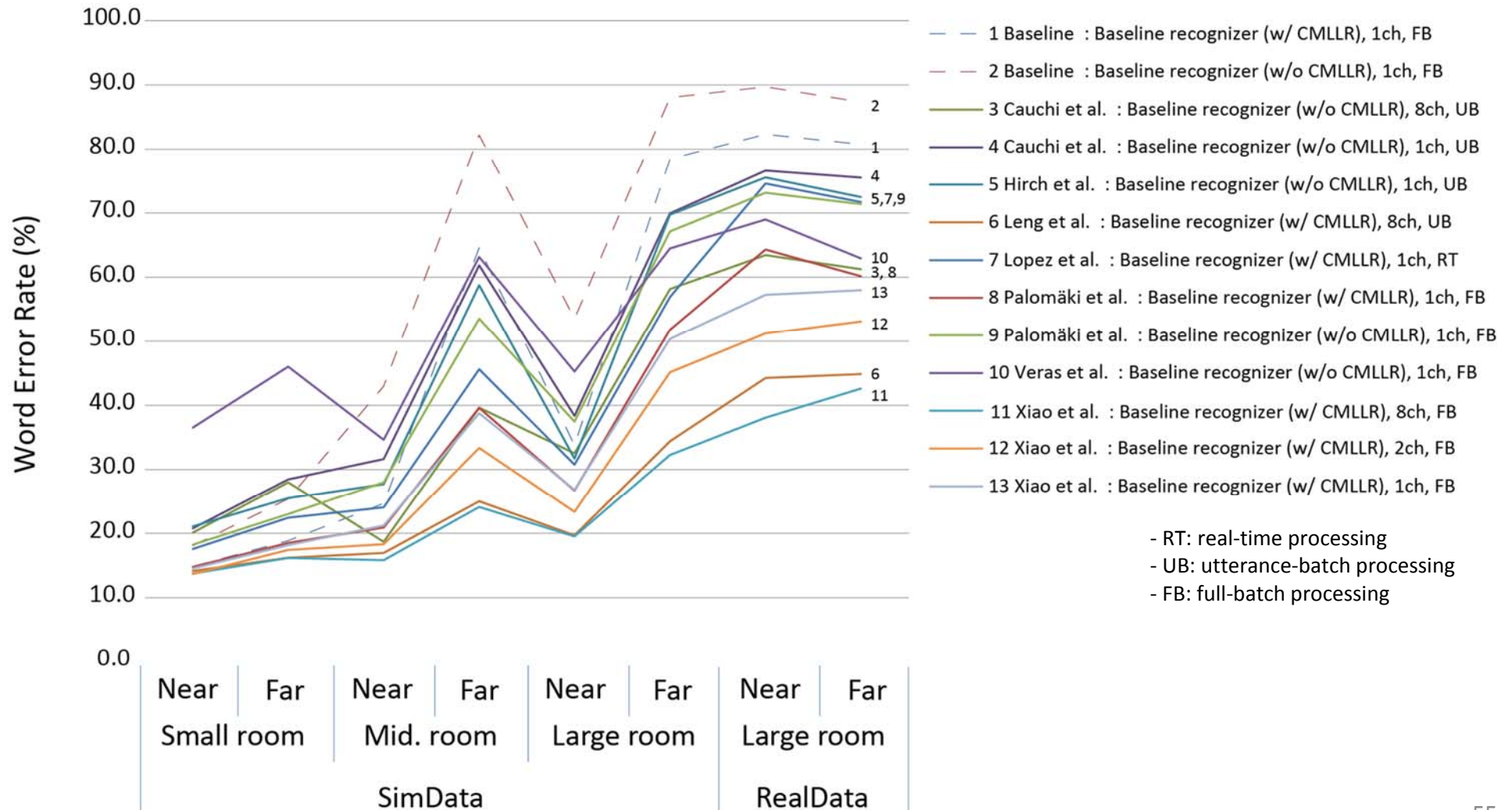
Notes:

- The differential score was calculated by subtracting the score for the unprocessed signal from all the scores to remove potential biases [1].
- It is not recommended to directly compare the numbers obtained with the different reverberation conditions.
- All mean scores are plotted with their associate 95% confidence intervals.



