3 AUDITORY AND AUDIOVISUAL SPEECH PERCEPTION MODELS

Speech perception has mostly been considered in unimodal, auditory terms. This impression might be based more on technological influences (e.g. telephone and radio) than on the effective nature of auditory speech perception (Rosenblum, 2005). The focus on the auditory aspects of speech perception is also reflected in the many efforts to determine the basic – *nota bene* - phonological unit of speech perception (whatever its size may be) and its generality. Consequently, there are numerous models of auditory speech perception that focus on such a phonological unit. Two of the many models of auditory speech processing (the TRACE model of speech perception by McClelland & Elman, 1986, and the Merge model by Norris, McQueen, and Cutler, 2000) are described in 3.1.

However, seeing a mouth producing speech provides characteristic visual speech information that can improve speech perception (e.g. Sumby & Pollack, 1954) or alter an acoustically clear speech percept (e.g. McGurk effect), together with the variability in the acoustic signal, challenging purely acoustic theories. Thus, in 3.2, the consequences for models of audiovisual speech perception are presented. The third section of this chapter summarizes evidence from recent neuroimaging studies for the process of audiovisual speech processing (3.3). Finally, a linguistic account of the McGurk effect, based on Optimality Theory, is presented in 3.4.

3.1 Models of (unimodal auditory) speech processing

The smallest and most general unit of speech perception that is postulated comprises a universal set of distinctive features (e.g. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Stevens & Blumstein, 1981; Stevens, 1989). Other accounts regard phoneme-sized units (e.g. Nearey, 1997) as the smallest phonological unit of speech perception, which is less general than distinctive features and more language-specific. Blumstein and Stevens (1981) report that the sound systems of natural languages are typically structured in terms of individual segments and features. Based on these two entities Halle (1964) defines a natural class of speech sounds as
CHAPTER 3

a set of speech sounds forms a natural class if fewer features are required to designate the class than to designate any individual sound in the class. (Halle, 1964, p. 328)

From the point of view of linguistic theory, linguistic generalizations (for example, general phonological principles) are assumed to apply to a natural class. One possible example is final devoicing; for example, in Dutch and German that applies to the natural class of voiced obstruents in word-final position. Psycholinguistic evidence from slips of the tongue (Fromkin, 1973) or aphasic errors in speech production (Blumstein, 1973) that frequently affect the segmental or featural level seem to demonstrate the relevance of distinctive or phonetic features and phonemes in the representation and processing of natural languages. However, there are also researchers who assume the syllable as the basic unit of speech perception (e.g. Mehler, Dommergues, Frauenfelder, & Segui, 1981).

Until now there has been no agreement on the phonological dimension of speech perception. This is probably due to the characteristic nature of speech, which is a continuous flow of quickly fading information, without any clear boundary with respect to whatever its smallest perceptual units may be. In addition, independent of the size of the perceptual units, the speech signal is highly variable and at no point only influenced by just one particular unit, but neighboring units influence each other by means of coarticulation. That is, for example, one and the same phoneme can sound quite differently in different phoneme contexts and/or syllable positions, as indicated by rather different formant transitions (Liberman, Delattre, Cooper, & Gerstman, 1954). Similarly, one and the same acoustic cue (for example, a formant transition) can point to different phonemes in different contexts (Liberman, Delattre, & Cooper, 1952). Also, in a study on French, one and the same (CV- or CVC-) syllable, for example /pa/, has been reported to be detected faster in words with the identical initial CV-syllable (e.g. PA#lace; # = syllable boundary) than in words in which it is part of an initial CVC-syllable (e.g. PA#mier palm tree). According to Mehler and colleagues (1981) this indicates that the syllable is a unit of speech perception.
3.1.1 Bottom-up versus Top-down Processing Models of Speech Perception

Whatever the basic unit of auditory speech perception, in everyday communication situations between two or more human beings, the auditory speech perception process passes into the spoken word recognition process with the two being almost inextricable. At this point the debate about whether speech is processed top-down or bottom-up comes into play again, an issue which, strictly speaking, emanates from the word recognition process. However, the many postulated models of auditory speech perception often do not make an explicit distinction between the processes of pure auditory speech perception and auditory word recognition.

The distinction between bottom-up and top-down speech processing theories is reflected in autonomous versus interactive models of speech processing. In autonomous models the information flow is strictly bottom-up and independent of lexical context, at least at the early stages prior to word recognition. In interactive models, however, as the name already suggests, the information flow is bidirectional: bottom-up and top-down. Lexical context can be used to constrain lower-level processes. To illustrate the debate of top-down versus bottom-up processing in auditory speech perception, two different versions of stage models are described here.

