Atten Percept Psychophys (2013) 75:557–575
DOI 10.3758/s13414-012-0407-8

Regional accent variation in the shadowing task: Evidence for a loose perception–action coupling in speech

Holger Mitterer · Jochen Müsseler

Published online: 24 January 2013

Abstract We investigated the relation between action and perception in speech processing, using the shadowing task, in which participants repeat words they hear. In support of a tight perception–action link, previous work has shown that phonetic details in the stimulus influence the shadowing response. On the other hand, latencies do not seem to suffer if stimulus and response differ in their articulatory properties. The present investigation tested how perception influences production when participants are confronted with regional variation. Results showed that participants often imitate a regional variation if it occurs in the stimulus set but tend to stick to their variant if the stimuli are consistent. Participants were forced or induced to correct by the experimental instructions. Articulatory stimulus–response differences do not lead to latency costs. These data indicate that speech perception does not necessarily recruit the production system.

Keywords Speech perception · Speech production · Psycholinguistics

What we hear influences how we speak. For instance, children almost always take over the accent of their peer group and are remarkably flexible in this respect. Baron-Cohen and Staunton (1994) reported a case in which a 4.5-year-old boy, who moved from London to Dublin, changed from a Hackney to an Irish accent of English within just 2 months. But even in adulthood, accents are flexible. Most readers are probably familiar with the fact that moving within one country often leads to accent changes in the direction of the new regional accent. Such slow changes are also documented on more than an anecdotal basis. For instance, an ingenious study by Harrington, Palethorpe, and Watson (2000) measured vowel acoustics in the Christmas address by the Queen of England from the sixties and eighties of the 20th century and compared them with the vowels of BBC broadcasters from the eighties. The latter group was taken as an indication of the state of "modern" (i.e., eighties) Received Pronunciation. As it turns out, the Queen's English—the actual and not the metaphorical one—had changed in the direction of the "modern" Received Pronunciation. Such changes in articulation patterns based on ambient input raises the question of how speech perception and speech production are coupled. In this study, we investigated this question by presenting participants with regional pronunciation variants in German, asking them to repeat them as quickly as possible. With this paradigm, we tested, first, to what extent participants imitate the pronunciation variants and, second, whether the failure to imitate has consequences for response latencies in the task.

Previous research has also investigated the relation between speech perception and production at a more microscopic level in experiments by using a shadowing task, in which participants are asked to repeat spoken utterances as quickly as possible. Mirroring the macroscopic language changes in the examples of the Queen's English, results from this task have indicated phonetic imitation on a microscopic level. Goldinger (1998) established the occurrence of imitation in the shadowing task. An initial group of participants produced words in a preexperimental recording session and later produced the same words as shadowing responses. A second group of participants then performed a similarity judgment task to test whether there was phonetic imitation. This second group performed an AXB task with the preexperimental recordings, the shadowing responses, and the stimuli used to elicit the shadowing responses. The

H. Mitterer (✉)
Max Planck Institute for Psycholinguistics,
P.O. Box 310, 6500 AH, Nijmegen, The Netherlands
e-mail: Holger.Mitterer@mpi.nl

J. Müsseler
RWTH Aachen University, Aachen, Germany

stimuli denoted as A and B were utterances of a given word by a given participant from both the preexperimental recording session and the shadowing task. Critically, the X stimulus was the stimulus used to elicit the shadowing response, and participants were asked to indicate whether the A or B stimulus sounded more similar to the X stimulus. Goldinger found that participants chose the shadowing response significantly more often than the token from the preexperimental recording session, showing that the shadowing response imitates, to some extent, the stimulus. Using the AXB task, Pardo and co-workers (Pardo, 2006; Pardo, Jay, & Krauss, 2010) showed that this form of phonetic imitation is not restricted to the somewhat unnatural shadowing task, in that the same pattern is also observed if 2 participants have a conversation. Tokens of the same word from two interlocutors sound more similar after the conversation when compared with preconversation recordings.

One disadvantage of the AXB method is, however, that it is unclear which aspect of the stimulus is imitated. Other studies therefore have varied specific phonetic properties of the stimulus and then tested, via acoustic or articulatory measurement, whether the stimulus differences affect the phonetic properties of the responses. With this approach, Fowler, Brown, Sabadini, and Welhing (2003) found that a longer voice onset time (VOT) in the stimulus leads to a longer VOT in the response for English voiceless stops. Imitation is, however, not ubiquitous. A shortening of VOT from the canonical (voiceless) value in English does not influence the responses (Nielsen, 2011). A similar selective pattern has been found in Dutch for voiced stops. Mitterer and Ernestus (2008) found that the presence versus absence of prevoicing was imitated, while the amount of prevoicing was inconsequential. Both findings indicate that imitation is selective.

On a theoretic level, findings of phonetic imitation have fueled the debate about whether the objects of speech perception are auditory or gestural (Diehl, Lotto, & Holt, 2004; Fowler, 1996; Ohala, 1996). Prima facie, the occurrence of imitation indicates that listeners are quite attuned to the speech gestures they hear. Indeed, gestural theories of speech perception assume that listeners recover (or directly perceive) the speech gestures they hear, which makes it easy to account for phonetic imitation (see, e.g., Sancier & Fowler, 1997).

Additional support for this assumption stems from several different sources. First, humans seem to have an innate bias to imitate (Meltzoff & Moore, 1977), and it is often assumed that language acquisition is based on this bias. Moreover, theoretical approaches in psychology and neuroscience argue that perception and action are intimately linked (Hommel, Müsseler, Aschersleben, & Prinz, 2001; Rizzolatti & Craighero, 2004). As empirical evidence, these approaches refer to studies in which the observation of another person's actions activates the same motor neural circuitry as is activated when the action itself is performed (for overviews, see Iacoboni & Dapretto, 2006). For instance, in a study by Iacoboni and co-workers (Iacoboni et al., 1999), participants were asked to observe and/or imitate a finger movement. Results showed evidence for a neural mechanism that directly matched the observed action onto an internal motor representation of that action.

Following this empirical and theoretical perspective, Galantucci, Fowler, and Goldstein (2009) showed stimulus–response compatibility effects. Participants were asked to produce a syllable in response to an arbitrary character string (e.g., ##) and heard spoken syllables as distractors. Matching distractors led to faster responses, while mismatching responses led to slower responses. An earlier experiment had already shown analogous findings with visual speech (Kerzel & Bekkering, 2000). A final supporting line of evidence stems from findings in neuroscience that the motor cortex seems to be involved in speech perception (S. M. Wilson, Saygin, Sereno, & Iacoboni, 2004). In a similar vein, Watkins, Strafella, and Paus (2003) found increased excitability of the muscles involved in speech production caused by listening or seeing speech. That is, listening to speech seems to activate the motor cortex and even the motor system. It has to be noted, however, that the effects of auditory speech are not consistent over studies. Sundara, Namasivayam, and Chen (2001) found only an effect of visual speech on motor excitability. Given the bias toward reporting (and creating) positive results (Simmons, Nelson, & Simonsohn, 2011), especially in behavioral sciences (Fanelli, 2010), the motor excitability caused by perceiving speech may be consistent phenomenon only with visual speech.

On the basis of such findings, it has been argued that the speech production system is recruited for speech perception. Such an involvement would be beneficial, since it would supply a common currency for perception and production. The need for such a common currency is evident both within and between speakers: Learning to speak one's native language obviously involves learning relations between phonetic categories in perception and production. One has to learn that the vowel with the largest F2–F1 difference and the vowel produced with a closed jaw and a front tongue position refer to the same category, which happens to be /i/. A gestural representation even in perception is able to provide such a common currency. Theories differ, however, in how this common currency is achieved. Motor theory (Liberman & Whalen, 2000) suggests an analysis-by-synthesis approach, in which the incoming signal is compared with the output of a synthesizer producing candidate gestures and choosing those gestures that account best for the incoming signal. An alternative approach is derived from the theory of direct perception (Gibson, 1979). In this

approach, no intermediate steps are necessary, but the perceptual system directly perceives the sound-producing gestures of the vocal tract. As was argued in Goldstein and Fowler (2003) and Fowler et al. (2003), both approaches provide an explanation of how language users can achieve parity, both within speaker and between speakers.

However, the view that speech perception relies on the perception of speech gestures is not generally accepted (Lotto, Hickok, & Holt, 2009), and there is also counterevidence to the arguments listed above. With regard to the imitative tendencies, Gergely, Bekkering, and Kiraly (2002) showed that imitation is not necessarily based on imitation of motor commands but, rather, is based on the imitation of action outcomes. Similarly, the role of the motor cortex for speech perception has also been questioned on the basis, for instance, of the observation that the response of the motor system often does not differ between speech and other complex acoustic signals (Scott, McGettigan, & Eisner, 2009).

