Cognition 109 (2008) 168–173

Brief article

The link between speech perception and production is phonological and abstract: Evidence from the shadowing task

Holger Mitterer [a,*], Mirjam Ernestus [a,b]

[a] Max-Planck-Institut für Psycholinguistik, Wundtlaan 1, Nijmegen, The Netherlands
[b] Radboud Universiteit Nijmegen, Faculteit der Letteren, Erasmusplein 1, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 21 September 2007
Revised 7 August 2008
Accepted 8 August 2008

Keywords:
Speech perception
Speech production

ABSTRACT

This study reports a shadowing experiment, in which one has to repeat a speech stimulus as fast as possible. We tested claims about a direct link between perception and production based on speech gestures, and obtained two types of counterevidence. First, shadowing is not slowed down by a gestural mismatch between stimulus and response. Second, phonetic detail is more likely to be imitated in a shadowing task if it is phonologically relevant. This is consistent with the idea that speech perception and speech production are only loosely coupled, on an abstract phonological level.

1. Introduction

The first author vividly recalls a journey to the north-eastern part of Germany, during which a travel companion used the uvular trill [R] (also used in French, produced with the tongue in the back of the mouth), but started using the alveolar trill [r] (also used in Spanish, with the tongue in the front of the mouth) on arrival when speaking to her mother. Many of us have experienced such phonetic imitations, suggesting a rather tight link between perception and production.

Gestural theories of speech perception such as Motor Theory (Liberman, 1957; Liberman & Whalen, 2000) or Direct-Perception Theory (Fowler, 1996; Fowler & Smith, 1986; Goldstein & Fowler, 2003) easily explain such a link. Motor theory assumes that listeners use special machinery to infer the intended speech gestures, while the direct-perception account assumes that the acoustic signal provides information about the gestures. If speech is perceived along gestural lines, it would not only explain why one would take over the gestures of one's interlocutors, but it also solves the invariance problem in speech perception to a great degree. The acoustic differences between, for instance, the vowel /i/ produced by a child and an adult male become irrelevant, as listeners will in both cases perceive a similar gesture.

One type of evidence for gestural theories of speech perception comes from the shadowing task, in which participants repeat a speech stimulus as fast as possible. Marslen-Wilson (1973, 1975) found that shadowing complete sentences is possible at very short latencies (about 350 ms for fast participants). If participants shadow simple stimuli from a limited stimulus set, average latencies are even as short as 180 ms. This is only slightly longer than the 150 ms observed for simple responses – saying "ba" on hearing any stimulus (Fowler, Brown, Sabadini, & Weihing, 2003; Porter & Castellanos, 1980; Porter & Lubker, 1980). In simple response tasks, there is also a congruency effect: Latencies are shorter if the stimulus is the same as the pre-assigned response. Fowler et al. argue that the congruency is gestural in nature so that the shadowing data support gestural theories of speech perception.

Alternatively, the stimulus–response congruency in the shadowing task may be due to learned associations between perceptual cues, phonological categories, and motor representations. Fowler et al. (2003) tested this alternative phonological account by analyzing not only the latency but

* Corresponding author. Tel.: +31 24 3521375.
E-mail address: holger.mitterer@mpi.nl (H. Mitterer).

also the phonetic properties of the shadowing response. They found that shadowers imitate sub-phonemic variation in the stimulus: They produce stops with more aspiration (i.e. longer voice-onset-time, VOT) if the stimulus also has more aspiration (see also Shockley, Sabadini, & Fowler, 2004). Fowler et al. argue that a phonological link between perception and production would not predict shadowing of phonetic detail, because phonemes behave categorically.

However, the degree of aspiration in unvoiced stops, the sub-phonemic variation studied by Fowler et al. (2003), is phonologically relevant, that is, it contributes to the decoding of the message. The amount of aspiration is a cue for lexical segmentation in English, because there is more aspiration in word-initial (e.g., "use *p*ies") than non-initial stops ("you s*p*ies"). A phonological account can thus accommodate imitation of aspiration.

Against this background, we report a shadowing experiment in Dutch to differentiate a phonological from a gestural account of the link between speech production and perception. We address two questions: (1) How is response latency affected if the shadower cannot imitate the gestures of the stimulus and produces a different but phonologically equivalent response? (2) Does phonological relevance matter for the imitation of phonetic detail?

