McGURK EFFECT IN GERMAN AND HUNGARIAN LISTENERS

H. Grassegger
Institute of Linguistics, Section of Phonetics, Graz, Austria

ABSTRACT

The goal of this study is to determine how bimodal speech with conflicting auditory and visual information is processed by German (more exactly: Austrian) and Hungarian subjects. This was tested by bimodal presentation of the syllables /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /fa/, /sa/. The results, analysed by confusion matrices for each visual stimulus, showed that the McGurk was less strong and widespread in German than in Hungarian.

INTRODUCTION

The well-known McGurk effect phenomenon demonstrates that visual information on place of articulation influences phonetic perception. Unlike normal audio-visual congruent information which helps auditory perception, lip-read information with audio-visual discrepancy on place of articulation (i.e. whether the place is labial or non-labial) misleads and biases auditory perception.

Although this visual biasing effect on speech perception has been replicated in many studies for English speaking subjects, it has hardly been examined for other languages, with a few exceptions, amongst these Japanese [1], Spanish [2] and a single study with German speaking subjects who identified English bimodal CV syllables [3].

In the present study German (more exactly: Austrian) and Hungarian subjects were tested as to the perceptual influence of bimodal speech with conflicting auditory and visual information. As the phonological inventories of these two languages differ with regard to the consonant categories used in the test syllables (see below) the outcome of bimodal speech perception was expected to be influenced by these differences as well.

TEST SYLLABLES

Ten syllables with consonants occurring in both languages were used: /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /fa/, /sa/. For recording the audio and video signals a male Austrian talker pronounced each syllable once while his face was videotaped with a camera located in front of him and the audio signal recorded seperately to provide highest quality recording. For presentation a random order sequence of 100 audio-visual stimuli was produced resulting from the combination of the 10 audio stimuli dubbed on each visual stimulus. To ensure precise synchronization between audio and video signals the dubbing timing was adjusted by a 25msec frame unit. Each audio-visual stimulus was embedded in a 7sec unit consisting of a 4sec talking face preceded by a 3sec black screen with the respective stimulus number inserted. With a viewing distance of about 1 m visual stimuli were presented on a color monitor showing the speaker's face in approximately life-size. Audio stimuli where presented through the two built-in loudspeakers at each side of the screen.

SUBJECTS

Ten native speakers of German and Hungarian participated in the experiment. All subjects had normal hearing and normal or corrected visions. The age of the German subjects ranged from 20 to 25, the age of the Hungarian subjects from 14 to 16.

PROCEDURE

Subjects were presented an audio-visual stimulus every 7 seconds and were asked to look at and to listen to each utterance. The subjects' task was to write down what they *heard* not what they saw. To make the subjects attend to the visual stimuli they were instructed to report any noticed perceptual discrepancy. It was also suggested that some people might hear syllables not existing in their mother tongue's phonological system, like /bga/ or /pta/.

As the tests for German and Hungarian subjects were carried out in Graz and Budapest respectively some equipment differences in the presentation of the stimuli was unavoidable. Care was taken to make possibly affecting factors consistent.

EXPERIMENTAL DESIGN

The subjects were required to follow the above instructions in five repetitions of trials for the audio-visual condition, thus yielding 50 observations (10 subjects x 5 repetitions) for each AV-stimulus in German and Hungarian.

To measure *auditory* intelligibility five repetitions of the ten audio stimuli were similarily randomized and administered to the subjects only once, thus yielding also in 50 observations (10 subjects x 5 items) for each audio-stimulus in both languages. To avoid the influence of hearing the audio-alone stimuli on the McGurk effect this test was done after the audio-visual session.

RESULTS

The results were analyzed by producing confusion matrices for each visual stimulus and one confusion matrix for the audio-alone condition.

Audio-alone condition

In the audio-alone task almost all of the auditory stimuli were identified as what the speaker intended to pronounce, by the German as well as by the Hungarian listeners. Some minor deviations in the identification of stimuli, amongst these /m/ twice heard as /n/ by Hungarian subjects and three times by German subjects, were not able to explain respective fused responses in the audio-visual condition as a consequence of reduced auditory intelligibility. This was confirmed by performing a chi-square test that compared the frequencies of the fused responses in both conditions (df=1, N=100).