The focus of the present article is on stimulus–response compatibility effects. Mitterer and Ernestus (2008) argued that most studies have confounded gestural and phonological compatibility of stimulus and response. They argued that learned stimulus–response associations, rather than a direct link between perception and action, can explain the findings of stimulus–response compatibility in speech production tasks. Consider, for example, the well-known Stroop effect. Most European participants would not find it difficult to name the ink color of the Mandarin character for the word red, thus demonstrating that stimulus–response compatibility effects can simply be explained by learned associations (Elsner & Hommel, 2001). In this context, it is important to note that written language is acquired much later than and only through formal schooling. Even though this represents a much less direct route to articulation than do early-acquired associations between spoken and heard words, these late-acquired associations are still powerful enough to produce interference effects. Hence, the associations necessary for language learning between acoustic and articulatory properties of speech sounds should also be able to generate compatibility effects. This provides an account for findings of stimulus–response compatibility without invoking a notion of a recruitment of the action system for production. As was already noted above, learning to speak one's native language obviously involves learning relations between phonetic categories in perception and production. Instead of this parity being supplied directly in perception, it is conceivable that the language user has to learn the relation between perception and production units. Boersma (1998), for instance, proposed completely different perception and production grammars with completely different vocabularies. In this view, a common currency is provided by phonological representations, which bridge the gap between these two domains. Fowler et al. (2003) also acknowledged the possibility of such an account: "The obvious common currency would be the covert phonetic categories that serve as the end point of phonetic perception and might serve as the starting point of phonetic production planning" (p. 397). In a similar vein, Plaut and Kello (1999) proposed a model of language learning in which perception and production are only indirectly linked via phonological representations. These learned phonological associations are then sufficient to explain the findings of stimulus–response compatibility effects.

To decide between these two accounts, it is necessary to find stimuli that differ in their speech gestures but are phonologically equivalent. With such stimuli, the predictions of a learning account and a gestural account differ. The gestural account predicts that stimulus–response compatibility effects should still be observed, while the learning account predicts that no compatibility effects will be observed in such cases.

The obvious problem is that gestural and phonological compatibility are often confounded. The gestural difference between a front vowel with rounded and spread lips translates into the phonological difference between the vowel categories /i/ and /y/. Mitterer and Ernestus (2008), however, managed to find a set of stimuli in Dutch that are gesturally different (in a categorical manner) but phonologically compatible. They exploited the variety of phonetic implementations for the phoneme /r/ in Dutch to address this question. In Dutch, the phoneme /r/ has many possible implementations (Van Bezooijen, 2005), including the alveolar trill [r], which is used in standard Spanish, and the uvular trill [ʀ], which is used in standard French. The two different trills involve quite different gestures but are phonologically equivalent in Dutch; the phonetic forms [ʀos] and [ros] both mean "rose" in Dutch. The alveolar trill is generated with the tongue tip close to the alveolar ridge. The trilled effect then arises as the passing air creates a Bernoulli effect so that the tongue tip periodically (at ±20 Hz) touches the alveolar ridge. The gesture for the uvular trill is radically different; here, the tongue body is moved to the back of the mouth, and the Bernoulli effect sets the uvula in motion to generate a trill. Despite their phonological equivalence, Dutch listeners are able to hear the difference between these variants (Van Bezooijen, 2005).

This leads to different predictions between a learning account and a gestural account for the relation between speech perception and production. If the production system is recruited in perception, hearing an alveolar trill should activate a tongue tip gesture approaching the alveolar ridge. This should make it difficult to produce an uvular gesture, with the tongue body being retracted. Accordingly, the gestural account predicts that the mismatch between input and output gestures should lead to longer shadowing

latencies, just as it is difficult to produce an alveolar stop when hearing a labial stop (Galantucci et al., 2009). A learning account, however, predicts that the phonological categories activated by an alveolar trill and an uvular trill are the same, so that no compatibility effects should be observed. The results were in line with the prediction from a learning account; shadowing latencies were just as fast when the stimulus and response gestures matched (alveolar–alveolar, uvular–uvular) as when they mismatched (alveolar–uvular, uvular–alveolar). This, in fact, supports the learning account for these compatibility effects. The phonological equivalence of [r] and [ʀ] in Dutch has to be learned, because there are languages in which [r] and [ʀ] are separate phonemes (e.g., Moghol, a Mongolian language spoken in Afghanistan). Dutch speakers must therefore have learned to treat different trills as equivalent due to exposure to speakers with different variants referring to the same referent. Hence, there is no incompatibility between stimulus and response, because both can be mapped onto the same phonological category.

Nevertheless, the point can be made that the efforts by Mitterer and Ernestus (2008) concern the special case of rhotic sounds. Ladefoged and Maddieson (1996), for instance, indicated that they could find no other reason for these sounds to be grouped in the same class other than that orthographies seem to choose the letter /r/ for these sounds. Moreover, trills are notoriously difficult to master in second-language acquisition, and most Dutch speakers are able to produce only one kind of trill. Mitterer and Ernestus thus focused on the latency effects for the ensuing mismatches in stimulus and response gestures, since it was expected that participants would not imitate the version of the /r/. One potential criticism of this work, then, is that Dutch speakers have acquired a special exception to deal with the fact that they have to interact with speakers that use speech gestures they are themselves not able to produce. The present study investigated this criticism by focusing on the effects of variants of speech gestures that every speaker of the language has mastered. The present work thus addressed two questions. First, how strongly is regional variation imitated in a shadowing task when it is easy for participants to do so? Second, is there a latency cost when participants use a different gesture in their response than the gesture used in the stimulus?

Experiment 1

This experiment made use of two cases of variation that occur in German and employed speech gestures that all German speakers have to master in order to be proficient speakers of German. The first variation is how fricative-stop clusters are produced. German allows only /st/ and /sp/ fricative-stop clusters in onset position—except for some loans, such as *scannen*—and the phonetic implementation of the fricative varies regionally. Standard German uses the postalveolar [ʃ] (as in English *she*) for onset clusters (e.g., *Stein* [ʃtɛɪn], Engl. ‘stone’), while some Northern German accents use the alveolar fricative [s] (as in English *sea*). Nevertheless, speakers of Standard German are able to produce [st], because they have to in coda position (e.g., *fast* [fast] and not *[faʃt],[1] Engl. ‘nearly’).

A second variation in German that stays within the basic phoneme inventory of all speakers is the phonetic implementation of the frequent orthographic word ending *-ig* (as in *König*, Engl. ‘king’), which can be produced as either [ɪk] or [ɪç]. Note that this is a phonetic implementation difference, since all speakers produce the plural *Könige* as [kønɪgə]. This shows that the underlying form contains a voiced velar stop, and the pronunciation variation is how this voiced velar stop [g] is implemented in the coda position. Again, the difference is regional, with the [ɪk]-variant more likely in the southern parts of Germany. Additionally, all speakers of German need to be able to produce both [ɪk] and [ɪç]. Words that end their orthographic form on *-ik* have to be produced with a final [ɪk] (e.g., *Plastik*, Engl. ‘plastic’, [plastɪk] and not *[plastɪç]), while words that end in *-ich* have to be produced with [ɪç] (e.g., *Kranich*, Engl. ‘crane’, [kra:nɪç] and not *[kra:nɪk]). We therefore tested to what extent such variation is imitated in a shadowing paradigm. We also used words ending on *-ik* and *-ich* to establish that the participants are indeed able to produce [ɪç] and [ɪk].

It should be noted, however, that the two forms of variation differ clearly in their markedness. The fricative-stop clusters immediately indicate that the speaker is from a (far) northern area. It is also undisputed that the [s]-variant is not the standard variant. Variants in *-ig* pronunciation are, first of all, more evenly distributed, and often speakers do not consciously know which one they are using. In fact, German speakers are often not sure what to consider the standard variant. This is reflected in the results of a Google search (March 11, 2011) for “Aussprache von *–ig*” (Engl. ‘pronunciation of *–ig*’), which produced among the first ten hits three for Internet fora that controversially discuss which variant is the “correct” one.

The main questions in this experiment were, first, how likely would participants be to imitate or correct the presented variants, and, second, whether imitation versus correction would have repercussions for the response latency. By “correction,” we mean that the stimulus [ɛsɪç] is “corrected” in the response of the participant to [ɛsɪk] (variants

[1] We follow the linguistic notation and use the “*” symbol to indicate forms that do not conform to the language norms. Note, however, that the [ʃt] variant in coda position is used in Southwestern German accents (e.g., *fast*, [faʃt], Engl. ‘nearly’).

of the word *Essig*, Engl. 'vinegar'). If hearing the stimulus variant [ɛsɪç] activates the speech gestures for this variant, corrections should be associated with slower responses and imitations with faster responses, because of the gestural stimulus–response congruency. Note that this means that the critical analysis has to be restricted to participants who produce both corrections and imitations.

Method

Participants

Twelve native speakers of German (9 female) from the student population of the RWTH University Aachen participated in the experiment for pay. Their mean age was 24.0 years (*SD* = 3.0).

Materials

Using the Celex lexical database (Baayen, Piepenbrock, & Gulikers, 1995), we selected 20 *-ig*-final words, 10 words each with an *st-* or *sp*-onset and 10 words each ending on /ɪk/ and /ɪç/. The selected sets had similar mean lexical frequencies (using the logarithm of the frequency per million plus one; *-ig*-final words, 2.66; /sC/ words, 2.55; *-ik*-final words, 2.65; *-ich*-final words, 2.67). For each *-ig*-final word and each /sC/ word, an analogous nonword was created with the same syllable structure (see the Appendix for the complete list of items). These words and nonwords were then recorded by a female native speaker of German in all variants. That is, the word *Essig* (Engl. 'vinegar') was recorded in the [ɛsɪç] and [ɛsɪk] variants, and the word *Spinne* (Engl. 'spider') was recorded in the standard [ʃpɪnə] and the **N**orthern German [spɪnə] variants. To verify that the pronunciation variants were correctly produced, we measured the spectral center of gravity for the fricative-onset stimuli, which showed a clear separation of /s/ onsets (<7.5 kHz) and /ʃ/ onsets (<4 kHz). For the *-ig* stimuli, we measured the maximal positive acceleration of the intensity curve after the offset of voicing. This differentiated the fricative versions, with a more or less constant fricative noise (mean maximal acceleration: 143 db/s), from the stop with a closure and a burst (mean maximal acceleration: 750 db/s).

