A constraint-based explanation of the McGurk effect

Paul Boersma, 3 July 2011

Abstract. This paper gives an Optimality-Theoretic explanation of the McGurk-effect, a robust phenomenon that illustrates the low-level interaction of visual and auditory cues in speech perception. Perception tableaus illustrate the interaction between cue constraints (which evaluate the relation between sensory and phonological representations) and language-specific structural constraints (which evaluate phonological representations). The procedure of lexicon-driven perceptual learning then explains how the constraints automatically come to be ranked as they are. Finally, it is shown that the same structural and cue constraints are used in production by the speaker.

The *McGurk effect* is a spectacular phenomenon that can arise with manipulated spoken language in the laboratory. It occurs when visual cues to phonological categories override auditory cues. McGurk & MacDonald (1976) devised a videotape on which the visual part of the recording was from a person saying [ɡɑ], whereas the auditory part of the recording was from a person saying [bɑ]. Listeners were then asked what consonant they heard when watching the video. Although all the auditory information pointed at /bɑ/, they in fact reported hearing /dɑ/.

The McGurk effect is *robust*, i.e. it tends to occur even in cases where the listener knows what is going on, i.e. that the sound is that of somebody saying [bɑ]: “[these effects] do not habituate over time, despite objective knowledge of the illusion involved. By merely closing the eyes, a previously heard /dɑ/ becomes /bɑ/ only to revert to /dɑ/ when the eyes are open again.” [McGurk & MacDonald 1976, p. 747][1]

McGurk & MacDonald’s interpretation was that the main piece of visual information (namely, the open lips when mouthing [ɡɑ]) was compatible with perceiving either /dɑ/ or /ɡɑ/, whereas some of the auditory information of a sounding [bɑ] was compatible only with /bɑ/ and /dɑ/ (and not with /ɡɑ/), so that “the unified percept” /dɑ/ is most compatible with the visual and auditory cues combined. McGurk & MacDonald did not specify what the common auditory cues for /bɑ/ and /dɑ/ could be.

McGurk & MacDonald’s observations were the starting point of a surge of interest among speech perception researchers, generally confirming the robustness that McGurk & MacDonald had anecdotally described, as well as corroborating McGurk & MacDonald’s speculative interpretation. As a result, the McGurk effect is nowadays generally seen as a case of low-level multimodal cue integration. The effect has turned out to be stronger in adults than in children (McGurk & MacDonald, 1976), and stronger in English than in several other languages, which also differ with respect to each other (Sekiyama & Tohkura 1991; Grassegger, 1995; Burnham, 1998).

The present paper provides a description and explanation of the McGurk effect within an integrative formal model of bidirectional phonology and phonetics, in which

[1] I adapted the notation in this quote to the one used elsewhere in the present paper, namely with slashes for phonological surface structures such as /bɑ/ and /dɑ/, and with square brackets for auditory or visual peripheral representations such as $[\text{bɑ}]_{\text{Aud}}$ and $[\text{ɡɑ}]_{\text{Vis}}$. McGurk & MacDonald used square brackets throughout, but do make a distinction in the text between auditory and visual on the one hand, and (response or phonological) categories on the other.

decisions for speaking and listening are made with the help of ranked constraints that evaluate phonological and phonetic representations and their relations.

1. Representations and constraints

Figure 1 shows a bidirectional model of phonology and phonetics (Boersma 1998, 2007), which contains four connected mental *representations*. The number of *phonological* representations is the minimum that phonologists regard as sensible, namely two: the Underlying Form, which is a sequence of discrete phonological structures stored in the lexicon, and the Surface Form, which is an equally discrete phonological structure consisting of features, segments, syllables, and feet. The number of *phonetic* representations is also minimal, namely two: the continuous Auditory Form (pitch, formants, duration, silence, noise) and the equally continuous Articulatory Form (muscle gestures).