3.1.1.1 The TRACE model (McClelland & Elman, 1986)

Already by 1980, Stanovich had suggested that, in the case of problems in one processing direction, the speech processing system can still draw (as kind of a last resort) on the opposite processing direction to accomplish processing of the auditory speech signal. This point of view supports the interactive activation architecture, suggesting bidirectional processing with mutual influences of the two processing routes. One model representing this point of view, that auditory speech processing proceeds in a connectionist or top-down way is, for example, the TRACE model proposed by McClelland and Elman (1986). It is a stage model representing a neural net that aims at the identification of single words. That is, TRACE has an interactive activation architecture with a set of interconnected nodes and comprises three stages or layers: the feature, phoneme, and word stage. The stages are interconnected, allowing for (an excitatory) bidirectional (ascending and descending) flow of information between stages. This means that activation flows not only in a feedforward
manner from the feature stage via the phoneme to the word stage but it is also fed back from the higher stages to the lower feature stage.

Furthermore, there are connections between the competing units at one stage. Hence, these intra-stage connections are assumed to be inhibitory. In addition, the TRACE model is provided with numerous copies of the entire network, in which connections between the units at one stage, and between the different stages, reflect different points in (processing) time.

By assuming that for frequently occurring items (usually words) the connections carry more weight, TRACE does not only account for a frequency effect but also for faster reaction times on words than on non-words. That is, via stronger inter-level connections and feedback, phonemes in words reach the activation threshold faster than phonemes in non-words.

TRACE accounts for lexical effects on (ambiguous) phoneme identification by assuming a gradual build-up of all possible interpretations of the first phoneme, followed by the same process for the subsequent phonemes. Unfolding activation at the phoneme stage in turn activates nodes at the word stage. As soon as an entry wins the competition at the word stage, feedback is sent back to the phoneme stage. This is how (in the case of ambiguous phonemes) the lexicon can influence decisions at the phoneme stage. McClelland and Elman (1986) claim that the TRACE model can also account for findings by Fox (1984). He found that lexical effects on ambiguous, word-initial phonemes vanish when participants are required to respond within 500 ms after the ambiguous segment in question. TRACE would account for these findings by assuming that, prior to identifying the ambiguous initial phoneme, more disambiguating information from subsequent phonemes has to be encountered to establish the lexical identity of the item in question. Since this disambiguation point can occur only two or three items later, 500 ms are easily exceeded. Unambiguous phonemes, however, are less dependent on the lexical status of the item of which they are a part; hence, responses can be given faster. In summary, this suggests that when the bottom-up process is pivotal, lexical effects are rendered redundant. However, phonemes that occur in lexical words are assumed to be recognized faster since they are part of the interactive activation feedback loop. It is just this top-down feedback loop from the lexical word stage to the phoneme stage that causes critics of interactive activation models to object to them because of the risk of hallucination (e.g. Norris, McQueen, & Cutler, 2000). That is, word-like non-words would be tentatively perceived as words and, accordingly, top-down activation from the higher lexical stage to the phoneme stage would cause the representations at the lower stage to be modified according to lexical
requirements to be in accordance with the top-down received lexical bias (cf. Massaro, 1989). The feedback between lexical and pre-lexical stages is actually one of the most criticized shortcomings of the interactive activation architectures and the TRACE model in particular. That is, the critics of the TRACE model consider this feedback mechanism, that is intended to stabilize perception in the TRACE model, as afunctional, since it predicts that the most activated word sends facilitatory top-down feedback to its lower level nodes (Norris et al., 2000; Mitterer & Cutler, 2006). This is considered redundant since once a word wins the competition (by being highly activated) it would not need further assistance from a lower stage. Another point of criticism concerns the limited set of features, phonemes, and words in the model and whether this limited set is robust enough to account for the great acoustic variability of speech.

3.1.1.2 The MERGE model (Norris, McQueen, & Cutler, 2000)

As an alternative, Norris and colleagues (2000) postulate a purely bottom-up model of speech perception: the Merge model. The core of this model is that the information flow is strictly bottom-up and that it declares feedback unnecessary. Rather, it assumes that the output of two autarkic processes is integrated to determine the most adequate response. Integration is executed by phoneme decision units at the decision stage. This is to ensure that phoneme decisions are never fully based on, or too strongly influenced by, lexical information alone; to avoid that, prelexical processing (by means of a bidirectional activation flow) is directly influenced by higher processing levels. The flow of information in this model is assumed to be as follows: strictly bottom-up from the prelexical level to the lexical level, enabling activation of lexical entries compatible with the input. In addition, both levels, the prelexical and the lexical level, feed the decision stage at which information from the two levels is merged and a decision is made about the phonemes actually present in the input. Characteristic for such a strictly bottom-up model is that there are no inhibitory connections at the perceptual prelexical level itself. At the lexical level and at the decision stage, however, inhibitory connections between entries are assumed. Inhibitory connections at the lexical level should allow for the correct modeling of spoken word recognition, whereas the inhibitory connections at the decision stage should allow for unambiguous phoneme decisions. In summary, by assuming a bottom-up information flow and a decision unit, the authors assume the Merge model to be maximally efficient in capturing the advantages of interaction and avoiding the disadvantages, and thus, guaranteeing prelexical processing that is independent from higher-level lexical influences.
Again, other researchers see bottom-up processing as the preferred route, but not as the only processing direction. For example, Eysenck and Keane (1990) in principle believe in bottom-up processing but they suggest that top-down processing comes into play when auditory speech stimuli are of mediocre quality.

