Predicting Automatic Speech Recognition Performance over Communication Channels from Instrumental Speech Quality and Intelligibility Scores

Laura Fernández Gallardo[1], Sebastian Möller[1], John Beerends[2]

[1]Quality and Usability Lab, Technische Universität Berlin, Germany

[2]TNO, The Hague, Netherlands

Introduction

- The ASR performance based on transmitted speech depends on the signal quality
  - channel impairments: bandwidth limitation (NB, WB, SWB) and codec

- Can WER be predicted using degraded speech?
  - from subjective intelligibility test scores
  - from POLQA-intelligibility
  - from speech quality (POLQA-MOS)

Speech Data Preparation

- ASR experiment: 322 words from 3 recording sessions of 37 speakers (16m, 21f) of the AusTalk database (Australian English). Clean speech sampled at 44.1 kHz
- Intelligibility test: 8 VCV logatomes: "ama", "aba", "afa", "ana", "apa", "asa", "awa", and "ascha“ from 4 German speakers (2m, 2f). Clean speech sampled at 48 kHz
- 19 transmission channel conditions

Required: fs ≥ 32 kHz for SWB transmission

- Simulated channels with ITU and ETSI tools:
  - Narrowband (NB, 300–3400 Hz)
  - Wideband (WB, 50–7000 Hz)
  - Super-wideband (SWB, 50–14000 Hz)
  - codecs of the corresponding bandwidth applied at different bitrates

ASR and Intelligibility Test

ASR Experiment using CMU Sphinx 4

- Separate context-dependent Hidden Markov Models (HMM) were trained and tested for each transmission condition
  - Train: words from recording sessions 1 and 2
  - Test: words from recording session 3

Intelligibility Listening Test

- 30 German listeners (15m, 15f)
- closed set of logatomes with 8 alternatives
- speech stimuli presented randomly

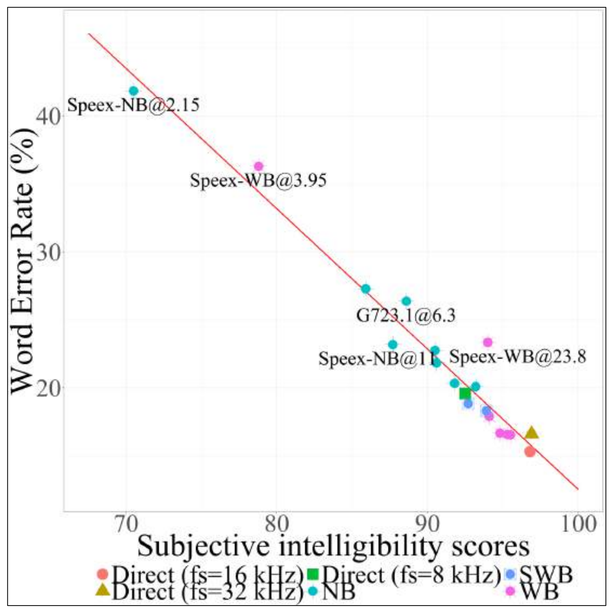
To the best of our knowledge, no other intelligibility model (e.g. SII, STOI) can satisfactorily predict subjective scores under the effects of communication channel degradations such as POLQA-intelligibility

POLQA MOS, the only ITU-T standard for estimating SWB speech quality

Distortion	WER	subj-intell	obj-intell	MOS
8kHz,nocodec	19.6	92.5	96.3	3.68
G.711@64	20.1	93.2	92.9	3.43
G.723.1@6.3	26.4	88.6	91.2	2.64
GSM-EFR@12.2	21.8	90.6	90.9	2.76
AMR-NB@4.75	27.3	85.9	85.1	2.03
AMR-NB@12.2	22.8	90.5	91.5	2.91
Speex-NB@2.15	41.8	70.5	82.0	1.86
Speex-NB@11	23.2	87.7	91.6	2.81
Speex-NB@24.6	20.3	91.8	92.2	3.05
16kHz,nocodec	15.3	96.8	99.8	4.43
G.722@64	16.6	95.5	97.7	4.14
AMR-WB@12.65	17.9	94.1	97.0	3.43
AMR-WB@23.05	16.7	94.8	97.5	3.78
Speex-WB@3.95	36.3	78.8	87.6	1.55
Speex-WB@23.8	23.4	94.0	95.2	3.79
Speex-WB@42.2	16.6	95.3	95.8	3.93
32kHz,nocodec	16.6	96.9	99.1	4.66
G.722.1C@24	18.9	92.7	97.2	3.46
G.722.1C@48	18.3	93.9	97.2	3.78

