AUDIOVISUALLY PERCEIVED "FUSIONS" WITHIN DIFFERENT VOWEL CONTEXTS

JÖRG SCHORRADT HANS G. PIROTH HANS G. TILLMANN

Institut für Phonetik und sprachliche Kommunikation
der Universität München
Schellingstr. 3, 8000 München 40, F.R.G.

ABSTRACT

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 /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.

INTRODUCTION

H. McGurk and J. MacDonald 1976 [3] first described the effects which result from presenting utterances of CV-syllables with conflicting visual and acoustic initial consonants, namely that subjects will hear the consonantal part of the syllable as a function of the distribution of the consonantal information over the two modalities. For example a visual ‹ba› combined with an acoustic [ga] will be heard as a syllable containing both consonants /b/ and /g/ but a visual ‹ga› combined with an acoustic [ba] will be heard as a /da/. The first effect was called "combination", the second effect was called "fusion". Both these effects still work even if subjects know how they are achieved and they work after many repetitions as well.
By systematic temporal desynchronisation of the visual and the acoustic component of the stimulus Tillmann et al. 1984 [7] looked for the characteristic temporal limitations of the effects.
In accordance with earlier findings of audio-visual desynchronisation (Dixon & Spitz 1980 [1]) it could be shown again, that a desynchronisation with the acoustic stimulus leading the visual stimulus was generally more sensitive for being perceived as non-contradictorilly, than a desynchronisation with the visual stimulus leading. This greater tolerance for integrating conflicting information in the latter order as a manifestation of a general principle was refined by further experiments leading to so called "phonological fusions", for example visual ‹ba› and acoustic [la] are heard as /bla/. Here it could be shown that perceived phonotactically regular combinations are much more sensitive to desynchronisation than "fusions" are.
The experiments presented in this paper are concerned with the influence of desynchronization in different vocalic environments.

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 one visual sequence for each vocalic context was arranged according to the randomization plan.
c) The acoustic component of the stimuli. Three realizations of /ba/, /bi/, /bu/ were taken for dubbing the pictures. They were recorded on Revox and digitalized and stored on a PDP 11/50 for later processing. Spectrographic measurements and auditory comparisons showed no perceptual differences. Therefore the three acoustic realizations could be counted as repetitions in the statistical evaluation.
d) 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 desynchronization using the first period of the sinosoidal 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 test tape received the sentence frames, the test syllables and the pauses in relation to the visible articulation and the randomizing plan.

PROCEDURE

In experiment I we had 90 stimuli randomized over 10 steps of desynchronisation (0 ms - 270 ms), three vocalic contexts (/a/, /u/, /i/) and three acoustic realisations of each context. Vocalic context and desynchronisation counted as factors, the realizations as repetitions in a two-factorial design.
Subjects were tested singly in the speaker room of the recording studio. They sat at a distance of 3 m from a Sony colour monitor with a tube of 68 cm in diameter. Only one of the monitor's loudspeaker was used and was placed on top of the monitor. Therefore we must take into account about 10 ms delay of acoustic information relative to visual information due to the velocity of sound.
The subjects read an instruction-sheet before the tests started. They were asked to make a binary forced-choice decision about the identity of the initial consonant. On an answering sheet they could choose to mark a /d/ answer, which we interpreted as "fusion" or to mark a /b,g/ answer which we interpreted as no "fusion"
It was pointed out to the subjects, that a correct fullfillment of the experimental task was entirely dependent on the fact that they looked at the speaker's lips and answered which syllable they heard.
In a first demonstration the subjects had to respond to 18 syllables. The three acoustic realisations were presented with a 30-ms- and a 270-ms-desynchronisation for each vowel context, first /a/, then /i/, and finally /u/. After a short break which could be used for questions the test was started.
With a small pause after each sentence and a larger pause after each block of 20 sentences the total duration of the test was 15 minutes.
At the end of this test-run subjects were asked in a short interview to describe their impressions according to the second twofold category.
Following the interview the subjects' purely auditory perception was checked by presenting 18 sentences for demonstration but without vision.

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

RESULTS

Fig. 1, 2, and 3 show the percentage of "fused" answers, that is /d/, for each of the ten steps of desynchronisation, for each of three acoustic realisations and for each vocalic context averaged about subjects.