We addressed the first question by presenting shadowers with two variants of the Dutch phoneme /r/, the alveolar and uvular trill. These are radically different gestures, which in other languages, such as Berber, represent different phonemes. For the alveolar trill, the tongue tip trills near the alveolar ridge, whereas it is the velum that trills for the uvular trill with the tongue moved to the back (Ladefoged & Maddieson, 1996). Because most Dutch speakers master only one of these two variants, they need to shadow an alveolar trill with the mismatching gestures of an uvular trill, or vice versa. If the incoming speech is perceived gesturally, this mismatch should lead to a latency cost. However, the phonological account predicts no latency cost, because both types of /r/ match the same phonological representation.

The role of phonological relevance was investigated based on work of van Alphen and McQueen (2006). They showed that the difference between no pre-voicing and some pre-voicing in voiced stops is phonologically relevant, while the *amount* of pre-voicing, even though audible, is irrelevant. The phonological account hence predicts that only the absence versus presence of pre-voicing is imitated in a shadowing task. The gestural account in contrast predicts that the amount of prevoicing is imitated as well (e.g., Fowler et al., 2003, p. 408).

2. Method

2.1. Participants

Eighteen native speakers of Dutch participated in the experiment for pay; half habitually using an alveolar, the other half an uvular variant of /r/. Participants were students aged 18–26 and reported no history of hearing or speech disorder.

2.2. Materials

All stimuli were Dutch nonwords, produced by a female Dutch speaker, who is special in being able to produce both alveolar and uvular trills.

As in Fowler et al. (2003), participants heard two syllables on each trial (see upper panel of Fig. 1). The fist syllable consisted of /p/, /t/, or /k/, followed by /a/, /i/, /y/, or /u/, with a very long vowel (0.8–1.6 s). This long syllable induces overlapping shadowing, in which the participant does not wait for the end of the stimulus to start responding (see the lower panel of Fig. 1). The duration of this syllable varied to make the trial structure less predictable, and keep the participants vigilant. These initial syllables were recorded in isolation so that they carried no cues to the identity of the second syllable. The speaker was instructed to produce these syllables with durations of 0.8, 1.2, and 1.6 s by use of an electronic LED metronome. The tokens used for the experiment were on average 0.86, 1.16, and 1.58 s long (SD $\approx$ 0.05 s).

The initial syllable was followed by 500 ms of silence and then by the second syllable. This was the experimental item of interest, for which latency and phonetic properties were coded. These were pseudowords: 10 CVVC (C = consonant, V = Vowel) pseudowords starting with /r/, 18 CVVC pseudowords starting with a voiced stop (/b/ or /d/), and 24 fillers. The speaker produced the /r/-initial words once with an alveolar and once with an uvular trill. All realizations of /r/ had only one closure except for four alveolar realizations. This (first) closure was achieved earlier in the alveolar trills (41 ms, SD = 11 ms) than in the uvular trills (66 ms, SD = 18 ms). We tested whether native listeners of Dutch can distinguish these alveolar and uvular versions of our stimuli. Twenty listeners were trained with feedback on half of the tokens to identify sequences of two stimuli as alveolar–uvular versus uvular–alveolar. This training was necessary, because phonetically naive listeners do not know the terms "alveolar" and "uvular". Then, the participants were tested without feedback on sequences from the other half of the tokens. The mean d' were 2.8 for the training and 3.3 for the test phase, indicating a high sensitivity. This shows that Dutch listeners do not assimilate both trills into one perceptual category.

For the stimuli with voiced stops, the speaker produced tokens with and without pre-voicing. Stops with six or twelve cycles of pre-voicing were generated by cutting cycles out of the middle of tokens with longer pre-voicing. The stops in the resulting stimuli had a pre-voicing duration of 0, 38 and 64 ms.

2.2.1. Procedure

Participants were seated in a sound-shielded booth in front of a computer screen. First, they read a list of existing Dutch words with initial /r/s among fillers to determine their preferred realizations. During the shadowing task, participants heard the stimuli over headphones, and their responses were recorded digitally with the stimuli. The instructions stressed speed and a visual warning signal appeared if no response was recorded after 600 ms.

