Perception of audiovisual speech produced by human and virtual speaker

SVEN ALLER (UNIVERSITY OF TARTU, ESTONIA)

EINAR MEISTER (TALLINN UNIVERSITY OF TECHNOLOGY, ESTONIA)

Baltic HLT 2016 | Riga, Latvia | October 6-7, 2016

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 /š/
- 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/

AV stimulus, SNR -12dB

/appa/

Audio stimulus, no noise

/unnu/

Audio stimulus, SNR -12dB

/appa/

Tajutestid

Test 1. Audiovisuaalne kõnesüntees: videotest (kontekst a)

Küsimus nr. 2/169

- affa
- ahha
- ajja
- akka
- alla
- amma
- anna
- appa
- arra
- ašša
- assa
- atta
- avva

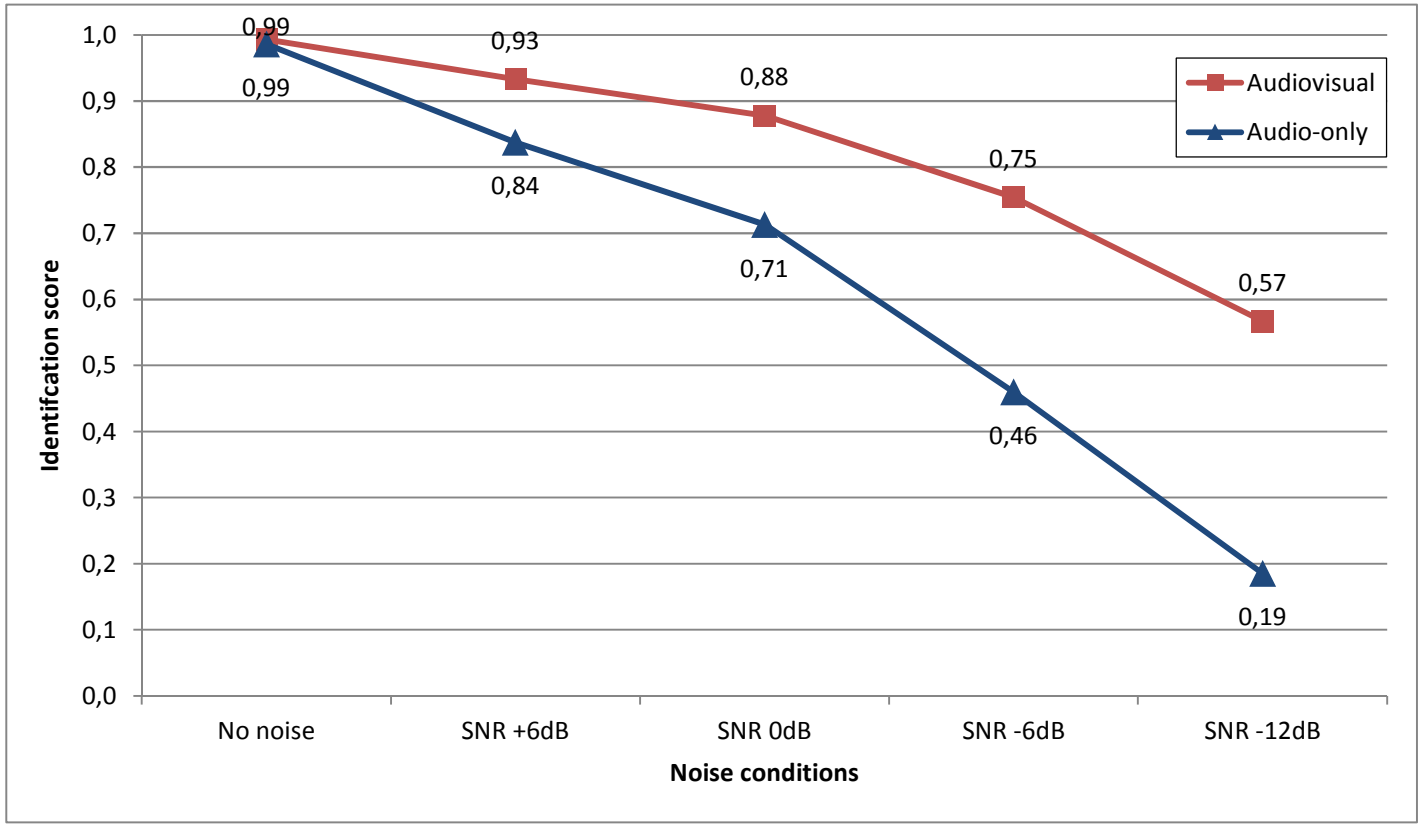
Järgmine

Kasutaja **PROOV**

Results (1)

