Open your eyes and listen carefully. Auditory and audiovisual speech perception and the McGurk effect in aphasia
Klitsch, Julia Ulrike

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2008

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
7 Summary, Discussion, and Prospect

This chapter summarizes the results of both experiments and discusses the main findings in the light of the theoretical background outlined in the introduction and in chapters 4 and 5. Finally, an outlook for further research is proposed, based on questions raised by the current study.

7.1 Summary

The two experiments assessed the auditory discrimination and identification skills of Dutch-speaking aphasic patients. The data obtained from the first experiment, two auditory discrimination tasks, suggest a ranking of perceptibility of the three distinguished phonetic features. The observed ranking suggests that manner of articulation is most easily discriminated and that place of articulation is most difficult to auditorily discriminate. In analogy to findings from studies with non-brain-damaged speakers, it was concluded from the first experiment (auditory discrimination of words and non-words) that aphasia may cause some kind of system-internal noise. Studies with non-brain-damaged speakers also showed that the perception of place of articulation is particularly vulnerable to noise (Grant et al., 1985; Miller & Nicely, 1955; Sumby & Pollack, 1954). Furthermore, it was shown that in these noisy listening conditions, it was the availability of visual articulatory cues that most enhanced the identification of the feature place of articulation (Summerfield, 1987). Hence, without equating the postulated system-internal aphasic noise with external background noise in the listening condition, it was suggested that the aphasic performance should improve by providing visual articulatory information in addition to auditory speech information.

This was assessed in the second experiment, a McGurk experiment that focused on identification of the feature place of articulation in four presentation conditions: an auditory-only condition (AO), a visual-only condition (VO), an audiovisual condition (AV), and a McGurk condition (MG). The McGurk experiment was set up as a three-alternative forced-choice identification task. Comparing the identification performance of the aphasic group on the auditory-only condition and the audiovisual condition revealed, as predicted, a significant improvement in the audiovisual condition. That is, as non-brain-damaged speakers in noisy listening conditions, the aphasic group, which is assumed to be exposed to system-inherent aphasic noise, profited from the availability of additional visual speech
information in the audiovisual condition. However, visual speech information can assuage
the effects of the system-inherent noise caused by the aphasia but cannot entirely
compensate for it.

As in the first experiment that focused on auditory discrimination of words and non-
words, the stimulus material of the McGurk experiment was controlled for lexical status. In
both experiments performance was better on words than on non-words. This finding
suggests that the (phonological) processing resources in both experiments are supported by
the lexicon if possible. That is, processes that are assumed to occur at an early processing
level are supported, if possible, by higher levels, for example the lexicon.

7.2 The influence of visual articulatory information on spoken
speech perception

The phonetic feature place of articulation (PoA) is acoustically characterized by rapid and
weak acoustic changes. Hence, background noise is particularly detrimental to the
perception of place of articulation contrasts (e.g. Sumby & Pollack, 1954; Miller & Nicely,
1955). However, the availability of visual articulatory information can compensate for this
background noise to a large extent. The current study supports these findings in a group of
aphasic speakers. With respect to the feature place of articulation (PoA), system-internal
aphasic noise seems to have comparable effects on auditory speech perception, as has
background noise on auditory speech perception of non-brain-damaged-speakers. Deficiencies have been shown to be – at least partly - remedied in the same way. Providing
visual articulatory information in the audiovisual condition (AV) of the second experiment
(the McGurk experiment) resulted in a significant increase of the identification scores
compared to the auditory identification scores. This was specifically true for the target
phoneme /p/, which is visually more salient than the other two target phonemes /t/ and /k/.

It is not claimed here that aphasic noise in the speech-processing system is to be
equated with background noise. The nature of the aphasic noise is unclear, whereas the
consequences of background noise added to the listening condition are relatively definite
and more or less predictable. Nonetheless, it is noteworthy that the aphasic speech
processing system is responsive to the same back-up mechanisms as is the unimpaired
speech processing system under noisy listening conditions. That is, providing visual
articulatory information can be seen as a factor contributing to the enhancement of aphasic
spoken speech perception.
7.3 Lexical effects on auditory discrimination and identification

Several studies have shown that lexical knowledge can improve and increase the speed of phoneme recognition (e.g. Cutler, Mehler, Norris, & Segui, 1987; Vitevitch & Luce, 1999; Mirman, McClelland, & Holt, 2005): phonemes in words are processed faster than phonemes in non-words. The current data point into the same direction. Especially in the aphasic group, auditory discrimination and identification performance was better when the stimuli were words than when they were non-words.

