Collection of re-transmitted data and impulse responses and remote ASR and speaker verification.

Igor Szoke, Lada Mosner (et al.)
BUT Speech@FIT

LISTEN Workshop, Bonn, 19.7.2018

Why

- DRAPAK project
  - To ship an ASR coping with distant and hidden mics (bugs).
- Gap between WER on ASR's trained on retransmitted data using real RIR or artificial RIR. It is few percents but still it is a gap (Ravanelli 2012).
- There is not such large dataset (regarding our goal of 50 environments).
  - According to AcouSP, and openairlib.net
  - Or is there?
- To support other R&D at BUT and also in the world.
- Status:
  - 8 rooms processed so far
  - Running verification experiments now
  - If OK then scale-up.

Microphones placement

Mics positions

- 1-8 - spherical mic array
- ~5 - table top close to the 1st speaker position (SPKID01)
- ~5 - hidden (in a shelf, AC, waste bin, under a table, in a drawer, ...)
- 2 - IoT
- ~5 - ceiling, light, etc.
- ~5 - table top on other places

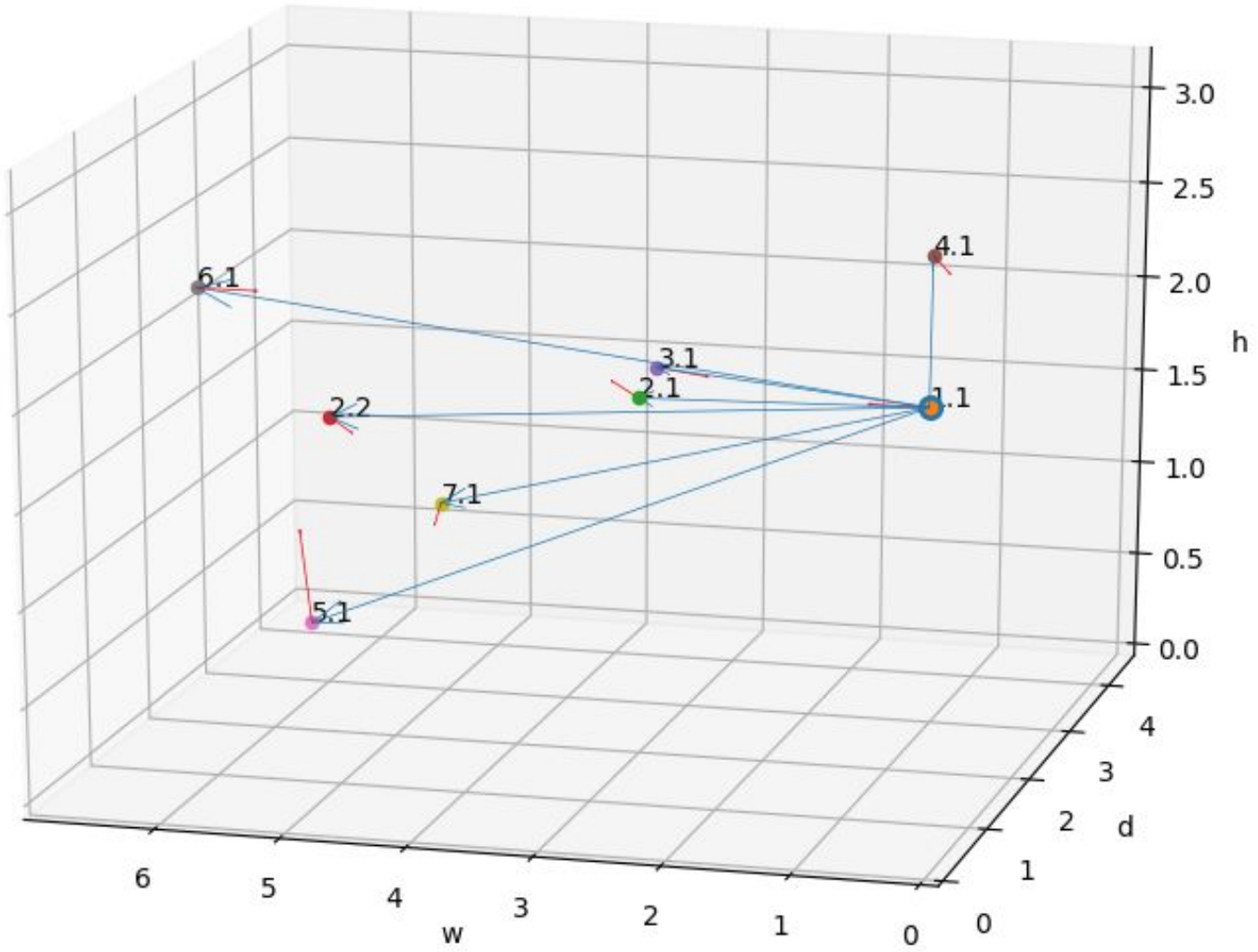
Speaker positions

- Sitting person
- Standing person
- Noise source (near wall)
- Non-standard position (rotated to ceiling, etc.)