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

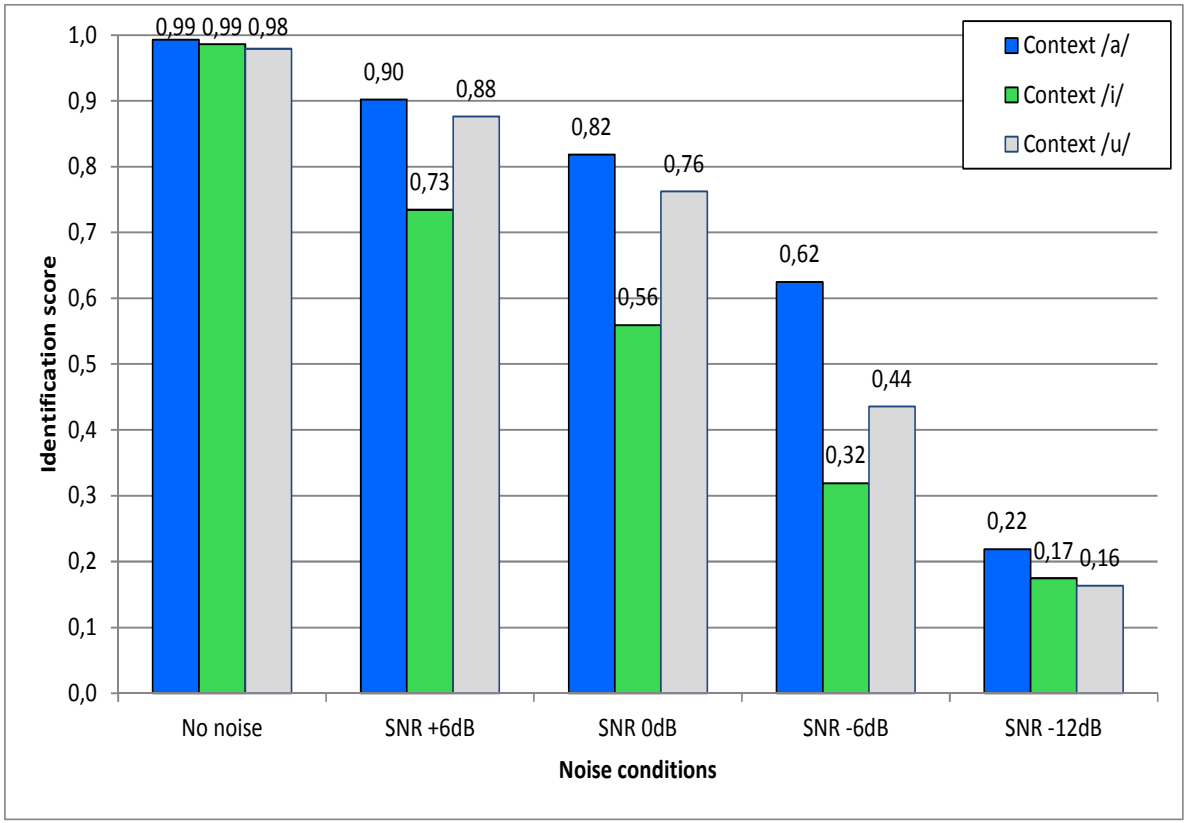
- Visual information is much more important in noisy environment



Results (2)

