This paper describes two experiments with video dubbing. Subjects had to identify a talker's utterances of CV-syllables within a sentence frame. The test syllables had conflicting bimodal information about place of articulation. By desynchronisation we wanted to examine the influence of timing phenomena with regard to different vocalic contexts. The results show a main effect of 'fused' answers only in an /a/-environment. A missing visual consonantal articulation as in the context of /u/ leads under certain conditions to a systematic elision of the initial acoustic stop consonant.

EXPERIMENTS

In two experiments on audio visual fusions we wished to test a 10-step desynchronisation continuum from 0 ms to 270 ms delay of the acoustic component following the visual component of the stimulus and the influence of the vowel contexts /i/, /a/, /u/. We used three acoustic realisations of each vocalic context for dubbing.

Subjects

39 untrained subjects participated in experiment I and 30 untrained subjects in experiment II.

Stimuli

a) Video recording. The recordings of a male speaker were done in a sound-treated studio using a Panasonic VHS system. Head and shoulders were visible on the monitor screen in a straight front picture. The talker was instructed not to move during recording. A 1000 Hz sinusoidal reference signal of 300 ms duration was generated periodically every 10 s by a PDP 11/50. Every time the speaker heard the signal, he had to utter the following sentence with one of the test syllables: "Ich habe /ba/, bi, bu, ga, gi, gu, gesagt". To avoid misleading information about the closing of the following stops, he was instructed not to close his lips after he had said "ich habe". Reference signal and utterance were recorded on the first soundtrack of the video tape.

b) The visual component of the stimuli. To obtain fusions we only used visual /ga/, /gi, /gu and acoustic /ba/, /bi, /bu utterances.

For the visual components of the stimuli we used only one realisation of each vocalic context. Cutting the tape was performed on Panasonic source and editing recorders with an editing controller unit.
in such a way that a visual sequence for each context was arranged according to the randomization plan.

(i) The acoustic component of the stimuli. Three realizations of /ba/ /wa/-/an/- were taken for dubbing the pictures. They were recorded on Revox and digitized and stored on a PDP 11/50 for later processing. Spectrographic measurements and auditory perceptions showed no perceptual difference. Therefore the three acoustic realizations could be counted as repetitions in the statistical evaluation.

(ii) Audio-visual dubbing of the stimuli. Editing the acoustic signals for dubbing was controlled on the PDP 11/50. The programs developed especially for experiments in the McGurk paradigm allowed the second soundtrack of the video tape to be dubbed with exact desynchronisation using the first period of the allophone reference signal and different segment files providing the necessary information. During the dubbing process the output of the reference signal was suppressed so that the tape tape received the sentence frames, the test syllables and the pauses in relation to the visible articulation and the randomizing plan.

**RESULTS**

![Fig. 1, 2, and 3 show the percentage of ‘fused’ answers, that is /dr/, for each of the ten steps of desynchronisation, for each of three acoustic realisations and for each vocalic context averaged over subjects.](image1)

In Experiment I we presented the same stimuli and used the same procedure as in Experiment II, but only one vocalic context /ba/, /wa/ or /an/ was presented in each of the three tests and no demonstration preceded the test. Here the total duration of one test was 5 minutes.

As already mentioned the acoustic realizations were counted as repetitions. The data were interpreted in a two-factorial design with the percentage of ‘fused’ answers as dependent variable. An analysis of variance with fixed effects and repeated measures showed that the interactions could be discounted (F18, 3680) = 1.37005, p = 0.153 > 0.05, that the influence of context is significant (F18, 3680) = 268.0048, p = 0.000 < 0.05, as well as desynchronisations (F18, 3680) = 2.5709, p = 0.007 < 0.05.

Within a Logit-Analysis a model without significant interactions but only main effects was found to be most suited to fit the experimental data.

(The Goodness-of-Fit Test: Likelihood Ratio Chi Square = 23.85962, DF = 18, p = 0.047 > 0.05. Pearson Chi Square = 23.57842, DF = 18, p = 0.169 > 0.05.)

