The Effects of Visually Presented Speech Movements on the Perception of Acoustically Encoded Speech Articulation as a Function of Acoustic Desynchronization

H.G.Tillmann, B. Pompino-Marschall, U. Porzig
Munich, Federal Republic of Germany

1. Introduction

Under certain conditions, visually presented speech movements have a strong influence on what is "auditorily" perceived when the acoustic signal contains the mapping of another speech movement. If, for instance, an acoustic [ga] is synchronized to a seen <ba> movement and not presented too clearly to the listener he perceives a heard "[da]". In a series of pretests we found that the different effects of VPSM (Visually Presented Speech Movements) on the perception of AESA (Acoustically Encoded Speech Articulation), which have been observed since the appearance of McGurk and Mac Donalds (1977), cannot be adequately accounted for by their division into fusions and combinations, e.g. in the respective cases of <ga> + [ba] = "[da]" and <ba> + [da] = "[bda]".

The 'winning eye' effects are to be judged quite differently depending on whether or not the subjects see that a labial articulation is taking place. For convenience we put the description of VPSM in angled brackets and if there is a difference between <VPSM> and [AESA] we enclose what is actually auditorily perceived in quotation marks. To express that it is true that the labial movement is visually present, we write <+L>, when it is not present <-L>.

Yet the nondominant auditory mode may also gain in influence depending on how clearly the true 'articulatory content' (cf. Tillmann 1980, 68ff, 244ff) of the given utterance is mapped onto the acoustic speech wave. Along these lines one finds an explanation for combinations such as "[bda]". Here we would like to make quite another observation. In a noisy computer room, looking at a not so clearly visible <ga>, which is presented with its original [ga] leads in most cases to a perceived "[da]".

We would like to distinguish three different VPSM/AESA effects. If the contradiction between <+L> and [+da] leads to the perception of "[bda]", we will speak of a phonetic combination. If in cases of <-L> the place of articulation of [+L] is moved from the lips into the mouth of the speaker (<ga> + [ba] = "[da]") we speak of the resulting "[-L]" as a phonetic fusion; two different phonetic categories fuse into a new category in between. But if in the case of <+L> there is also a transfer of labial manner of articulation, say "[p-]" or "[b-]", we would prefer to call the resulting

combination a phonemic fusion, because this effect strongly resembles the effect of phonological fusion found in dichotic listening experiments (Cutting 1976). This is clearly the case with <+ba> + [la] = "[bla]".

In the experiments described below we are less interested in producing the different VPSM/AESA-effects but rather in destroying them by systematically desynchronizing the temporal coincidence of <VPSM> and [AESA].

2. Experiment I

In our first experiment we tested the phonetic fusion using the two German words "Gier" and "Bier", taking the VPSM from the first word and the AESA from the second one. We expected that <Gier> + [Bier] would result in "[dir]", which is also a German word. To prepare a test tape for the Sony-Umatic-recorder a female speaker was filmed uttering the two sentences "Ich habe - Gier - gesagt", "Ich habe - Bier - gesagt". The hyphens indicate a pause of nearly 1 s. As the lips were closed during the pause before "Gier" any preparatory tongue movements were masked and only the <g>-release could be seen. The plosive of the word "Bier" was produced with a short noticeable lip pressing.

The test tape had six blocks each consisting of eight copies of the "Gier"-sentence and of two copies of the "Bier"-sentence randomly inserted. The original speech signals from the test tape were digitally recorded and properly segmented. Based on this segmentation (for the details of this procedure cf. Tillmann, 1983) the second track of the test tape received the sentence frame via direct AD/DA-conversion from the original track whereas the desynchronized [Bier]-signals came from the computer memory. The desynchronizations of the acoustic signals had the values of 200, 250, 300, 350, 400 and 500 ms (deviation 0.05 ms) in positive and negative direction, respectively, and each of these items occurred four times (the zero condition was omitted because in pretests we had found that the effect of phonetic fusion is very resistant to small desynchronizations). The dummy "Bier"-copies received the original acoustic signal without desynchronization via direct AD/DA-conversion.