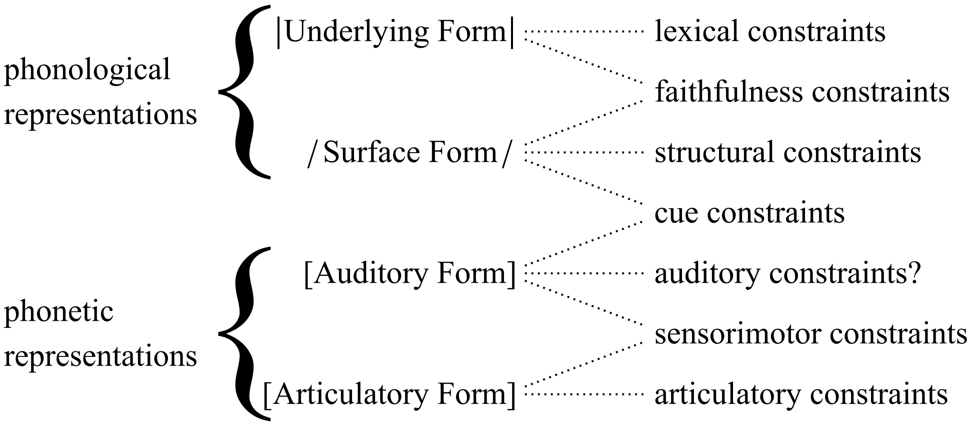

Fig. 1 Representations and constraints (sound only).

Several *processes* can be defined on the representations in the figure. In the process of *comprehension*, the listener is given an auditory form and ultimately has to find underlying forms in the lexicon (which again connect to meaning and world knowledge, not represented in Figure 1). In the process of *production*, the speaker starts with an underlying form (which is fed itself by an intended meaning) and ultimately has to decide on an articulatory form that will generate the sound of speech.

The decisions made in the processes of comprehension and production are guided by *constraints*, which are the elements of the grammar.

The four representations are connected by three types of constraints, which express the speaker-listener's knowledge of the relations between the representations. The faithfulness constraints favour similarity of underlying and surface form in production (McCarthy & Prince 1995) as well as in comprehension (Smolensky 1996). The cue constraints (Escudero & Boersma 2003) express the speaker-listener's knowledge of the relation between phonological features and auditory cues. The sensorimotor constraints express the speaker's knowledge of the relation between muscle commands and sound, and are only needed in production.

The four representations themselves are evaluated by two more kinds of constraints. The articulatory constraints militate against articulatory effort and are used in production alone. The structural constraints disfavour selected surface structures and are used in production (Prince & Smolensky 1993) as well as in comprehension (Tesar 1997). The auditory constraints, if they exist, will militate against loud and otherwise unpleasant sounds.

If visual cues have to be included in the model of Figure 1, it is the Auditory Form that will have to be generalized. The result is in Figure 2, which now contains a general Sensory Form. The cue constraints now express the speaker-listener's knowledge of the relation between phonological features and both auditory and visual cues. The sensorimotor constraints now express the speaker's knowledge of the relation between muscle commands on the one hand and sound and vision (e.g. visible lip closure) on the other. The sensory constraints, if they exist, will now militate as well against flashing and otherwise unpleasant sights.

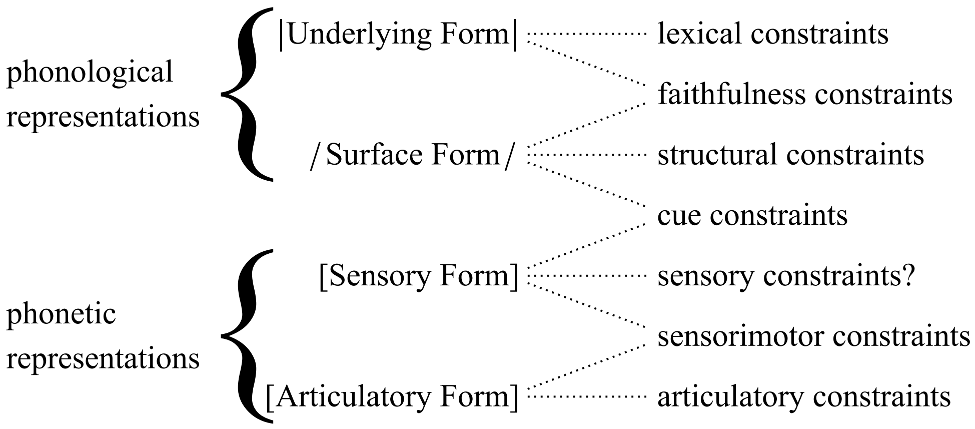

Fig. 2 Representations and constraints (both sound and vision).