There was only one exception to the almost perfect intelligibility of the auditory stimuli: /b/, which yielded 46% /v/-responses with German and 38% with Hungarian listeners. The chi-square test for the respective responses in the audio-visual condition consequently revealed most of the deviant /v/-responses for /b/ as not significantly different from the audio-alone results and thus not visually biased.

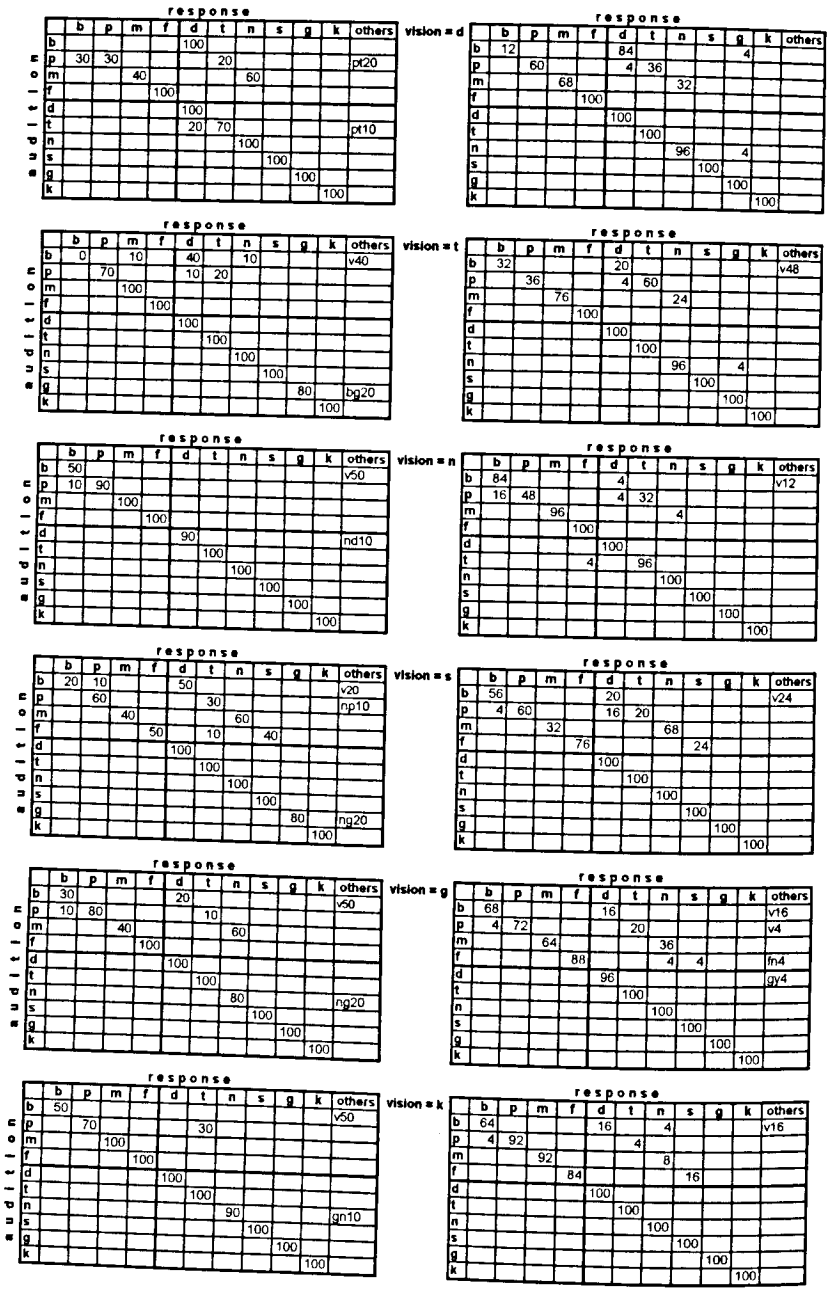
Visual labials

For visually presented labials, i.e. visual /b, p, m, f/, the confusion matrices for both languages show high rates in the diagonal cells, indicating that most of the auditory stimuli were perceived correctly and visual biasing effects were fairly weak. There are only two exceptions.