Different context in **audio-only** tests

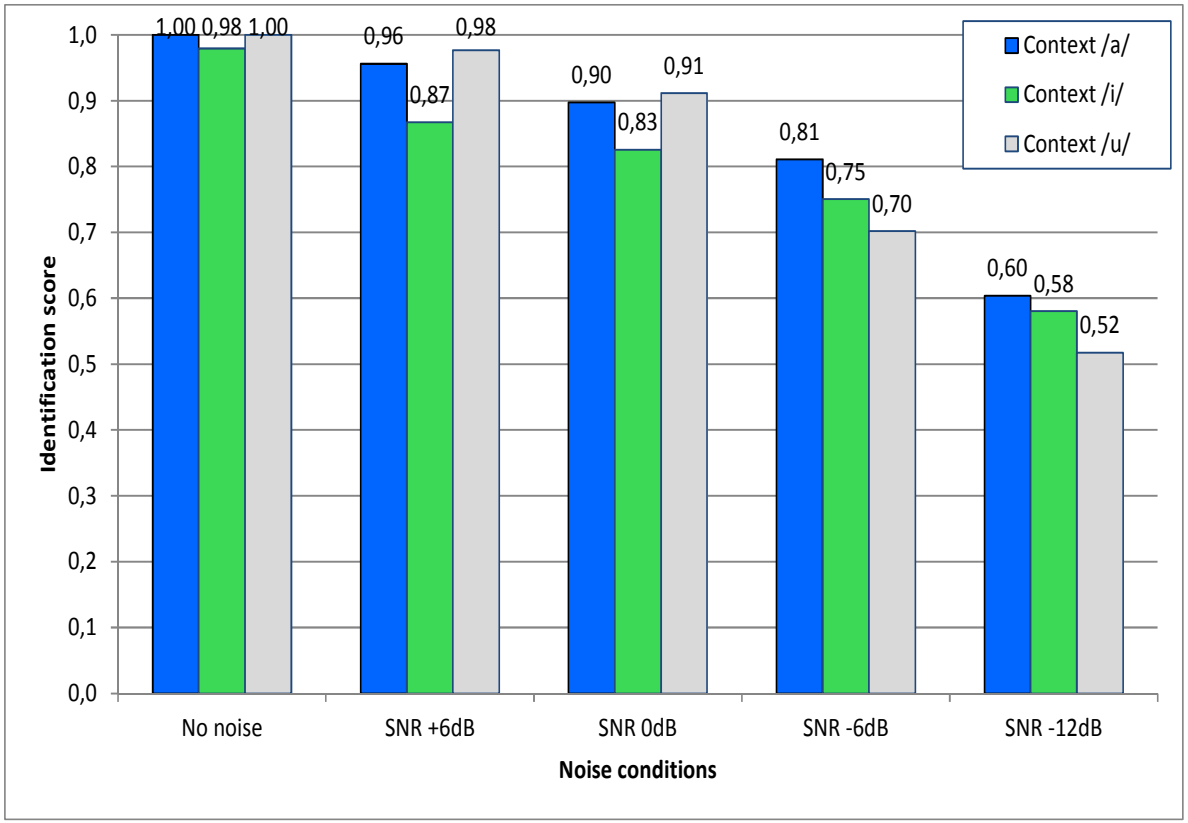
- More difficult is recognize consonants in the context of closed vowels, especially /i/



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

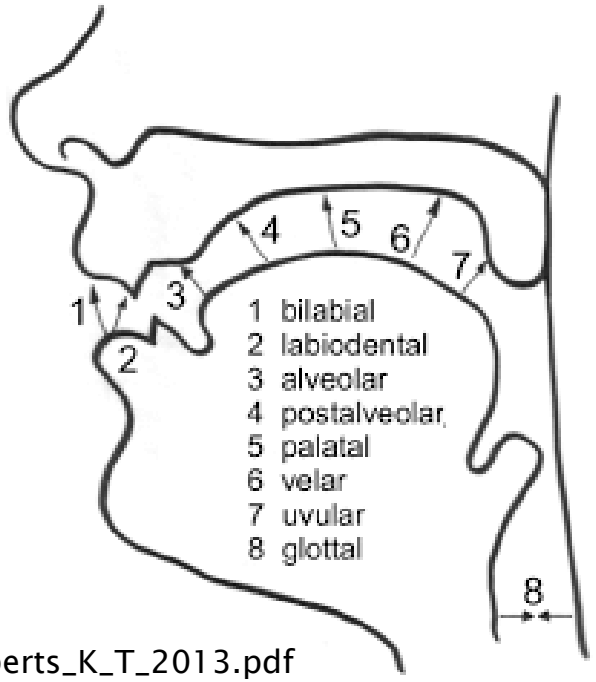
labiodentals /f, v/ - lower lip against upper teeth

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



https://vtechworks.lib.vt.edu/bitstream/handle/10919/23752/McRoberts_K_T_2013.pdf

Results (4)

Confusion matrices (noisy environment (SNR -12dB)):

- audio-only (top)
- audiovisual (bottom)