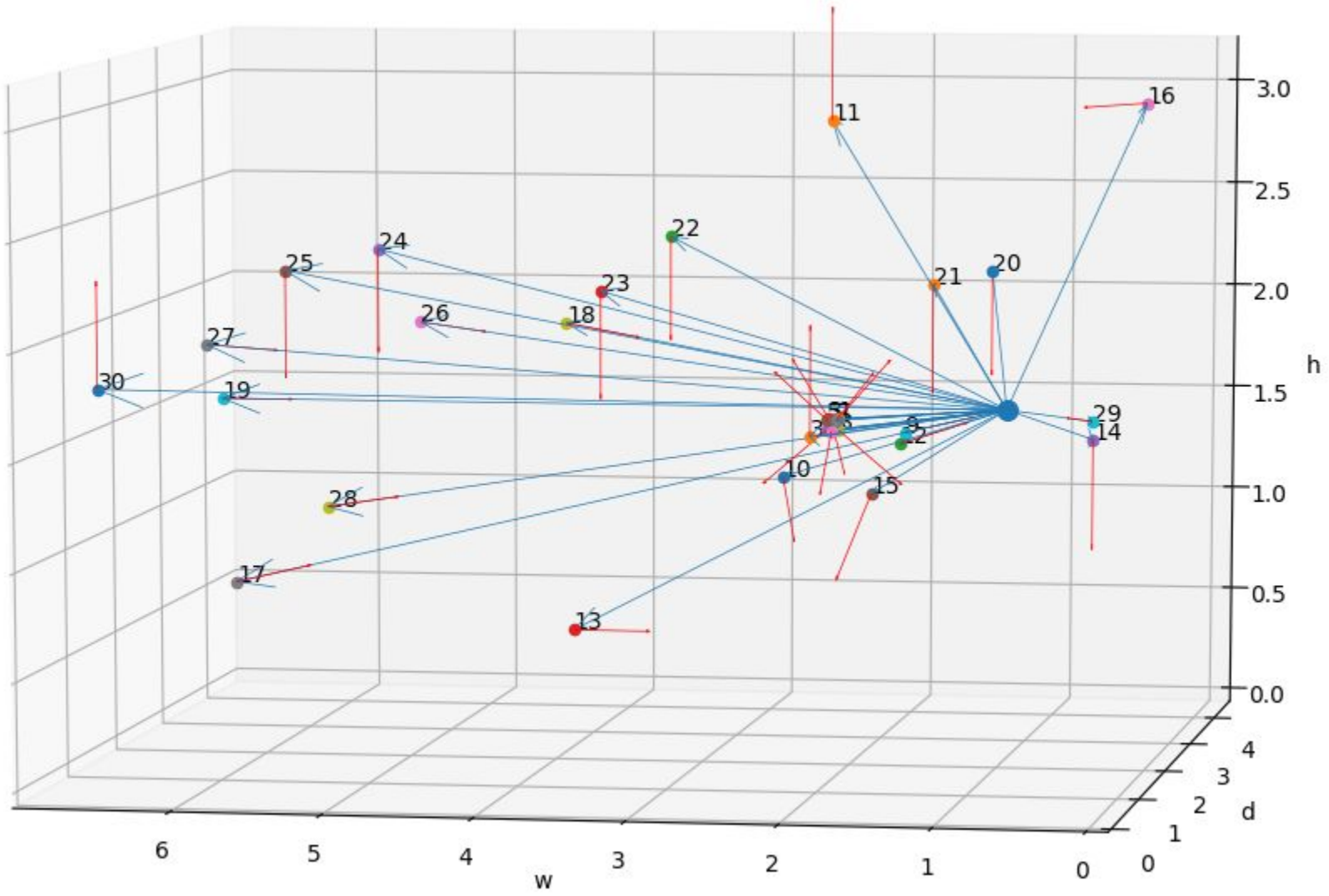
How

- To take it seriously we made a recording “protocol”
  - Measure the room size, material, etc
  - Position of the speaker
  - Position of microphones (delay compensation)
  - Set mic gains
  - Take photos
- Visualisation tool
- Absolute & relative coord.
- It takes ~5 hrs to setup a room
- And ~3 hrs to dismount.

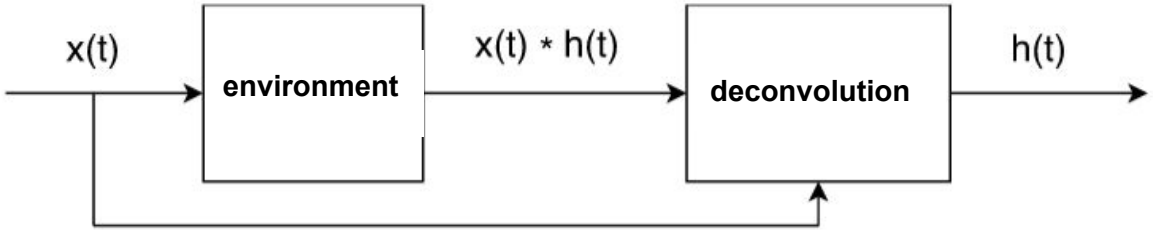
L207 - Speaker positions



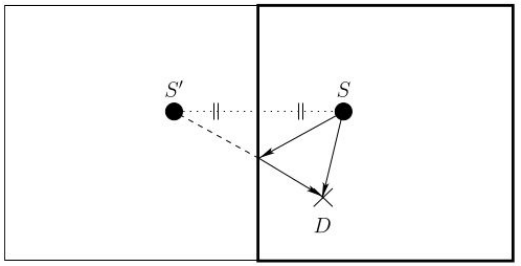
L207 - Microphones positions to SpkID01



RIR estimation



- Maximum length sequence (MLS) - “real RIR”
  - White noise like
  - h(t) is product of circular cross-correlation of y(t) and x(t)
  - Expects the same clocks (synchronized input and output) - bad for our case
- Exponential sine sweep (ESS) - “real RIR”
  - Sine with increasing frequency (exponentially to overcome some distortions)
  - h(t) is product of convolution of y(t) and inverse filter
  - It works fine for our case
- Image source model (ISM) - “artificial RIR”
  - Numerical way how to calculate a RIR given room dimensions, spk+mic coord., reflec. coef.
  - Cannot simulate obstacles
  - Can get “infinite” number of them