2. The McGurk effect as low-level perception

In the case of the McGurk effect, the input to the process is the sensory form that consists of [bɑ]-like auditory cues and [gɑ]-like visual cues. The output of the process is a nonsense syllable chosen from the candidates /bɑ/, /dɑ/, and /gɑ/. Since these are not lexical items, they are just instances of a Surface Form. Therefore, to describe the McGurk effect as it happens in the laboratory, we only need to consider the Sensory Form and the Surface Form in the figure; the Underlying Form and Articulatory Form are irrelevant (in §3, where an *explanation* for the McGurk effect is given, the Underlying Form will turn out to be relevant as well; and in §5, where *production* is considered, the Articulatory Form may play a role).

In the mapping from Sensory Form to Surface Form, two kinds of constraints could be relevant according to Figure 2, at least if an Underlying Form does not have to be constructed in parallel. These two constraints are the cue constraints and the structural constraints (any sensory constraints would evaluate the input only, but this could never decide between the output candidates). Of these, I propose that the structural constraints are irrelevant, because /bɑ/, /dɑ/, and /gɑ/ are all fully legitimate syllables in English. This leaves the cue constraints as the only kind of constraints relevant to the description of the McGurk effect.

Cue constraints tend to be formulated negatively (Boersma 2007). To see this in the McGurk case, consider what the visually present open lips tell the viewer about the possible sounds. A positively formulated constraint would say “[open lips] can be /d/ **or** /t/ **or** /n/ **or** /g/ **or** /k/ **or** /ŋ/”, whereas a negatively formulated constraint would say “[open lips] is **not** /b/ **and not** /p/ **and not** /m/”. As shown by Boersma & Escudero (2008), the negative formulation can be split into its parts, the positive formulation cannot. The relevant high-ranked visual cue constraints, therefore, are “[open lips] is not /b/”, “[open lips] is not /p/”, and so on, or in a more symmetric notation (which can be used for production as well, see §6): “*/b/[open lips]” and “*/p/[open lips]”. The arbitrariness of the cue constraints (Escudero & Boersma 2003) demands the additional existence of constraints like “*/d/[open lips]”, and “*/g/[open lips]”, but these must be low ranked in order to describe the McGurk effect (§3 explains how the constraints have become ranked as they are).

A high ranking of “*/b/[open lips]”, then, describes the fact that listener-viewers of the McGurk movie are reluctant to perceive /bɑ/. But what makes them perceive /dɑ/ rather than /gɑ/? The answer cannot be a visual cue constraint, because the motion was that of somebody mouthing, in fact, [gɑ]. So the answer must be an auditory cue constraint. An auditory cue that would have favoured the perception of /gɑ/ would have been a close approach of the second and third formants during the transition from the plosive to the vowel (Ladefoged & Maddieson 1996, Stevens 1998). In an auditory [bɑ], F2 and F3 are fairly separated, as they are in [dɑ]. Hence, a high-ranked “*/g/[separated F2 & F3]” could eliminate the candidate /gɑ/. Apparently, the information about F2–F3 separation is more important than the information about F2 alone, which must have been low for the auditory [bɑ] and disfavours the /dɑ/ perception more than the /gɑ/ perception. All these considerations are summarized in perception tableau (1), which shows the formalization of this phenomenon within the decision-making framework of Optimality Theory.

(1) *The McGurk effect (eyes open)*

[open lips, separated F2 & F3, low F2]	*/b/ [open lips]	*/g/ [separated F2 & F3]	*/d/ [low F2]	*/g/ [low F2]	*/b/ [low F2]	*/d/ [open lips]	*/g/ [open lips]
/bɑ/	*!				*		
☞ /dɑ/			*			*	
/gɑ/		*!		*			*

Tableau (2) shows with the same grammar (constraint ranking) that if the listener has her eyes closed, she will perceive /b/. This is because the visual cue is no longer present.

(2) *The McGurk effect (eyes closed)*

[separated F2 & F3, low F2]	*/b/ [open lips]	*/g/ [separated F2 & F3]	*/d/ [low F2]	*/g/ [low F2]	*/b/ [low F2]	*/d/ [open lips]	*/g/ [open lips]
☞ /bɑ/					*		
/dɑ/			*!				
/gɑ/		*!		*			

Tableaus (1) and (2) together describe the phenomenon that listener-viewers who alternatingly close and open their eyes while repeatedly watching a McGurk movie alternate between perceiving /bɑ/ and /dɑ/.