This resulted in a stimulus set of 180 sound files based on 100 word forms: 20 *-ig*-final words and nonwords in two variants, 20 /sC/-initial words and nonwords in two variants, and one token for each of the 10 *-ik-* and *-ich*-final control words.

Apparatus and procedure

Participants were seated in a sound-proof booth with headphones and a microphone. They were instructed that they should repeat the words they heard over the headphones as quickly as possible. The instruction simply focused on response speed and mentioned neither pronunciation variation nor what to do with such variation. Stimulus presentation was controlled by an Apple Macintosh computer using the MATLAB-based Psychophysics Toolbox-3 (Kleiner, Brainard, & Pelli, 2007). The experiment script initiated a recording, which was time-locked to the onset of the sound file. Using an audio mixer, the input to the sound card was such that the recording contained the stimulus on one channel and the response on the other.

Each participant repeated each of the 100 word forms once in each of four blocks. Variants were blocked, so that a participant heard only [ɪç] variants of *-ig* words and [ʃC] variants for the fricative-stop clusters in one block. The order of the variants was counterbalanced over participants.

Data coding and analysis

The resulting 12 * 400 sound files were analyzed semiautomatically using the PRAAT software (Boersma, 2001). Response latencies were estimated using the silence estimation method in PRAAT, which estimates sounding and silent parts of a sound file. The automatically estimated response latencies were checked by visual inspection, and the response variant was coded as [ɪç] or [ɪk] for *-ik*, *-ich*, and *-ig* words and as [ʃ] and [s] for fricative-stop onset. If the response contained another variant, the response was coded as error. Additionally, responses on *-ik-* and *-ich*-final words —for which no variation is allowed in German—were counted as correct only if response contained [ɪk] or [ɪç], respectively. That is, a response such as [tɛpɪk] to the stimulus *Teppich* /tɛpɪç/ (Engl. "carpet") was counted as an error.

The response times (RTs) were coded from stimulus onset to response onset for trials on which the fricative-stop clusters were critical. For trials on which the word-final consonant was critical ([ç] vs. [k]), RTs were measured from onset of the final consonant in the stimulus to the onset of the final consonant in the response. In order to exclude a labeling bias, the onset of the final consonant was estimated automatically using the pitch estimation function in PRAAT. Both consonants are voiceless, so the offset of the pitch contour was taken as the onset of the final consonant in both the stimulus and response.

The analyses below make use of linear mixed-effect models, which allows us to simultaneously account for participant and item variability within the same linear model (Baayen, 2008). In all analyses, participant and item were used as random factors. Random slopes were included for all fixed factors that varied over participants and/or items. For analysis of categorical outcomes, such as "Is the response correct or not?" or "Does the response variant match

the stimulus variant?" we used a logistic linking function (as suggested by Dixon, 2008). For the analysis of response latencies, a linear link between predictors and dependent variable was used. Next to the experimental factors, trial number was used as a predictor. Trial numbers were mean centered (ranging from −0.5 to 0.5); thus, the regression weights indicate the amount of change over the experiment. Data analysis started with a full model with all interactions. Insignificant interactions were pruned. Only the final models are discussed.

Results

Control trials

For the *-ik* and *-ich* control words, the error rates were low (0.8 % and 1.3 %, respectively) and did not differ significantly ($p > .2$).[2] There were also no effects of trial number, either in an interaction with type of word or as a main effect ($p_{max} > .2$). Mean RTs were also similar in both conditions (*−ik*, 667 ms; *-ich*, 684 ms), and an analogue linear mixed effect model proved to be not significant ($p > .1$). Responses got significantly faster over the course of the experiment, but this effect was not significant, $b_{\text{Trial Number}} = -62$, $t = -1.69$.

-ig words

Figure 1 summarizes the relevant aspects of the data for trials with *-ig*-final words. Fig. 1a shows the proportion of correct responses. Note that the variation of the final consonant was not coded as an error; that is, a response such as [ɛsɪç] for the stimulus [ɛsɪk] (or vice versa) was counted as correct. To reiterate, both versions are attested in German, while no variation is allowed for the control words. Participants responded within these limits on nearly all trials with words (98.7 %) but made some errors on nonword trials (91.6 % correct). As Fig. 1a indicates, the effect of lexical status was stable over the course of the experiment, although there was a stronger improvement for [ɪç]-final nonwords. The initial model contained the fixed factors variant, lexical status, and (scaled) trial number, with all their interactions. Model selection was used to remove nonsignificant interactions, and the final model contained only one interaction of variant and trial number, $b_{\text{Variant = [ɪç]} \times \text{Trial}} = 2.9$, $p < .01$. Note that there was no three-way interaction, so that the apparent specific improvement for [ɪç]-final nonwords leads to a statistically significant overall improvement only for all [ɪç]-final stimuli. The model additionally contained a main effect of lexical status, $b_{\text{Word}} = 1.76$, $p < .01$, with fewer errors on words than on nonwords.

The critical question in this experiment is which variant is used on correct trials and how imitation versus correction is related to response latencies. Thus, we first assessed the overall rate of imitation. Figure 1b shows that participants had a strong tendency to imitate, overall (86.3 %). That is, on most trials, participants used the same variant as the stimulus. The final model contained no interactions but three significant main effects. There was less imitation of the [ɪç] variant, $b_{\text{Variant = [ɪç]}} = -3.4$, $p < .01$, less imitation of words, $b_{\text{Word}} = -1.55$, $p < .05$, and more imitation over the course of the experiment, $b_{\text{Trial Number}} = 0.68$, $p < .05$.

Given the overall higher likelihood of imitation, it is not straightforward to test the relation of imitation or "correction" and the ensuing gestural match or mismatch on response latencies for all conditions and subjects. On the one hand, some participants hardly ever produced a response that mismatched with the stimulus, and on the other hand, there were no corrections of [−ɪk] stimuli to [−ɪç] responses. Therefore, we focused first on the [ɪç] stimuli and second on the 5 participants who corrected more than 20 % of the 80 [ɪç] stimuli they heard over the course of the experiment. Note that this selection is necessary because the data from participants who always imitated or always corrected do not allow us to assess whether, within a participant, stimulus–response mismatch influences the latencies. Because it has been suggested that imitation is especially likely with fast responses (Honorof, Weihing, & Fowler, 2011), we tested whether the participants who responded quickly were more likely to imitate. However, we found that this was not the case, since there was no correlation between the likelihood to imitate and the average response latency over participants, $r(10) = .19$, $p = .54$ (imitation likelihood was transformed into logOdds for this analysis; cf. Dixon, 2008, for the necessity of the transformation).

To further investigate under which circumstances imitation or correction occur, we ran linear mixed effect models with lexical status, experiment half, and response latency as predictors for those participants who sometimes produced a different variant than the stimulus.[3] Since the dependent variable was binary (imitation: yes/no), a logistic linking function was used. For this analysis, response latencies were normalized by subtracting the average RT of the participant. Three outliers with a deviation larger than 400 ms from the individual averages were deleted from the data set. Figure 1c shows

[2] The *p*-values are estimated on the assumption of 20 degrees of freedom. This results in conservative estimates. SPSS, for instance, used degrees of freedom in mixed-effect models that are close to the number of trials minus the number of parameters, which would lead to more than 100 degrees of freedom in all cases. However, in R, the method for estimating *p*-values (pvals.fnc) does not function for models with a maximal random effect structure, which is advisable here (Quene & van den Bergh, 2008).

[3] Experiment half was chosen rather than trial number because it is difficult to visualize the results for two continuous variables; an analysis with trial number gives essentially the same results.

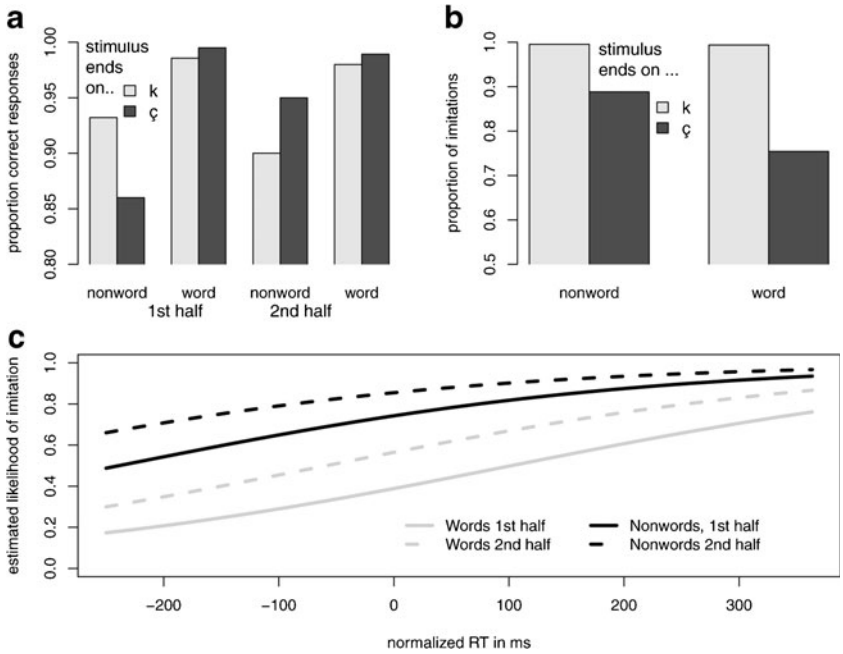

Fig. 1 Results for *-ig*-final stimuli in Experiment 1. **a** Accuracy with which participants shadowed German *-ig*-final words depending on lexical status, part of the experiment, and stimulus variant. **b** Proportion of trials on which stimulus and response matched. **c** Determinants of imitation versus correction. Imitation was more likely with longer response latencies, more likely in the second half of the experiment, and more likely with nonwords

the results of this analysis. While there were no interactions, there were three significant main effects. Imitation was more likely in the second half of the experiment, $b_{2ndHalf} = 0.77$, $p < .01$, less likely with words, $b_{Word} = -1.88$, $p = .08$, and more likely for slower responses, $b_{Normalized\ RT} = 0.0035$, $p < .05$ (note that the range of this predictor variable is much larger, which explains the numerically smaller regression weight).