Every participant completed 864 trials spread over two sessions on different days, with 12 presentations of

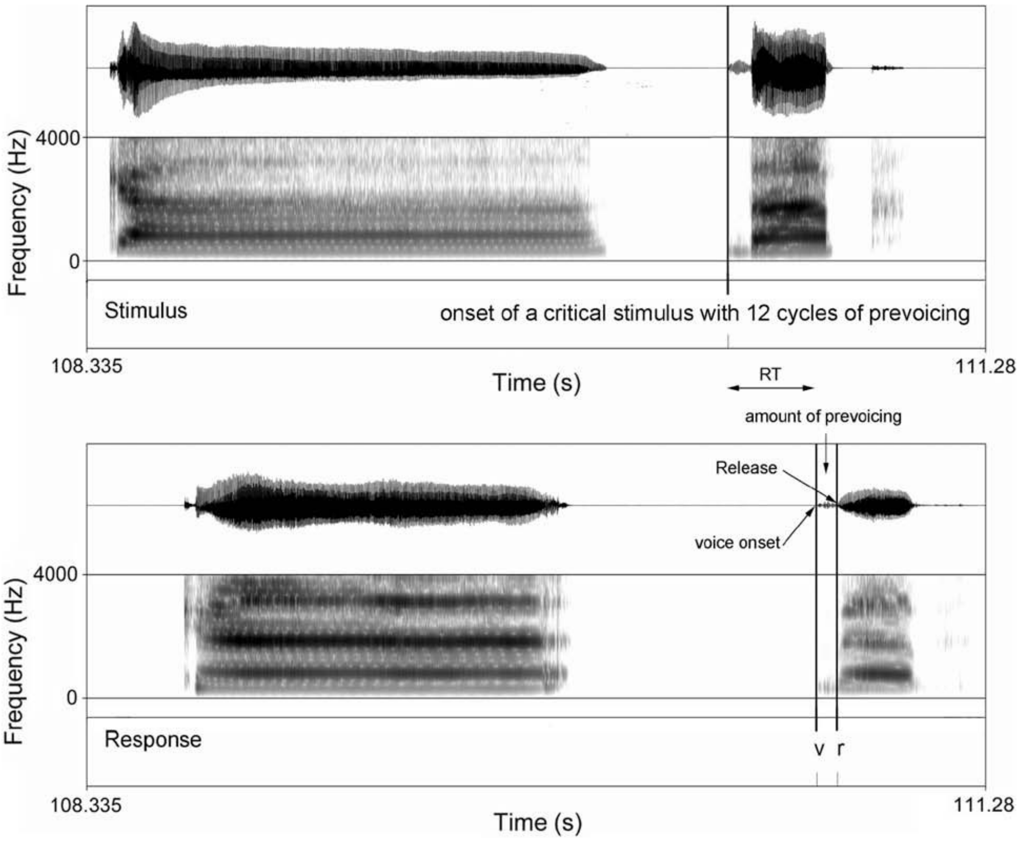

Fig. 1. An example of a coded experimental trial. The upper panel shows the stimulus, with a stop-vowel precursor syllable and an experimental stimulus starting with a voiced stop with 12 cycles of pre-voicing. The lower panel shows the time-locked response and the coded onset of voicing (v) and release (r). The reaction time was 291 ms with 68 ms pre-voicing (VOT = −68 ms).

the 10 /r/-initial pseudowords with both /r/s (240 trials), 8 presentations of the 18 initial-stop pseudowords with three degrees of prevoicing (432 trials), and fillers. Pre-cursor syllables were randomly assigned to these trials. Every participant was presented with a different randomized order of the stimuli. Within each session, participants had the opportunity to take a break after every block of 50 trials.

2.3. Results

The recordings of one block were lost for three subjects. For the remaining recordings, the stimulus and response onsets were hand-marked, and place of articulation of the /r/ was coded for /r/-trials and VOT for stop-trials by research assistants with no knowledge of the actual stimulus. The reliability of the /r/-coding was tested by presenting 108 randomly selected responses to another phonetically-trained research assistant (six items from each speaker, each presented twice in a random order). The agreement with the original coding was nearly as high (88%) as the internal consistency (89%) of this second coder, which is similar to consistencies for other phonetic features (Ernestus, Lahey, Verhees, & Baayen, 2006).