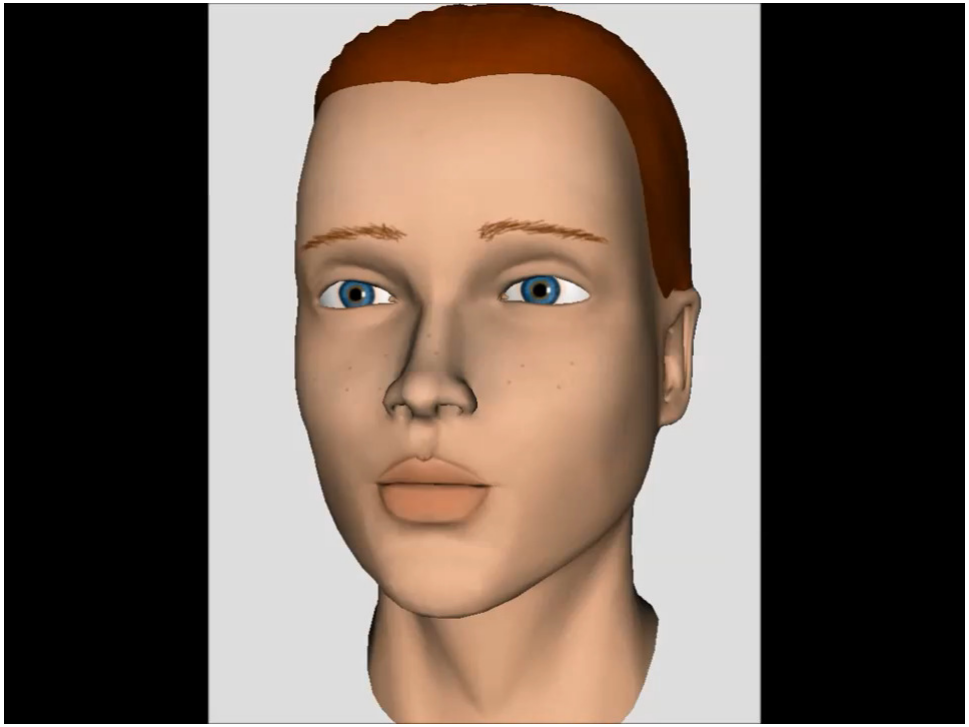
Stimuli are in rows, responses in columns

	m	p	f	v	l	n	r	s	t	š	j	k	h
m	**0,14**	0,05	0,02	0,14	0,19	0,08	0,14	0,03	0,01	0,02	0,09	0,01	0,07
p	0,07	**0,19**	0,03	0,22	0,1	0,02	0,03	0,03	0,08	0,05	0,03	0,04	0,1
f	0	0,09	**0,06**	0,07	0,04	0,03	0,03	0,09	0,24	0,13	0,01	0,14	0,06
v	0,03	0,15	0,02	**0,24**	0,08	0,01	0,09	0,11	0,07	0,04	0,02	0,04	0,09
l	0,06	0,03	0	0,12	**0,33**	0,04	0,05	0,03	0	0,05	0,26	0,01	0,01
n	0,21	0,04	0	0,05	0,18	**0,3**	0,01	0,01	0,03	0,03	0,11	0,01	0,01
r	0,04	0,08	0,02	0,13	0,06	0,04	**0,1**	0,06	0,06	0,12	0,18	0,04	0,06
s	0	0,05	0,09	0,03	0,01	0,01	0,02	**0,16**	0,09	0,25	0,11	0,05	0,12
t	0	0,09	0	0,03	0,01	0,03	0,06	0,22	**0,21**	0,2	0,02	0,08	0,04
š	0,05	0,1	0,04	0,05	0,01	0,05	0,03	0,18	0,1	**0,22**	0	0,04	0,12
j	0,08	0,06	0,02	0,14	0,14	0,05	0,05	0,05	0,03	0,03	**0,23**	0,05	0,06
k	0,02	0,09	0,02	0,06	0,09	0,04	0,05	0,1	0,17	0,1	0,09	**0,09**	0,07
h	0,02	0,07	0,09	0,17	0,04	0,03	0,09	0,05	0,06	0,05	0,07	0,13	**0,12**

	m	p	f	v	l	n	r	s	t	š	j	k	h
m	**0,66**	0,33	0	0,01	0	0	0	0	0	0	0	0	0
p	0,31	**0,68**	0	0,01	0	0	0	0	0	0	0	0	0
f	0	0	**0,69**	0,3	0	0	0	0	0,01	0	0	0	0
v	0	0	0,23	**0,77**	0	0	0	0	0	0	0	0	0
l	0	0	0	0	**0,67**	0,01	0,06	0	0	0	0,25	0,01	0
n	0	0	0	0	0,36	**0,41**	0,01	0	0	0	0,21	0	0
r	0	0	0	0,01	0,42	0,02	**0,21**	0,01	0,02	0	0,15	0,11	0,04
s	0	0	0	0	0	0	0	**0,55**	0,01	0,41	0	0,02	0,01
t	0	0	0	0,01	0,02	0,04	0,04	0,29	**0,48**	0,08	0	0,02	0,01
š	0	0	0	0	0	0	0	0,12	0	**0,88**	0	0	0
j	0	0,02	0	0,01	0,19	0,11	0,02	0,04	0,01	0	**0,56**	0,04	0
k	0	0,01	0	0	0,11	0	0,02	0,03	0,08	0,01	0	**0,51**	0,23
h	0	0	0	0	0,1	0,03	0,12	0,04	0,05	0	0	0,33	**0,32**