3. The explanation of the McGurk effect

The previous section presented a description of the McGurk effect in terms of constraint ranking, but did not provide an explanation of how the constraints have come to be ranked as they are. This section explains the ranking as a result of lexicon-driven acquisition of Optimality-Theoretic perception (Boersma 1998).

The lexicon-driven acquisition algorithm by Boersma (1998) is capable of explaining how constraints for more reliable cues become higher ranked than constraints for less reliable cues. The only assumption needed for explaining the McGurk effect is then that the [open lips] cue, *if present*, is more reliable than the [F2 & F3 separated] cue. This assumption is quite plausible, because acoustic noise is omnipresent.

Now suppose that the visual cue constraints are ranked at a height where they do not contribute much to the perception decision, as in tableau (3). This ranking means that the choice among candidates like /bɑːn/, /dɑːn/, and /gɑːn/ will usually be made on the basis of the formant cues alone. Assume, moreover, that the [open lips] cue is more reliable than any of the formant cues. This greater reliability means that cases with incorrectly available formant cues are more common than cases with incorrectly available lip cues. One of these more common cases is shown in tableau (3). In this example, an intended underlying |dɑːn| is incorrectly transmitted as having a [low F2] cue, while the lip cue (open lips) and the [separated F2 & F3] cue are transmitted correctly. The perceived structure is /bɑːn/.

(3) *Acquiring the McGurk effect*

[open lips, separated F2 & F3, low F2] (intended |dɑːn|)	*/g/ [separated F2 & F3]	*/d/ [low F2]	*/b/ [open lips]	*/d/ [open lips]	*/g/ [open lips]	*/g/ [low F2]	*/b/ [low F2]
☞ /bɑːn/			←*				←*
√ /dɑːn/		*!→		*→			
/gɑːn/	*!				*	*	

Now suppose that the listener subsequently accesses meaning in the lexicon, and the lexicon tells here, informed by semantic considerations, that the speaker's intended word was |dɑːn| 'darn'. The proposal of lexicon-driven acquisition of perception (Boersma 1998: 338) now implies that the listener will consider the candidate /dɑːn/ to be the *correct candidate* in tableau (3), i.e. the candidate that she should have perceived but didn't. This lexicon-informed knowledge is depicted in tableau (3) by supplying the candidate /dɑːn/ with a check mark.

The fact that the listener's perceived candidate in tableau (3), namely /bɑːn/, differs from the correct candidate means that she has made a *perceptual error*. The fact that the lexicon has told the listener what the correct candidate was, namely /dɑːn/, implies that the listener "knows" that she has made this error, so that she is "aware" that her grammar (constraint ranking) may be in need of modification. The fact that the correct candidate occurs in the tableau implies that the listener's Gradual Learning Algorithm "knows" how the constraint ranking has to be modified. The required modification is that the constraints that prefer the correct candidate, namely "*/b/[open lips]" and "*/b/[low F2]" will have to be raised a bit, and that the constraints that prefer the learner's incorrect winner, namely "*/d/[low F2]" and "*/d/[open lips]", will have to be lowered a bit. These raisings and lowerings are indicated in the table by arrows.

In tableau (3) we can see that the raisings and lowerings indicated by the arrows will ultimately cause the visual cue constraint "*/b/[open lips]" to rise above the auditory cue constraint "*/d/[low F2]", which is enough to produce the McGurk effect, which has therefore now been explained, although two minor issues have to be resolved.

One minor issue is a possibly unwanted side effect predicted by tableau (3), namely the rise of "*/b/[low F2]". This side effect will be counteracted by the lexicon as soon as a too high ranking of this constraint will cause spurious perceptions of intended |b| (with correct [low F2]) as the incorrect categories /d/ or /g/. In the end, the most reliable constraints will emerge on top.

The other minor issue is the question how far "*/b/[open lips]" will end up being ranked above "*/d/[low F2]". The answer according to Boersma (1998: 339) is *probability matching*. If incorrect instances of the auditory cue [low F2] appear 20 times more often than incorrect instances of the visual cue [open lips], the listener's grammar will in the end favour the cue [open lips] over the cue [low F2] 20 times more often than the reverse. The strong reproducibility of the McGurk effect suggests that such high factors are indeed involved. The probability matching property of the learning algorithm does predict that the McGurk effect is less strong for people who are used to watching dubbed movies and have therefore learned to ignore visual cues to some extent.

