AUDITORY-VISUAL INTEGRATION IN THE PERCEPTION OF AGE IN SPEECH

Sascha Fagel

Berlin University of Technology
sascha.fagel@tu-berlin.de

ABSTRACT

The present experiment investigates the bimodal perception of age in audiovisual speech. Five female speakers of different age uttering one short sentence were recorded audiovisually. Audiovisual stimuli were created where auditory and visual information are coherent (i.e. from the same speaker) as well as incoherent (i.e. combinations of audio track from one speaker and video track from another speaker); furthermore the channels were split to create all unimodal auditory and visual stimuli. 60 subjects rated the speakers' age. The subjects were split into four subgroups. Three of them were presented with the audiovisual stimuli. One of these subgroups was instructed to rate the overall perceived age of each presented stimulus, the second subgroup should rate the age only by the voice they hear while still looking at the face but ignoring it, the third subgroup was instructed to rate the age only by the face they see while the voice was audible but should be ignored. The fourth subgroup as a reference for the perceived age of the single channels rated the unimodal auditory and visual stimuli.

Results reveal that subjects integrate both modalities if available in all three tasks, i.e. ratings of audiovisual stimuli clearly show correlations to each of the unimodal ratings even if one of the channels should be ignored. Additionally it could be shown that a) this effect is stronger (compared to the opposite case) if visual information should be ignored, b) in coherent stimuli the subjects rely more on the visual information, and c) the robustness of the visual modality exceeds that one of the auditory modality. Overall results give evidence for vision as the leading modality with respect to age perception in audiovisual speech.

Keywords: audiovisual speech, age perception, sensory integration.

1. INTRODUCTION

Speech production and speech perception is bimodal in nature, i.e. humans process both auditory and visual information – if present – when perceiving speech. It is known for decades that audiovisual speech leads to better recognition compared to pure audio speech [2] [9]. This effect is due to the fact that audition and vision contain partly complementary information cues that are jointly used when modalities are combined. Strong and well-known evidence for the sensory integration was found by McGurk & MacDonald [7] who showed that a visual syllable /ga/ combined with an audible syllable /ba/ mainly leads to the overall auditory perception of /da/, the so-called McGurk effect, and hence auditory and visual cues are both integrated into one percept even in incoherent stimuli. The integration process takes place whether or not the subject is aware about the effect. Regarding the perception of prominence of syllables, [11] showed that prosodic cues are also subject to auditory-visual integration in coherent and incoherent stimuli, where auditory information turned out to be more important than visual information. Regarding emotion perception, [2] showed that if both modalities are available subjects cannot ignore one of the modalities, and [4] showed that audition and vision transmit different cues that can be combined to an emotional percept neither present in the auditory nor in the visual channel (a real "emotional McGurk effect"). As information about a speaker's age is contained in the visual channel [8] as well as in the audio channel [6] it is interesting and has not been investigated yet how subjects integrate these information, if one of the channels plays a leading role and whether subjects can willingly ignore one of the information channels. In the present experiment unimodal as well as coherent and incoherent bimodal speech stimuli are used to investigate the auditory-visual integration in age perception.

Figure 1: The five female speakers at post-phonatoric rest position.

