Speaker selection for the SmartKom diphone database

BITS workshop Munich

Antje Schweitzer

Institute of Natural Language Processing

University of Stuttgart

Overview

- "Iterative" speaker selection
- Speaker selection for the SmartKom diphone database
  - steps
  - results
- Summary

Voice requirements

- users' expectations
  - agreeable, pleasant, and natural voice
  - intelligibility?
  - adequateness for specific application
- additional requirements
  - experienced speaker?
  - multilinguality?
  - speaker availability, contract issues

Naturalness, pleasantness

- unpredictable from original voice
- subjective
- influenced by
  - spectral consistency
  - constancy of voice quality

Adequateness

- Is there a specific target application?
- Is the voice suitable for this application?
  - male or female?
  - young or old?
  - e.g., does the voice "fit" the visual appearance of a given avatar?

Iterative selection process

- from speakers' demo material:
  - subjective impression of (original) voice
  - (multilinguality)
  - (level of experience)
- from recording sessions with the speaker
  - level of experience, adaptability

Iterative selection process

- from analysis after recordings
  - better comparison for subjective impression (same material, same recording environment)
  - spectral consistency
  - voice quality
- from evaluation of test synthesis voice
  - robustness to concatenation, signal manipulation
  - naturalness, pleasantness

SK speaker selection: 1st step

- collect demo material from 40 speakers (11 male, 29 female)
- demo material contained
  - 3 diphones embedded in nonsense words
  - all German vowels
  - very short dialogue containing English names
  - excerpt from a movie critique

SK speaker selection: 2nd step

- listening test with 13 phonetically trained participants
- dialogue containing English names
- excerpt from movie critique
- subjective ratings on a 5−point scale (very good to very bad)
- additional free comments

SK speaker selection: 3rd step

- record test database for best 10 candidates (4 male, 6 female) and build synthetic voice
- evaluation with listeners (20 expert listeners with experience in speech technology, 37 naive listeners)
- audio−only and audio−visual stimuli
- original and rule−based prosody
- different synthesis techniques (MBROLA, PSOLA, waveform interpolation)