As an additional test of whether imitations are indeed associated with longer latencies in this data set, we ran a linear mixed effect model using response latency as the dependent variable. As predictors, we used response match versus mismatch as a fixed factor and subject and item as a random factor. This analysis confirmed that imitation—with a gestural match between stimulus and response—was significantly slower than responses in which the participants used a different gesture than the stimulus (match, 696 ms; mismatch, 648 ms), $t = -2.60$.

Fricative-stop clusters

Figure 2a shows that participants produced more correct responses for words than for nonwords. This was also the only significant effect in the linear mixed effect model, $b_{Word} = -1.14$, $p < .005$. Figure 2b shows the likelihood of imitation, indicating the unsurprising finding that participants never corrected the standard variant. Moreover, frequent correction of [s] to [ʃ] was observed only by 2 participants. The other 10 participants imitated the model in more than 97 % of the cases. Moreover, the 2 participants who produced corrections behaved quite differently from each other. One participant corrected words and nonwords alike on about 70 % of the trials. A simple logistic regression showed that neither response latency nor trial number nor lexical status was related to the likelihood of imitation. The other participant corrected nearly all [s] to [ʃ] for words on 92 % of the trials but imitated all nonwords. Due to this complete separation in the data set, it is not possible to fit a logistic regression model. There was thus too much imitation in this condition to estimate the effects of gestural mismatch on response latencies.

Finally, Fig. 2c shows the estimated RT functions over the course of the experiment for [s] and [ʃ] stimuli. A linear mixed effect model showed significant interactions of trial number with variant and lexical status with variant. The regression weights for main effects are hence valid only for the levels mapped on the intercept ([s]-initial nonword). Participants were faster to react to [ʃ]-initial nonwords than to [s]-initial nonwords, $b_{Variant=[ʃ]} = -50$, $t = -2.7$, faster with words than with nonwords, $b_{Word} = -80$, $t = -3.5$, and faster over the course of the experiment, $b_{Trial\ Number} = -176$, $t = -4.6$. Additionally, the effect of trial number was smaller for [ʃ]-variants, $b_{Variant=[ʃ]*Trial} = 94$, $t = 3.23$; note that this means that the participants were 176 ms faster at the end than at the start of the experiment for the [s]-variants, while the latency decrease for [ʃ]-variants is obtained by adding the two regression weights (−176 + 94), leading to a −82-ms effect.

Discussion

The questions addressed in this experiment were (1) how likely participant are to imitate or correct the presented variants and (2) whether imitation versus correction has repercussion for the response latency. The answer to the

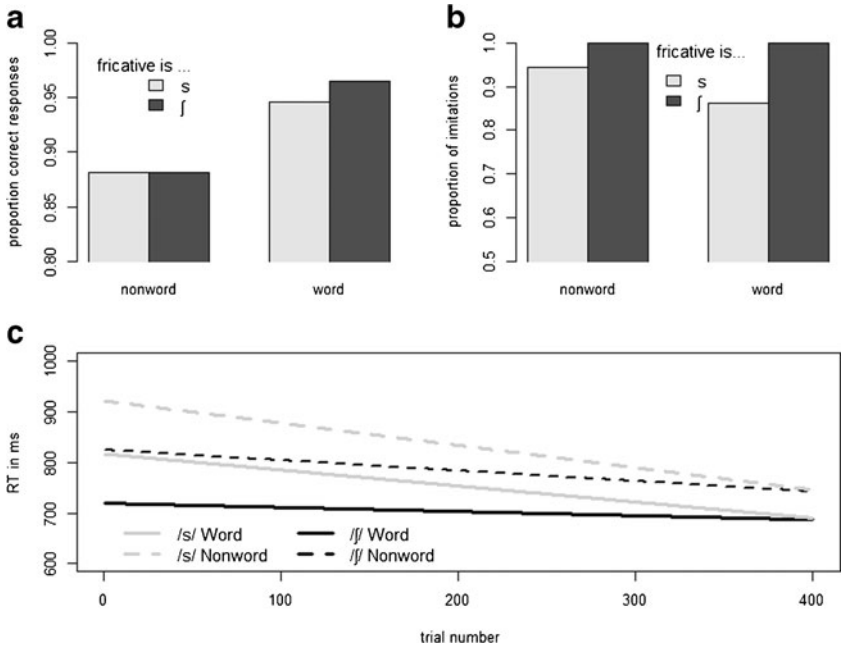

Fig. 2 Results for fricative-stop cluster stimuli in Experiment 1. **a** Accuracy data. **b** Likelihood of imitation. Note that only 2 of the 12 participants produced a sizable number of corrections. **c** Response latencies depending on lexical status and trial number, showing a stable effect of lexical status and a dissipating effect of stimulus variant. The infrequent [st] variants led to longer response latencies at the beginning of the experiment, but this effect disappeared over the course of the experiment

first question is clear: When able, participants are quite likely to shadow regional variation in a shadowing task. This would be in line with the assumption that there is an intimate link between perception and action in speech.

Given the overall high likelihood to imitate, however, it is difficult to answer the second question convincingly, given that there are very few relevant data points. Nevertheless, the evidence from the *-ig* trials indicates that corrections are not associated with longer response latencies, despite the ensuing gestural mismatch between stimulus and response. In fact, corrections seem to be faster than imitations. It is important to note, however, that there is a confound here: Since corrections were observed only with [ɪç] versions, all corrections are [ɪk] responses, and all imitations are [ɪç] responses. Given that our participant group was more likely to use [ik], it is likely that this may simply be a production effect as a result of learning, where [ik] may be easier and faster to produce than [ɪç].

It is somewhat surprising that the more marked variation in the fricative-stop cluster did, in fact, lead to more imitation, with 10 of the 12 participants nearly always imitating the stimulus. A priori, it seems more likely that the more generally accepted variation for *-ig*-final words should be more easily imitated than the more marked [st] pronunciation of fricative-stop clusters. This seems to indicate that the more salient the variation, the more likely it is to be imitated.

Finally, it is worth noting not only that participants got faster over the course of the experiment (a typical finding), but that this effect was larger for the unfamiliar [s]-versions of the clusters. This in line with recent studies showing that listeners can adapt to a given speaker's idiosyncrasies or local accent (McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011; Mitterer & McQueen, 2009; Norris, McQueen, & Cutler, 2003). The fact that the difference between the two variants disappears at the end of the experiment shows that the stimuli with the regional variant were not inherently less intelligible; otherwise, an RT difference should have persisted throughout the experiment. They were simply unusual for our participants, and it is well established that this influences the efficiency of word recognition (Connine, 2004). However, participants seemed to be able to adapt to this over the course of the experiment.

To sum up, the data on the amount of imitation speak for an intimate link between perception and production, while the latency data for mismatches between perception and action indicate a more loose connection, although the latency data rest on a somewhat sparse database. Therefore, we ran Experiment 2 with several changes to get a larger database of shadowing responses in which the gestures of stimulus and response mismatch.

Experiment 2

The results of Experiment 1 show that the salient variation of fricative-stop clusters is nearly always imitated. It hence seems unlikely that participants would spontaneously correct those in a shadowing task. Therefore, we decided to instruct participants to "correct" the [s] pronunciation to the Standard German pronunciation [ʃ]. With this instruction, we are certain to obtain responses in which the response with the standard pronunciation mismatches the stimulus with the regional variant. The gestural mismatch should lead to longer RTs to the regional variant. Note, however, that the regional variant was responded to more slowly already in Experiment 1 (without a stimulus–response mismatch). The

critical question is, therefore, whether the difference between the standard stimulus and the regional variant is larger and more resistant to training than in Experiment 1. Note that, at the end of the experiment, the regional variant is responded to just as quickly as the standard variant. Stimulus–response incompatibility effects tend to be more resistant against training (cf. MacLeod, 1991), so that a gestural account predicts that the longer latencies to the regional variant should persist throughout the experiment. A learning account, in contrast, predicts that the participants will learn that the speakers treats [st] and [ʃt] as phonologically equivalent, so that the difference between the conditions should dissipate over the course of the experiment.

As in Experiment 1, we also used the *-ig* words in this experiment. In Experiment 1, spontaneous correction was already observed for the *-ig* words. For these, we took two measures to increase the likelihood of correction. First of all, we decided to use only words, because Experiment 1 showed fewer corrections with nonwords. Second, our aim was to make the variation less salient by varying it only between participants. That is, a given participant heard only [ɪç] or [ɪk] variants of *-ig* words.