The tape was presented to 17 subjects who were instructed to watch the screen and to report whether they had heard "Bier" or "dir". It appeared that 10 of the subjects always heard the original acoustically given word "Bier", which means that they did not show any phonetic fusion. According to our pretests this was probably the case because no zero delayed <Gier> + [Bier] items were used in this test. The results of the seven fusioners are given in Fig. 1.

The same tape was presented a second time to the subjects who were now instructed to judge the quality of the synchronization as good or incorrect. The results of the seven fusioners are given in Fig. 2.

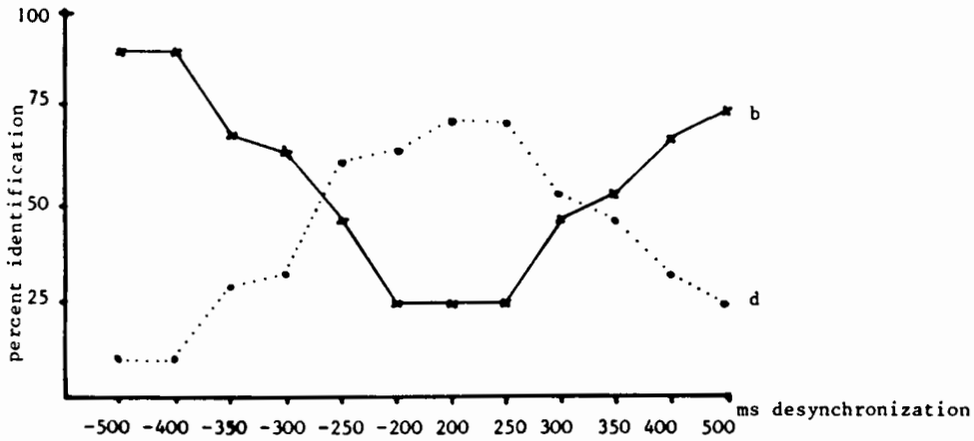

Figure 1. Identification results of experiment I ('undecided' responses omitted).

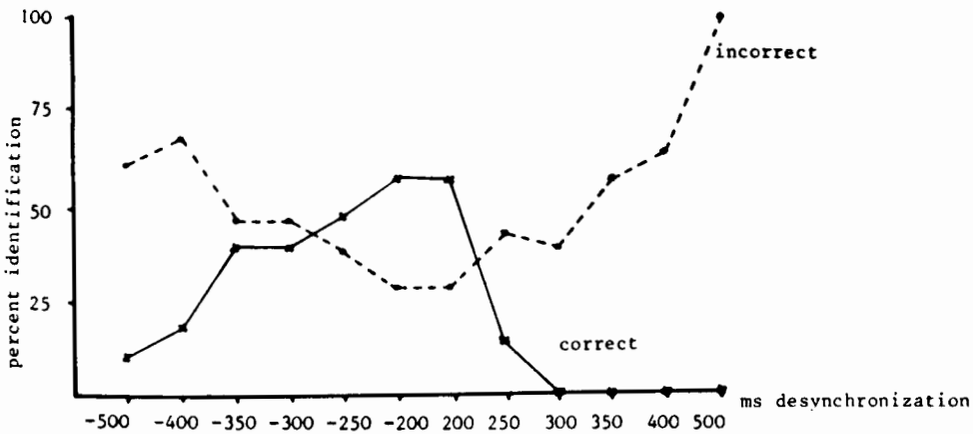

Figure 2. Results of quality judgement in experiment I ('undecided' responses omitted).

3. Experiment II

Phonemic fusion was tested in Exp. II where we used the German sentences "Ich habe - ba - gesagt", "Ich habe - la - gesagt". We expected that seen <ba> and acoustic [la] would give the fusion "[bla]". The tape was prepared as in Exp. I but we introduced only positive desynchronizations (acoustic delays) in this case: 0, 100, 200, 300, 400 and 500 ms. The tape was presented to 6 subjects who were instructed to report whether they had heard "bla" or "la". The results are shown in Fig. 3.