Results: ranking of original voices

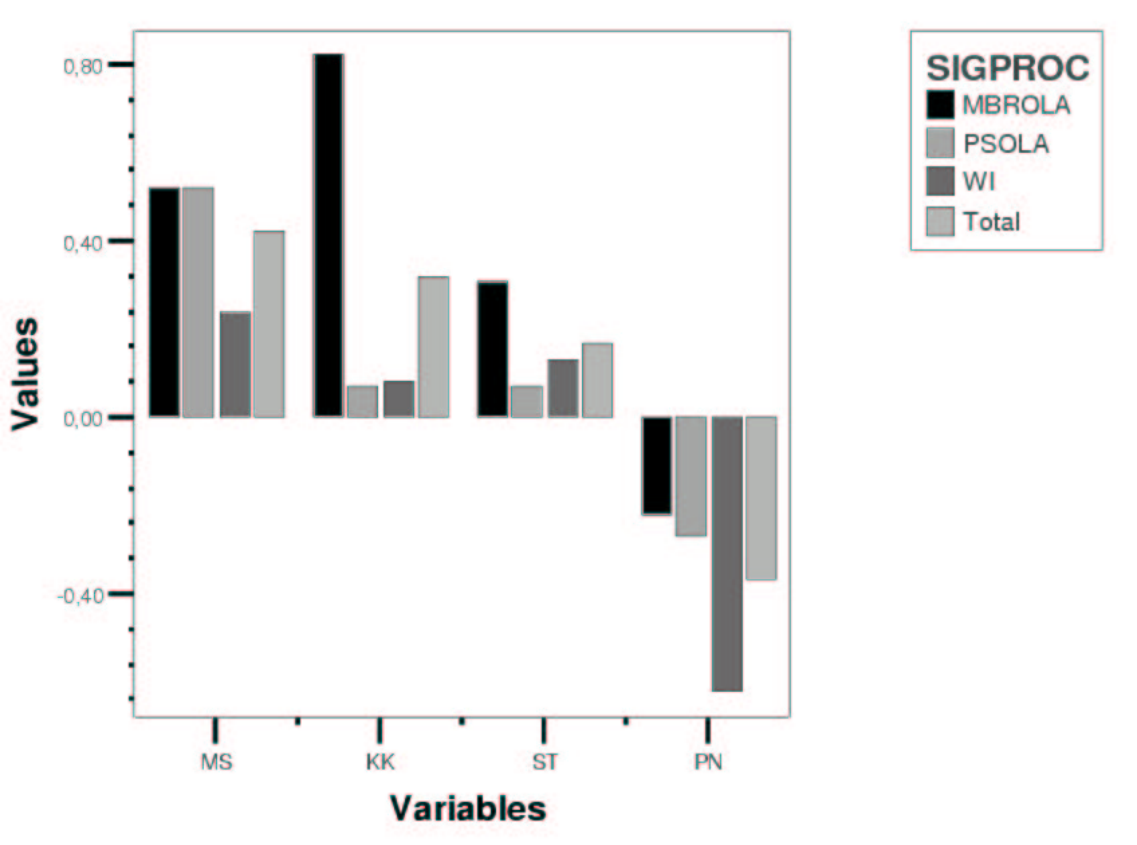
male voices

ST
PN
KK
MS

female voices

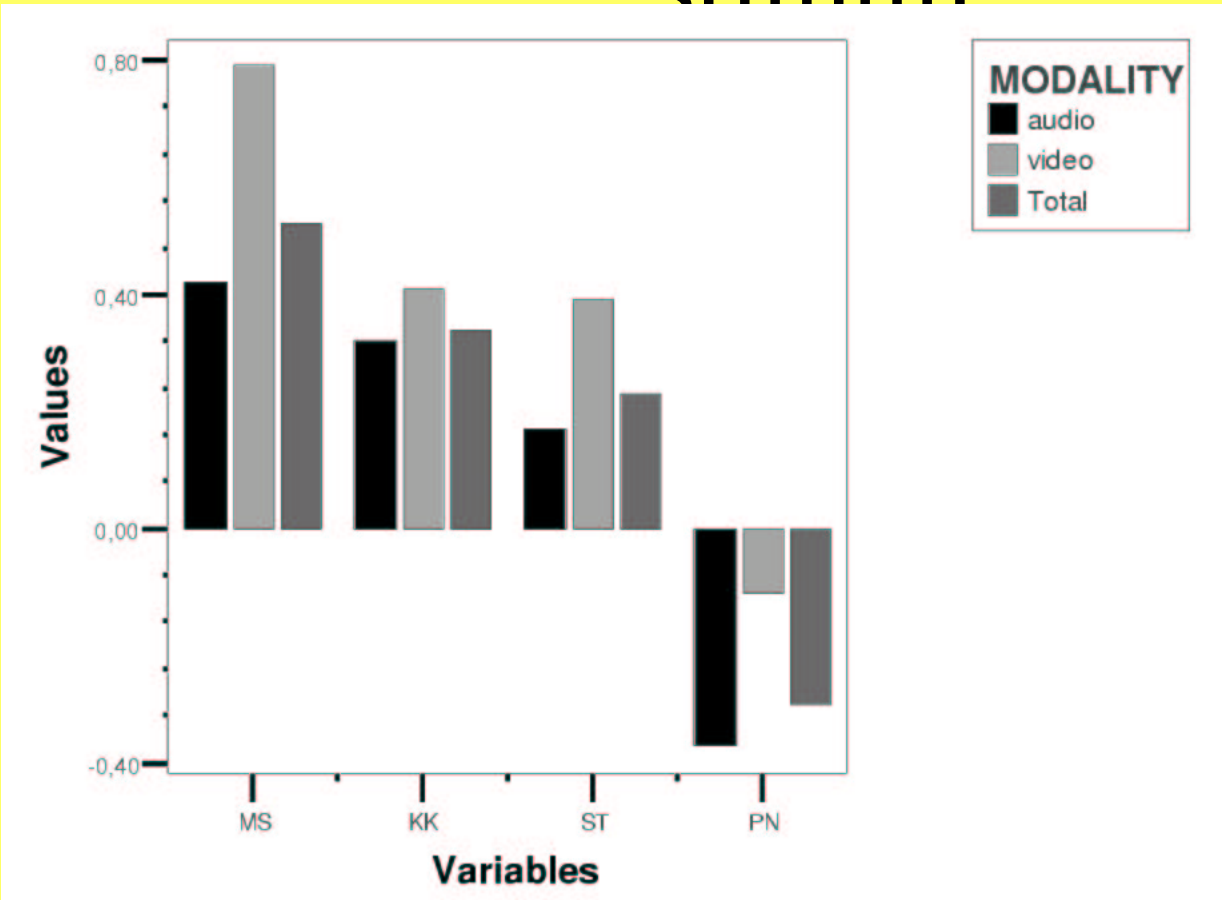
EO
KS
UM
BQ
BC
MT

Ranking of male synthetic voices



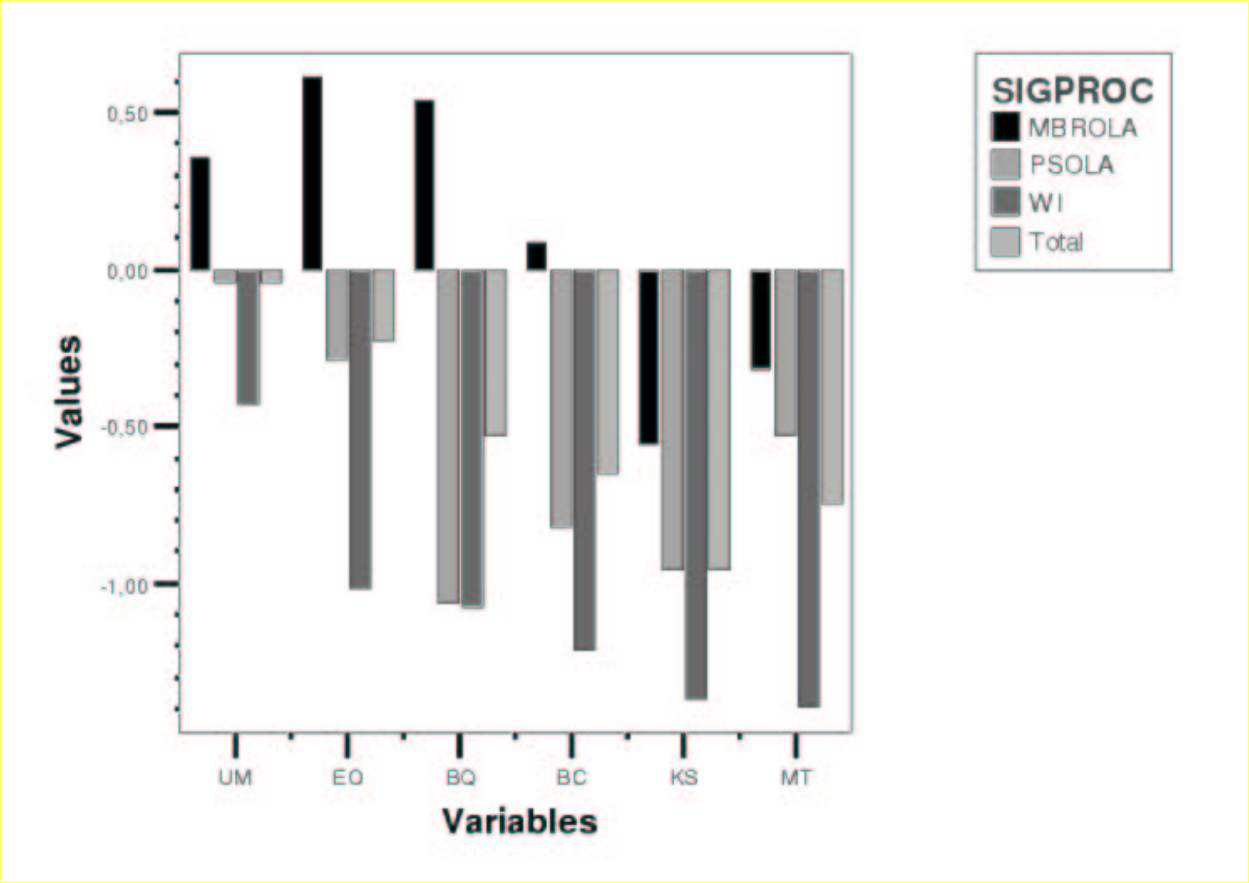
4 3 1 2 original rank