Method

Participants

Sixteen native speakers of German (15 female) from the student population of the RWTH University Aachen participated in the experiment for pay. Their mean age was 21.7 years (*SD* = 3.7).

Materials

This experiment used a subset of the materials used in Experiment 1: The 20 words with initial fricative-stop clusters were used in both versions, the 20 *-ig*-final words in both versions, and the 20 control words ending on [ɪç] or [ɪk] with no variation. This resulted in a stimulus set of 100 sound files.

Apparatus and procedure

The apparatus was the same as in the previous experiment. Each participant heard 80 stimuli over the course of the experiment: the 20 fricative-stop cluster words in both versions, the 20 *-ig*-final words in either the [ɪç] or [ɪk] variant, and the 20 control words ending on [ɪç] or [ɪk] with no variation. A given participant heard the stimuli four times in four blocks of 80 stimuli. The stimulus order was permutated randomly in each block. Half of the participants heard the *-ig*-final words in the [ɪç] variant, and the other half heard the -ig words in the [ɪk] variant.

Each participant was instructed to repeat the word heard as quickly as possible. The instruction mentioned that the speaker came from a Northern German background and would sometimes produce [ʃt] clusters as [st] and that they should nevertheless use the standard version of this cluster.

Data coding and analysis

The resulting 16 * 320 sound files were analyzed semiautomatically as in Experiment 1. Again, the analyses make use of linear mixed effect models, as in Experiment 1.

Results

Control trials

For the *-ik* and *-ich* control words, the error rates were somewhat higher than in Experiment 1 (1.6 % for *-ik* words and 6.1 %, for *-ich* words). The difference between the two types of words was significant, $b_{\text{-ich Word}} = -2.78$, $p < .001$. However, no other effects were significant.

An analysis of the RTs showed an effect of trial number, $b_{\text{Trial Number}} = -95$, $p < .05$, that also interacted with type of word, $b_{\text{-ich Word} \times \text{Trial Number}} = -52$, $p_{\text{MCMC}} < .05$. This indicates that the latencies decreased over the course of the experiment by 95 ms for *-ik* words and by 147 ms (−97 + −52) for *-ich* words. Overall, latencies were also shorter than in Experiment 1, with an overall mean of 529 ms.

-ig words

Participants shadowed the *-ig* words with a high accuracy (>99 %), and statistical testing revealed no effects of experimental variables on accuracy rates. Response latencies were also not different between variants, $b_{\text{-ich Variant}} = 35$, $p > .2$, but decreased over the course of the experiment for both variants/groups, $b_{\text{Trial Number}} = -83$, $p < .01$ (note that variant is a between-subjects manipulation here).

Analysis of the amount of imitation started with an overview of how often individual participants imitated. Figure 3a shows that all but 1 participant in the [ɪç] condition produced a sizable amount of correction, while the majority of the participants in the [ɪk] group produced the same response variant as that used in the stimuli. This was confirmed by a linear mixed effect model with imitation as a dependent variable and group, trial number, and normalized RT as covariates. Normalization here means, as in Experiment 1, that response latencies were corrected by the individual mean. After pruning of all interactions due to their lack of statistical significance, the model with only main effects shows a main effect of group, $b_{\text{[ɪç] Group}} = -6.87$, $p < .001$. There was no significant relation between response latency and the likelihood of imitation, $b_{\text{normalized RT}} = 0.001$, $p > .2$.

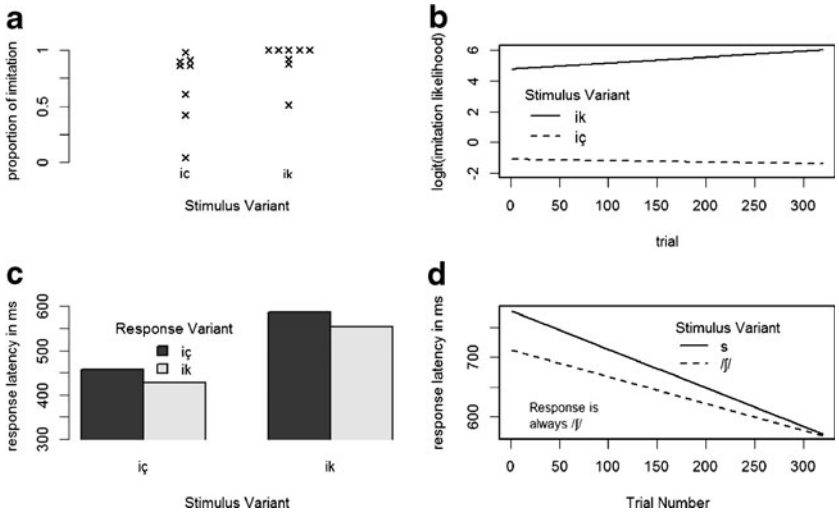

Fig. 3 Results for *-ig*-final and fricative-stop cluster stimuli in Experiment 2. **a** How often individual participants produced the same variant as the stimulus for *-ig*-final words. Note that participants heard only one variant over the course of the experiment. **b** Likelihood of imitation depending on variant and trial number. **c** Absence of an effect of imitation versus correction on response latencies. **d** Response latency data for the fricative-stop trials on which participants were asked to correct the nonstandard [st] variant. The data show an initial latency cost for correcting the infrequent [st] variant, which disappeared at the end of the experiment

To further investigate the possible relation between imitation and response latency, we focus on those participants who produced enough data points with both imitations and corrections. For this analysis, data from the participants with more than 90 % and less than 10 % imitation were disregarded. This data set includes 3 participants from the [ɪç] group and 2 from the [ɪk] group. Note that this selection is necessary because the data from participants who always imitated or always corrected do not allow us to assess whether, within a participant, stimulus–response mismatch influences the latencies. To estimate whether this would select especially slow or fast participants, we first correlated the likelihood of imitation with mean RT. This analysis showed no relation between response speed and imitation likelihood, either for the whole group, $r(14) = .05$, $p > .2$, or for either of the experimental groups separately [[ɪç] group, $r(6) = .05$, $p > .2$; [ɪk] group, $r(6) = .09$, $p > .2$]. This shows that the overall likelihood to imitate is not associated with a participant's overall speed of responding.

In a next step, we analyzed whether a stimulus–response match in this data set leads to shorter RTs. If there is an effect of stimulus–response compatibility, this should lead to an interaction of stimulus and response variant, because an [ɪç] response should be faster than an [ɪk] response following an [ɪç] stimulus, while the opposite should be observed for [ɪk] stimuli. Figure 3c shows the relevant data aggregated over participants. In line with what the figure suggests, the critical interaction between stimulus and response variant was not significant, $b_{\text{Stimulus = [ɪç]} \times \text{Response = [ɪç]}} = 28$, $p > .2$. After pruning of insignificant interactions, the final model contained only a significant effect of trial number, $b_{\text{Trial Number}} = -116$, $p < .001$, and an effect of response variant with slower response for [ɪç] responses, $b_{\text{Response = [ɪç]}} = 34$, $p < .01$.

Fricative-stop clusters

For the fricative-stop clusters, participants were instructed to correct the regional variant [st] to Standard German [ʃt]. Participants had no problems following the instruction, and there was little difference in error rates between stimuli with the standard variant (2.9 %) and stimuli with the [st] variant (3.2 %). Statistical testing showed that neither stimulus variant nor trial number influenced accuracy rates (ps > .2).

For the RT analysis, response latencies with outlier values below 100 ms and above 1,300 ms were disregarded (11 out of 2,480 cases). Analyses revealed significant main effect of both stimulus variant, $b_{\text{Stimulus = [ʃt]}} = -32$, $p < .001$, and trial number, $b_{\text{Trial Number}} = -196$, $p < .001$, as well as a significant interaction, $b_{\text{Stimulus = [ʃt]} \times \text{Trial Number}} = 56$, $p < .01$. Figure 3d shows the predicted latencies arising from these parameters. The figure indicates that the effect of stimulus–response incompatibility for the [st] stimuli with enforced [ʃt] responses dissipated over the course of the experiment. To confirm that the effect really disappeared over the course of the experiment, we ran separate analysis for the four blocks. Table 1 shows that in the first three blocks, there is a significant effect of stimulus variant but that in the last block, the effect is numerically very weak and statistically insignificant ($p > .1$).

Discussion

The purpose of Experiment 2 was to gather more data points in which there was a stimulus–response incompatibility in

Table 1 Effect of stimulus–response compatibility in the fricative-stop cluster trials in Experiment 2

Block	Mean RT [st] Stimuli	Mean RT [ʃT] stimuli	$b_{\text{stimulus = [ʃt]}}$	$b_{\text{Trial Number}}$	$b_{\text{Trial Number × stimulus = [ʃt]}}$
1	770	713	−58**	−145**	–
2	667	637	−37**	−28*	–
3	640	615	−23*	1	–
4	620	607	13	−26	64*

Note. Participants were asked to shadow [st] variants with the standard variant [ʃt]. The predictor trial number was centered around zero and scaled to range from −0.5 to 0.5 for each analysis. The interaction term was not significant for the first three blocks and, hence, was pruned from the model

* $p < .05$

** $p < .01$

terms of the involved speech gestures. In trials with fricative-stop clusters, these were enforced by the instruction. The results showed that the ensuing gestural mismatch led to longer response latencies in the beginning of the experiment, but not at the end of the experiment. It is premature, however, to attribute this effect solely to the stimulus–response incompatibility, because the stimulus–response incompatibility is confounded with stimulus type, given that we asked participants to correct the variant [st] to the standard variant [ʃt]. The difference in RT might simply be due to the unfamiliarity of the participants with infrequent [st] variant.