### 3.2 Models of audiovisual speech processing/integration

The theories presented above are only two exemplars of numerous models on spoken-word recognition, to illustrate both the close connection between speech perception and word recognition and that there is no general agreement on which (type of) model architecture best captures the speech perception process. With respect to the McGurk effect\(^1\) (chapter 5), however, yet another problem is encountered. Models that are attempting to account for the speech perception processes usually only consider phonological information – that is, the auditory component of the speech signal. This is the case even though there is evidence that speech perception is not a purely auditory process but that visual phonetic information also influences or alters phoneme and/or word identification, even when acoustically unambiguous (e.g. Brancazio, 2004; Erber, 1969; McGurk & MacDonald, 1976; Sumby & Pollack, 1954).

Models that account for audiovisual speech perception do not only have to account for the temporal aspect of the integration of auditory and visual information (early or late), but also for the common denominator of the two quite different sensory input sources (auditory versus visual articulatory). That is, audiovisual speech perception can be assumed to amalgamate the auditory and visual inputs right away or, alternatively, auditory and visual inputs are first analyzed separately before being integrated into a unified percept. In this latter case, it is still conceivable that one modality is more pivotal than the other. Based on four of the “five metrics for audiovisual speech integration” by Summerfield (1987)\(^2\),

---

\(^1\) The McGurk effect is obtained when incongruent audiovisual input is fused (McGurk & MacDonald, 1976). For example, when auditory /ba/ is presented in combination with visual /ga/, participants perceive /da/. A detailed description of the McGurk effect is given in chapter 5.

\(^2\) The “five metrics for audiovisual integration” distinguished by Summerfield (1987; p. 4) are:

1. Phonetic features
2. Filter function of the vocal tract
3. Vectors representing the magnitudes of independent acoustical and optical parameters
4. Successive static vocal tract configurations
5. Time-varying kinematic patterns providing evidence of articulatory dynamics.
Schwartz, Robert-Ribes, and Escudier (1998) distinguish four basic architectures for audiovisual speech perception.

3.2.1 Direct identification model (DI)

The first architecture corresponds to Summerfield’s third metric (Schwartz et al., 1998) and suggests an audiovisual speech perception model in which fusion occurs immediately, meaning that fusion and audiovisual identification coincide without prior unimodal identification. That is, the bimodal input signal is directly classified by the amodal classifier that determines the closest prototype to both input sources. That is, the core of such a model is that there is no intermediate representation.

However, findings like the McGurk effect, especially when subjects are fully aware of the incongruence in the audiovisual signal and yet report the McGurk effect (fusion responses), seem to be incompatible with this model architecture. Rather, the awareness of the audiovisual incongruence suggests that the two input sources have first been analyzed separately and could still be fused. This seems to rule out DI-models, while strongly preferring models which assume integration of the audiovisual input at some point after unimodal identification – either at an early or late integration stage (Schwartz et al., 1998).

3.2.2 Separate identification model (SI)

Audiovisual speech perception models with separate identification architectures correspond to Summerfield’s first metric (Schwartz et al., 1998). This model assumes that integration occurs only at a late level, after comparison with prototypes. The core assumption of late identification models is that separate recognition mechanisms operate in parallel on the respective auditory and visual speech inputs. An example for a model representing this architecture is the Fuzzy Logical Model of Perception (FLMP) proposed by Oden and Massaro (1978) and Massaro (1987), assuming the integration of independently activated information. This model consists of three processing stages (Figure 3.1): the feature evaluation, the feature integration, and the decision stage. At the evaluation stage, the different sources of input are transformed into prototypes in memory which are compared to the input signal. During the integration stage, the degree to which each alternative prototype matches the input is determined. Finally, the decision stage determines the relative quality of the fit of the outputs from the integration stage; this can either be done by taking a concrete decision or by rating among response alternatives.
The Fuzzy Logical Model of Speech Perception (FLMP) is characterized by four basic assumptions (cf. Massaro, 2001):

1. Each source of information evaluates a specific degree of potential response alternatives.
2. Each source of information is analyzed independently of the other sources of information.
3. Integration of information sources renders a ranking of the overall appropriateness for each response alternative.
4. The response candidate that received the highest degree of support is the final percept.
3.2.3 Early integration models
Models of the early integration family assume that there is a common pre-linguistic representation. This common representation could either be amodal or compatible with either one of the two input modalities. An issue that remains with models of this type is the dimension or unit of the input at the moment fusion takes place. Schwartz and colleagues (1998) report two formats of early integration models that are presented in the following sections.