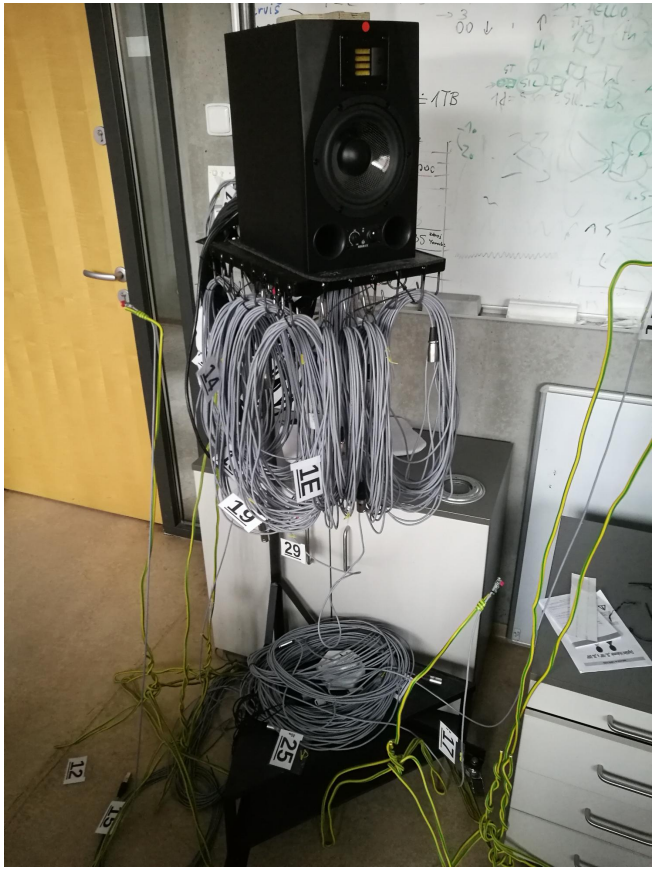


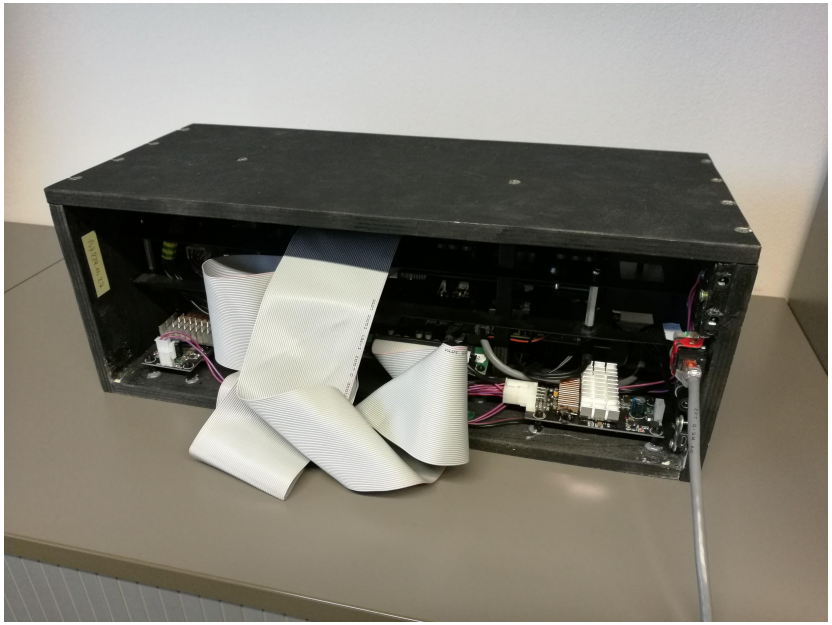
What to record

- Everything (we do not want to re-setup the room again)
- Real-RIR
  - MLS - Maximum length sequence - (bad) - few of them
  - ESS - Exponential sine sweep - good - 1s to 30s
- Silence :) = environmental noise
- Speech data
  - A Czech test set (not public :( ) - few hours
  - English Librispeech Test Clean (public :)) - few hours
  - English NIST SRE 2010 subset (not public :( ) - 2 days :(
  - The Czech train set - to fill space if possible
- Any other ideas???