2. EXPERIMENTAL SETUP

2.1. Stimulus generation

[9] have shown that the presentation of isolated sentences lead to smaller estimation errors and smaller assimilation effects compared to isolated words. Hence, the short but phonetically rich sentence "Gustav kennt alle" ("Gustav knows them all") was created (59% coverage of German phonemes following the counting of German phonemes in written text in [5]). The sentence was uttered without special expression by five non-smoking female speakers (Fig. 1) aged 24, 39, 45, 53, and 61 years, respectively. The utterances were recorded audiovisually on a MiniDV camera. For each audio track a synchronized video of each of the speakers should be created. Therefore the audio tracks were manually aligned to the phone labels /gʊstafkɛntʔalə/ with praat [1]. For each of the five recorded utterances one video frame per phone was extracted (except for the glottal stop). That one nearest to the temporal center of the aligned phone was chosen. For the duration of the utterance about half of the frames remained in each video which preserved reasonably detailed motion. Fig. 2 shows the lip regions of the chosen frames for speaker 5. In a video editor these frames were lengthened, i.e. repeated, to best fill the duration of the respective phone. One neutral pre- and one neutral post-phonatoric frame were added. This was done for the five original combinations of audio and video tracks as well as for all 20 possible combinations of one speaker's video track with another speaker's audio track. This procedure ensures synchronized audiovisual stimuli where the original combinations are manipulated in the same way as the incoherent (audiovisually mixed) stimuli and hence coherent and incoherent stimuli should not differ in audiovisual synchrony. Audio alone and visual alone stimuli were created using the audio track of the five original recordings and the 25 created image sequences, respectively. The stimuli can be found on the ICPhS'07 DVD or at http://avspeech.info/avAgePerception.html.

2.2. Procedure

60 undergraduate students participated voluntarily in the experiment. Pre-tests had shown that a subject's rating of a presented face or voice depends on the rating of that face or voice in earlier presentations during the test. As faces and voices should be presented in different combinations within the experiment, the whole set of stimuli had to be distributed among subjects to present each face and each voice only once to a subject. Each subset contained one coherent and four incoherent stimuli in a per subject different random order. The subjects were asked to rate the age in an open choice after each stimulus presentation. One subgroup of 15 subjects should rate the speaker's age without special instructions, another subgroup was instructed to rate the age only by the voice they hear and to ignore the face while still looking at it, and a third subgroup was instructed to rate the age only by the face and to ignore the audible voice. After a subject of these three subgroups had rated the age of the five audio-visual stimuli it was mentioned that stimuli might be pairings of a face and a voice that did not originate from the same speaker. Then the same stimuli were played once more in the same order aiming at monitoring the subjects' awareness of discrepancy. This time the subject had to rate the coherence of the voice-face combination on a scale from 1 ("does not match at all") to 5 ("matches completely").

Figure 2: The frames representing the phones of the utterance (lip region only) of the fifth female speaker.

A fourth subgroup was presented with five audio only and five visual only stimuli and had to rate the age of the voice or the face, respectively. Stimuli of each modality were played in a per subject different random order. Answers were given in an open choice.

3. RESULTS

All ratings of a stimulus are averaged before analysis. The used method to present each face and each voice only once to a subject leads to only 3 ratings per stimulus at 60 subjects. Furthermore, comparisons between results for the audio only, visual only and audio-visual conditions and between the tasks to rate either the person, the face or the voice could only be made across subjects.

3.1. Perceived age

At first the ratings of the unimodal stimuli were analyzed to obtain references for the audiovisual ratings. The ratings in the audio only condition and the visual only condition are correlated at r=.90. The voice of speaker 3 is rated much younger (31.1 years) than her face (43.6 years). Linear regression shows that the speakers are rated younger in voice than in face by 4.8 years (Fig. 3). The correlation of the age ratings between audiovisual condition (coherent stimuli) and visual only condition is higher (r=.98) than between audiovisual and audio only condition (r=.90) which shows that subjects rely more on visual information than on auditory information when both modalities are available. Furthermore for four of the five speakers the standard deviations of the ratings are higher in the audio only condition than in the visual only condition (Tab. 1).

3.2. Rating tasks

Details of the results on the three rating tasks are displayed in Tab. 2. When the task was to rate the whole speaker's age subjects mainly relied on the visual information which is reflected by a very high correlation between the ratings in audiovisual condition and in visual only condition (.94≤r≤.98). Nevertheless the subjects integrated the auditory information to a high degree: the correlation between ratings in audiovisual and audio only condition was .49≤r≤.86.

The task to rate the face also leads to a very high correlation between ratings in audiovisual condition and in visual only condition (.86≤r≤.99). But the subjects still integrated the auditory information to a very high degree in case of the youngest and the oldest face (r=.97 and r=.79, respectively). An influence of the voice to the rating of a face is also present for the second oldest face (r=.33); for the other faces there is no relevant influence of the audio.