In each factor (step) two steps could be evaluated in the analysis with the restriction it can be shown that in each of three vocalic contexts /ba/, /wa/ and /an/ the first and the second step of desynchronisation are significant at the 5% level. /wa/-context: r-value = 8.14265, lower 95% confidence interval = 0.15816, upper 95% confidence interval = 0.25463, /wa/-context: z-value = -19.92128, lower 95% CI = -0.65215, upper 95% CI = 0.35332, 0 ms desynchronisation r-value = 2.35506, lower 95% CI = 0.22155, upper 95% CI = 0.37371, 240 ms desynchronisation z-value = -19.92128, lower 95% CI = -0.65215, upper 95% CI = 0.35332.

A contingency analysis for each vocalic context showed that only in the /wa/-context had an effect be indicated. Because of the previously run over the same data, significant differences were found between the 1 and 2 level (1: -0.051 < 0.016). (Chi Square = 27.02, DF = 18, p = 0.004 < 0.016. For /ba/-context: Chi Square = 12.29, DF = 18, p = 0.051 < 0.016). This explains the missing interactions betwen desynchronisation and context. The Logit-Analysis of desynchronisations was confirmed by a different analysis using orthogonal contrasts. The differences between the average of the "fused" responses between step 1 and the other steps is significant (F1, 3680) = 4.76418, p = 0.019 < 0.05 as well as between step 9 and 10 (F1, 3680) = 6.39973, p = 0.007 < 0.05.

Summing these results it can be stated, that the variable "fusion" is dependent on the variability within the first and eighth step of the variable desynchronisation and on the within all three steps of the variable context. The influence of temporal shifts between visual and acoustic information is only relevant for the /ba/-context. After the test 33 subjects were interviewed. They were asked whether they had always heard an initially consonant, and which alternative of the second response category they preferred in each vocalic context.

Only three subjects heard mostly /ba/ in /ba/-context. no subjects heard /dr/, but 30 subjects heard with very few exceptions only the vowel /u/. Some subjects mentioned a hard glottal attack. The vowel alone was heard almost exclusively in /wa/-context but never in /an/-context.

Experiment II was a control experiment with regard to the mixed vowel condition in experiment I. Fig. 4, 5, and 6 show ![Fig. 4, 5, and 6 show](image2)
Fig. 6 (gibbi)

the percentage of "fused" answers of 30 subjects (10 subjects for each test)

The results agree broadly with the earlier test. But based on the following interview only one subject had heard only the vowel /u/, whereas 9 subjects heard /bu/. No differences occurred concerning the /i/- and /a/-contexts.

DISCUSSION

In the /a/-context subjects could observe a downward shift of the tongue body, which was clearly different from tongue movement in the /i/-environment ending at the teeth and therefore enforcing information about a dental place of articulation, which led to highly "fused" responses. As the statistical evaluation shows, the desynchronisation effect is only significant in the /a/-context. This might be explained by the specific tongue movement being clearly visible. In the /u/-context the protrusion of the lips totally masked all visible consonantal information. Therefore "fusions" did not occur. In the first experiment the /a/- and /i/-contexts provided visual consonantal information for fusions. In comparison to these stimuli the visual information in the /u/-context had a special effect: subjects heard no oral consonant at all. We call this effect "elision".

In the second experiment the different result of the interview concerning the /u/-context indicates that the visual component of the stimulus loses its influence, if no consonantal articulation can be seen. Here integration of the bimodal information did not happen and subjects realized the contradiction of the information presented by both modalities.

The effects of fusion and of elision in the /i/- and /u/-context have in common that they are not affected by desynchronisation within the temporal domain tested in the experiments. The effect of desynchronization is specific only to the /a/-environment. Further experiments are necessary to investigate the bimodal temporal relationships and the special effect of elision.

REFERENCES


ACKNOWLEDGEMENTS

Statistical advice was given by Dr. Alexander Yeassouridis, Department of Biostatistics, Max-Planck-Institute for Psychiatry, Munich F. R. G.