Shadowing reduced speech and alignment

Susanne Brouwer[a] **and Holger Mitterer**
Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
susanne.brouwer@mpi.nl, holger.mitterer@mpi.nl

Falk Huettig
Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands and Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands
falk.huettig@mpi.nl

Abstract: This study examined whether listeners align to reduced speech. Participants were asked to shadow sentences from a casual speech corpus containing canonical and reduced targets. Participants' productions showed alignment: durations of canonical targets were longer than durations of reduced targets; and participants often imitated the segment types (canonical versus reduced) in both targets. The effect sizes were similar to previous work on alignment. In addition, shadowed productions were overall longer in duration than the original stimuli and this effect was larger for reduced than canonical targets. A possible explanation for this finding is that listeners reconstruct canonical forms from reduced forms.

PACS numbers: 43.70.Mn, 43.71.Sy [AL]
Date Received: March 12, 2010 **Date Accepted:** May 14, 2010

1. Introduction

Although speech production is highly variable, listeners are most of the time able to understand what a speaker intended to say. Recently, more attention has been paid to the nature of the connection between production and perception. Pickering and Garrod (2004), for example, have argued that people align unconsciously and spontaneously to the person to whom they are speaking. Interlocutors tend to converge on a common speaking style in natural conversations (see Giles *et al.*, 1991, for a review). Characteristic of such natural conversations is that words are often reduced (Johnson, 2004). Such reductions may deviate on multiple segments from their citation form (e.g., [pjutər] for the Dutch word 'computer' [kɔmpjutər]). The present study examines whether listeners align their productions when listening to reduced speech.

Two main lines of research have investigated this production-perception link. One type of research mainly uses the shadowing task, in which participants are asked to listen and quickly repeat a speech stimulus. The type of material used in this task is typically careful speech read from a previously prepared script. Porter and Castellanos (1980), for example, used the shadowing task to measure the latency between stimulus and response onsets. In a simple version of this task, participants shadowed an extended /a/ from a model speaker and always had to switch to /ba/. In a choice version of this task, participants again shadowed the long vowel /a/, but had to switch to an unexpected CV. In both tasks, participants shadowed the targets surprisingly quickly. Porter and Castellanos argue that listeners perceive the articulations of a speaker, so that perception delivers—as a byproduct—a blueprint for production.

Fowler *et al.* (2003) also used the simple and choice task to investigate *what* exactly is imitated. Stop consonants were presented with short and long voice onset times (VOTs). The results showed that listeners produced longer VOTs in their shadowing responses to long VOT stimuli. This supports the idea that perceived gestures guide participants' responses and that alignment may occur at the phonetic level. However, Mitterer and Ernestus (2008) argue that a

[a]Author to whom correspondence should be addressed.

phonological approach can account for the findings of Fowler and colleagues. In their study, two variants of the Dutch /r/ were presented: uvular /r/ and alveolar /r/. These phonemes represent different gestures, but are mapped onto a similar phonological representation. The results from a shadowing task showed that participants hardly imitated the two types of /r/-stimuli but responded with their preferred variant. Further, no latency costs were found if there was a gestural mismatch between the stimulus and the response. In the same experiment, stops without or with six or twelve prevoicing cycles were presented. The gestural account predicts that the degree of prevoicing should be imitated, whereas the phonological account predicts that only the phonologically relevant presence of prevoicing should be shadowed while the amount of prevoicing, which is phonologically irrelevant in Dutch, should be ignored. The results supported the prediction of the phonological account.

In the context of the debate on the nature of lexical representations, several studies tested imitation in shadowing isolated words (e.g., Goldinger, 1998). In these studies imitation was assessed using the AXB task. In this task, listeners hear three versions of the same word and are asked to judge whether the production of stimulus A or B by a given participant is more similar to that of the model talker, X. The two stimuli from the participant were a pre-experimental baseline recording and a shadowing response. Goldinger (1998) found that listeners judged the shadowing responses to be more similar to the model talker than the productions in the baseline recording, indicating that listeners imitate the speech they hear. This study therefore provides evidence for a link between perception and production of lexical items.

A second line of research investigates whether alignment between speaker and listener also occurs in more natural communicative situations. Pickering and Garrod (2004), for example, focused on a natural form of language: the dialogue. They argue that "interlocutors align their linguistic representations at many levels ranging from the phonological to the syntactic to the semantic. This interactive alignment process is automatic and only depends on simple priming mechanisms that operate at the different levels, together with an assumption of parity of representation for production and comprehension" (p. 188). Evidence for the interactive alignment model focuses mainly on lexical and syntactic levels. Interlocutors will use the same words and syntactic structures (e.g., Branigan *et al.*, 2000). Pardo (2006) examined alignment on a phonetic level. Different talkers had to produce similar lexical items before, during and after a conversational exchange using a map task. A different set of participants then performed the AXB task, and they judged later realizations as more similar. This indicates that participants perceived increased similarity in pronunciation between talkers over the course of the conversational interaction.

