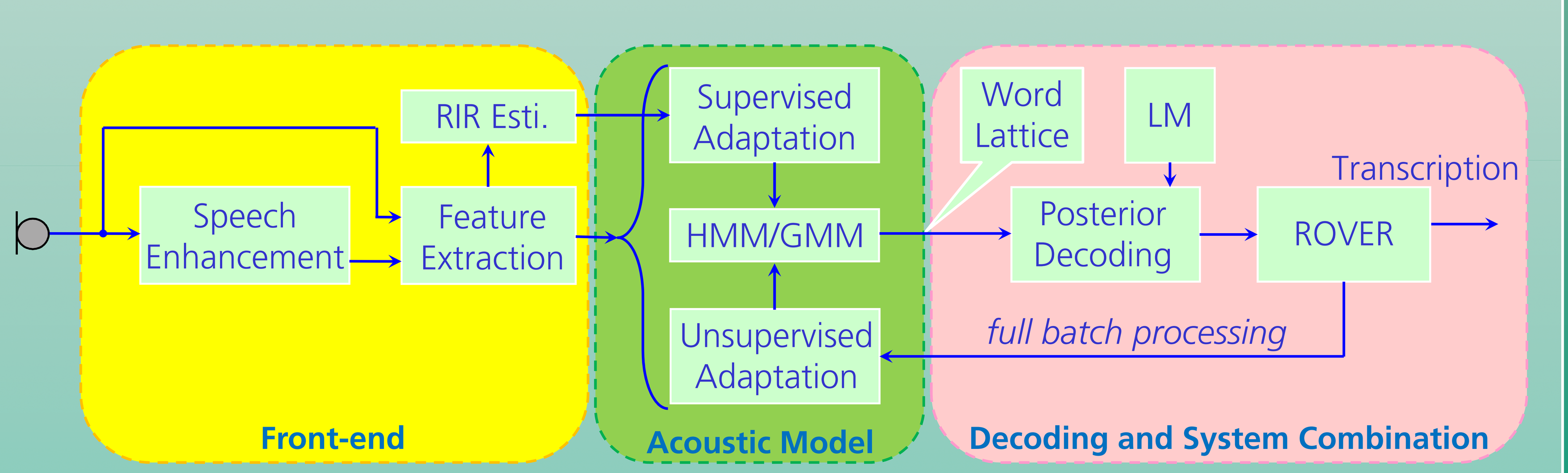
ROBUST ASR IN REVERBERANT ENVIRONMENTS USING TEMPORAL CEPSTRUM SMOOTHING FOR SPEECH ENHANCEMENT AND AN AMPLITUDE MODULATION FILTERBANK FOR FEATURE EXTRACTION