Numerous accounts for this phenomenon can be found in the literature. Posner and Snyder (1975), for example, accounted for different percepts in word and non-word contexts by suggesting expectancy-based strategies. In the case of ambiguous phonemes, a preference is expected for the phoneme rendering a (context-appropriate) word. Also, phonological short-term memory (pSTM; Baddeley, 1984) can be involved in the origin of lexical effects. Yet, the pSTM is not in the focus of most theories on speech processing (Jacquemot & Scott, 2006). However, in some models (e.g. Logogen model by Morton, 1969) it is represented in terms of a buffer system. When words are to be processed, the buffer system can be supported by the higher lexical level via feedback, whereas in the case of non-words no such support is expected. The architecture of the TRACE model (McClelland & Elman, 1986) can account for the observed processing preference of words by the feedback connections from the lexical to the prelexical level. Non-words, however, do not get such feedback from the lexical level. The architecture of the Merge model (Norris et al., 2000) does not allow feedback from the lexical to the pre-lexical level. Rather, in the case of words (and to a certain degree for non-words that are phonological neighbors of words) this model assumes flow of activation from the lexical level to the phoneme decision level. In this way, the Merge model tries to explain lexical influences on phoneme decisions. However, since the phoneme decision level is, in addition, fed in a purely bottom-up manner by the prelexical level, the extent of lexical influence on phoneme decisions is quite restricted in the Merge architecture.

According to the critics of interactive theories (e.g. Norris et al., 2000), such as the TRACE model, these theories are further challenged with regard to an attentional modulation mechanism. The attentional modulation mechanism is consulted to answer for lexical effects in phoneme processing (e.g. Mirman, McClelland, Holt, & Magnuson, in press). It is based on observations that the strength of lexical effects can vary depending on the task demands (e.g. Vitevitch & Luce, 1999) and the nature of the stimuli (e.g. Cutler et al., 1987). These observations are interpreted to represent cases of implicit attention.
(Mirman at al., in press). The core of implicit attention modulation is that attention is focused on task-relevant information. For example, when the stimulus material consists of few or no words, lexical attention is reduced since it is not relevant for the task (it may even disrupt it). At this point, the criticism of the opponents of interactive models comes into play. They argue that in interactive models lexical feedback is mandatory and, hence, interactive models cannot properly handle attentional modulation of lexical effects (e.g. Norris et al., 2000). However, supporters of interactive models counter that the TRACE model is able to explain lexical effects in terms of attentional modulation. Under the assumption that attentional modulation occurs at the lexical stage itself, the dynamics of lexical feedback are modulated, leaving the interactive activation architecture intact. Mirman and colleagues (in press) show that the TRACE model, extended by implementations of this attentional modulation mechanism, can account for attentional modulation of the lexical effects. In summary, the data of the current study do not clearly favor any of the models since both architectures, the interactive TRACE model as well as the autonomous Merge model, explain lexical effects in speech perception.

However, the account of the attentional modulation of lexical feedback effects (Mirman et al., in press) is interesting. In the following (paragraph 7.4), age-related decline of cognitive functioning, including attentional factors, especially divided attention, is suggested as a possible reason for the observed age-related differences in the magnitude of the McGurk effect. With respect to the first experiment, the account of attentional modulation predicts enhanced lexical attention in the auditory word discrimination task and reduced lexical attention in the non-word discrimination task.