In sum, the second research line shows that convergence not only occurs in laboratory settings, but also in more natural settings. One of the main differences between the two settings is that the speech in laboratory settings is often carefully pronounced, whereas the speech we are exposed to in our daily encounters is full of reductions. Segments or even whole syllables may be deleted and/or changed into different sounds. Listeners are, however, able to understand conversational speech with ease despite these reductions. It is yet unknown whether people imitate exactly what they perceive if speech is reduced.

The present study takes an intermediate position between the two research lines: using the shadowing task to investigate the perception and the subsequent production of *conversational* speech. Participants were asked to repeat back sentences extracted from a spontaneous speech corpus. Each sentence contained one target word. Crucially, half of the target words were produced in their citation forms whereas the other half were reduced forms. If production and perception are strongly linked at the phonetic level, participants should produce exact copies of the reduced forms (e.g., listening to the Dutch pronunciation [pjutər] should produce [pjutər]). If the connection between production and perception is weak, listeners should produce similar renditions of the target words regardless of the input form (e.g., listening to the Dutch pronunciation [pjutər] may produce [kɔmpjutər]). As dependent variables, we use target

word duration and the realization of the target words' segments rather than global measures of similarity as measured with an AXB task.

2. Method

Participants. Sixteen members of the Max Planck Institute's subject pool were paid to participate. All participants were native speakers of Dutch and reported no (history of) hearing or speech impairments.

Materials. Sixty-four sentences were extracted from the spontaneous speech subcorpora of the Spoken Dutch Corpus (Oostdijk, 2000). This corpus contains approximately 900 h of speech of standard Dutch (ca. 9 million words) of which 225 h are spontaneous, face-to-face conversations. All recordings have been aligned with orthographic transcriptions. We searched the corpus for recordings of mid-to high-frequency words in full or in reduced form. Recordings with background noise or overlapping speech were excluded. The test materials were composed of 64 target sentences uttered by 59 different speakers. Each stimulus sentence contained one target word. Half of the target words was produced canonically (e.g., [bənedə] for *beneden* 'downwards') and the other half was produced in a reduced way (e.g., [məneə]). The average duration of both target types are presented in Table 1. Canonical targets ($M=490$ ms; range $=329-773$ ms) were significantly longer than reduced targets ($M=364$ ms; range $=195-588$ ms; $\beta_{Word\ Form}=-125.5$, $p=0.0001$). Note that the context for a canonical target was never identical to that of a reduced target, because they occurred in different natural corpus utterances.

Procedure. Participants were tested individually, seated in a sound-attenuated booth in front of a computer screen. Stimuli were presented over headphones at a comfortable listening level. Participants received written instructions on the screen. They had to perform a shadowing task. They were instructed to listen to Dutch sentences and asked to repeat back the sentence as fast as possible. If they were not able to repeat back the whole fragment, they were requested to report individual words. Participants could listen to each sentence only once. Their responses were recorded digitally. The next trial initiated after 1.5 times the total duration of the fragment. For example, if the duration of the sentence was four seconds, participants had six seconds to repeat this particular sentence. A visual warning signal (a cross) appeared when the next trial initiated. Participants were presented with the 64 experimental items. The order of the items was randomized, so that each participant received a different order of presentation.

Design and analysis. The dependent measures were error rate, duration of the shadowed target responses (for correct responses only) and type of segmental response (canonical versus reduced) to the original stimuli. For all statistical analyses, we used linear mixed effects models (Baayen *et al.*, 2008), with participants and items as random effects. Word form was coded as a numeric contrast (−0.5 and 0.5), in which canonical forms were coded as −0.5 and reduced forms as 0.5. A logistic linking function was used for the error pattern.

Table 1. Segmental responses split by stimulus and response type.

Stimulus type	Canonical target (mean duration: 490 ms)		Reduced target (mean duration: 364 ms)	
Response duration	501 ms		480 ms	
	Target phonemes realized as		Target phonemes realized as	
Response phoneme realized as	Canonical (%)	Reduced (%)	Canonical (%)	Reduced (%)
Canonical	88 (2017)	0	93 (1050)	68 (493)
Reduced	12 (280)	0	7 (78)	32 (230)

Note: Frequencies between parentheses.