second-order curve fit: R^2 = .870, RMSE = 2.10

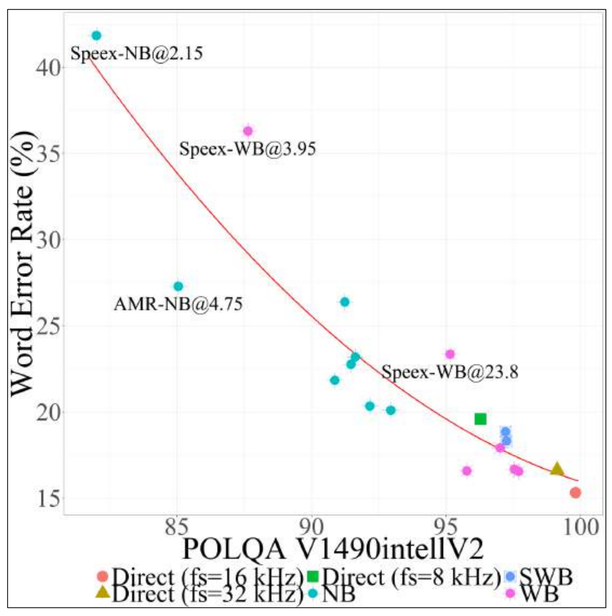
Predicting WER from Intelligibility and Quality



Linear fit to predict WER using subjective intelligibility scores

$$l(x) = 115.7 - x$$

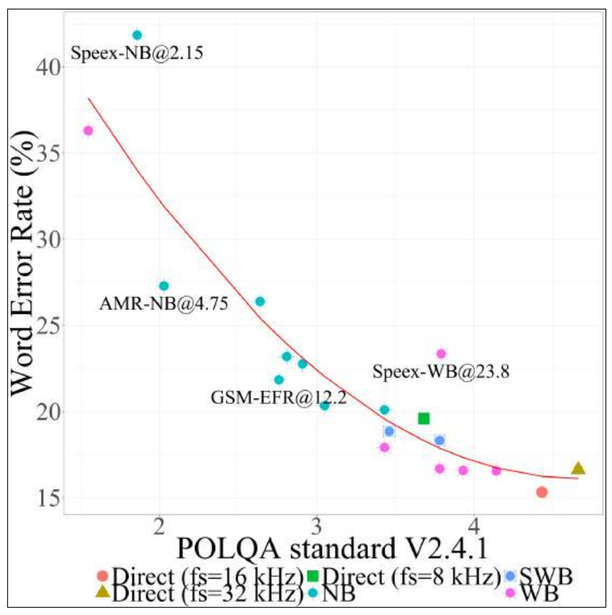
R^2 = .951, RMSE = 1.49



Second-order polynomial fit to predict WER using POLQA-intelligibility values

$q_{Pinte}(x) = 22.1 - 26.3x + 5.1x^2$

R^2 = .834, RMSE = 2.75



Second-order polynomial fit to predict WER using POLQA MOS values

$q_{MOS}(x) = 65.7 - 21.3x + 2.3x^2$

R^2 = .853, RMSE = 2.67

Conclusions

- ASR, MOS, and intelligibility improve in the transition from NB to WB, but no clear advantage of SWB
- Subjective intelligibility scores can be satisfactorily predicted from the ASR performance (reducing costs of listening tests)
- Polynomial fits can be useful to evaluate the merits of a channel codec for communications (reducing efforts of conducting ASR experiments)
- Future work: create more precise estimation models with state-of-the-art ASR systems and more SWB codecs (subject to the availability of clean speech databases of sufficient bandwidth, fs >= 32 kHz)

Quality & Usability Lab

Technische Universität Berlin