4. The interaction of structural and cue constraints

Audio-visual perception seems not to be handled by cue constraints alone. Language-specific structural constraints also seem to play a role. This is predicted by Figure 2, where the output of the mapping from Sensory Form to Surface Form can be evaluated by the same structural constraints that phonologists use to model production. Evidence is found in what English viewers-listeners do when the video mouths [bɑ], but sounds [gɑ], i.e. the opposite combination of the main McGurk effect described above. McGurk & MacDonald (1976) report that people will often hear /bɑgbɑ/, where the initial consonant cluster is only labial and the second consonant cluster is both velar and labial. My interpretation is that the visual labiality and the auditory velarity do not conflict intervocalically, because in that position a cluster of two plosives is allowed phonotactically in English (as in the word *rugby*). By contrast, such a cluster is not allowed in initial position, so that viewers-listeners can only decide for a single consonant. Tableau (4) summarizes, and includes syllable boundaries for explicitness. The constraint "[closed lips]⇒/lab/" reads as "*if* the sensory form has closed lips, *then* there must be a labial" (and can be used bidirectionally, as we will see in §6).

(4) *The reverse McGurk effect*

[closed lips]$_{\text{Vis}}$ [close F2 & F3]$_{\text{Aud}}$	*/.labvel/	*/.C./	[closed lips]$_{\text{Vis}}$ ⇒/lab/	[close F2 & F3]$_{\text{Aud}}$ ⇒/vel/
/.gɑ.gɑ./			*!*	
☞ /.bɑg.bɑ./				*
/.gbɑg.bɑ./	*!			
/.gbɑ.gbɑ./	*!*			
/.g.bɑg.bɑ./		*!		
/.bɑ.bɑ./				**!

The constraint "*/.labvel/" is an abbreviation for "no labial-velar sequences in onset". It can be seen that the tableau requires the additional structural constraint "/.C./", which militates against syllables without vowels, which are not allowed in English.

The conclusion is that the asymmetry between initial /bɑ/ and medial /gbɑ/ is caused by language-specific structural constraints.

5. Why OT and not neural nets? The case of phonological production

The McGurk effect and its acquisition were modelled successfully by using the decision mechanism of Optimality Theory. The question naturally arises why the tried and tested decision mechanism of neural net classification was not used instead. The answer is: because the perceptual decision is influenced by language-specifically ranked structural constraints. These constraints are linguistic because (1) they are language-specific or language-specifically ranked, and (2) they are also used in production, where they interact with faithfulness constraints.

The language-specificity of the constraints and/or their ranking follows both from the language-specificity of the strength of the McGurk effect itself and from the language-specificity of the ranking of */.labvel/ in (4). After all, listeners of a Slavic language like Czech should have no trouble perceiving a /bg/ or /gb/ cluster, given the existence of phrases like /.gbr̩.nu./ 'to Brno', and could therefore favour a winner like /.gbag.ba./, very similar to the third candidate in (4). Likewise, listeners of a Gbe language like Ewe would have no trouble perceiving labial-velar plosives like /g͡b/, and might therefore favour a winner like /.g͡ba.g͡ba./, somewhat similar to the fourth candidate in (4) (the third candidate would be ruled out by a constraint against coda consonants). To see whether this prediction is true, my colleague Kateřina Chládková manufactured a video of herself saying [baba] visually and [gaga] auditorily.[2] Nine Czech listeners and one Yoruba listener, when asked to write down what they "heard", all reported hearing both a velar and a labial consonant, sometimes with the labial first (6 times "bgabga", 2 times "mgamga"), sometimes with the velar first (2 times "gbagba"); a Gbe (Fongbe) listener heard "bgaga". By contrast, 25 Dutch listeners displayed a variety of strategies: next to 6 "integrating" perceptions (4 times "bgabga", 1 time "mgamga", 1 time "gbagba"), they failed to report one of the labials or velars in 9 cases (2 times "gaabgaa", 1 time "gaapga", 1 time "gabga", 2 times "mgaga", 1 time "mgaagaa", 1 time "bgaga", 1 time "bkaka") and failed to report two labials or velars in 10 cases (8 times "gaga", 1 time "ganggang", 1 time "mama").[3] A chi-square test on whether listeners fully integrate or not ([[10, 1], [6, 19]], *df* = 1), yields a two-tailed *p* value of 0.0008, indicating that the Dutch group performed differently from the Czech-African group.[4]