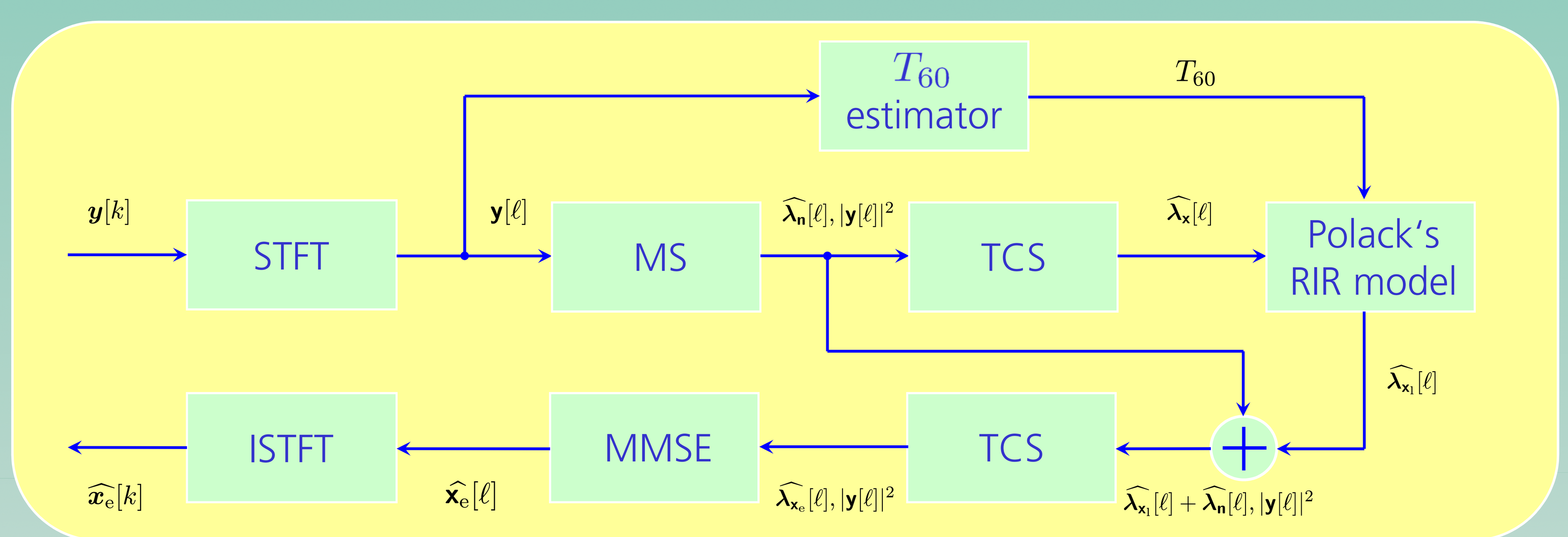
Feifei Xiong[13], Niko Moritz[13], Robert Rehr[23], Jörn Anemüller[23], Bernd T. Meyer[23], Timo Gerkmann[23], Simon Doclo[123], Stefan Goetze[13]
[1]Fraunhofer Institute for Digital Media Technology IDMT, Project Group Hearing-, Speech- and Audio-Technology (HSA), Oldenburg, Germany
[2]University of Oldenburg, Department of Medical Physics and Acoustics, Oldenburg, Germany
[3]University of Oldenburg, Cluster of Excellence Hearing4All, Oldenburg, Germany

ABSTRACT

- Improving ***ASR*** in ***1ch*** scenario of the REVERB Challenge
- Temporal cepstrum smoothing (TCS) noise reduction technique is applied to enhance the reverberant speech signal at moderate noise levels
- Robust feature extraction is performed by amplitude modulation filtering of the cepstrogram to extract temporal modulation information
- The acoustic models are adopted using different RIRs and a RIR selection scheme based on a multi-layer perceptron (MLP) and Gabor features
- ROVER-based system combination is employed to obtain a jointly optimal recognized transcription
- An overall average absolute improvement of **11%** is obtained
  - ***utterance-based batch processing (12.43%)***
  - ***full batch processing (9.41%)***



SPEECH ENHANCEMENT (SE)



- Minimum statistics (MS) → estimate noise PSD (3 s window)
- Temporal cepstrum smoothing (TCS) → estimate speech signal PSD
- Parameterized MMSE spectral magnitude estimator → weighting function
  - minimal gain chosen to -10dB
- For ASR, preserving the fundamental frequency is not crucial
  - smoothing coefficients in TCS

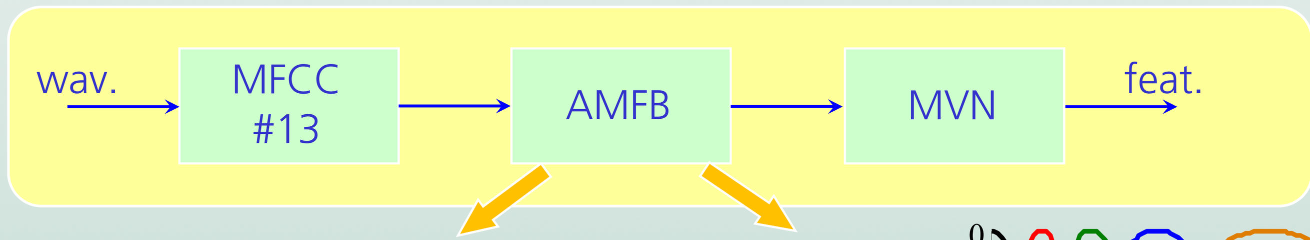
$$\boldsymbol{\alpha}^{c}[\ell_c] = \begin{cases} 0.0 & \ell_c = 0, \ldots, \lceil f_s \cdot 0.5\,\mathrm{ms} \rceil - 1 \\ 0.5 & \ell_c = \lceil f_s \cdot 0.5\,\mathrm{ms} \rceil, \ldots, \lceil f_s \cdot 1\,\mathrm{ms} \rceil - 1 \\ 0.9 & \text{otherwise} \end{cases}$$