3D talking head

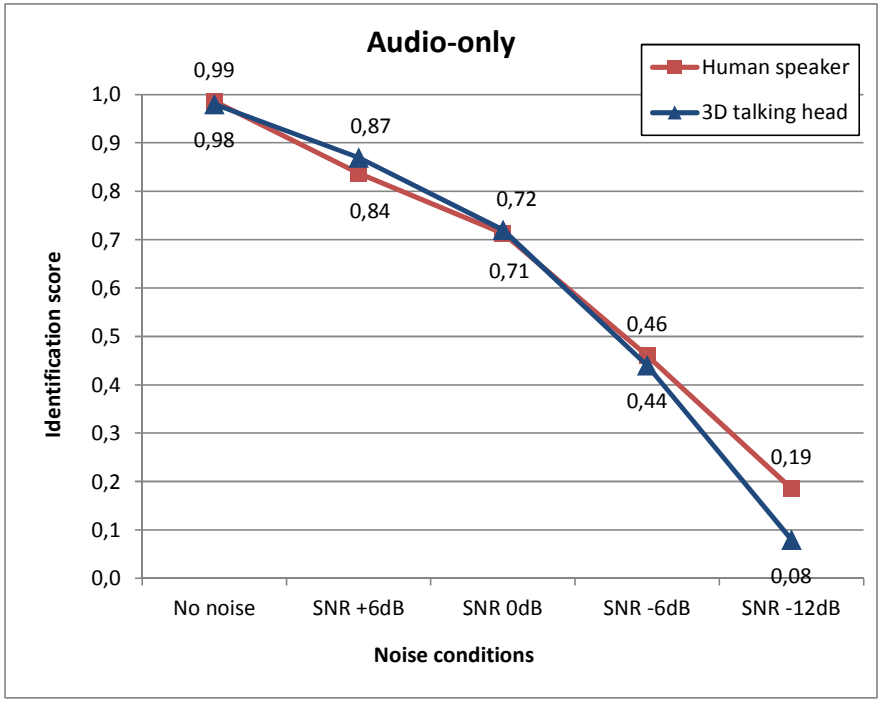
Prototype (made in Massy, only in Internet Explorer, needs Cortona3D Viewer):
http://massy-est.phon.ioc.ee/MASSY/peamudel.php

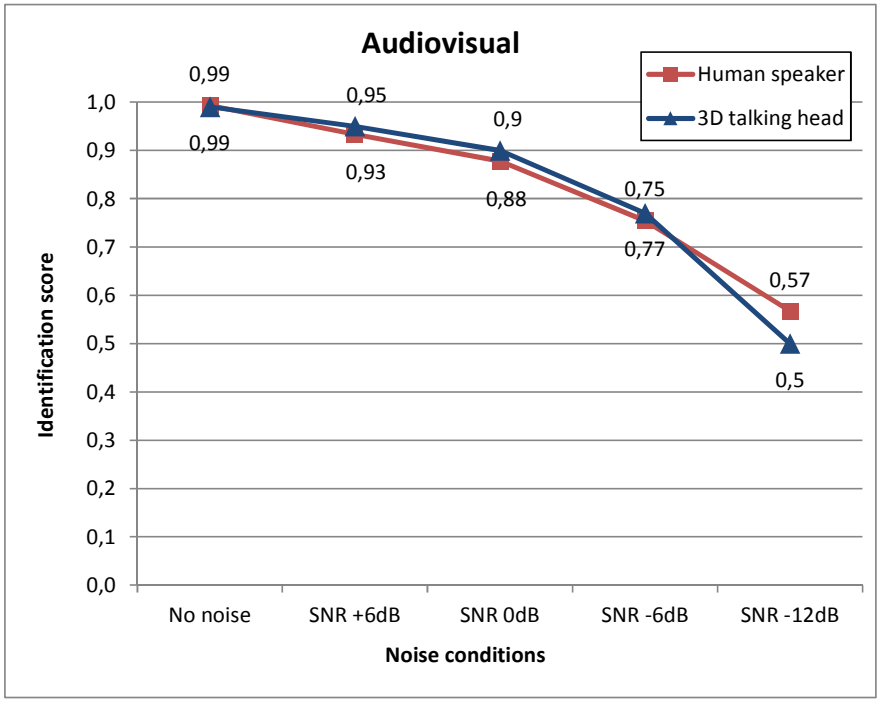


Results (5)