Fig. 1 <ga>[ba]

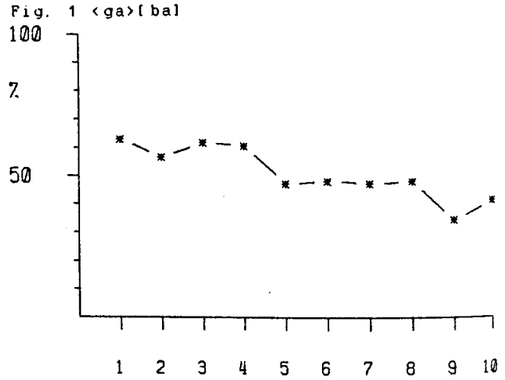


Fig. 2 <gu>[bu]

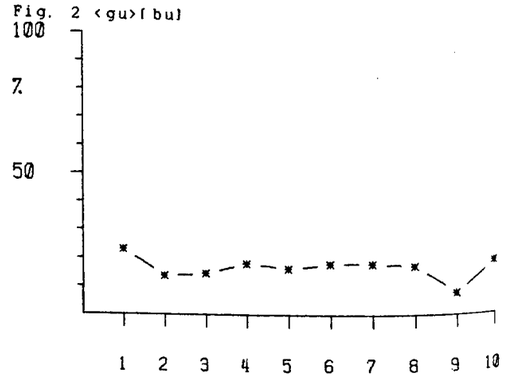
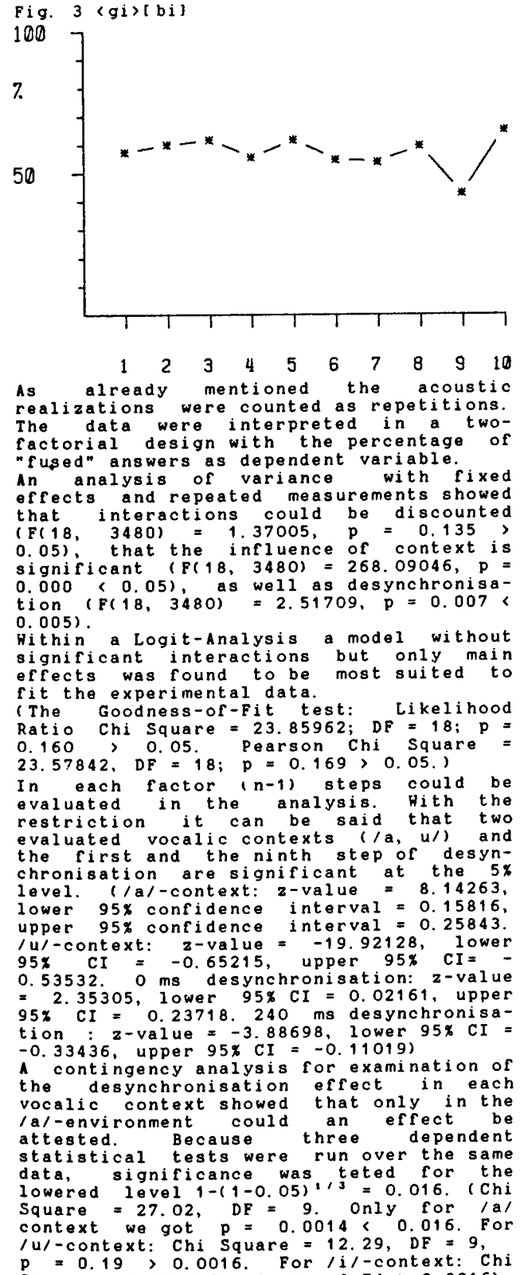


Fig. 3 <gi>[bi]



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 measurements showed that interactions could be discounted ($F(18, 3480) = 1.37005$, $p = 0.135 > 0.05$), that the influence of context is significant ($F(18, 3480) = 268.09046$, $p = 0.000 < 0.05$), as well as desynchronisation ($F(18, 3480) = 2.51709$, $p = 0.007 < 0.005$).
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 = 0.160 > 0.05. Pearson Chi Square = 23.57842, DF = 18; p = 0.169 > 0.05.)
In each factor (n-1) steps could be evaluated in the analysis. With the restriction it can be said that two evaluated vocalic contexts (/a, u/) and the first and the ninth step of desynchronisation are significant at the 5% level. (/a/-context: z-value = 8.14263, lower 95% confidence interval = 0.15816, upper 95% confidence interval = 0.25843. /u/-context: z-value = -19.92128, lower 95% CI = -0.65215, upper 95% CI= -0.53532. 0 ms desynchronisation: z-value = 2.35305, lower 95% CI = 0.02161, upper 95% CI = 0.23718. 240 ms desynchronisation : z-value = -3.88698, lower 95% CI = -0.33436, upper 95% CI = -0.11019)
A contingency analysis for examination of the desynchronisation effect in each vocalic context showed that only in the /a/-environment could an effect be attested. Because three dependent statistical tests were run over the same data, significance was teted for the lowered level $1-(1-0.05)^{1/3} = 0.016$. (Chi Square = 27.02, DF = 9. Only for /a/ context we got $p = 0.0014 < 0.016$. For /u/-context: Chi Square = 12.29, DF = 9, $p = 0.19 > 0.0016$. For /i/-context: Chi Square = 6.00, DF = 9, $p = 0.74 > 0.0016$) This explains the missing interactions between desynchronisation and context.
The Logit-Analysis of desynchronisation was confirmed by a different analysis using orthogonal contrasts. The differences between the average of the "fused" responses between step 1 and all the other steps is significant ($F(1, 3480) = 4.76418$, $p = 0.019 < 0.05$) as well as between step 9 and 10 ($F(1, 3480) = 6.39973$, $p = 0.007 < 0.05$).
Summing up these results it can be stated, that the variable "fusion" is dependent on the variability within the first and ninth step of the variable desynchronisation and on the variability within all three steps of the variable context. The influence of temporal shifts between visual and acoustic information is only relevant for the /a/-context.
After the test 33 subjects were interviewed. They were asked whether they had always heard an initial consonant, and which alternative of the second response category they preferred in each vocalic context.
Only three subjects heard mostly /b/ in /u/-environment, no subjects heard /g/, but 30 subjects heard with very few exeptions only the vowel /u/; some subjects mentioned a hard glottal attack.
The vowel alone was heard almost exclusively in /u/-context but never in /i/-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 <ga>[ba]

