Perception of audiovisual speech produced by human and virtual speaker

SVEN ALLER (UNIVERSITY OF TARTU, ESTONIA)

EINAR MEISTER (TALLINN UNIVERSITY OF TECHNOLOGY, ESTONIA)
Background

In human face-to-face communication both auditory and visual channels are important.

Particularly important is visual channel in noisy environment and for hearing-impaired people.

Increasing use of virtual talking heads has incurred the necessity to evaluate the intelligibility of audiovisual synthetic speech.

Previous study (Meister et al. 2014) explored the perception of the Estonian 3D talking head, current study explores the perception of human audiovisual speech and compares results with previous study.
Questions

How big is the impact of visual information on speech understanding in Estonian?
◦ In different conditions?
◦ By different phonemes?

What are differences between synthetic audiovisual speech (3D head, previous tests) and the human speaker?

The results of the study will be necessary for better animation of the articulatory movements of the Estonian 3D talking head
Perception tests (1)

Nonsense words in the form of **vowel-consonant-vowel** (e.g. /amma/, /inni/, /ukku/...)

3 different vowels:
- unrounded open back vowel /a/
- unrounded close front vowel /i/
- rounded close back vowel /u/

13 different consonants in 7 classes:
- bilabials /m, p/
- labiodentals /f, v/
- alveolars /l, n, r, s, t/
- postalveolar /ʃ/
- palatal /j/
- velar /k/
- glottal /h/
Perception tests (2)

6 tests
- 3 audio-only (one for each context vowel)
- 3 audiovisual (one for each context vowel)

In every test 169 stimuli (all together 6 x 169 = 1014) in random order

Stimuli with no background noise and with four pink noise levels with signal-to-noise ratio (SNR) +6dB, 0dB, -6dB, -12dB

11 people took part
- Age 24-58
- 6 female, 5 male
Perception tests (3)

Examples

AV stimulus, no noise
/illi/

Audio stimulus, no noise
/unnu/

AV stimulus, SNR −12dB
/appa/

Audio stimulus, SNR −12dB
/appa/
Tajutestid

Test 1. Audiovisuaalne könesüntees: videotest (kontekst a)

Küsimus nr. 2/169

© Sven Aller 2015-2016
Results (1)

Comparing **audiovisual** and **audio-only** tests

- Visual information is much more important in noisy environment

![Graph showing the comparison of audiovisual and audio-only tests under various noise conditions. The graph illustrates the identification score under different signal-to-noise ratios (SNR) and noise conditions. The scores decrease as the noise level increases, with audiovisual tests generally performing better than audio-only tests.](image)
Results (2)

Different context in audio-only tests

- More difficult is recognize consonants in the context of closed vowels, especially /i/
Results (3)

Different context in audiovisual tests
- The context of closed vowels is still a problem: closed lips do not allow to see the tongue
Place of articulation of consonants

bilabials /m, p/ - both lips
labiodentals /f, v/ - lower lip against upper teeth
alveolars /l, n, r, s, t/ and postalveolar /š/ - tongue tip against teeth ridge
palatal /j/ - tongue blade against hard palate
velar /k/ - back of tongue against soft palate
glottal /h/ - vocal fold closure in larynx

Results (4)

Confusion matrices (noisy environment (SNR -12dB)):
- audio-only (top)
- audiovisual (bottom)

Stimuli are in rows, responses in columns
3D talking head

Prototype (made in Massy, only in Internet Explorer, needs Cortona3D Viewer):
Results (5)

3D talking head vs. human speaker

- Better intelligibility with human speaker, especially in auditory tests
Results (6)

3D talking head vs. human speaker: comparison of consonant classes
Conclusions

Visual information in conversation is very important
- Mean: 0.83 (audiovisual) vs. 0.63 (audio-only)

Particularly useful is visual channel in noisy environment
- Audiovisual: 0.99 (no noise) vs. 0.57 (SNR -12dB)
- Audio-only: 0.99 (no noise) vs. 0.19 (SNR -12dB)

Listeners perceive better natural stimuli than synthetic stimuli
- Audio-only: 0.19 vs. 0.08 (SNR -12dB)
- Audiovisual: 0.57 vs. 0.50 (SNR -12dB)

Consonant classes alveolars, postalveolars and labidentals need better animation in the 3D head model
Paldies!
Ačiū!
Tänan!