Tools

- Synced 32 chs
- 48kHz 24 bit
- Soft gain
- Run for 3 days



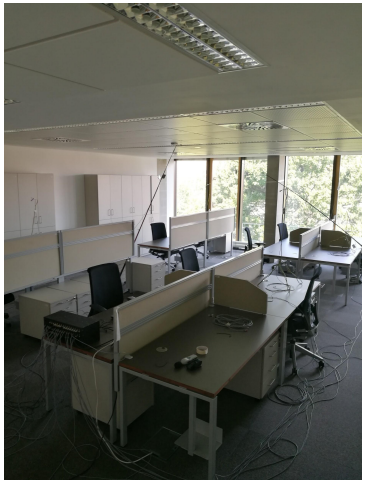
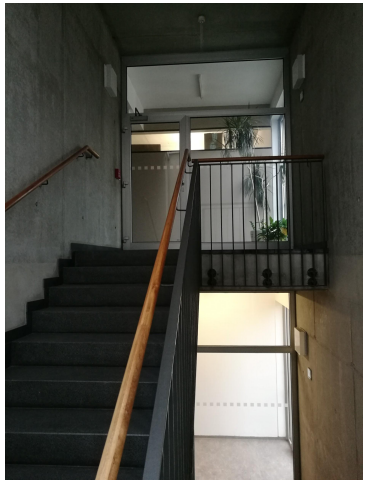


^^ BUT Workshop on Room ^^
Acoustics Measurement

Stojan Jakotytch






RIRs collected (so far)

- 8 rooms
- 14 test sets
- 50 RIRs
- times 31 microphones
- = 1550 RIRs

Room	Size	#Tests	#RIRs	#Mics
VUT_FIT_L212	middle	2	5	31
VUT_FIT_Q301	middle	4	6	31
VUT_FIT_D105	large	2	7	31
VUT_FIT_E112	large	1	3	31
Hotel_SD_R112	small	0	5	31
Hotel_SD_ConferenceRoom 2	large	0	4	31
VUT_FIT_L207	middle	3	9	31
VUT_FIT_L227	large	2	11	31

ASR Experiments

Czech - retransmission experiment

- Decent DNN based ASR, trained on 400hrs, incl. reverb and noises
- Scoring uses fixed segmentation
- Baseline 75.5% WAC