For all statistical analyses, we used linear mixed-effect models with subject and item as random factors, Stimulus and Response type and their interaction as the predictors of interest, and Session, Trial number, and first syllable duration as numerical co-variates (Pinheiro & Bates, 2000). Overall, latencies decreased over trials, but less so in the second session.

2.3.1. Initial /r/

Of the 4287 responses to /r/-stimuli, 10.4% were errors, since they contained other segments than the stimulus. On correct trials, shadowers mostly used their habitual /r/ (97.2%, see Table 1). Only two participants deviated from their preferred (alveolar) pronunciation of /r/ on more than 8 trials (<5%) and together account for 86% of the imitative responses.

Two statistical analyses were performed on the latencies of correct trials. First, we analyzed for the trials in which participants used their habitual /r/ whether their responses were faster if the stimulus matched this habitual /r/. Fig. 2A displays the relevant means. There was a tendency for faster responses to alveolar stimuli ($t(3708) = 2.4$, $d = 0.2$, $p < 0.05$), which is unsurprising given that the (first) closure for the /r/ occurred earlier for alveolar than for uvular trills. Fig. 2A suggests that uvular responses were faster than alveolar responses, but this difference was not significant ($t(3708) = -0.67$). If a match between stimulus and response is beneficial, there should be – independent of any main effects – an interaction of Stimulus by Response Type, because alveolar responses match with alveolar but mismatch with uvular stimuli. This interaction was, however, not significant ($t(3703) = -0.74$).

Table 1
Frequencies and mean latencies of /r/-responses broken down by stimulus and response type, as well as habitual /r/ type

Response type	Uvular				Alveolar			
	Preferred /r/				Preferred /r/			
	Uvular		Alveolar		Uvular		Alveolar	
	N	RT (SD)	N	RT (SD)	N	RT (SD)	N	RT (SD)
Stimulus								
Error	149	–	89	–	121	–	86	–
Alv. /r/	0	–	900	414 (129)	1	358	981	414 (133)
Uvul. /r/	917	399 (121)	91	543 (88)	939	381 (114)	13	510 (100)

Note: SD = Standard deviation.

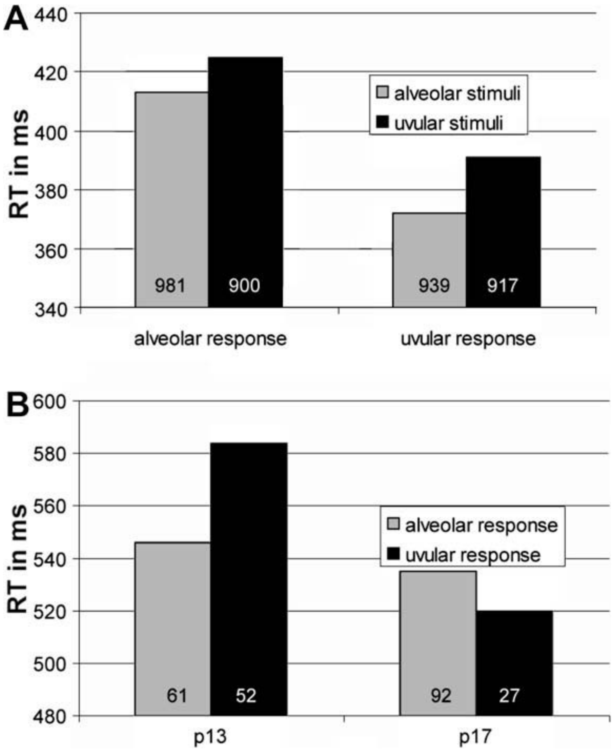

Fig. 2. Estimated marginal means for the r-stimuli. Panel A shows the latencies for the responses with participants' habitual /r/ (so that response type is a between-subject variable). Panel B shows the latencies to uvular stimuli for the two participants with a habitual alveolar /r/ who sometimes imitated the uvular /r/. The numbers in the bars indicate the numbers of observations on which the means are based.