The Dutch results may be explained by the idea that people either report an analytical perception ("gbagba"), a true phonological perception ("gabga"), a phonological perception with incorrect localisation of the labial ("bgaga"), or a phonological perception influenced by an idea of repetition ("gaga"); the analytical perception may arise from a listener's individual ranking of [close F2 & F3]$_{\text{Aud}}$⇒/vel/ above */.labvel/, the other report from a listener's individual high ranking of */.labvel/, as in tableau (4). By contrast, the Czech results are never influenced by phonotactic restrictions such */.labvel/, so that the Czech always report "bgabga", which is both analytically and phonologically correct. This finding has to qualify the speculation by Mills & Thiem (1980) that "It might be expected that the perception of combinations [i.e. things like /bg/, PB] would be governed by phonotactic rules, but this is not at all the case." Mills & Thiem based their speculation on the analysis of their results with a single language (German, which is very similar to Dutch in this respect), and although my experiment finds that some Dutch listeners do hear "combinations" like "bgabga", their number is much smaller than for Czech listeners; thus, language-specific phonotactic restrictions do influence the perceived structures at least probabilistically.

The strongest argument in favour of a linguistic analysis of McGurk perception is the fact that the same structural constraint */.labvel/ is used in production. Suppose that a speaker of English knows the name of the language group Gbe. Her underlying

[2] The recording has her saying the word three times. A listener's response reported in this paper was constructed by noting the listener's most frequently occurring response for the first and second syllable separately. For instance, if a listener wrote "bgaga gaga gamga", this is counted as "gaga". This worked because a listener never reported three different perceptions for the first (or second) syllable.

[3] In addition we tested two teachers of phonetic transcription, including Norval Smith. Reassuringly, both of them reported hearing "gbagba" or "bgabga".

[4] The failure of the Gbe listener to perceive a labial-velar plosive can be explained by the fact that the labial-velar [g͡b] does *not* lie auditorily close to [g]; in fact, viewed from [g] it lies even beyond [b]: it has by far the lowest F2 locus of all stops, just like the labial-velar vowel [u] has a lower F2 than the exclusively labial [y] and the exclusively velar [ɯ]. Ladefoged & Maddieson (1996) show spectrograms.

form, partly based on the orthography, will be |gbei| (at least if she has the vowel right). When asked to produce this word, however, she will say [gəbei], suggesting a phonological surface form /.gə.bei./. The tableau that describes this schwa insertion is given in (5). In this tableau, MAX is the usual faithfulness constraint against having underlying segments that do not correspond to anything in the surface form, and DEP is the usual faithfulness constraint against having segments in the surface form that do not correspond to anything in the underlying form (McCarthy & Prince 1995).

(5) *Schwa insertion in English: the same structural constraints as in perception*

|gbei|	*/.labvel/	*/.C./	MAX	DEP
/.gbei./	*!			
/.g.bei./		*!		
/.bei./			*!	
☞ /.gə.bei./				*

Production tableau (5), therefore, makes it plausible that the candidates that are most faithful to the underlying form (namely, the first and second candidates), are ruled out by the very same constraints that rule out candidates 3, 4, and 5 in perception tableau (4). In Figure 1, we can indeed see that the surface form is the output of both prelexical perception and phonological production, so that the constraints that evaluate this surface form (namely, the structural constraints) must be able to restrict the outputs of both prelexical perception and phonological production. Whereas Prince & Smolensky (1993) and most of the OT literature since stressed the use of structural constraints in production, and Boersma (1998 et seq.) stressed their use in perception, the bidirectional use of these constraints in comprehension as well as production was stressed by Tesar (1997) and Pater (2004), and in (4) and (5) we see another example of this bidirectional use (for a detailed example of this bidirectionality in the phonology of a single language, see Boersma & Hamann 2009).