3.2.3.1 Dominant recoding model (DR)
This model architecture corresponds to Summerfield’s second metric (Summerfield, 1987). It owes its name to the assumption that there is one modality, probably the auditory modality, which dominates the audiovisual speech perception process. The common pre-linguistic representation, the core of early integration models, is determined by the dominant modality. The input to the other, less specialized, modality is assumed to be converted into a representation of the dominant modality and subsequently the estimations from both modalities are fused.

A chink in models of this type might be that findings by, for example, Sumby and Pollack (1954) show that the auditory modality is not as all-dominant as might be expected. At least not for the perception of certain phonetic features, especially place features, under noisy conditions, as they show that visual information significantly improves speech perception (see also Miller & Nicely, 1955; Summerfield, 1987). In addition, Schwartz and colleagues (1998) give a more technical argument against this model architecture. Their argument contradicts the prediction of an audiovisual speech perception model of the DR-type that visual input only modifies an auditory percept when the auditory signal is noisy (also see the language-specific account of the McGurk effect of Sekiyama & Tohkura, 1991). As counter-evidence for the DR-architecture they mention a study by Lisker and Rossi (1992) that showed that visual speech information can also modify a clear, corresponding auditory speech input. In their study, the rounding category of the vowel [u] had to be judged. Presented visually, this vowel was considered to be [+round] in 1% of the responses; when presented auditorily it was judged to be rounded. Audiovisual presentation of this vowel caused a drop of [+round]-judgments to 25%. This finding suggested that the auditory impression was dominated by the visual speech information,
casting some doubt on the assumption of the DR-architecture. In addition, psychophysical modeling never built on this type of architecture (Schwartz et al., 1998).

3.2.3.2 Motor recoding model (MR)

This architecture corresponds to Summerfield’s fourth and fifth metric (Schwartz et al., 1998). A model of this type assumes fusion of the audiovisual input by a common, amodal processor that is linked to articulatory gestures. As the name suggests, the Motor Theory of Speech Perception (Liberman et al., 1967; Liberman & Mattingly, 1985) constitutes an example of this category. This theory is based on the lack of invariance problem. As described earlier, Liberman and colleagues (1967) observe that there is no one-to-one mapping between acoustic speech signals and phonetic categories. This variability in the acoustic signal led to the assumption that there is a cognitive module specialized for speech, enabling the listener to extract articulatory gestures that cause the perceived speech signal rather than perceiving phonemes (cf. Liberman et al., 1967; Liberman & Mattingly, 1985). As one option of how their theory could function, Liberman and colleagues referred to the analysis-by-synthesis account, as was put forward by, for example, Stevens and Halle (1967). According to this account the acoustic speech signal is analyzed in terms of motor speech gestures; that is, sensory information is used to generate motor hypotheses about how the speech signal was produced by the speaker. On the basis of this motor guess an acoustic hypothesis is computed which, in turn, is again matched to the actual sensory information. Perception then means establishing an adequate match between the invariant motor gestures that evoked the virtual signal and the actual acoustic signal.\(^3\) That is, according to Liberman and colleagues (1967) speech is perceived in terms of invariant motor gestures. In the revised version of the Motor Theory of Speech Perception, Liberman and Mattingly (1985) maintain the assumption that the objects of speech perception are motor gestures, and postulate that phonemic perception does not require auditory processing at all. However, inter alia, due to context sensitivity during articulation, the nature of these objects of speech perception was adjusted. That is, in the revised version of the Motor Theory of Speech Perception, it is no longer the actually perceived motor gestures (vocal tract gestures) that are the in the center of speech perception but the phonetic gestures

\(^3\) In the ‘analysis-by-synthesis’-model suggested by Stevens and Halle (1967) hypotheses about the speech gestures intended by the speaker are generated on the basis of sensory information. However, motor information is only resorted to when there is phonetic ambiguity or lack of invariance (cf. Nusbaum & Magnuson, 1997).
intended by the speaker at a prevocalic, linguistic level (cf. Galantucci, Fowler, & Turvey, 2006). This notion is in line with an analysis of audiovisual speech by Summerfield (1987), who suggests that the central object of speech perception is not a specific acoustic signal but rather the gesture itself. The motor, acoustic, visible characteristics of the central speech gesture are extracted by the respectively responsible modalities.