Fortunately, the data from Experiment 1 elucidate the issue. Figure 4 shows how quickly participants reacted to the different variants in both experiments. While the unfamiliarity of the participants with the [st] variant did not vary over the experiments, responses in Experiment 1 generally imitated the stimulus, while responses in Experiment 2 corrected the [st] variant. The first obvious difference is that participants in Experiment 1 were slower, overall. This is probably due to the fact that the stimuli in Experiment 1 included nonwords, which participants found more difficult to shadow; that is, response latencies were consistently longer for nonwords than for words. It is well known that response latencies are influenced not only by the stimulus on a given trial, but also by the overall difficulty of the trials (Stone & Orden, 1993; Van der Heijden, Hagenaar, & Bloem, 1984). Having accounted for the overall latency difference, the time course is remarkably similar in both experiments. Over four blocks of item repetitions, the difference in response latency between [st] and [ʃt] variants disappears. This result indicates that there seems to be little effect of gestural mismatch.

In Experiment 2, we also aimed to generate more corrections of *-ig*-final words by presenting the variants as between-subjects manipulations, and this manipulation was successful. In Experiment 1, the pronunciation variant of *-ig* words was imitated in 87 % of the cases versus only 60 % in Experiment 2, $t(22) = 2.2$, $p < .05$. Note that "imitation" can be a somewhat misleading term here because, at a conceptual level, a numerical value of 50 % of imitation means no imitation at all. If the participants respond at random, they will still use the same variant as the stimulus on 50 % of the trials. This indicates that the amount of imitation in Experiment 2 (60 %) is only slightly above chance (50 %). Using the ensuing data set of matching and mismatch responses from the same participants, we found again that a gestural stimulus–response mismatch did not lead to longer RTs. Nevertheless, we found that participants were faster to produce the [ɪk] response. This is an important data point for narrowing down the interpretation of Experiment 1, where imitation was associated with longer RTs. However, Experiment 1 allowed testing the effect of imitation only for [ɪç] stimuli, so that corrections always were [ɪk] responses. Experiment 2 indicated that [ɪk] responses are faster to produce, so that the effect in Experiment 1 can be attributed to a response effect. Nevertheless, the overall picture emerging from these data is that stimulus–response mismatches have little effect in a shadowing task. This pattern is observed for enforced

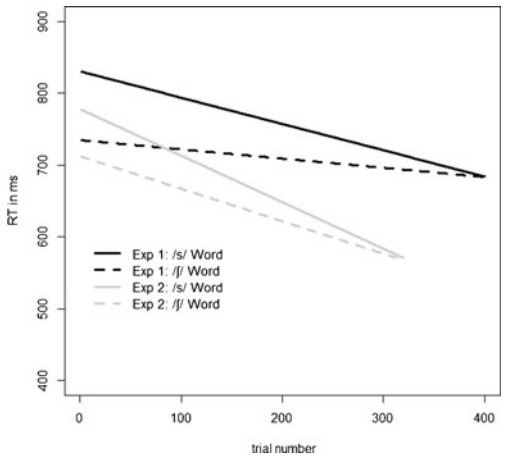

Fig. 4 Response latencies to fricative-stop clusters in Experiments 1 and 2. In Experiment 1, participants responded with the same gestures as the stimuli; in Experiment 2, the [st] variant had to be corrected to the standard variant [ʃt]. Both experiments show, however, a similar pattern. Initially, responses were slower to the infrequent variant, but this effect disappeared over the course of the experiment

mismatches in this experiment and for spontaneously arising mismatches between stimulus and response in Experiment 1.

Experiment 3

Experiment 2 introduced the instruction to "correct" a regional variant to the standard pronunciation. The results showed that the ensuing mismatch between stimulus and response does not lead to an RT cost at the end of the experiment. However, the instruction to "correct" is slightly unusual. In this experiment, our aim was to gather converging evidence with a different method to induce gestural mismatches between stimulus and response. To this end, we changed the instructions and added a block of word reading to the experiment. Participants were instructed that the purpose of the experiment was to compare the efficiency of written and spoken word recognition. This focuses the attention on the lexical properties of the stimuli, and Experiment 1 had shown that words are more likely to induce corrections than are nonwords. Similarly, the focus on word recognition may make participants more likely to say the words in the way they usually would.

To achieve this focus on word recognition, they were first asked to read out loud the same written words they later had to shadow. The instruction mentioned that the spoken words could be produced in different regional variants but did not ask participants to "correct" those to the Standard German form. Rather, the focus was on how fast the words could be recognized.

Method

Participants

Sixteen native speakers of German (12 female) from the student population of the RWTH University Aachen participated in the experiment for pay. Their mean age was 22.4 years (*SD* = 3.8).

Materials

This experiment used the same materials as Experiment 2, plus written forms of those target words. This resulted in a stimulus set of 100 sound files and 100 bitmaps.

Apparatus and procedure

The apparatus was the same as in the previous experiment. Each participant heard and saw 80 stimuli over the course of the experiment. For the written part, there was no variation in the stimulus form. Written words were presented on a 22-in. color CRT monitor (100-Hz refresh rate; 1,024 × 768 pixels) in Helvetica font with a size of 15. The participant's distance to the monitor was about 500 mm.

For the spoken words, the 20 fricative-stop cluster words were presented in both versions, the standard [ʃt] and the regional variant [st], the 20 *-ig*-final words in either the [ɪç] or [ɪk] variant, and the 20 control words ending on [ɪç] or [ɪk] with no variation. A given participant heard the stimuli four times in four blocks of 80 stimuli. The stimulus order was permutated randomly in each block. The experimenter determined the version of the presented *-ig* words on the basis of the responses to the written words. If a given participant read out loud the *-ig* words with the [ɪç] variant, the [ɪk] variant was presented in the auditory blocks. Vice versa, if a given participant read out loud the *-ig* words with the [ɪk] variant, the [ɪç] variant was presented in the auditory blocks.

The instructions were as follows. Participants were informed that the experiment tested the efficiency of word recognition in the written and spoken modality. To that end, they had to read out loud or repeat the word they heard as quickly as possible. Given that the variation of the /st/ onset in German is rather marked, the instruction mentioned that efficiency of word recognition was also tested for different regional variants. All participants first did the reading task and then the shadowing task.

Data coding and analysis

The resulting 16 * 320 sound files were analyzed semiautomatically, as in Experiment 1. Again, the analyses make use of linear mixed effect models as in Experiment 1.

Results

Control trials

For the *-ik* and *-ich* control words, the error rates were quite low (1.2 % for *-ik* words and 0.5 % for *-ich* words). The difference between the two types of words was not significant, $b_{\text{-ich Word}} = 0.84$, $p > .2$, nor was the effect of trial number, $b = 0.007$, $p = .08$.

An analysis of the RTs showed an effect of trial number, $b_{\text{Trial Number}} = -160.5$, $p < .001$, but no effect of type of word, $b_{\text{-ich Word}} = -25$, $p > .1$. Overall, latencies were also shorter than in the earlier experiments, with an overall mean of 480 ms.

-ig words

There were 8 participants each who responded with the [ɪk] and [ɪç] variant, respectively, in the reading task. Note that these participants were presented with the other variant in the shadowing blocks. Participants shadowed the *-ig* words with a high accuracy (>98 %), and statistical testing revealed no effects of experimental variables on accuracy

rates. Response latencies were also marginally different between variants, $b_{\text{-}ich\ \text{Variant}} = -99$, $p = .07$, but decreased over the course of the experiment for both variants/groups, $b_{\text{Trial Number}} = -160$, $p < .001$ (note that variant is a between-subjects manipulation here).

Analysis of the amount of imitation started with an overview of how often individual participants imitated. Figure 5a shows that the majority of the participants produced mostly imitations, but 1 participant in each condition predominately corrected the presented version, and 6 participants (2 in the [ɪç] condition) sometimes produced corrections. A first test of whether imitation is associated with shorter RTs was a correlation of overall proportions of matches with mean RT over participants. This correlation was not significant, $r = .15$, $p > .2$.

A linear mixed effect model with imitation as a dependent variable and group, trial number, and normalized RT as covariates tested whether imitation was associated with fast reactions. As in the corresponding analysis in the two previous experiments, a logistic linking function was used to account for the categorical nature of the dependent variable. Normalization of the RTs was achieved by correcting response latencies with the individual mean. After pruning of all interactions due to their lack of statistical significance, the model with only main effects shows a main effect of normalized RT, $b_{\text{normalized RT}} = 0.004$, $p < .01$. Note that the positive regression weight means that imitations are associated with larger (i.e., longer) RTs.