3. Results

Error rate. Errors consisted of target misidentifications (e.g., shadowing *presentatie* 'presentation' as a response to the stimulus *prestatie* 'performance') or no target response at all. Six participants were omitted from the final analysis, because they made more than 25% of errors in the reduced form condition. The 10 remaining participants made on average 2.8% errors (0.9/32) in the canonical form condition and 22% errors in the reduced form condition (7.1/32). The statistical analysis revealed that this differences was significant ($\beta_{Word\ Form}=2.66, p<0.0001$). The positive beta indicates that participants made more errors in shadowing reduced targets than in shadowing canonical targets.

Duration alignment. Table 1 presents the average duration of the shadowed target responses. All erroneous responses were excluded from the analysis. The duration of participants' shadowed responses to the canonical forms ($M=501$ ms; range=344–673 ms) were significantly longer than to the reduced targets ($M=480$ ms; range=294–731 ms; $\beta_{Word\ Form}=24.7$, $p=0.0001$).

A comparison between the average duration of the canonical targets and the corresponding shadowed responses showed a significant difference ($\beta_{Stim/Resp}=7.7, p<0.01$), indicating that the shadowed responses were longer than the presented canonical stimuli. A similar statistical difference was found for the average duration of the reduced targets and their shadowed responses ($\beta_{Stim/Resp}=78, p=0.0001$). Importantly, this effect was much larger for the reduced targets than for the canonical targets, and a combined analysis showed a significant interaction effect ($\beta_{Stim/Resp*Word\ Form}=71, p=0.0001$).

Alignment to segment realizations. As a next step, we examined specific participant responses to the canonical and the reduced stimuli. The first author transcribed the target words by observing each target word in auditory and visual spectrographic form using the software package PRAAT (Boersma, 2001). We examined whether a canonical or a reduced segment in the original stimuli remained a canonical or reduced segment in participants' responses. For example, the reduced form [məneə] consists of two reduced segments (the [m] and the [d]) and four canonical segments (the [ə], the [n], the [e], and the [ə]), whereas the canonical form [bənedə] only consists of canonical segments. We calculated how often listeners produced these segments in their original form or in another form (i.e., canonical or reduced, see Table 1).

The results show that participants produced in 88% of the cases a canonical realization and in 12% of the cases a reduced realization when listening to canonical targets. The canonical segments in the reduced targets also often remained in tact (93% of the cases). Importantly, however, participants produced 68% of the time a canonical segment when reduced segments of reduced targets were presented. We used a mixed effect logistic regression model to test whether a reduced response was more likely if the stimulus was reduced as well. This was the case ($\beta_{Word\ Form}=-1.92, p<0.0001$).[1]

4. Discussion

We examined whether listeners align their productions when listening to reduced speech. In a shadowing task participants had to repeat sentences from a casual speech corpus containing canonical and reduced forms. The error pattern showed that canonical forms are easier to recognize than reduced forms. This is convergent with previous offline findings (e.g., Kemps *et al.*, 2004).

Our results provide further evidence for alignment between speaker and listener. The duration data showed that participants' responses to canonical targets were longer than to reduced targets, indicating that listeners accommodate to the duration of the original form of the target. The size of the effect was similar to the results reported by Fowler *et al.* (2003). In Fowler *et al.* the VOTs in the stimuli were extended with approximately 78% (from 73 to 130 ms). Participants' responses to the extended VOTs were significantly longer than to the original VOTs, but the difference in the stimuli of a factor of about 1.78 was reduced to a difference in the responses of a factor of 1.10 (in Experiment 4A: from 61 to 69 ms; and in Experiment 4B: from 53 to 57 ms). Similarly, in our study the canonical and reduced targets differed by a factor

of 1.35, but responses only differed by a factor of 1.04. The shadowed responses were approximately 4% longer for the canonical forms than for the reduced forms. Thus, both studies showed that the amount of alignment between the original extension and the shadowed extension was around 10%.

Another type of evidence consistent with the previous work on alignment comes from the analysis on segment realization. Branigan *et al.* (2000) showed that the syntactic structure of the confederate strongly influenced the syntactic structure of the participants, especially when participants had to use the same verb. In these cases, participants produced 55% more syntactically equivalent responses than different responses. However, when participants were asked to use a different verb than the confederate, participants produced 26% more syntactically similar responses than dissimilar responses. In a similar way, our results showed that participants produce 25% more canonical segments in response to canonical than reduced segments, when listening to reduced targets. This demonstrates that the degree of alignment in our study is of a similar size as previous work on alignment of syntax.

However, our findings also show that the shadowed target responses were overall significantly longer than the duration of the original target stimuli. Critically, this effect was much bigger for the reduced targets than for the canonical targets, indicating that participants' productions show a bias toward the canonical forms. Apparently, people imitate canonical forms more closely than reduced forms. A possible explanation for the misalignment in reduced speech is that listeners reconstruct canonical forms from their reduced forms (e.g., Kemps *et al.*, 2004). As a result, much longer responses are produced.