BACK-END

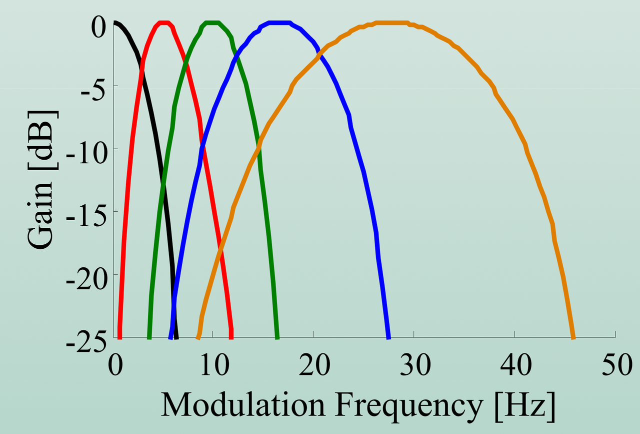
- Baseline HTK framework
- Cluster-based supervised adaptation (CSA) with MLLRMEAN
  - a set of 24 models ←→ 24 different RIRs
  - model selection based on an MLP classifier with 2D Gabor features
- Unsupervised adaptation with MLLRMEAN in the *full batch processing*
  - MLLRMEAN performs better then CMLLR
- Lattice-based posterior decoding (SRILM toolkit)
- ROVER with confidence scores for system combination