If, now, the interaction between structural and faithfulness constraints is uncontroversially linguistic and therefore has to be modelled with OT (and not with neural nets), then the interaction between structural and cue constraints must also be linguistic and has to be modelled with OT as well. Otherwise, the strength of the same entities (namely, the structural constraints) would at the same time be measured in terms of ranking (in production) and in terms of weighting (in comprehension), an unwanted duplication of theoretical elements. Of course, language is ultimately performed by the brain, so the ultimately correct theory of language processing *will* involve neural networks, but these will then have to implement structural constraints as well as an OT-like decision mechanism (if that is how language works).

6. OT in phonetic production

If structural constraints can be used bidirectionally, then perhaps the cue constraints can be used bidirectionally as well, namely to specify what auditory cues the speaker should produce for a given underlying form.

This turns out to be correct. Suppose that the speaker wants to produce the underlying form |bɑgɑ|. The cue constraints will explain why she pronounces this as [bɑgɑ] (i.e. [closed lips, separated F2 & F3, low F2] followed by [ɑ] followed by [open lips, close F2 & F3] followed by [ɑ]) rather than as [bɑbɑ] or [gɑbɑ]. Tableau (6) gives all 16 relevant candidates, assuming that the surface form, the auditory form, and the articulatory form are evaluated in parallel, i.e., that every output candidate is a triplet of surface, auditory, and articulatory forms. The constraint IDENT is the usual faithfulness constraint that evaluates the identity of a pair of corresponding Underlying and Surface segments (McCarthy & Prince 1995); the subscript Sens is short for both Vis and Aud; and the articulatory representations look very similar to the sensory representations because I assume that sensorimotor knowledge is perfect. An example of the workings of cue constraints in production is that the fifth candidate, /.bɑ.gɑ./ [bɑbɑ]$_{\text{Sens}}$, violates [closed lips]$_{\text{Vis}}$⇒/lab/ because the second sensory [b]$_{\text{Sens}}$ must have been pronounced with visibly closed lips, although the corresponding surface segment /g/ (the onset of the second syllable) is not labial.

(6) *Phonetic production of plosives: the same cue constraints as in perception*

|bɑgɑ|	IDENT	[closed lips]$_{\text{Vis}}$ ⇒/lab/	[close F2 & F3]$_{\text{Aud}}$ ⇒/vel/
/.bɑ.bɑ./ [bɑbɑ]$_{\text{Sens}}$ [bɑbɑ]$_{\text{Art}}$	*!		
/.bɑ.bɑ./ [bɑgɑ]$_{\text{Sens}}$ [bɑgɑ]$_{\text{Art}}$	*!		*
/.bɑ.bɑ./ [gɑbɑ]$_{\text{Sens}}$ [gɑbɑ]$_{\text{Art}}$	*!		*
/.bɑ.bɑ./ [gɑgɑ]$_{\text{Sens}}$ [gɑgɑ]$_{\text{Art}}$	*!		**
/.bɑ.gɑ./ [bɑbɑ]$_{\text{Sens}}$ [bɑbɑ]$_{\text{Art}}$		*!	
☞ /.bɑ.gɑ./ [bɑgɑ]$_{\text{Sens}}$ [bɑgɑ]$_{\text{Art}}$			
/.bɑ.gɑ./ [gɑbɑ]$_{\text{Sens}}$ [gɑbɑ]$_{\text{Art}}$		*!	*
/.bɑ.gɑ./ [gɑgɑ]$_{\text{Sens}}$ [gɑgɑ]$_{\text{Art}}$			*!
/.gɑ.bɑ./ [bɑbɑ]$_{\text{Sens}}$ [bɑbɑ]$_{\text{Art}}$	*!*	*	
/.gɑ.bɑ./ [bɑgɑ]$_{\text{Sens}}$ [bɑgɑ]$_{\text{Art}}$	*!*	*	*
/.gɑ.bɑ./ [gɑbɑ]$_{\text{Sens}}$ [gɑbɑ]$_{\text{Art}}$	*!*		
/.gɑ.bɑ./ [gɑgɑ]$_{\text{Sens}}$ [gɑgɑ]$_{\text{Art}}$	*!*		*
/.gɑ.gɑ./ [bɑbɑ]$_{\text{Sens}}$ [bɑbɑ]$_{\text{Art}}$	*!	**	
/.gɑ.gɑ./ [bɑgɑ]$_{\text{Sens}}$ [bɑgɑ]$_{\text{Art}}$	*!	*	
/.gɑ.gɑ./ [gɑbɑ]$_{\text{Sens}}$ [gɑbɑ]$_{\text{Art}}$	*!	*	
/.gɑ.gɑ./ [gɑgɑ]$_{\text{Sens}}$ [gɑgɑ]$_{\text{Art}}$	*!		