Also, the implicit attention modulation by means of the stimuli was supplemented by the instructions given that explicitly modulated the lexical attention. The instructions clearly announced that in one task only lexical minimal pairs had to be discriminated, whereas the other task fully consisted of non-lexical minimal pairs. This suggests that even though not explicitly required for the metalinguistic auditory discrimination task, there is evidence for the processing of lexical information. If this were the case, it would account for why modulation of lexical attention has an influence on auditory discrimination. As for auditory non-word discrimination, attention was explicitly drawn away from the lexicon by means of the instructions provided. Thus, the auditory non-word discrimination task should be more or less free from interfering, task-irrelevant lexical information. However, in the aphasic group, auditory discrimination was better for words than for non-words. There are at least two conceivable reasons for this finding.
First, even seemingly irrelevant (attention to) lexical information somehow contributes to the metalinguistic phonological auditory discrimination task. The exact nature of this supportive function, however, remains questionable. One possibility is that it is all about the phonological short-term memory (pSTM). As spoken speech is very transient, the pSTM is assumed to be a temporary storage that assists the back-up of incoming spoken speech information. When in the aphasic group this temporary storage is reduced due to the aphasia, lexical feedback is assumed to back up the limited pSTM. However, in the case of non-words lexical assistance is impossible and later incoming information may overwrite or interfere with previously encountered information in the pSTM. This is detrimental for the metalinguistic auditory discrimination task.

Second, even though the instruction explicitly announced whether words or non-words were to be discriminated, the aphasic speech processing system could not simply switch attention away from the lexicon. In the case of non-words, this persistence involves a vain search for lexical matches that attracts too much attention, distracting attention from the actually task-relevant phonological domain. In addition, it is conceivable that aphasic participants (unconsciously or not) stick to lexical information as a last resort. As the aphasic speech processing system has trouble with quickly fading speech, it tries to attach meaning to it as quickly as possible to be better able to process speech. However, even if not applicable, as in the case of non-words, lexical attention is not easily suppressed. This suggests that attention is not simply directed to another source of information (which could be suggestive of an executive functioning impairment), so that limited processing resources are wasted, which is reflected in the lower auditory discrimination scores with non-words.

In the current study, the lexicon was also shown to influence the outcome of the identification task, and as such the perception of the McGurk effect. As for the two lexically consistent McGurk subconditions, in which the incongruent audiovisual input and the McGurk response were either all words or all non-words, there was no difference with respect to the magnitude of the McGurk effect. This suggests that the lexicon itself does not necessarily interfere with the perception of the McGurk effect (contrary to predictions by, for example, Easton & Basala, 1982).

In the two lexically inconsistent McGurk subconditions, the audiovisually incongruent input was lexical and the expected McGurk response was non-lexical, or vice versa. These two conditions yielded a lexicality effect. The lexically inconsistent conditions can be assumed to shift the focus to lexical attention. In McGurk subcondition 2, in which the incongruent audiovisual input is lexical and the expected McGurk response is a non-word, cognitive factors are assumed to interfere with the early perceptual level. That is, the
actual phoneme identification (by means of audiovisual integration) appears to occur at an early, prelexical level. By means of top-down processing, lexical attention is considered to modulate the response, preferring lexical responses and, thus, possibly revising some of the earliest perceptual McGurk responses. This can account for the small number of McGurk responses in subcondition 2. Subcondition 3, however, consists of non-lexical input that is expected to evoke a lexical McGurk response. Thus, in this subcondition, early perceptions of the McGurk effect might be supported by lexical attention. In addition, some input that initially seems a likely candidate for a McGurk response might turn into an actual McGurk response once lexical attention kicks in. This accounts for the rather large McGurk effect in subcondition 3, in which non-lexical audiovisual input evokes a lexical McGurk response. This effect was particularly distinct in the aphasic group, suggesting that the influence of lexical attention is strongest for the aphasic participants.

7.4 The McGurk effect in Dutch

Neither of the earlier described models of audiovisual speech perception readily accounts for the McGurk effect observed in the current study. However, there are pros and cons for each of the model architectures suggested by Schwartz and colleagues (1998) on the basis of the five metrics by Summerfield (1987). As mentioned under 3.2.1, the direct identification model (DI) seems rather implausible since the McGurk effect has been reported to be a quite robust effect that can occur even when participants are aware of the audiovisual incongruency. This observation is incompatible with a model that postulates that fusion of the audiovisual (incongruent) input occurs prior to unimodal identification. Also in the current study there were some participants who reported the audiovisual incongruency and yet they were able to fuse. Rather, the robustness of the McGurk effect to occur even after detection of the audiovisual incongruency suggests compatibility with the Motor Theory of Speech Perception (Liberman et al., 1967; Liberman & Mattingly, 1985). This approach assumes a speech-processing module in the Fodorian sense,\textsuperscript{11} which is assumed to automatically convert the audiovisual speech information into intended articulatory gestures and, hence, regards the McGurk effect as compulsory phenomenon. Those instances in which participants were aware of the audiovisual incongruency and yet showed audiovisual speech integration are in accordance with the Motor Theory of Speech