In the context of imitation of gestures, Galantucci and colleagues (2006) cite a study by Fowler, Brown, Sabadini, and Weihsing (2003) as being supportive of the view that phonetic gestures rather than acoustic forms are the object of speech perception. Fowler and colleagues had a model speaker producing a target vowel /a/ that varied in duration and that at a certain point was superseded by one of three CV-syllables: /pa/, /ta/, or /ka/. Participants had to shadow the initial vowel and respond as soon as they realized that the model had shifted to one of the CV-syllables. The stimuli were distributed across two tasks; a simple and a choice task, each of which had its specific response requirements. In the simple task, subjects had to indicate that they noticed the change into a CV-syllable by giving a predetermined uniform response (either one of the CV-syllables), independent of the exact identity of the perceived CV-syllable. In the choice task, they had to imitate the respective CV-syllable.

There were two major findings suggesting that the objects of speech perception are indeed phonetic gestures. The first finding concerned the simple task. Reaction times were faster when the actually perceived syllable and the allotted response syllable were identical. The second finding was made with regard to the reaction time difference of only 26 ms between the simple and the choice task. This time difference is much smaller than the difference reported by, for example, Luce (1986) who observed reaction time differences between simple and choice (non-speech) tasks of 100 ms to 150 ms. To account for the results by Fowler and colleagues in the light of phonetic gestures as the objects of speech perception, Galantucci and colleagues further cite studies that also reported a reduction of reaction time in choice tasks (e.g. Umiltà, Rubichi, & Nicoletti, 1999), when targets contained relevant information with respect to the required response. Take, for example, an object identification task, in which one target object (e.g. a rectangle) requires a left-hand response while another object (a square) asks for a right-hand response. If the target object requiring a left-hand response is presented to the left and the object calling for a right-hand response is presented to the right, reaction times can be reduced, as opposed to the condition in which the objects are presented in the center of the screen. Correspondingly, the perception of speech in terms of phonetic gestures (rather than an acoustic signal) accounts for the small reaction time difference between the simple and choice tasks in the study by
Fowler and colleagues. In the choice task, the perception of phonetic gestures virtually limited the number of choices to one by providing clear instructions for the required response – that is, imitation of the perceived gestures. In contrast, speech perception in terms of acoustic objects would not reduce the number of choices. The acoustic input would have to be transformed into one of the relevant response gestures: /pa/, /ta/, or /ka/; however, this is assumed to require time.

However, the claim of the Motor Theory of Speech Perception that speech perception is special, in that there is a specialized module for speech (perception), seems somewhat challenged by the parallel or comparable results of the findings by linearity or direct compatibility between results with speech- and non-speech stimuli. That is, the studies by Fowler and colleagues (2003) and Umiltà and colleagues (1999), cited by Galantucci and colleagues, seem to cast some doubt on the claim that speech is special or modular.

Thus, the speech input, be it unimodal or bimodal, is assumed to be mapped onto phonetic gestures. These phonetic gestures serve as the common denominator for the audiovisual speech input.

3.3 Models of AudioVisual Speech Perception from NeuroImaging Studies

Accounts such as the Motor Theory of Speech Perception recently got some support from neuroimaging data, particularly with regard to one of its core aspects, namely the idea that phonetic gestures are the object of speech perception. Support for this view mainly comes from studies on audiovisual speech perception.

In these studies audiovisual speech stimuli activate a network of cortical motor areas and the cerebellum; these areas are known to be involved in the planning and execution of speech gestures during the speech production process (e.g. Skipper, Nusbaum, & Small, 2005; Skipper, Van Wassenhove, Nusbaum, & Small, 2007; Van Wassenhove, Grant, & Poeppel, 2005).

Skipper and colleagues (2005) showed that this activity in the cortical motor areas was mainly due to and modulated by the visual saliency of audiovisual speech. Increasing the visual distinctiveness of an audiovisually presented story concurred with an increased activity in the motor system. However, this was not the case in the auditory-only condition. Hence, this increased motor activity was attributed to the visually submitted information about the phonological information in the speech signal and not to the phonetic-
phonological information or lexical information. Analogous to the Motor Theory of Speech Perception (Liberman et al., 1967; Liberman & Mattingly, 1985), Skipper and colleagues conclude that processing of audiovisually presented speech implements the activation of motor programs that are associated with the perceived motor gestures.