English - retransmission experiment

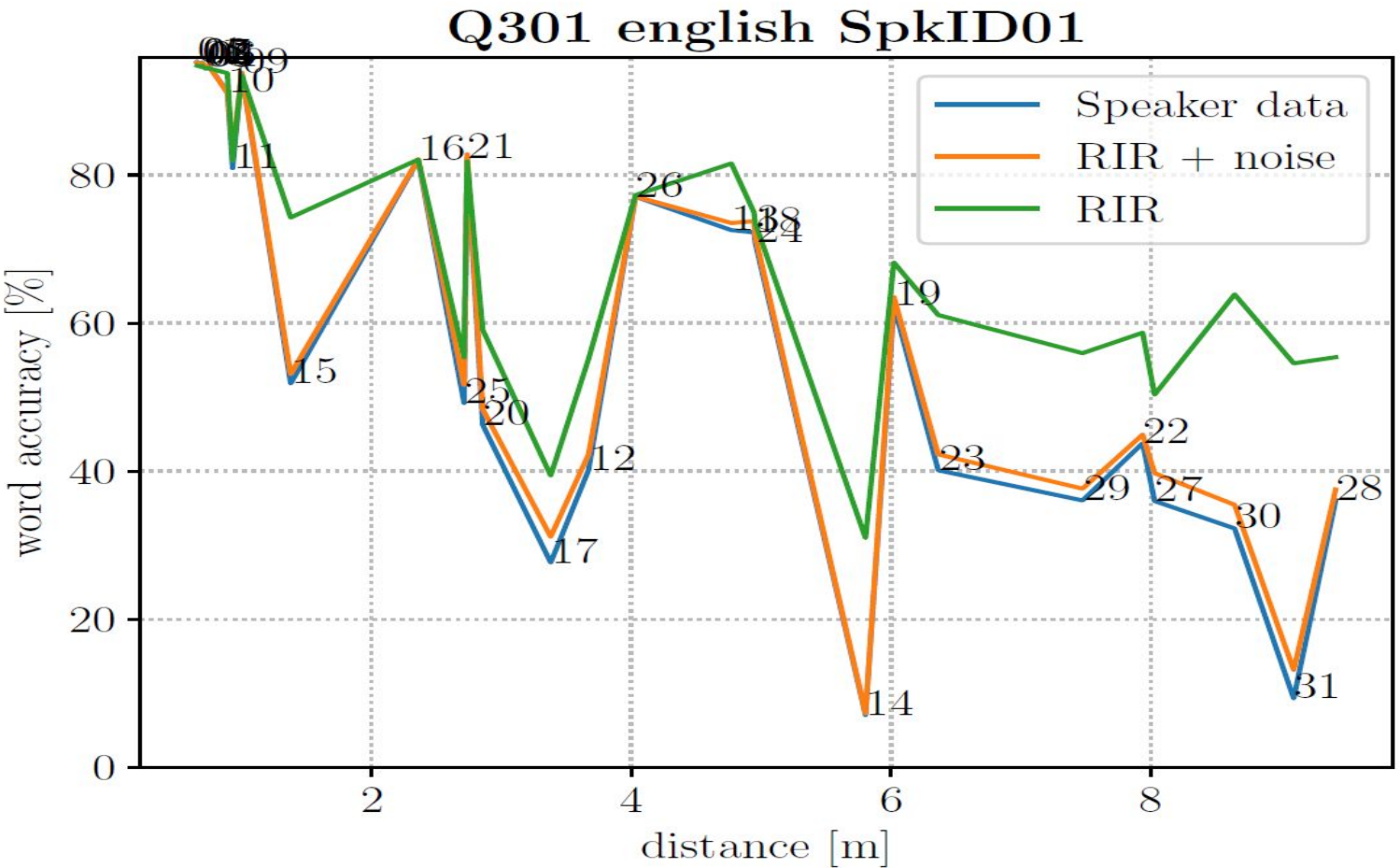
- Librispeech - Standard Kaldi recipe
- Baseline 95.86% WAC