\textsuperscript{11} Fodor (1983) claims that modular systems have to fulfill a number of requirements; for example, (1) domain specificity; (2) genetically specified neuronal architecture; (3) fast, autonomous, and mandatory; (4) informationally encapsulated; (5) stimulus-driven and automatic.
Perception (cf. Sekiyama & Tohkura, 1993; Colin, Radeau, & Deltenre, 2005). Awareness of the audiovisual incongruency does not prevent audiovisual integration, therefore reflecting a mandatory, automatic process, as suggested by the Motor Theory of Speech Perception. However, if audiovisual speech integration was as automatic, phonetic, and compulsory as proposed by the Motor Theory of Speech Perception, then it would be difficult to account for the observed influences of the lexicon and the phonetic effects. Yet, it is conceivable that the lexical influence on the McGurk effect occurs at a later stage and, thus, does not modulate the integration mechanism itself (cf. Colin et al., 2005).

At first glance, the dominant recoding architecture (DR) is compatible with the observations of Sekiyama and Tohkura (1991), who suggested that the McGurk effect only occurs with those stimuli of which the auditory component is not well identified. In addition, Sekiyama and Tohkura (1991) only obtained the McGurk effect to some degree when adding noise of various intensities to the auditory component of the McGurk stimuli (also see Fixmer & Hawkins, 1998; Sekiyama, 1998). Furthermore, manipulating the saliency of the auditory component of the McGurk stimulus, by varying its intensity, has been reported to influence the magnitude of the McGurk effect. That is, the greater the auditory intensity of a given McGurk stimulus, the smaller the McGurk effect (e.g. Colin et al., 2002; 2005; for some counter-intuitive results, see Kuhl, Green, and Meltzoff, 1988). All these findings are in line with the claims of an audiovisual speech-processing model with DR-architecture. In the current study (and in many other McGurk studies), however, the auditory component of a McGurk stimulus was pure and yet, the presumably dominant auditory modality was influenced by the interfering, mismatching visual articulatory information. That is, every single McGurk response and every single visual alternative response are interpreted to cast doubt on the validity of audiovisual speech processing models of the DR-architecture. In summary, the current data make the models with DR-architecture appear rather implausible.

Audiovisual speech perception models that postulate separate identification (SI), such as the Fuzzy Logical Model of Speech Perception (FLMP) suggested by Oden and Massaro (1978) and Massaro (1987; 1998), assume that the sensory speech input is processed by mechanisms that are ordered in three stages. In the case of audiovisually incongruent input, each modality is initially individually evaluated (evaluation stage). At the next stage (integration stage), auditory and visual information are integrated and finally, the most appropriate response is given (decision stage). In his paper from 1998, Massaro makes the prediction that the influence of one modality will increase to the extent that the other modality is ambiguous or neutral (p. 241). This suggests that in the age-
matched and aphasic group the auditory modality was more ambiguous or neutral than in
the student group. Hence, the importance of the visual modality increased, resulting in the
observed increase in the number of McGurk responses. Even though this seems acceptable,
in the current study it was not the stimulus material that was manipulated, neither were the
testing conditions impoverished. Rather, factors internal to the human perceptual system are
suggested to have caused the differences in the readiness to perceive the McGurk effect.
Furthermore, it is not clear how the FLMP-model accounts for the fact that in the absence
of stimulus-internal noise or manipulations the stimuli sometimes evoked McGurk
responses and sometimes did not. Also, the differences in the alternative response patterns
can only be accounted for by adjusting the truth values on the basis of the obtained results
rather than generating true verifiable predictions (cf. Vroomen & De Gelder, 1999). In
addition, it is unclear how and at which stage lexical effects interfere.