On the basis of their findings on how the availability of visual speech influences auditory ERP measures, Van Wassenhove and colleagues (2005) postulated a model. This model does not only account for congruent audiovisual speech perception but it also explains the occurrence of the McGurk effect (Figures 3.2 and 3.3). The model is geared to synthesis-by-analysis models (cf. Halle, 2002), a central assumption of which is that an internal representation of an event directs input and output processes. These early generated internal representations make predictions about the characteristics of the sensory input. Hence, in models of this kind the sensory input is not simply processed in a feedforward manner, but early generated internal representations predict and restrict the interpretation of the nature of the sensory input. Three ERP-experiments were conducted with auditory stimuli, with visual stimuli, with audiovisual congruent and audiovisually incongruent (McGurk-type) stimuli. Overall, comparing the audiovisual condition with the auditory-only condition, even in the earliest time frequencies, the auditory ERP-measures suggested the following conclusions. In contrast to earlier studies that proposed the supra-additivity of modalities in audiovisual processing, Van Wassenhove and colleagues reported a general decrease in amplitude and latency reduction that was a function of stimulus identity (p > t > k). In audiovisual speech perception, the auditory signal is preceded by phonetic gestures by a few hundred milliseconds and, thus, visual speech information may be predictive for (parts of) the auditory signal.
Figure 3.2 Analysis by synthesis in audiovisual speech integration.
Model by Van Wassenhove, Grant, & Poeppel (2005); the general scheme.

Figure 3.3 Analysis by synthesis in audiovisual speech integration.
Model by Van Wassenhove, Grant, & Poeppel (2005); the specification that accounts for the McGurk effect (the original version in PNAS, 102; is corrected here for one error in the output).
Hence, one core assumption of the model put forward by Van Wassenhove and colleagues is that the higher the auditory predictability from the visual signal, the greater the temporal facilitation (as compared to auditorily presented speech). Since /p/ is visually much more salient (95% correct identification in the visual-only condition) than, for example /k/ (65% correct visual identification), the latency reduction is greater for /p/ (ca. 10 ms at N1 and ca. 25 ms at P2) than for /k/ (ca. 5-10 ms at N1 and P2, respectively). The amplitude reduction, however, is not differentially influenced by the degree of visual predictability.

For the McGurk-type stimuli with visual /k/, no or only little temporal facilitation is expected, and indeed this is what the authors observed. The amplitude reduction, however, was comparable to that of a congruent audiovisual /p/. According to the authors this suggests that the availability of visual speech information distributes the attention more evenly across modalities. In favor of this assumption are the results from experiment 3, in which the participants were presented with McGurk-type stimuli but were explicitly asked to pay attention to the visual modality (in the previous experiments instructions tended to shift attention to the auditory modality). Independent of the focus of attention the amplitude of the auditory ERP-response was reduced, suggesting an automatic process.

In summary, the model proposed on the basis of these findings incorporates efficient interaction between the natural dynamics of audiovisual speech (the asynchrony between modalities as a transmitter of information) and phonological knowledge. The visual component enables the speech processing system to constrain and anticipate the nature of the auditory signal, as reflected by the temporal facilitation and amplitude reduction of the ERP-response to the auditory signal. Since these two effects are spread over time the authors further suggest at least two computational stages of multisensory integration during audiovisual speech processing. During the first stage, a featural stage, visual speech information makes predictions about the nature of the anticipated acoustic signal. This stage is dependent on the saliency of the visual speech information and covers the 25 ms range during which temporal facilitation can be observed. The more salient, that is, the more disambiguating the visual stimulus, the more predictive the auditory portion of the signal. Consequently, auditory processing is facilitated to a higher degree, as is reflected by a decrease in auditory ERP latency facilitation.

During the second stage, a perceptual unit stage, the speech processing system has shifted to the bimodal processing modus. Thus, this stage operates independently of the attended modality and, accordingly, independent of (featural) information in either of these modalities. This stage is electrophysiologically assumed to be reflected by amplitude
reduction and, according to Van Wassenhove and colleagues, it may relate to segmental-based analysis and syllabic processing.