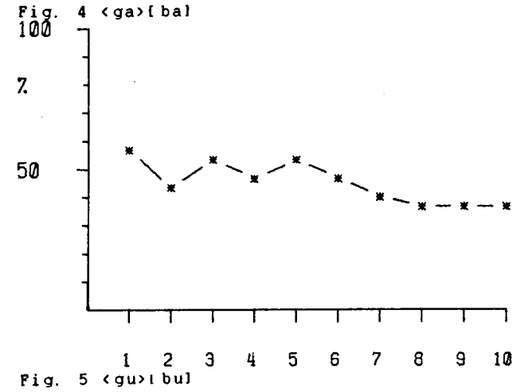


Fig. 5 <gu>[bu]

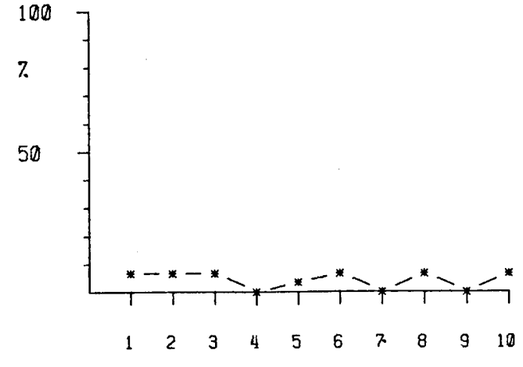
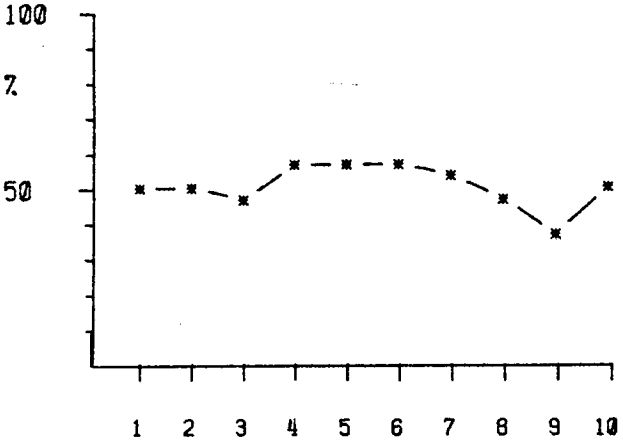


Fig. 6 <gi>[bi]



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 occured 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 looses 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 desynchronization 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

[1] N. F. Dixon, L. Spitz, "The detection of auditory visual desynchrony", Perception, 9, p. 719 - 721, 1980.

[2] M. McGrath, Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults", J. Acoust. Soc. Am. 77 (2), p. 678 - 685, 1985.

[3] H. McGurk, J. MacDonald, "Hearing lips and seeing voices", Nature, 264, p. 746 - 748, 1976.

[4] J. MacDonald, H. McGurk, "Visual influences on speech perception processes", Perc. & Psyphy., 24 (3) p. 253 - 257, 1978.

[5] B. H. Repp, S. Y. Manuel, A. M. Liberman, M. Studdert-Kennedy, "Exploring the 'McGurk Effect'", paper presented at the 24th annual meeting of the Psychonomic Society, San Diego, 1983.

[6] Q. Summerfield, "Use of visual information for phonetic perception", Phonetica Vol. 36 No. 4 - 5, S. 314 - 331, 1979.

[7] H. G. Tillmann, B. Pompino-Marschall, U. Porzig, "The effects of visually presented speech movements on the perception of acoustically encoded speech articulation as a function of acoustic desynchronization", Proc. of the 10th ICPhS, p. 496 - 473, 1984.

ACKNOWLEDGEMENTS

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