Results of the voice rating task nearly show the opposite: little influence of the visual information in case of the youngest face (r=.19), moderate influence in case of the oldest face (r=.42) and high to very high influence of the other three faces (.75≤r≤.90), while the correlation between the ratings in the audiovisual condition and the audio only condition are very high (.88≤r≤.99).

Figure 3: Age ratings in unimodal conditions.

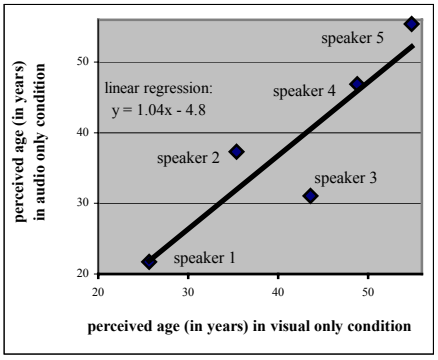


Table 1: Standard deviations of the perceived age in unimodal conditions sorted by speaker.

speaker	visual only	audio only
1	5.34	4.11
2	3.83	8.92
3	4.22	4.77
4	4.14	6.48
5	6.94	9.15
average	4.90	6.69

Table 2: Correlations between audiovisual and visual only and between audiovisual and audio only conditions for the three tasks sorted by speaker.

rating task	whole speaker		only the face (ignore voice)		only the voice (ignore face)	
conditions	av–v	av–a	av–v	av–a	av–v	av–a
Speaker 1	0.94	0.76	0.99	0.97	0.19	0.95
Speaker 2	0.95	0.49	0.96	0.11	0.75	0.88
Speaker 3	0.98	0.86	0.94	0.09	0.9	0.99
Speaker 4	0.98	0.74	0.89	0.33	0.83	0.97
Speaker 5	0.97	0.61	0.86	0.79	0.42	0.95

3.3. Awareness of discrepancy

As a measure of difference in age between the two channels for each audiovisual stimulus the absolute differences of age ratings of the according stimuli in both unimodal presentations were calculated. Fig. 4 shows that subjects were clearly able to rate the discrepancy between the voice and the face. The correlation between the ratings of discrepancy on the scale from 1 (no match) to 5 (full match) and the absolute difference of perceived age in the unimodal conditions was -.85.

3.4. Ratings and discrepancies

Fig. 5 shows the shift of rating of a face's age due to the presence of voice, i.e. how much older or younger a face is rated when the added audio is older or younger. The rating of the audiovisual stimulus is clearly shifted into the direction of the auditory age. A high correlation of r=.66 can be seen even though for two of the five faces (No. 2 and 3: open triangle

and open circle) no effect occurs (see Section 3.2). Linear regression leads to a 20% contribution of the auditory information to the rating of the face. Fig. 6 shows similar results for the voice rating task. The correlation is even higher (r=.74) than for the face rating task. As can be seen the contribution of the visual information to the rating of the voice in audiovisual presentation is 24%.

Figure 4: Absolute age differences between the two modalities of the audiovisual stimuli plotted against the ratings of discrepancy.

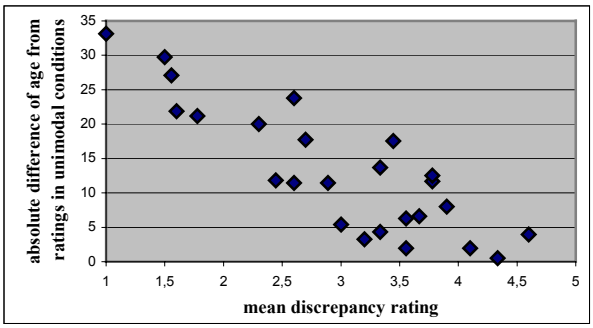


Figure 5: Relation between the age differences of the two channels of each stimulus and the resulting shift in age rating in the task to rate the face.