As it is possible that congruency effects show up only for fast responses (Heijden, Hagenaar, & Bloem, 1984), we investigated the stability of our results over the reaction time range by creating five RT bins. We ordered the RTs for every participant in every condition and the first RT bin was then filled with the 20% fastest responses in every condition. The second bin contained the RT from the percentiles 20–40, and so on. Fig. 3A shows no congruency effect for any bin – with RTs around 300 ms in the fastest bin, and, statistically, no interaction of Stimulus by Response Type by Bin ($t(3704) = -0.62$).

A second analysis was performed on the data of the two participants with an habitual alveolar /r/ who regularly imitated uvular /r/ (Fig. 2B). We tested whether reactions to uvular stimuli were faster if the responses also contained uvular trills. For one participant, who imitated the uvular /r/ on 46% of the trials, imitation led to a significant latency *cost* ($t(111) = -3.5$, $d = 0.6$, $p < 0.001$). For the other participant – with 22.5% imitation – there was no significant difference ($t(117) = 1.1$, $d = 0.25$, $p > 0.1$).

To summarize, the /r/-stimuli induced hardly any imitation. The resulting gestural mismatch between stimulus and response did not lead to longer response latencies. A latency cost was observed for one participant when she *imitated* the (unpreferred) uvular /r/, so that a gestural match between stimulus and response was associated with slower instead of faster responses.

2.3.2. Voiced stops

Answers were counted as correct if the initial stop had the correct place of articulation and all other segments matched, which led to an overall error rate of 8.6%. For statistical analysis, we recoded the three-level pre-voicing variable into two linearly independent contrasts, a "phonological contrast" contrasting none versus any amount of pre-voicing, and a "timing contrast" between 6 and 12 cycles of pre-voicing.

Fig. 4 shows the mean latencies and amounts of pre-voicing of the responses. There are clear differences between the responses to stimuli with and without pre-voicing in both the duration of pre-voicing ($t(7022) = 6.27$, $d = 0.2$, $p < 0.001$) and the response latency ($t(7021) = 7.82$, $d = 0.4$, $p < 0.001$). The responses to stimuli with six and twelve cycles of pre-voicing do not differ (t^2s < 1). The phonologically relevant difference between presence and absence of pre-voicing was imitated, while the phonologically irrelevant amount of pre-voicing was not. There is also a (theoretically unimportant) RT advantage for stimuli without prevoicing, probably because the prevoicing provides no information about place of articulation, but the burst does.

As for the initial-/r/ results, we investigated whether the patterns of results co-vary with response speed. Fig. 3B shows the amount of pre-voicing for 5 RT bins. Statistical analysis revealed no interaction of bin neither with the timing contrast ($t(4697) = -1.47$, $p > 0.1$) nor the phonological contrast ($t(4697) = -1.27$, $p > 0.1$). This shows that neither the imitation of the phonological contrast nor the non-imitation of the timing contrast covary with response latency. Because Fig. 3B nevertheless suggests an effect of the timing contrast for the fastest bin, we also analyzed just these data, which revealed the same pattern as the overall data: a significant effect of the phonological contrast ($t(1368) = 4.41$, $p < 0.001$), but no effect of the timing contrast ($t(901) = 1.16$, $p > 0.2$).