Skipper and colleagues (2007) elaborated further on this theoretical model of audiovisual speech perception of the analysis-by-synthesis family. To verify and refine particular assumptions of the postulated model, McGurk stimuli were used in an event-related fMRI-study. Since for stimuli of this type the sensory identities (auditory and visual) are not concordant with the perceptual experience of the listener, these stimuli seem appropriate to assess the role of motor system activity during AV speech perception. According to the predictions of the postulated model, activity in the (frontal) motor areas is assumed to be an early hypothesis about the nature of the sensory patterns. That is, with respect to McGurk stimuli, the prediction is that at an early stage the observed motor activity should be more readily comparable to the listeners’ perceptual experience (i.e. /ta/) than to the sensory aspects of the actually presented stimuli (either /pa/ or /ka/). Another closely related aspect of the model concerned the assumption that the hypotheses about the phonetic identity of the stimulus, as reflected by activity in the (frontal) motor system, influences activity in the sensory cortices. That is, focusing on the sensory cortices, activity in these areas should initially resemble the activation patterns corresponding to the sensory characteristics of the input stimuli (that is, either /pa/ or /ka/). However, it is assumed that the motor commands as predictors for the sensory input can modulate the processing of the actual sensory input (cf. Webb, 2004). That is, by means of efference copy (cf. von Holst & Mittelstaedt, 1950) the activation pattern in the sensory cortices is assumed to be shifted, from an activation pattern characteristic for either one of the physical input stimuli (/pa/ or /ka/), to an activation consistent with that generated by the motor hypothesis (i.e. /ta/). The results show that in the group (n=13) that predominantly perceived the incongruent audiovisual McGurk-type stimuli as /ta/ (83%), the evoked activation pattern, especially in the frontal motor areas, was more comparable to that generated by a veridical AV /ta/. In the other group (n=8) the response was less consistent; the subjects perceived the McGurk stimulus as /ka/ in 61.5% of the cases. For this group there was a trend towards a motor activation pattern during the perception of the McGurk stimuli that was more consistent with the activation pattern evoked by a true AV /ka/ than either a true AV /pa/ or a true AV /ta/. According to Skipper and colleagues these findings suggest that the frontal motor areas that are activated during speech production, as well as during audiovisual speech perception, represent hypotheses about the early integration of AV-information. This is reflected by the fact that different hypotheses (as to whether the McGurk stimulus is
perceived as /ka/ or /ta/) seem to evoke activation in different motor areas and, hence, result in different perceptions.

For the group that perceived the McGurk-type stimuli predominantly as /ta/, activity in the cortical motor areas was at all measuring points most consistent with the activation pattern observed for a congruent AV /ta/. However, in the auditory and somatosensory cortices, during the first 1.5s of the hemodynamic response, the activity evoked by the McGurk-stimulus correlates most closely with the activation pattern elicited by a congruent AV /pa/. Thereafter, a shift towards the activation pattern of a veridical /ta/ occurs in these cortices. After the first 4.5s of the hemodynamic response, the correlation of the evoked activation patterns of the McGurk stimulus with a true AV /ta/ does not differ anymore between the motor and the somatosensory cortices. That is, initially, activity in the cortical sensory areas is more consistent with the activation pattern of the congruent AV syllable matching the equivalent sensory part of the McGurk stimulus (auditory /pa/). At later stages, the hemodynamic response in the sensory areas is more comparable with the true AV syllable that matches the finally reported perception of the subject (/ta/ in the group that predominantly perceived the McGurk stimuli as /ta/). That is, the results are suggestive that motor system activity (co)determines the final phonetic perception of AV stimuli. In summary, the activity in the motor cortex that is associated with both speech production and AV speech perception generates a multisensory hypothesis about the encountered speech input. Via efference copy, the acoustic and somatosensory consequences of executing these hypothetical articulatory movements are predicted. Hence, the motor system determines the final perception of the (congruent or incongruent) AV syllable.

The ‘analysis-by-synthesis’-based and neurobiologically specified model outlined by Skipper and colleagues (2007) suggests an interaction of sensory input with feedback or efference copy from the motor system. With respect to this, the postulated model differs from the Motor Theory of Speech Perception (Liberman et al., 1967; Liberman & Mattingly, 1985) that postulates direct mapping between perceived speech and the appropriate invariant motor gestures, as well as from the purely sensory theories of speech perception (e.g. Blumstein & Stevens, 1981). Instead, in their neurobiologically-specified model, the motor system predicts the sensory consequences of the generated hypotheses and, thus, limits the interpretation of the actually encountered sensory information.

---

4 For the visual portion of the McGurk stimulus a similar trend was observed in the visual cortical areas. Initially, the activation pattern resembled more that of a true AV ‘ka’. That is, it seemed to be sensory-based on the visual portion of the McGurk stimulus ‘ka’. Subsequently, it was shifted to one more consistent with the observed activity in the motor areas – that is, the actual percept of the listener (/ta/).
3.4 The McGurk Effect in terms of Optimality Theory

Optimality theory (OT) was developed by Prince and Smolensky (1993) within the framework of generative phonology. This theoretical approach is based on universal constraints which are part of the grammar and, as such, they are ordered language-specifically. Two main categories of constraints operate on the two basic mental representation categories: the underlying form (a stored sequence of discrete phonological structures) and the surface form (a discrete phonological structure assembled from features, segments, syllables, and feet).