Male voices – audio–visual stimuli



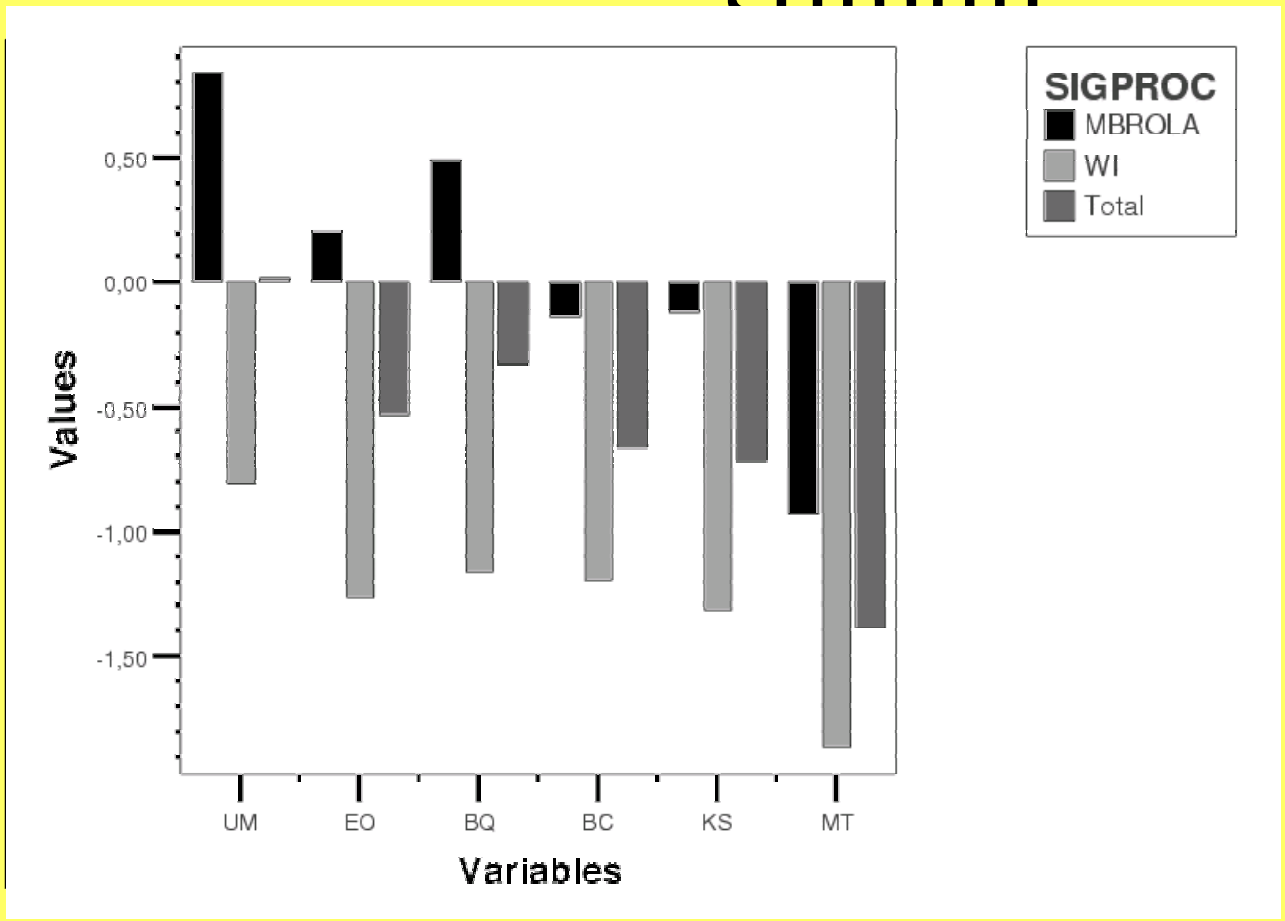
4 3 1 2 original rank

Ranking of female synthetic voices



3 1 4 5 2 6 original rank

Female voices – audio–visual stimuli



3 1 4 5 2 6 original rank

Summary

- It is currently impossible to predict the quality of a diphone voice from acoustic parameters
- The selection process should involve recording a small inventory for some sample sentences
- Some voices are not equally good for different synthesis methods
- Not every "good" synthetic voice is suitable for a given avatar