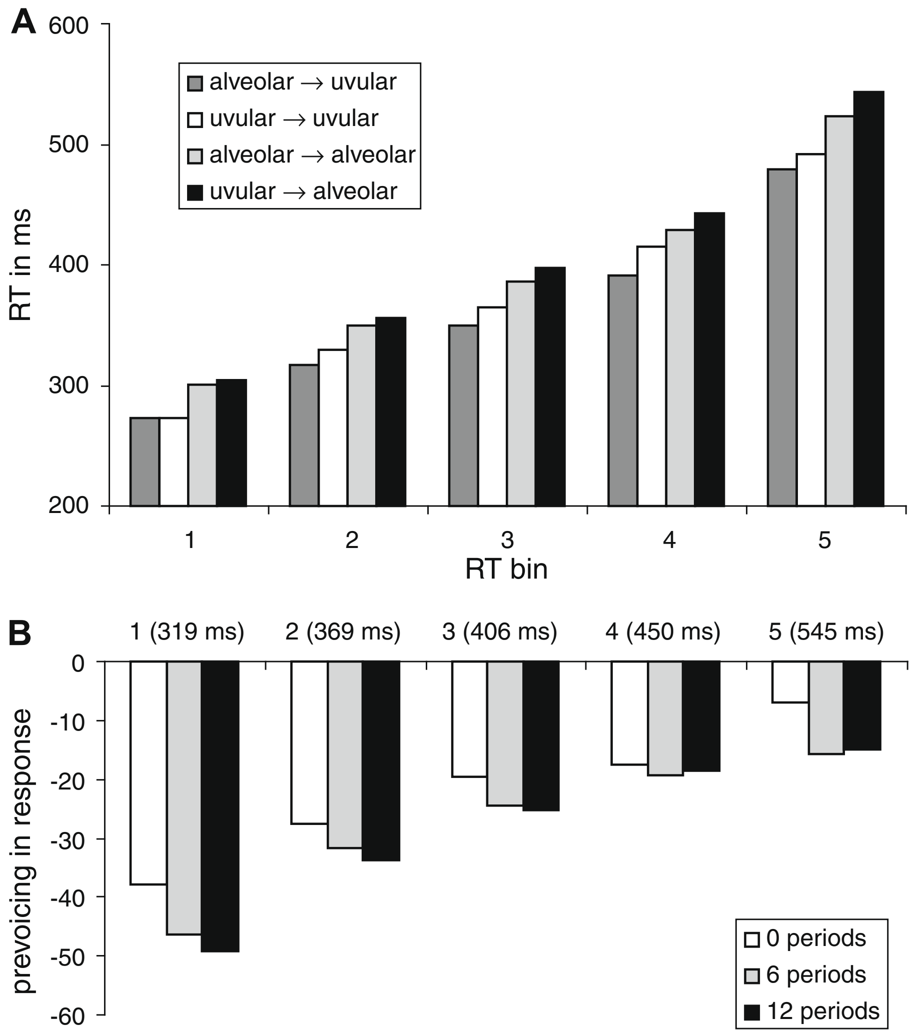

Fig. 3. Estimated marginal means in five RT bins for (panel A) the latency of the /r/-responses for each type of stimulus and response combination and (panel B) the amount of pre-voicing in the response depending on the amount of pre-voicing in the stimulus. The values in brackets show the mean RT for the five bins for the VOT data.

3. Discussion

We studied the relation between speech perception and production in shadowing responses. First, we investigated whether stimulus–response congruencies facilitate shadowing. This was not the case: Response latency is not necessarily affected if the shadower uses a different gesture for /r/ than the one provided by the stimulus. If a gestural percept is fed forward to the motor system, then responding with an uvular /r/ to an alveolar /r/ should lead to similar incongruency effects as observed in the Stroop task (MacLeod, 1991). The lack of such an incongruency effect is, therefore, unexpected for the gestural account.

Might the absence of an incongruency effect be due to the lack of one of the gestures in our shadowers' gestural inventory? Our pretest showed that Dutch listeners are clearly able to distinguish the two types of /r/, and the distinction is also socio-linguistically relevant (Bezooijen, 2005). If speech is perceived in gestures, listeners need to activate the relevant speech gestures to make the distinction. Under the gestural account, both variants are therefore in the listener's gestural inventory.

Another caveat might be that effects of gesture perception occur only with very short latencies. Our latencies were longer than in experiments with a more limited stimulus set (e.g., Fowler et al., 2003), which is expected given that shadowing latencies increase as the stimuli become less predictable (Fowler & Nye, 2003). It should be noted that if the effect is restricted to short RTs this would also mean that other mechanisms than perception of gestures need to be proposed to explain phonetic imitation based on massive long-term exposure without fast repetition (Sancier & Fowler, 1997). This obviously remains an issue for further investigation.

Importantly, we found some suggestive evidence that imitation of the stimulus' gestures may lead to longer latencies than the use of the preferred, and better practiced pronunciation. Imitation may therefore be an effortful process and is therefore more likely to be caused by social factors than by processing mechanisms.

