YRRSDS 2009
Session III
Metrics & Standards

Metrics

- **Normalized** count of indicators on a scale (min/max, over time)
- Types
  - **Objective**: task completion (reach a certain dialog step), call length, turn counts, turn length barge-ins, timeouts (lack of user input)
    - some **counter-indicative**: short length always good? hang-ups after a good call good, after a bad call bad? turn-length low for yes no, long for credit card collections.
    - correlation between "poor-mans" indicators and gold-standard metrics (True Total, CROUT+CAIN) can make evaluation cheaper
  - **Subjective** : people who call back are happy
    - **non-user judgement** of dialogue
    - some **counter-indicative**: task-perception, users didn't know that they were faster/slower, expert users may be faster than (still happy) novice users
  - **Physical** (e.g. face muscle activity)
  - **Macro-business** measures: number of calls/hits, entry points (where did the call terminate)

Standards

- **Industry** largely
- ITU-R recommendations for system evaluation
- voice-community.de - industry-lobby initiative.
- **How/what to log** to arrive at standard evaluation?
  - Log **everything**! (audio, system/ASR logs, version of system, grammars, n-best lists, caller-id).
- **Design/implementation standards**
  - VoiceXML (limited)
  - No usability standards exist (very vague)