Two earlier studies also found evidence for "online" repair by testing how mispronunciations were shadowed (Marslen-Wilson and Welsh, 1978; Small and Bond, 1986). Misarticulated three-syllable words and words with deleted segments respectively were reconstructed on the fly by participants in a shadowing task. Despite the clear difference between the spontaneous reductions in our study and the artificially created mispronunciations and deletions in those earlier studies, the results converge on the assumption that listeners actively "reconstruct" the input in a shadowing task.

Another indication for reconstruction is that participants' responses often do not mirror the exact reductions that occurred in the original stimuli. The majority of the reduced segments in the stimuli became canonical segments in the responses. Similarly, Gaskell (2003) showed that assimilation (e.g., producing 'leam bacon' for 'lean bacon') is undone prelexically in perception on the basis of fine phonetic detail in the signal and phonological context.

What remains an open question is *to what* people align when they listen to casual utterances from a spontaneous speech corpus. There are two possibilities. First, speech may be perceived along gestural lines (Fowler *et al.*, 2003). Participants' responses are guided by their perception of the speakers' articulatory gestures. A second possibility is that participants do not imitate gestures but rather the speech style. In this case, alignment does not target the exact phonetic properties of the input, but rather more global properties such as speaking rate, pitch range and the amount of hypo-and hyperarticulation. Both explanations do not require conscious effort due to automatic alignment.

In conclusion, our results indicate that the extent of alignment to phonological reductions is similar to the effects found in previous work on phonetic alignment (Fowler *et al.*, 2003) and on syntactic alignment (Branigan *et al.*, 2000). Importantly, however, our findings also suggest that the link between perception and production is weaker for reduced speech, because listeners seem to reconstruct canonical forms from their reduced forms. Our study indicates that varying the amount of phonological reductions in the input is a promising avenue to further explore the relation between perception and production.

References and links

[1]No differences in the extent of alignment were found between the first half and the second half of the experiment, indicating that participants did not align more to the speech stimuli over the course of the experiment.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**, 390–412.
Boersma, P. (**2001**). "PRAAT, a system for doing phonetic by computer," Glot International **5**, 341–345.
Branigan, H. P., Pickering, M. J., and Cleland, A. A. (**2000**). "Syntactic co-ordination in dialogue," Cognition **75**, B13–B25.
Fowler, C. A., Brown, J., Sabadini, L., and Weihing, J. (**2003**). "Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks," J. Mem. Lang. **49**, 396–413.
Gaskell, M. G. (**2003**). "Modelling regressive and progressive effects of assimilation in speech perception," J. Phonetics **31**, 447–463.
Giles, H., Coupland, J., and Coupland, N. (**1991**). "Accommodation theory: Communication, context and consequences," in Contexts of Accommodation: Developments in Applied Sociolinguistics, edited by H. Giles, J. Coupland, and N. Coupland (Cambridge University Press, Cambridge), pp. 1–68.
Goldinger, S. D. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psychol. Rev. **105**, 251–279.
Johnson, K. (**2004**). "Massive reduction in conversational American English," in *Casual Speech: Data and Analysis*, Proceedings of the First Session of the 10th International Symposium, edited by K. Yoneyama and K. Maekawa (The National International Institute for Japanese Language, Tokyo, Japan), pp. 29–54.
Kemps, R., Ernestus, M., Schreuder, R., and Baayen, H. (**2004**). "Processing reduced word forms: The suffix restoration effect," Brain Lang **90**, 117–127.
Marslen-Wilson, W. D., and Welsh, A. (**1978**). "Processing interactions and lexical access during spoken word recognition in continuous speech," Cogn. Psychol. **10**, 29–63.
Mitterer, H., and Ernestus, M. (**2008**). "The link between speech perception and production is phonological and abstract: Evidence from the shadowing task," Cognition **109**, 168–173.
Oostdijk, N. (**2000**). "The spoken dutch corpus project," The ELRA Newsletter **5**, 4–8.
Pardo, J. S. (**2006**). "On phonetic convergence during conversational interaction," J. Acoust. Soc. Am. **119**, 2382–2393.
Pickering, M. J., and Garrod, S. (**2004**). "Toward a mechanistic psychology of dialogue," Behav. Brain Sci. **27**, 169–225.
Porter, R., and Castellanos, F. (**1980**). "Speech production measures of speech perception: Rapid shadowing of VCV syllables," J. Acoust. Soc. Am. **67**, 1349–1356.
Small, L. H., and Bond, Z. S. (**1986**). "Distortions and deletions: Word-initial consonant specificity in fluent speech," Percept. Psychophys. **40**, 20–26.