To further investigate the possible relation between imitation and response latency, we focus on those participants who produced enough data points with both imitations and corrections. For this analysis, data from the participants with more than 98 % and less than 2 % imitation were disregarded. This data set includes 3 participants from the [ɪç] group and 6 from the [ɪk] group. In a next step, we analyzed whether a stimulus–response match in this data set leads to shorter RTs. If there is an effect of stimulus–response compatibility, this should lead to an interaction of stimulus and response variant, because an [ɪç] response should be faster than an [ɪk] response following an [ɪç] stimulus, while the opposite should be observed for [ɪk] stimuli. Figure 5b shows the relevant data aggregated over participants. In line with what the figure suggests, the critical interaction between stimulus and response variant was significant, $b_{\text{Stimulus = [ɪç] × Response = [ɪç]}} = 83$, $p < .01$. Note that the positive regression weight indicates that a match between stimulus and response is associated with slower responses.

Fricative-stop clusters

There were little differences in error rates between stimuli with the standard variant (4.6 %) and stimuli with the [st] variant (5.1 %). Statistical testing showed that neither stimulus variant nor trial number influences accuracy rates (*p*s > .2).

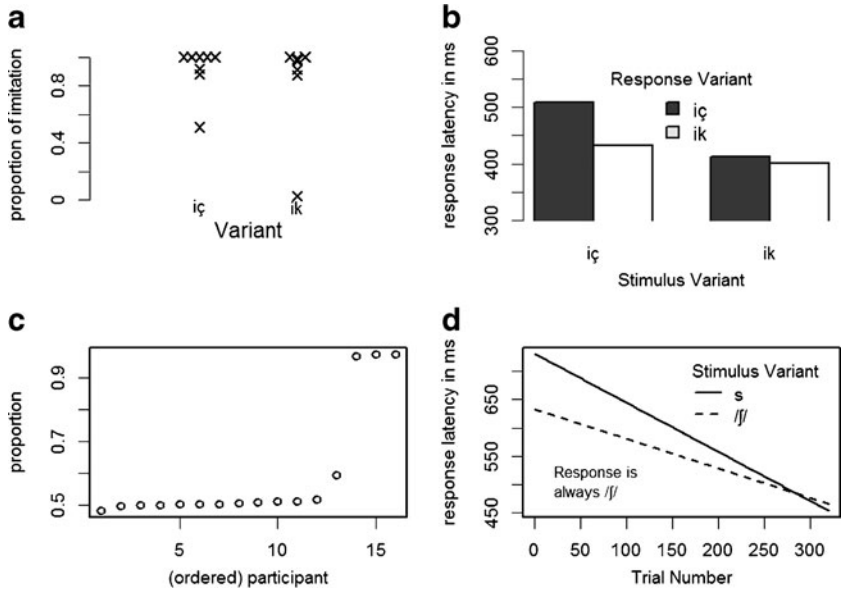

Fig. 5 Results for *-ig*-final and fricative-stop cluster stimuli in Experiment 3. **a** How often individual participants produced the same variant as the stimulus for *-ig*-final words. Note that participants heard only one variant over the course of the experiment. **b** Absence of an effect of imitation versus correction on response latencies. **c** How often individual participants produced the same variant as the stimulus for /st/ words. **d** Response latency data for the fricative-stop trials on which participants corrected the nonstandard [st] variant. The data show an initial latency cost for correcting the infrequent [st] variant, which disappeared at the end of the experiment

Participants had been presented with both the standard variant [ʃt] and the regional variant [st] in this experiment, with no particular instruction to imitate or correct a version. Note, however, that they had already produced the words with the standard variant in the reading task. The critical question was whether the focus on word recognition introduced "spontaneous corrections" rather than the "forced corrections" that were obtained in Experiment 2. Figure 5c shows the relevant data as the (ordered) proportions of matching responses to the nonstandard, regional variant [st]. (Responses to the standard variant were always matching.) This shows that most participants were responding mostly categorically, either producing the variant (3 participants) or mostly correcting it to the standard variant (12 participants). Only 1 participant produced a mixture of responses. As previously, we calculated the correlation of proportion of imitations and response latency, which again was not significant, $r(14) = -.26$, $p < .2$.

Given this data set, the best way to evaluate the impact of a gestural stimulus–response match within participants is to compare the speed of standard responses [ʃt] between standard and variant stimuli for those 13 participants who mostly used standard responses. The question is whether the responses that mismatch the input stimulus remain slower over the course of the experiment than responses that match the input stimulus. For this RT analysis, errors and response latencies with outlier values above 1,300 ms and below 100 ms (23 out of 2,560 cases) were disregarded. Analyses revealed significant main effects of both stimulus variant, $b_{\text{Stimulus = [ʃt]}} = -32$, $p < .001$, and trial number, $b_{\text{Trial Number}} = -196$, $p < .001$, as well as a significant interaction, $b_{\text{Stimulus = [ʃt] × Trial Number}} = 56$, $p < .01$. Figure 5d shows the predicted latencies arising from these parameters. The figure indicates that the effect of stimulus–response incompatibility for the [st] stimuli with enforced [ʃt] responses dissipated over the course of the experiment. To confirm that the effect really disappeared over the course of the experiment, we ran separate analysis for the four blocks. Table 2 shows that in the first two blocks, there is a significant effect of stimulus variant but that, in the last blocks, the effect is numerically very weak and statistically insignificant ($p > .2$).

Discussion

The purpose of this experiment was to test whether the consequences of a stimulus–response mismatch in terms of the articulatory gestures would be similar when the correction was not enforced by an instruction (as in Experiment 2). To this end, we instructed participants that the main focus of the experiment was on how quickly words are recognized when presented in spoken or written form and when there is a variation in the spoken form.

This instruction resulted in an intermediate amount of imitation, as compared with the previous experiments, especially with regard to the fricative-stop clusters. In Experiment 1, there were hardly any corrections; in Experiment 2, many corrections were obtained, because participants were instructed to correct the regional to the standard variant; and in Experiment 3, 13 out of 16 participants mostly corrected the standard to the regional variant. Apparently, the focus on word recognition in the instructions helped to reduce the tendency to produce regional variants, so that we again observed a large amount of corrections.

In these corrections, there is a clear difference in the articulatory gestures used to produce the stimulus and the response. Nevertheless, the patterns in the response latencies were surprisingly similar to those in Experiment 1, where there was little gestural mismatch between stimulus and response. In both cases, the regional variant was responded to more slowly than the standard variant in the beginning of the experiment, but this effect disappeared over the course of the experiment.

For the *-ig* words, the present experiment produced a smaller amount of corrections than in the previous experiment. Apparently, the instruction to correct the fricative-stop

Table 2 Effect of stimulus–response compatibility in the fricative-stop cluster trials in Experiment 3

Block	Mean RT [st] Stimuli	Mean RT [ʃt] stimuli	$b_{\text{Stimulus = [ʃt]}}$	$b_{\text{Trial Number}}$	$b_{\text{Trial Number × Stimulus = [ʃt]}}$
1	722	644	−174**	−3.8**	2.25**
2	597	535	−58**	–	–
3	564	539	−19	–	–
4	522	514	−5.4	–	–

Note. The analyses focused on those participants who responded with the standard variant [ʃt] for stimuli with both the regional variant [st] and the standard variant [ʃt]. The predictor trial number was centered around zero and scaled to range from −0.5 to 0.5 for each analysis. The interaction term was not significant for the last three blocks and, hence, was pruned from the model

** $p < .01$

clusters to the standard variant also increased the likelihood of corrections for the *-ig* words in Experiment 2. Nevertheless, the results suggest that for the corrections observed, there was no latency cost associated with a mismatch in terms of articulatory gestures between stimulus and response.

General discussion

In the present study, we set out to test how close the relationship between perception and action is for speech. As we argued in the introduction, critical evidence can be obtained with stimuli that are gesturally different yet phonologically equivalent. To do so, we tested the consequences of regional accent variation in a shadowing task. Regional accent variation leads to the same words (hence, phonological equivalence) to be produced with different speech gestures. A similar approach had already been presented by Mitterer and Ernestus (2008). Expanding this previous study, the variation in the present investigation stayed within the general gestural inventory of the standard variant. That is, variants were chosen such that any proficient speaker of German should be able to produce both variants. Within this setup, we asked two questions: First, to what extent is the regional accent variation imitated in a shadowing task? Second, what are the consequences for the response latency if stimulus and response use different gestures?

Assuming a tight coupling of perception and action, the prediction was that imitation should be ubiquitous and that the failure to imitate and the ensuing gestural stimulus–response mismatch should lead to a latency cost. While our data provide clear answer to these questions, these answers point in different theoretical directions. First of all, we found a strong tendency to imitate the stimulus variation at least under some conditions. However, if participants did not imitate, either spontaneously or by instruction, there were no latency costs for the ensuing stimulus–response incompatibility.

The likelihood of imitation varied with the type of variation and the experimental setup. In Experiment 1, variants of fricative-stop cluster led to more imitation than did variants in the pronunciation of *-ig*-final words. This variation in the pronunciation of *-ig*-final words was imitated more often in Experiment 1, in which participants heard both variants, than in Experiment 2, in which they heard only one variant. A possible account for these differences is that imitation is linked to the saliency of the variation. The variation in the fricative-stop clusters is clearly more marked than variation in the pronunciation of *-ig*-final words and, hence, also more salient. Moreover, hearing the German word for vinegar produced as [ɛsɪç] and [ɛsɪk] in the same experiment also makes this variation more salient. It therefore seems likely that the shadowing task itself generates a demand characteristic to imitate obvious variation in the input. This assumption also allows us to explain some surprising differences in the amount of imitation in another recent shadowing experiment. Recently, Honorof et al. (2011) measured the amount of imitation in a shadowing task using dark and light /l/ in American English, which is mainly cued by the difference between the first two formants. They found that the amount of imitation was nonlinearly related to the difference in the stimuli. In their Experiment 1, a difference of 200 Hz in the formant distance in the stimuli led to an imitation difference of only 20 Hz; in their Experiment 2, a 260-Hz difference in the stimuli led to a 66-Hz difference in the responses. If one relates the stimulus to the response differences, there is 10 % imitation in Experiment 1 but 25 % in Experiment 2. This difference can easily be attributed to salience: With a clearer difference between the stimuli, the amount of imitation increases. This would also explain the difference in prevoicing imitation found in Mitterer and Ernestus (2008). They found more imitation of presence versus absence of prevoicing than of the amount of prevoicing. Van Alphen and McQueen (2006) showed that the presence versus absence of prevoicing is more salient than differences in the amount of prevoicing. While Mitterer and Ernestus suggested that the difference is due to the phonological relevance of the difference, the present data show that salience may be an alternative interpretation.