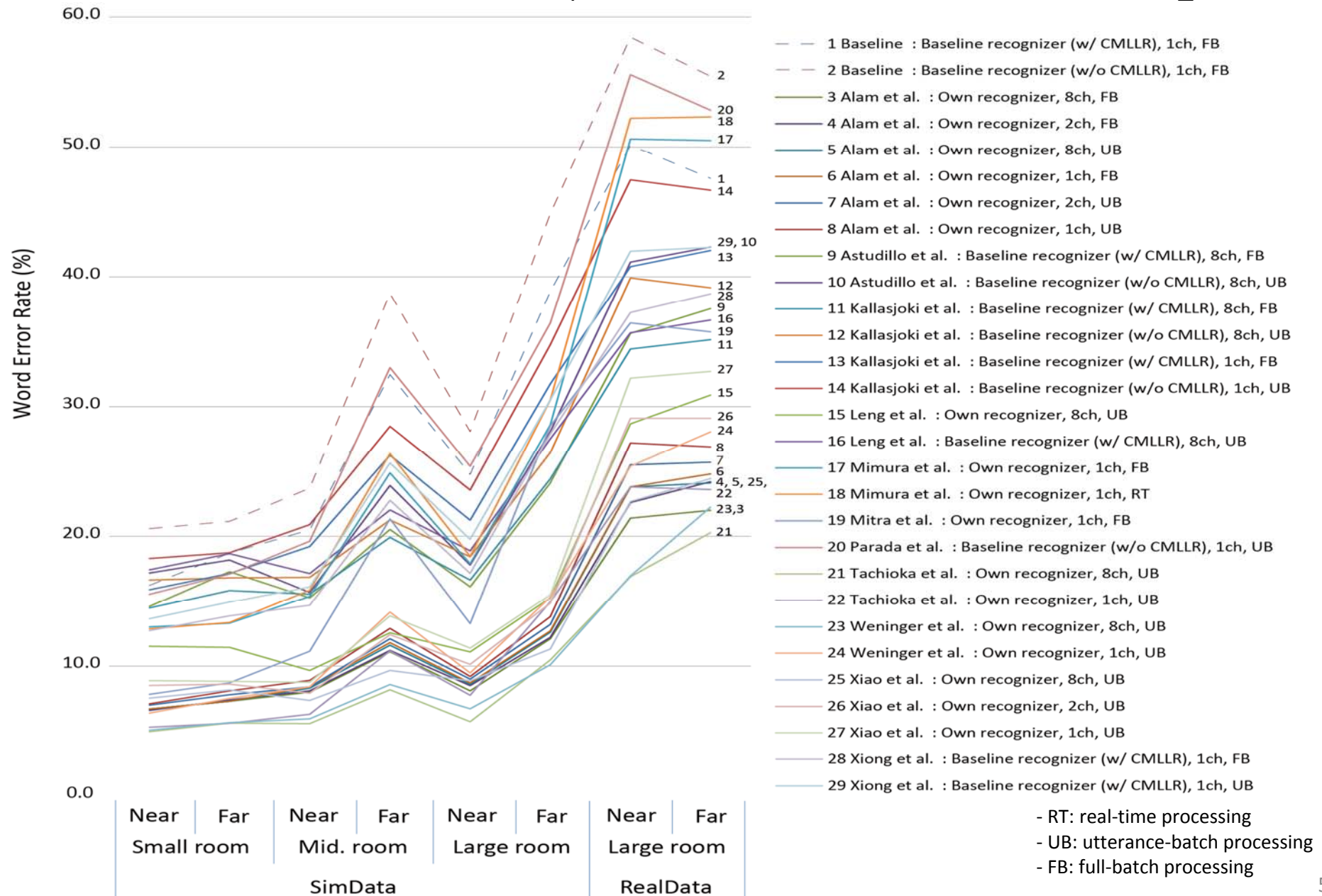
ASR result for the systems trained on clean data

Details of the ASR results are available at http://www.reverb2014.dereverberation.com/result_asr.html



ASR result for the systems trained on multi-condition data

Details of the ASR results are available at http://www.reverb2014.dereverberation.com/result_asr.html



ASR result for the systems trained on own data

Details of the ASR results are available at http://www.reverb2014.dereverberation.com/result_asr.html