1) *Markedness constraints* (e.g. *VOI or NoCoda) are applied on the output form (the surface form) and reject certain unwanted configurations of the generated output.

2) *Faithfulness constraints* watch over input and output forms, requiring exact replication of the input. An example for such a constraint is ‘MAX’, which prohibits deletion of segments in the output form that are present in the input. Analogously, the constraint ‘DEP’ rejects output forms that contain an extra segment compared to the input.

The higher-ranked a constraint, the more pivotal it is in determining the appropriate candidate. That is, for a given target form (in either perception or production) a number of possible candidates are generated; this is called a candidate set. Such a candidate set is evaluated by a language-specifically ranked set of constraints. Finally, the optimal candidate is chosen from the candidate set, meaning that it is more appropriate than all other candidates because prioritized by the most highly-ranked constraint that distinguishes the optimal candidate from an alternative candidate.

Boersma (2006) attempted to explain the McGurk effect in terms of Optimality Theory, suggesting a bidirectional model with four representation levels. Two of these levels, the underlying and the surface form, belong to the phonological representation level, whereas the other two postulated levels, the sensory form and the articulatory form, are assigned to the phonetic representation level (Figure 3.4).

---

5 Sometimes it is assumed that not only the ordering of universal constraints is language-specific but also that there are language-specific constraints (Prince & Smolensky, 1993).
At and between the distinct levels, various constraints are applied to regulate speech perception and speech production.

Between levels, three constraints are applied for evaluation: the faithfulness constraints, the cue constraints, and the sensorimotor constraints. Between the two phonological forms *faithfulness constraints* are applied, preferring similarity of underlying and surface form in perception (Smolensky, 1996) and production (McCarthy & Prince, 1995). The surface form at the phonological representation level and the sensory form at the phonetic representation level are linked by *cue constraints* (Escudero & Boersma, 2003), relating phonological features and auditory and visual cues. According to Boersma (2005), cue constraints are mostly stated negatively. *Sensorimotor constraints* are implemented between the two levels of phonetic representation by associating muscle commands with sound and associated visual aspects (e.g. lip rounding/closure).

At least two more constraints evaluate representations at one level. The surface form is evaluated by *structural constraints* which are applied in perception and production and which refuse (non-optimal) selected surface forms that violate phonological and
phonotactic rules of a particular language. Articulatory constraints are applied only during the production process at the articulatory level to avoid articulatory effort.

To account for the McGurk effect at syllable-level, Boersma (2006) assumes that it is an effect of low-level perception and that it occurs at the interface of the phonetic and phonological levels; that is, it occurs between the levels of sensory form and the surface form. In order to perceive the McGurk effect, the applying or relevant constraints have to be appropriately ranked to favor /da/ over the auditory portion of a McGurk stimulus /ba/, and over the visual portion /ga/. Cue constraints are assumed to be the most pivotal ones to the perception of the McGurk effect.

The perception of auditory /ba/ is avoided by the high-ranked constraint “*/b/ [open lips]”. During the articulation of bilabial /b/ the lips are by definition not open; hence this constraint is violated. For the sake of completeness the constraints “*/d/ [open lips]” and “*/g/ [open lips]” exist as well. Since for the articulation of these phonemes the lips are open, the constraints would make wrong predictions. Hence, they are low-ranked.

The negative and phoneme-specific formulation of an auditory cue such as [separated F2 and F3] is appropriate for /b/ and /d/; hence the constraints “*/b/ [separated F2 and F3]” and “*/d/ [separated F2 and F3]” have to be ranked low. High-ranking of the unfounded constraint “*/g/ [separated F2 and F3]”, however, can rule out perception of the observed articulation of /ga/.

According to Boersma (2006), the lexicon-driven acquisition algorithm (cf. Boersma, 1998) determines the appropriate ranking of the relevant constraints, giving rise to the McGurk percept. Then, the visual cue [open lips] is assumed to be ranked higher than the auditory cue [F2 & F3 separated] since the former is more reliable in the light of the omnipresent acoustic noise (Boersma, 2006). The degree of preference of one constraint over the other is determined by the awareness of perceptual errors and their origin. If perceptual errors are more often due to an incorrect ranking of a particular (acoustic) cue than to a particular (visual) cue, the grammar will end up ranking this visual cue proportionately higher (probability matching; Boersma, 1998: 339).

In summary, according to Boersma (2006), the McGurk effect occurs at the prelexical level and the ranking of particular cue (and structural) constraints are the decisive factors in its perception. The lexicon is only assumed to play a role in the acquisition process when a speaker or listener learns to select the relevant constraints and to rank them appropriately.