ETSI

World Class Standards

STF 294 Phase I:

Subjective Test Plan

ETSI Workshop on Speech and Noise in Wideband Communication

Laetitia Gros — France Telecom

Jan Holub — MESAQIN.com

Requirements

- Subjective tests should provide a database with a wide range of possible impairments, both for objective algorithm training and validation
- The original sample database recorded by STF 294 contains 4320 conditions → not feasible to run listening tests on all of them
- → Additional rules for qualified database size reduction identified and confirmed by listening tests:
  - only one speaker per gender is considered
  - AMR WB for noises perceived in mobile use (roads, crossroads and car) and G. 722 for office and cafeteria.
  - for road and cafeteria noises, speech signals are either inaudible or unintelligible with hands-free recording → removed
- → 432 samples per language

Methodology – Adopted from ITU-T P.835

- Appropriate for noisy (de-noised) speech
- Each trial contains three presentations of one sample + silent voting period of 4 sec. Each sample is 4 s in duration (about 1 s of background noise alone, 2 s of speech + noise, 1 s of background noise)
- Total duration of single trial: 24 seconds
- First two presentations: Listeners rate either the signal or the background depending on the rating scale order specified for that trial. For the signal, subjects are instructed to attend *only* to the *speech signal* and rate the speech. For the background, subjects are instructed to attend *only* to the *background* and rate the background
- The third presentation: Subjects are instructed to listen to the speech + background and rate it
- The order of the rating scales is balanced across the experiment

Methodology – Questionnaires for Listeners

Session 1 Block 1 Trial 1

Attending **ONLY to the SPEECH SIGNAL**, select the category which best describes the sample you just heard.

the **SPEECH SIGNAL** in this sample was

5 – NOT DISTORTED

4 – SLIGHTLY DISTORTED

3 - SOMEWHAT DISTORTED

2 – FAIRLY DISTORTED

1 – VERY DISTORTED

Session 1 Block 1 Tr ial 1

Attending **ONLY to the BACKGROUND** , select the category which best describes the sample you just heard.

the **BACKGROUND** in this sample was

5 – NOT PERCEPTIBLE

4 – PERCEPTIBLE BUT NOT ANNOYING

3 – SLIGHTLY ANNOYING

2 – ANNOYING

1 – VERY ANNOYING

Select the category which best describes the sample you just heard for purposes of everyday speech communication.

the **OVERALL SPEECH SAMPLE** was

5 – EXCELLENT

4 – GOOD

3 - FAIR

2 – POOR

1 – BAD

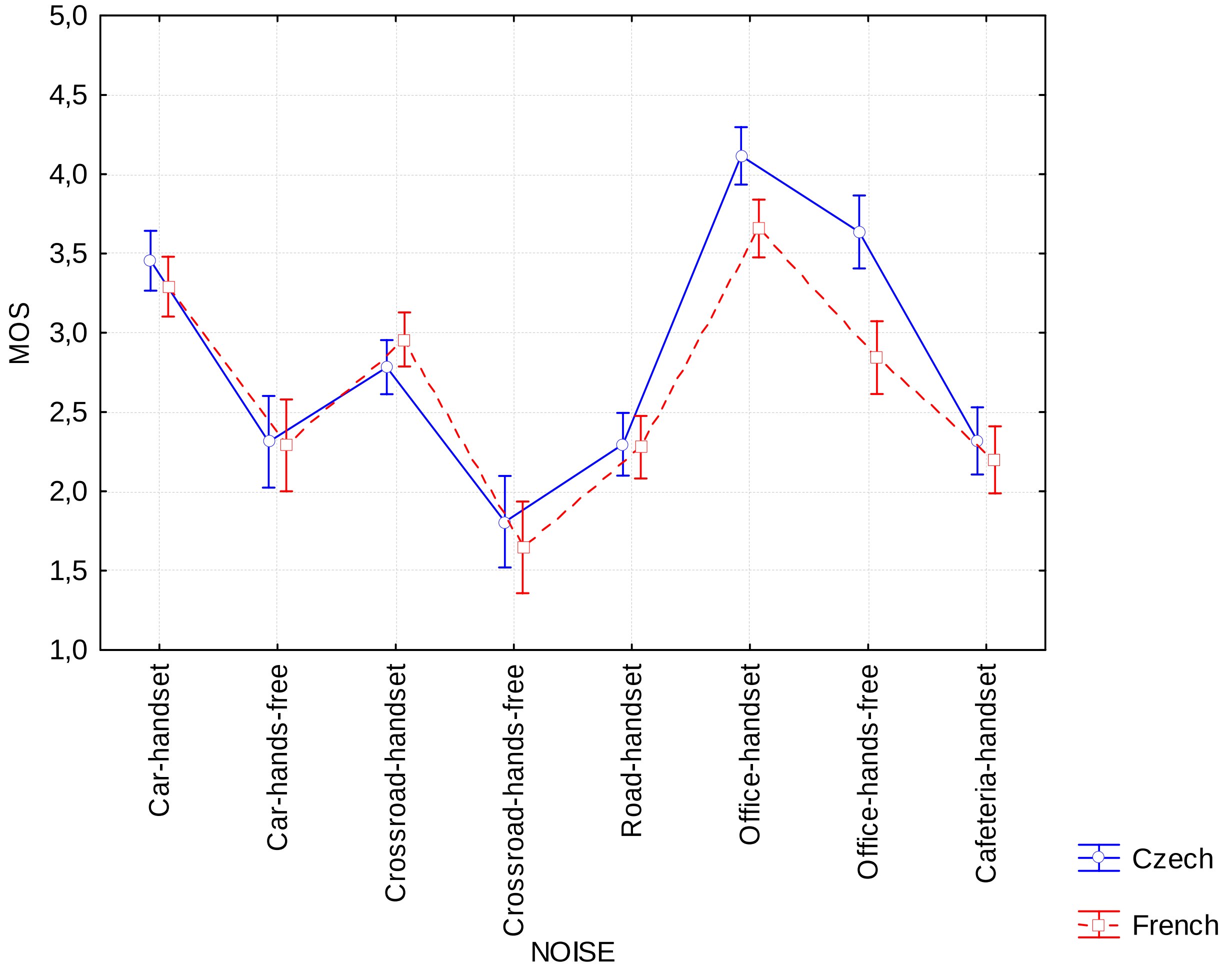
Subjective Test Conditions

- 48 naive listeners (24 listeners per language)
- High-quality headphones with low-noise digital playback system
- MOS scores must then be split into two parts: for algorithm training (about 70%) and for the algorithm validation (30%)
- The distribution between training and validation part tested by STF 294 to verify the equality of MOS and standard deviation distributions with respect to the entire database, according to a Kolmogorov-Smirnov test. The proportionality of occurrence of each parameter is tested to be as close as possible to 70:30 ratio
- Differences in test methodologies: speech content (8x three sentences vs. 24 x one sentence per condition), listening levels

Subjective Tests – Result Evaluation

- Overall MOS and the associated 95% CI calculated for each noise and for each language
- For each noise environment, the handset and hands-free recording has been analyzed separately
- The results match well between the two tested languages: The maximum differences occur for both Office noise conditions and do not exceed 0.8 MOS
- The subjective scores between app. 1.5 to 4.3 cover the typical quality range

Subjective Tests – Result Example



Thank you for your attention !

Questions ?