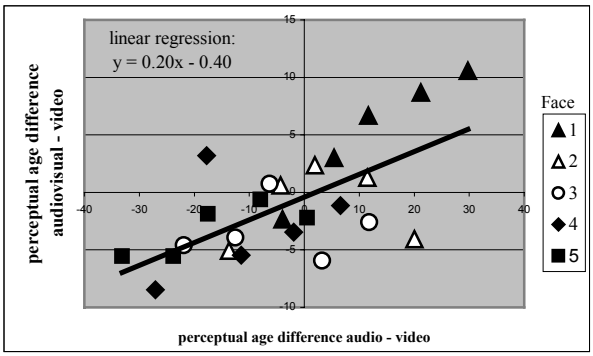
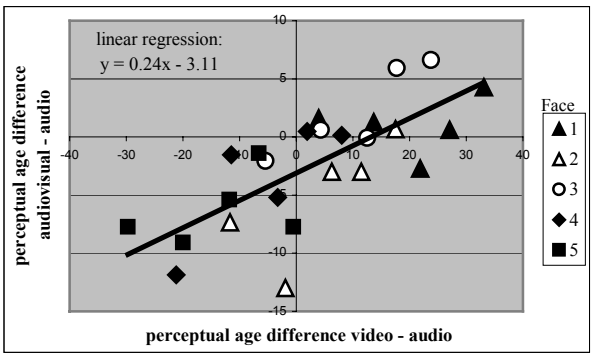


Figure 6: Relation between the age differences of the two channels of each stimulus and the resulting shift in age rating in the task to rate the voice.



4. CONCLUSIONS

In the present study speech stimuli were displayed to subjects where the age information in the auditory and the visual modality varied. It could be shown that subjects cannot avoid integrating the visual information if available even when instructed to rate only the voice: the higher (and the lower) the age of an additionally displayed face the higher (and the lower, respectively) is the perceived age of the voice. In the opposite case – when rating only the face – auditory information is strongly integrated in the same way for 3 of 5 faces (where the visual information is least reliable reflected in rather high standard deviation in the ratings of the visual only display). This also shows that subjects rely more on the visual than on the auditory modality in audiovisual speech perception when rating the age. These effects occur even though the subjects are well able to rate the discrepancy of age between the auditory and the visual information. The task to rate the whole person's age also yields consistent results: ratings are correlated to a higher degree with visual only ratings than with audio only ratings. An indication that the visual modality provides more robust information on age is shown by lower standard deviations in the visual only condition than in the audio only condition. The present experiment shows that age perception in audiovisual speech is a highly integrating process where the results provide evidence for the vision (compared to audition) as the leading modality.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2005. *Praat: doing phonetics by computer.* Retrieved 5/26/2005, http://www.praat.org/

[2] de Gelder, B., Vroomen, J. 2000. Bimodal Emotion Perception: Integration Across Separate Modalities, Cross-Modal Perceptual Grouping or Perception of Multimodal Events? *Cognition and Emotion* 14(3), 321-324.

[3] Erber, N. 1969. Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *JSHR* 12, 423-425.

[4] Fagel, S. 2006. Emotional McGurk Effect. *Proceedings of the Speech Prosody conference*, Dresden.

[5] Kohler, K.J. 1995. *Einführung in die Phonetik des Deutschen*. Erich Schmidt Verlag Berlin.

[6] Linville, S.E. 2001. Vocal Aging. Singular, San Diego.

[7] McGurk, H., MacDonald, I. 1976. Hearing Lips and Seeing Voices. *Nature* 264, 746-748.

[8] O'Toole, A.J., Vetter, T., Volz, H., Salter, E.M. 1997. Three-dimensional Caricatures of Human Heads: Distinctiveness and the Perception of Facial Age. *Perception* 26, 719-732.

[9] Ralston, J.V., Tse, M., Campbell, E.R., Wright, A.D., Fisher, T.L., McCall, M. 1994. Age Perception of Speakers of Isolated Words and Sentences. *JASA* 95(5), 3016.

[10] Sumby, W., Pollack, I. 1954. Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America* 26, 212-215.

[11] Swerts, M., Krahmer, E. 2004. Coherent and incoherent audiovisual cues to prominence. *Proceedings of the Speech Prosody conference*, Nara, 69-72.