The first one is auditory /b/, which - evidently due to the above mentioned poor intelligibility - even with visual labials was most frequently heard as /v/, more so by German than by Hungarian listeners. Visual /f/ most effectively supports the obviously inherent labio-dental information of the intended auditory /b/: with visual /f/ Hungarian listeners judged /b/ only 40% of the time as /b/, 4% as /p/ and 56% (!) as /v/; in the same visual condition German listeners never (!) recognized auditory /b/, fused responses being /m/ with 10% and /v/ with 90% (in this latter case significantly different from the audio-alone condition, thus showing high visual biasing effect).

The second exception is auditory /n/, which yielded (its complete auditory

Table 1. Confusion matrices for the stimuli with visual non-labials, indicated in % in 50 observations for German (left column) and Hungarian (right column).



the time with visual /b/ and 60% (!) with visual /m/ for German subjects, 16% with visual /p/ and 12% with visual /m/ for Hungarian subjects.

Visual non-labials

For the visual non-labials, i.e. visual /d, t, n, s, g, k/, it was mainly with the auditory labials that visual effects occurred. This can easily be interpreted from the confusion matrices in Table 1, where in the diagonal cells of the lower right section (i.e. the *non*-labial section) 100%-values indicating absence of visual biasing effects predominate.

For auditory labials the influence of visual non-labials is least prominent with /f/. Fused responses in both languages only occur, when auditory /f/ is combined with visual /s/, yielding erroneous /s/-responses (40% with German, 24% with Hungarian subjects) or /t/-responses (10% with German subjects). With Hungarian subjects the perception of /f/ is - to a lesser degree - also biased by visual /g/ and /k/.

Visual biasing effects for auditory /b/ are evidently enhanced by its rather poor intelligibility (see above), but nevertheless it is noteworthy that combined with visual non-labials there is a significant amount of erroneous /d/-responses (even 100% with visual /d/ for German subjects) in both languages. The same holds true for auditory /p/, which in spite of its full intelligibility in the audio-alone test shows a remarkable frequency of /t/-responses.

The perception of auditory /m/ is visually biased by visual /d/, /s/ and /g/ in both languages, whereas with visual /t/, /n/ and /k/ visual biasing effects only occur with Hungarian subjects. In all cases, however, the erroneous responses are restricted to /n/, thus retaining the auditory information on the manner of articulation.

As already mentioned above the outcome of bimodal speech perception was expected to be influenced by the difference of the phonological inventories of the two languages tested. As Hungarian with a voiced and voiceless palatal plosive and with a palatal nasal has - as far as the categories of the test syllables are concerned - a more complete series of consonants than was offered by the test stimuli, it is most striking that Hungarian subjects only once produced a fused response *gy* (which orthographically stands for the voiced palatal plosive) 4% of the time when auditory /d/ was combined with visual /g/ (see "others" in the response column).

CONCLUSION

This study showed that the McGurk effect occurs in both languages investigated, slightly more easily to induce in Hungarian than in German. However, the visual biasing effects seem to be not symmetrically distributed for labial and non-labial articulation. In either language visual labials do not highly influence the perception of auditory non-labials except for the (dental) nasal, whereas visual non-labials produce a fairly strong visual biasing effect on auditory labial stimuli.

Comparison of the results of the audio-visual condition with the audio-alone condition indicated that the McGurk effect was more easily induced with poorer intelligibility (as in our case for /b/), but was not eliminable for stimuli of 100%-auditory intelligibility.

REFERENCES

[1] Sekiyama, K. & Tohkura, Y. (1993), "Inter-language differences in the influence of visual cues in speech perception", *Journal of Phonetics* 21, pp. 427-444.

[2] Massaro, D. et al. (1993), "Bimodal speech perception: an examination across languages", *Journal of Phonetics* 21, pp. 445-478.

[3] Mills, A. E. & Theim, R. (1980), "Auditory-visual fusions and illusions in speech perception", *Linguistische Berichte* 68, pp. 85-108