Analogously to Exp. I the second quality judging test was run. The results are given in Fig. 4.

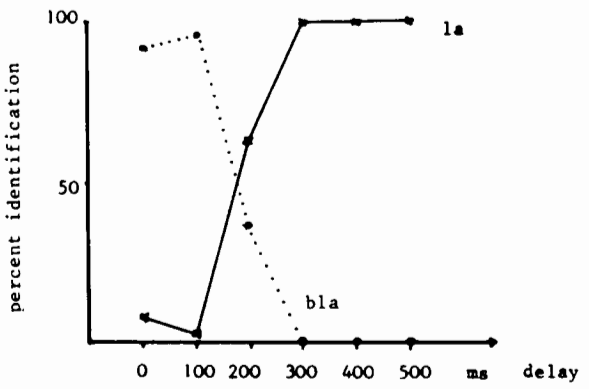

Figure 3. Identification of experiment II ('undecided' responses omitted).

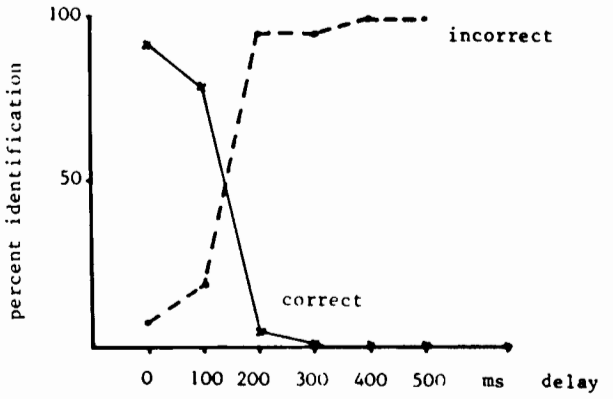

Figure 4. Results of quality judgement in experiment II ('undecided' responses omitted).

4. Discussion

The data of the fusioners in Exp. I (cf. Fig. 1) show that for one group of subjects there is a wide range of desynchronization where phonetic fusions stay predominant, the range going from -250 to +300 ms. We had already found this kind of asymmetry in our pretests. The data of the phonemic fusion experiment (cf. Fig. 3) show that in this case the timing relations are much more critical. The influence of the dominating eye breaks down as soon as the delay of the acoustic signal is more than 100 ms. Of interest is also the fact that exactly synchronized <ba> + [la] result in fewer "bla"-responses than in the case of the first desynchronized stimulus pair. This indicates that for phonemic fusion the timing relations of natural speech productions play a more critical role than in the case of phonetic fusion. A corresponding effect can also be seen in the judgements of the quality of the synchronization (cf. Fig. 4).

In general, the quality judgements are in agreement with the identification results. We find it very interesting that subjects react so much more critically to desynchronization in the case of phonemic fusion than in the case of phonetic fusion.

Further experiments are planned to determine more closely those articulatorily different conditions that influence the decreasing VPSM/AESA-effects as a function of desynchronization between VPSM and AESA.

The next experiment which has been prepared but not yet conducted concerns a situation where we have an amalgam of phonetic combination and phonemic fusion. As we have seen in the first pretests, <Bier> + [Gier] produces the phonetic combination "[bg]". We would like to see whether under the respective desynchronization conditions the phonemic fusions "[B'gier]" and "[G'bier]" result, since these could be understood by the subjects as allegroforms of the German words "Begier", "gebier".

Acknowledgement

We thank Prof. Müller, Münchner Hochschule für Fernsehen und Film, and his staff, especially Dr. Bamberg and Mrs. Schumann for helping us in cutting the original tapes.

References

Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening. *Psychol. Rev.* **83**, 114-140.

McGurk, H., Mac Donald, J. (1976). Hearing lips and seeing voices. *Nature* **264**, p. 746ff.

Tillmann, H.G. (1983, in press). Intra- und heteromodale Isochronie und temporale Koinzidenz von kategorialen Kontinua. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* **17**.