Secondly, we asked whether phonological relevance moderates imitative tendencies. This appeared indeed to be the case: The phonologically relevant difference be-

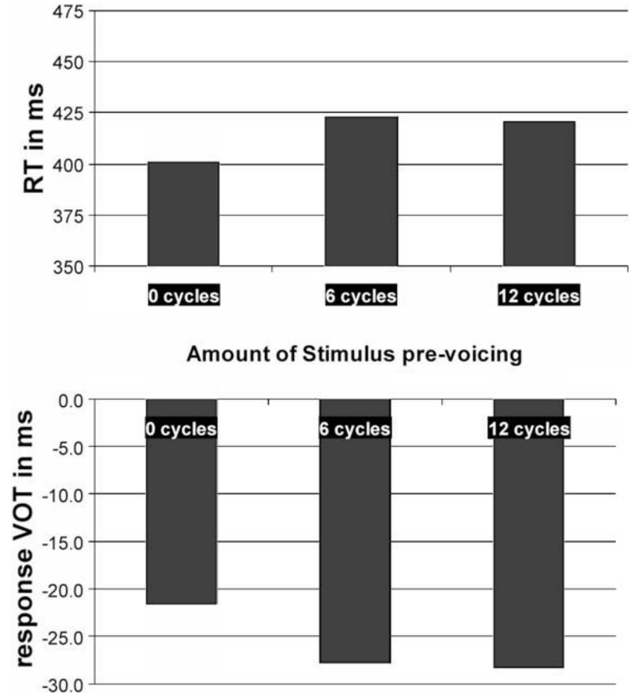

Fig. 4. Estimated marginal means for the latency and amount of pre-voicing of shadowed voiced stops, depending on the pre-voicing in the stimulus.

tween no versus some pre-voicing was imitated, while the exact amount of pre-voicing, which is phonologically irrelevant, was not.

Finally, our data suggest that speakers of Dutch can easily map alveolar and uvular trills, which occur as free allophones in Dutch (and other languages), onto one phonological category. These trills involve two categorically dissimilar gestures with similar acoustic structures, which are nevertheless distinct for Dutch listeners. This phenomenon provides a new answer to Liberman's famous question (1957, p. 121) "when articulation and sound waves go their separate ways, which way does perception go?" Perception appears to treat different gestures with similar acoustics as equivalent (see also Guenther et al., 1999).

In conclusion, our data suggest that speech perception and production are indeed linked, but the link is at an abstract level and governed by phonological relevance. This is in line with the finding by Kraljic, Brennan, and Samuel (2008), showing that listeners can adapt their perception to the idiosyncrasies of another speaker without changing their production. This view can be incorporated in phonological theories that see the function of phonology as bridging the gap between sound-based perception and motor-based production (e.g., Boersma, 1998).

Acknowledgements

The authors wish to acknowledge Marloes van der Goot for her help in running and coding the experiment, and Anne Cutler for comments on a previous version of this manuscript.

References

Bezooijen, R. V. (2005). Approximant r in Dutch. Routes and feelings. *Speech Communication, 47*, 15–31.

Boersma, P. (1998). *Functional phonology. Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Ernestus, M., Lahey, M., Verhees, F., & Baayen, H. R. (2006). Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America, 120*, 1040–1051.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99*, 1730–1741.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49*, 296–314.

Fowler, C. A., & Nye, P. W. (2003). Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Journal of Phonetics, 31*, 63–79.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123–136). Hillsdale, NJ: Lawrence Earlbaum Associates.

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 159–207). Berlin: Mouton de Gruyter.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., & Perkell, J. S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America, 105*, 2854–2865.

Heijden, A. H. C. V. d., Hagenaar, R., & Bloem, W. (1984). Two stages in postcategorial filtering and selection. *Memory and Cognition, 12*, 458–469.

Kraljic, T., Brennan, S., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*, 51–81.

Ladefoged, P., & Maddieson, I. (1996). *Sounds of the world's languages*. Oxford: Blackwell Publishers.

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29*, 117–123.

Liberman, A. M., & Whalen, D. W. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*, 187–196.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*, 522–523.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189*, 226–228.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Porter, R., & Castellanos, F. X. (1980). Speech-production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustic Society of America, 67*, 1349–1356.

Porter, R., & Lubker, J. F. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research, 23*, 593–602.

Sancier, M., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics, 25*, 421–436.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66*, 422–429.

van Alphen, P. M., & McQueen, J. M. (2006). The effect of Voice Onset Time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 187–196.