English - simulation experiment

- AMI - Standard Kaldi TDNN recipe
- SDM Baseline 39.6% WER

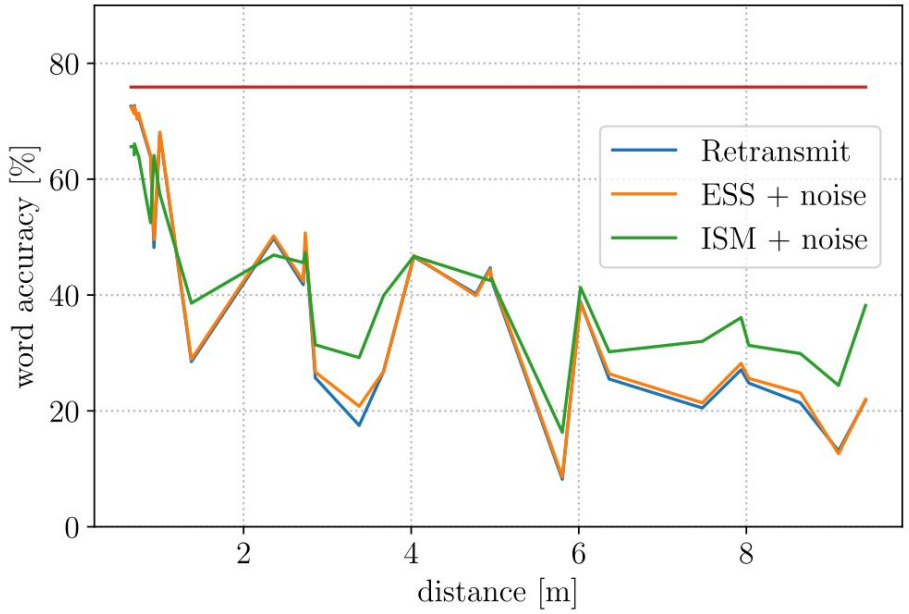
The Experiment on English

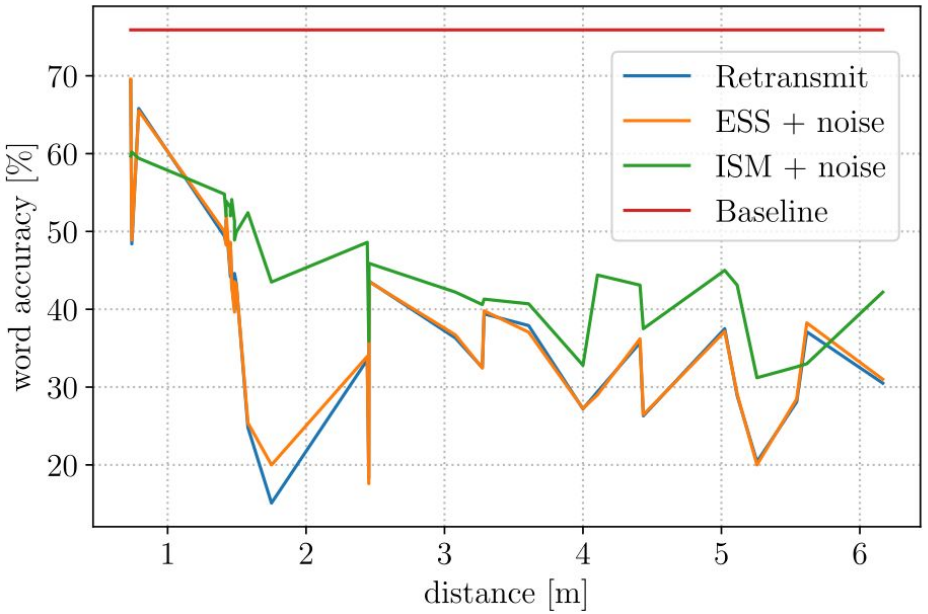
- Speaker data -> Retransmitted data
- RIR -> Exponential Sine Sweep (“real” RIR)



ESS to ISM comparison (Q301, L207)

- The experiment on Czech
- ISM -> Image Source Model (“artificial RIR”)
- ESS -> Exponential Sine Sweep (“real RIR”)
- Baseline on playback data (headset)





The AMI Experiment - still running

- Standard Kaldi TDNN recipe
- ISM -> Image Source Model (“artificial RIR”) 450x (RND 2-5 x 2-5 x 2-6)m
- ESS -> Exponential Sine Sweep (“real RIR”) 190x
  - 4 rooms (3.1x4.6x6.9, 2.6x6.9x10.8, 2.6x2.8x4.4, 3.1x4.6x7.5)m
- Noise -> Environmental noise recorded in the 4 rooms

Train \ Test	SDM Eval (WER)
IHM	70.9
IHM+ISM	57.7
IHM+ISM+Noise	**49.0**
IHM+ESS	55.1
IHM+ESS+Noise	**49.2**
SDM	39.6

Conclusion

- It works (the laboratory setup)
  - We can get close to real retransmitted data using RIR + noise
  - It is stable
- It is comparable to artificial RIRs (ISM method) so far
  - But not fully comparable setups
- We are using it in SID (Odyssey and ICASSP 2018 papers)
- The big question
  - **Are you interested? Any suggestions?**

2 rooms freely available here:

http://speech.fit.vutbr.cz/software/but-speech-fit-reverb-database