AMPLITUDE MODULATION FILTERBANK (AMFB) FEATURES

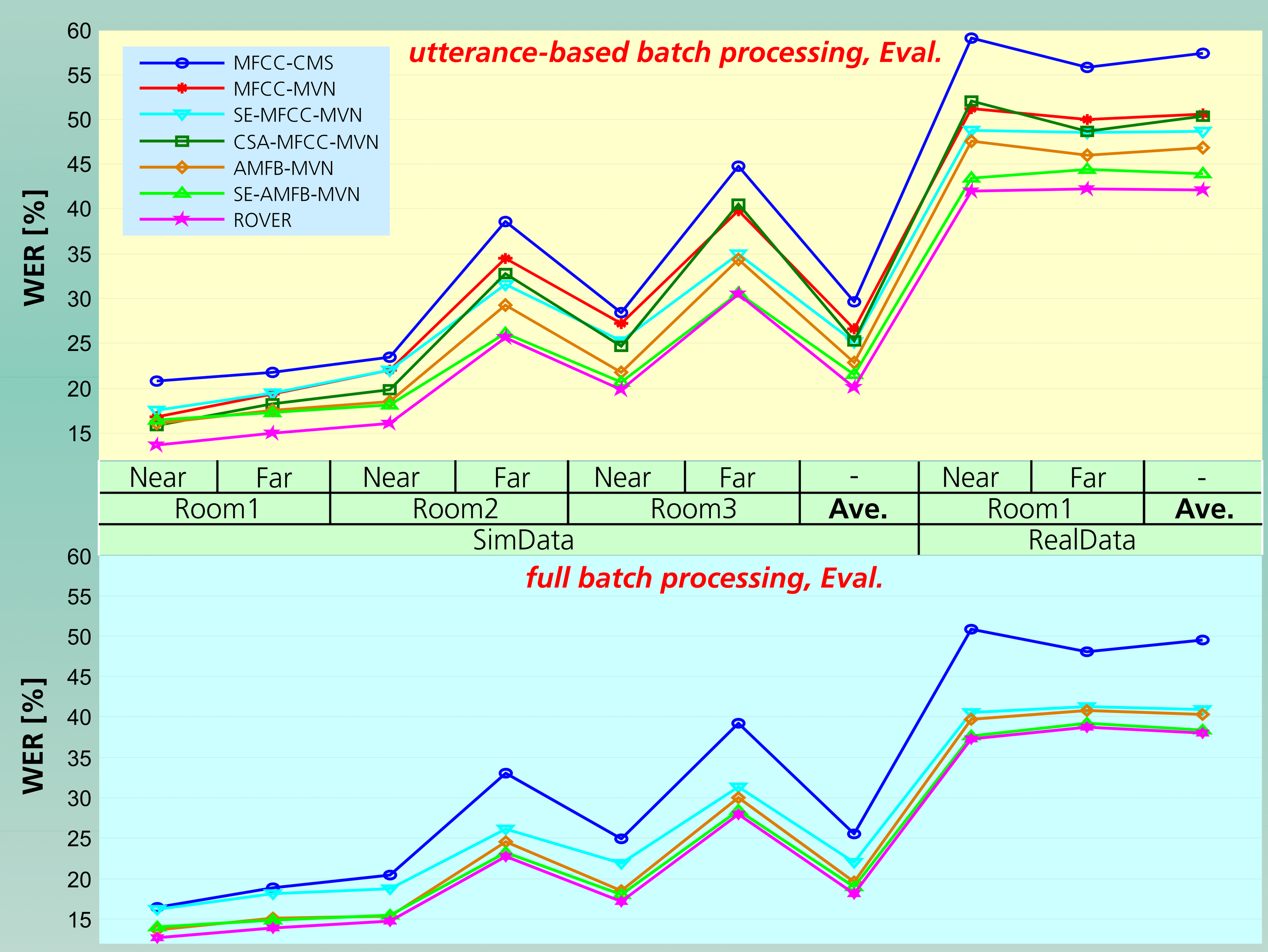
- To extract the temporal dynamics of cepstral coefficients
- 5 AM filters were selected; @{0, 5, 10, 16.67, 27.78} Hz



$$\begin{aligned} \mathbf{q}[\ell_c] &= \mathbf{s}_{\mathrm{carr}}[\ell_c] \odot \mathbf{h}_{\mathrm{env}}[\ell_c] \\ \mathbf{s}_{\mathrm{carr}}[\ell_c] &= \exp\left(i\omega(\ell_c - \ell_{c_0})\right) \\ \mathbf{h}_{\mathrm{env}}[\ell_c] &= 0.5 - 0.5 \cdot \cos\left(\frac{2\pi(\ell_c - \ell_{c_0})}{W_{\ell_c} + 1}\right) \end{aligned}$$



RESULTS



- Similar trend of the WERs reduction for the Dev. and Eval. test sets
- Constant 1~1.5% absolute WER reduction can be observed by the proposed SE algorithm
- AMFB features achieve an average absolute WER reduction of more than 4% compared to MFCCs
- Additional average absolute WER reduction of 1~2% is achieved by ROVER with complementary systems ← fine-tuning
- A better recognition transcription assists MLLR to better adapt the model to match more to the test set condition

CONCLUSIONS

- ***1ch*** combined ASR system consisting of speech enhancement, robust feature extraction, acoustic model adaptation, posterior decoding and ROVER-based system combination
- SE based on TCS is proven to be advantageous to cope with the reverberation effect to ASR systems
- Capturing the temporal modulation information is crucial for feature extraction when facing the reverberant speech for ASR systems

The research leading to these results has received funding from the EU Seventh Framework Programme project Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS) under grant agreement ITN-GA-2012-316969, from the DFG-Cluster of Excellence EXC 1077/1 "Hearing4all", and from the MWK PhD Program "Signals and Cognition".