This ranking, then, makes sure that an Underlying |b| is realized as a Surface /b/ because of the faithfulness constraints, and as a Sensory [b] because of the cue constraints (at least if there is no high ranked articulatory constraint, i.e. *[b]$_{\text{Art}}$, against producing labials). Note that in case some phonological rule had turned an underlying |g| into a Surface /b/, the cue constraints would have made sure that the

Sensory form would have been pronounced as [b], as most phonologists would expect.

What tableau (6) shows, then, is that OT can handle both phonological and phonetic production, by using the same cue constraints as in perception.

7. Conclusion

The McGurk effect in prelexical ('phonetic') perception can be described as an interaction of the same structural and cue constraints that also regulate phonological and phonetic production, respectively.

Acknowledgments

Thanks go to Kateřina Chládková for creating the video and testing the Czech listeners, and for Jan-Willem van Leussen for testing most of the Dutch listeners.

References

Boersma, Paul (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, Paul (2007). Some listener-oriented accounts of *h*-aspiré in French. *Lingua* **117**: 1989–2054.

Boersma, Paul, and Paola Escudero (2008). Learning to perceive a smaller L2 vowel inventory: an Optimality Theory account. In Peter Avery, Elan Dresher & Keren Rice (eds.) *Contrast in phonology: theory, perception, acquisition.* Berlin & New York: Mouton de Gruyter. 271–301.

Boersma, Paul, and Silke Hamann (2009). Loanword adaptation as first-language phonological perception. In Andrea Calabrese & W. Leo Wetzels (eds.) *Loanword phonology*. Amsterdam: John Benjamins. 11–58.

Burnham, Denis (1998). Language specificity in the development of auditory-visual speech perception. In R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*. Hove, UK: Psychology Press. 27–60.

Escudero, Paola, and Paul Boersma (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. In Sudha Arunachalam, Elsi Kaiser, and Alexander Williams (eds.), *Proceedings of the 25th Annual Penn Linguistics Colloquium. Penn Working Papers in Linguistics* **8.1**: 71–85.

Grassegger, H. (1995). McGurk effect in German and Hungarian listeners. *13th International Congress of Phonetic Sciences*, Vol. 3, pp. 210–213.

Ladefoged, Peter & Ian Maddieson (1996). *The sounds of the world's languages*. Oxford: Blackwell.

McCarthy, John, and Alan Prince (1995). Faithfulness and reduplicative identity. In Jill Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.) *Papers in Optimality Theory*. University of Massachusetts Occasional Papers **18**. Amherst, Mass.: Graduate Linguistic Student Association. 249–384. [Rutgers Optimality Archive **60**, http://roa.rutgers.edu]

McGurk, Harry, and John MacDonald (1976). Hearing lips and seeing voices. *Nature* **264**: 746–748.

Mills, Anne E., and Rudolf Thiem (1980). Auditory–visual fusions and illusions in speech perception. *Linguistische Berichte* **68**: 85–108.

Pater, Joe (2004). Bridging the gap between receptive and productive development with minimally violable constraints. In René Kager, Joe Pater & Wim Zonneveld (eds.) *Constraints in phonological acquisition*. Cambridge: Cambridge University Press. 219–244.

Prince, Alan, and Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report TR-**2**, Rutgers University Center for Cognitive Science.

Sekiyama, K., & Y. Tohkura (1991). McGurk effect in non-English listeners. *Journal of the Acoustical Society of America* **90**: 1797–1805.

Smolensky, Paul (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry* **27**: 720–731.

Stevens, Kenneth N. (1998): *Acoustic Phonetics*. Cambridge, Mass. & London: MIT Press.

Tesar, Bruce (1997). An iterative strategy for learning metrical stress in Optimality Theory. In Elizabeth Hughes, Mary Hughes & Annabel Greenhill (eds.), *Proceedings of the 21st Annual Boston University Conference on Language Development*, 615–626. Somerville, Mass.: Cascadilla.