7.5 Age, cognition, and the McGurk effect

In comparison to the McGurk effect obtained with voiceless stimuli and native speakers of
English, the number of McGurk responses given by the Dutch participants in the current
study was moderate. This was true for all three participant groups – the student, the age-
matched, and the aphasic group. However, there were between-group differences with
respect to the magnitude of the McGurk effect. The large McGurk effects in the age-
matched and aphasic groups can have two (possibly combined) reasons.

First, the participants of these two groups face an age-related decline of their
hearing abilities. Hence, for these participants the auditory modality is less reliable. To
compensate for this, the participants of these two groups can (consciously or not) apply a
lipreading strategy. In this way the visual component gains influence, explaining the larger
McGurk effect in the two older groups.

Secondly, cognitive factors can be at play; for example, an age-related cognitive
decline (e.g. Li & Lindenberger, 2006) may be present, also affecting attentional factors.
Since the instructions did not direct attention to one or the other modality, attention had to
be divided between the auditory and the visual modality. If it is assumed that in the two
older participant groups divided attention is attenuated due to age, this would account for
the larger McGurk effect in these two groups than in the student group. Whereas undamped
divided attention enables the student-age participants to detect the audiovisual
incongruency, attenuated divided attention prevents the elderly participants from doing so.
Consequently, the student participants more easily selectively attend to the dominant
auditory modality, so that they more often isolate the auditory component (cf. Sekiyama & Tohkura, 1993), ignoring the mismatching visual articulatory information. Due to attenuated divided attention, the age-matched and the aphasic groups are less likely to detect the audiovisual mismatch. Hence, they are more likely to integrate the incongruent audiovisual McGurk stimuli, resulting in more McGurk responses. It is also conceivable that a combination of age-related hearing loss and age-related cognitive decline caused the differences in the magnitude of the McGurk effect.

Attention is not only to be divided between the auditory and visual input modalities, but lexical attention may play a role in the McGurk experiment as well. This is especially the case in the two lexically inconsistent McGurk subconditions and most strongly in the aphasic group. This is reflected by the rather strong lexical effects in McGurk subcondition 2 (lexical input; non-lexical McGurk response; small McGurk effect) and McGurk subcondition 3 (non-lexical input; lexical McGurk response, large McGurk effect) and in the aphasic group (strongest lexical bias).

7.6 Prospect for future research
The focus is now on the future research that is needed to gain a better understanding of aphasic and unimpaired spoken speech perception and the nature of the McGurk effect. With respect to the first experiment, the auditory discrimination task of the Dutch version of the PALPA is well controlled for a number of variables, possibly influencing the discrimination performance. However, to obtain more reliable results that enable a strong conclusion to be drawn on which of the two features, voicing or place of articulation, is more susceptible to aphasic noise, an additional study is required. In addition to the variables that are controlled in the PALPA, this yet-to-be-developed test (at least for Dutch, as far as I know) has to control the distinguished feature contrasts in such a way that they are more readily comparable by assuring that they occur in the same word position across the board. That is, with respect to the distinguished phonetic features, manner of articulation, voicing, and place of articulation, care should be taken that, in addition to all other distinguishing positions, they all occur in word-initial position. Due to final devoicing in Dutch, this is the only word-position in which they can all be contrasted. This renders the feature contrasts more readily comparable and allows for the drawing of even stronger conclusions.
With respect to the second experiment, the lexical stimulus material should be
controlled for frequency. This should be done in order to exclude possible interaction with
and distortions by this lexical parameter.

Another variable that should be controlled in future research is the hearing of the
participants. In addition, and particularly with regard to the second experiment, cognitive
functioning should be assessed. Only when these two variables are controlled, can stronger
conclusions be drawn with respect to the cause of the differences in magnitude of the
McGurk effect in the student and the other two participant groups. Furthermore, it would be
interesting to assess how an open response set influences the magnitude of the McGurk
effect. In this case, no alternative answers are provided, so that no possibly biasing clues
are given and participants are to say or write down what they have perceived. Colin and
colleagues (2005) showed that in a multiple choice condition more McGurk responses were
given than in the open response condition. In addition, response time should be controlled
and strictly limited in order to control distorting effects from a later decision stage at which
expectancies are generated that do not reflect integration processes. However, such an
experimental setup, with open responses and strict response time limits, is unlikely to be
achieved with aphasic participants.