3D talking head vs. human speaker

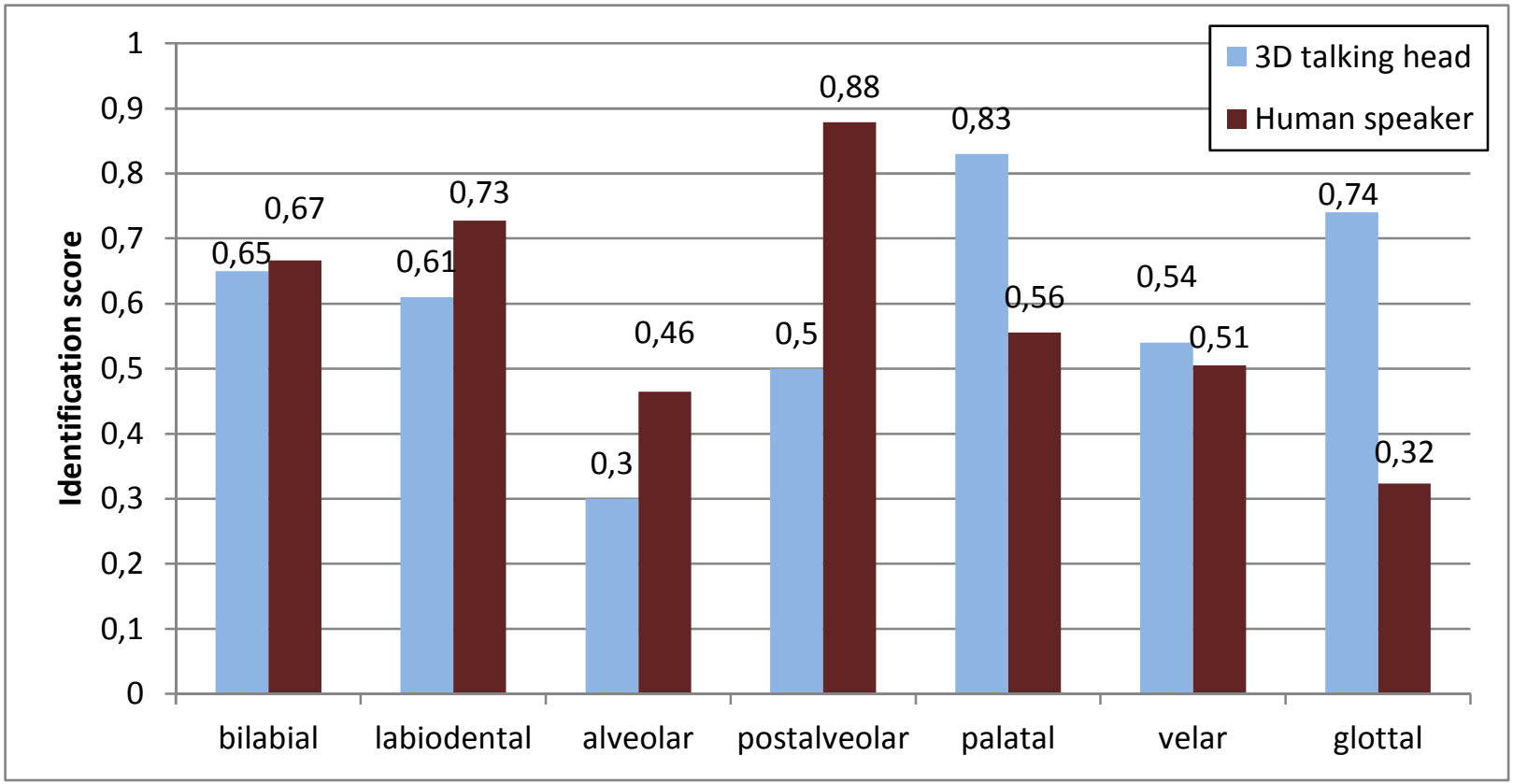
- Better intelligibility with human speaker, especially in auditory tests





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!