It is also noteworthy that the amount of imitation was much stronger in the present Experiment 1 than in Experiment 3, although both did not include an instruction to produce standard variants (as Experiment 2 did). Experiment 3 capitalized on the fact that lexical status seems to modify the tendency to imitate. In Experiment 1, we had already observed that words led to more corrections than did nonwords. Experiment 3 hence focused on the lexical properties of the items, with an instruction that focused on word recognition. Moreover, an initial block of reading responses established how the words should be pronounced, since the participants there responded with the standard variant.

As this discussion indicates, the strong tendency to imitate might simply be the default in a shadowing task, especially when nonwords are presented. In fact, answering why participants imitate the stimulus in a shadowing task may be more of a question for social psychology than for theories of speech processing (Giles, Coupland, & Coupland, 1991). There is, in fact, a long tradition in social psychology of viewing phonetic imitation as a socially rather than linguistically driven process (Gregory & Webster, 1996). The fact that the amount of imitation can be modified from nearly complete alignment for marked variants (fricative-stop clusters in Experiment 1) to near chance level for unobtrusive cases (between-subjects variation of German *-ig*-final words in Experiment 2) indicates that the tendency to imitate the phonetic properties in the

shadowing task is probably not the consequence of an automatic and tight perception–action coupling.

This conclusion is buttressed by our findings on the latency effects of stimulus–response incompatibilities. In short, there are none. For *-ig*-final words, stimulus–response mismatches occurred in both experiments and never led to latency costs. Experiment 2 and Experiment 3 seem to show a latency cost at first sight. Participants found it initially more difficult to respond with the standard variant [ʃt] when hearing the [st] variant than when the stimulus also used the standard variant. However, a comparison with Experiment 1 shows that this is simply an effect of the markedness of the [st] stimulus. The same latency difference that is observed between [ʃt] and [st] stimuli in Experiment 2 is observed in Experiment 1, where participants mainly imitated the variation. Moreover, the latency difference disappears over the course of the experiment in both cases. Usually, compatibility effects persist and are resistant to training (cf. MacLeod, 1991). This suggests that the gestural stimulus–response (in)compatibility does not influence latencies at all.

This conclusion seems at odds with two recent reports that argue that hearing a given speech gesture activates the relevant motor command. Galantucci et al. (2009) found that hearing a speech sound as a distractor has an impact on vocal choice RTs. In their experiments, participants had to react to visual stimuli (e.g., "##") with a spoken syllable (e.g., "ba"). Auditory distractors led to facilitation or inhibition when they matched or mismatched the intended response. In a similar vein, Yuen, Davis, Brysbaert, and Rastle (2010) showed that auditory distractors influence the exact position of the tongue, so that a /t/-distractor leads to more alveolar contact on nonalveolar target segments (e.g., /k/). However, in both cases, the authors confounded gestural and phonological stimulus–response compatibility. Some theories of phonology see the main task of phonological representation as a means of bridging the gap between perception and production (Boersma, 1998), so that these results can be attributed to phonological association, rather than to an automatic activation of speech gestures by hearing speech. As we highlighted in the introduction, it is necessary to find stimulus–response combinations that differ in their speech gestures but are phonologically equivalent. Only with such stimuli do the predictions of a learning account and a gestural account differ. Even though gestural and phonological compatibility are often confounded, it is possible (Mitterer & Ernestus, 2008) and necessary to deconfound them.

To do so in the present study, we used the instruction to correct in Experiment 2, and in Experiment 3, we had to mention the fact that regional variation was to be expected, given the markedness of the variation used in this experiment. In this way, we were able to elicit stimulus–response pairings that were phonologically compatible but gesturally incompatible. However, it remains possible that these instructions may have had some unexpected effect. Another potential way forward would be to simply "measure" to what extent the fine-phonetic details of stimulus and response match and whether this match is related to response latency. Instead of introducing relatively strong stimulus–response compatibilities by experimental measures and instructions, this approach would make use of natural variation in stimulus–response overlap (i.e., on a given trial, the response may be more or less similar to the stimulus). However, this approach, although theoretically promising, faces the problem of measuring the amount of gestural alignment of stimulus and response, which is difficult from the acoustic record for two necessarily different speakers (the model speaker and the participant). If these problems were solved, this would provide additional critical evidence.

The conclusion of the present work seems to be that what we hear influences how we speak, but probably not by a direct activation of speech gestures when hearing speech sounds. With this point in mind, it is worthwhile to consider how speech perception and production are linked in language use outside of the laboratory. One potential area for research is turn-taking in a dialogue, in which people take turns at being speakers and listeners. This phenomenon has attracted some attention in the psychological as well as the linguistic literature (Caspers, 1998; De Ruiter, Mitterer, & Enfield, 2006; Sacks, Scheglof, & Jefferson, 1974; Stivers et al., 2009; M. Wilson & Wilson, 2005). One aspect that has attracted considerable attention is the timing of turn-taking, often defined as the floor-transfer offset (FTO), defined as the time difference between the end of one speaker's turn and the onset of the following turn. A negative FTO thus indicates overlap, whereas a positive FTO is a silent gap. An interesting finding is that, cross-culturally, interlocutors seem to aim for a zero FTO, where there is neither an overlap nor a gap. Moreover, chronometric psycholinguistic work indicates that it takes about 600 ms to plan an utterance (cf. Indefrey & Levelt, 2004). To function in a dialogue, we therefore need to plan an utterance while still listening to speech. Automatic activation of the motor cortex would be quite unhelpful in the endeavor to produce a correct reply to our interlocutor's turn. In the psychological literature, the perception–action link is often investigated as "perception-for-action." Given how perception and action are linked in language use, automatic activation of speech gestures would, in fact, be perception against action, because the activated speech gestures would interfere with the planning of one's utterance. Any theoretical viewpoint that argues for a more tight coupling between perception and action than we do here needs to address how we can function in a dialogue if our perception engages the motor system, while the same motor system is generating an utterance plan of its own at the same time.

To sum up, the present study found that variation in the shadowing task tends to be imitated if participants are able to do so and if the variation is salient. Somewhat surprisingly, variation is more likely to be imitated when the speaker varies from trial to trial than when the speaker is consistent. Note that the latter option is the more ecologically valid one, since speakers tend to be consistent in their regional accents. Future research needs to address to what extent imitation may be due to demand characteristics (Durgin et al., 2009) of the shadowing task. Next to the strong tendency to imitate, we also gathered data in which participants did not use the same gestures as the stimulus model, either by instruction or by spontaneous imitation. For both cases, the gestural stimulus–response mismatches did not lead to latency costs. This confirms the finding in Mitterer and Ernestus (2008), in which the stimulus–response gestural mismatches were a consequence of the participants' inability to produce the stimulus gestures. The present data show that the same results are obtained when participants are able to produce the stimulus gestures.

Author note Holger Mitterer, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. Jochen Müsseler, Psychology Department, RWTH Aachen University, Germany.

The authors gratefully acknowledge the help of Laura Makowksi for running the experiments and the help of Laura Makowski and Michael Wiechers in coding the responses. We thank Dan Acheson for comments on a previous version of this article.

Appendix

Table 3 *-ig* words and nonwords

-ig Word	English translation	log (frequency + 1)	Anologous nonword
billig	cheap	4.16	zallig
deftig	hearty	0.29	gostig
emsig	assiduous	1.10	onsig
Essig	vinegar	0.61	dossig
ewig	eternal	3.73	lowig
fähig	capable	2.94	suhig
fertig	ready	4.26	schurtig
gültig	valid	3.19	doltig
heftig	fierce	4.07	schastig
Honig	honey	2.08	tänig
Käfig	cage	1.10	tofig
König	king	4.68	fonig
lässig	casual	1.61	wüssig
mickrig	pathetic	0.00	hukrig
mollig	chubby	0.00	rallig
schäbig	battered	0.69	heibig
übrig	residual	5.05	latrig
wenig	few	6.44	lünig
winzig	tiny	2.81	lonzig
zwanzig	twenty	4.40	pfonzig

Table 4 Fricative-stop cluster words and nonwords

Word	English translation	log (frequency + 1)	Anologous nonword
Spiegel	mirror	3.63	spieder
spülen	to do the dishes	1.25	spümen
Spektrum	spectrum	1.34	spetgon
sperren	to block	2.93	spellen
Spende	donation	2.60	spemke
spielen	to play	5.79	spieren
Spinne	spider	1.70	spimme
spitzen	to sharpen	1.99	spikfen
Spritze	syringe	2.08	sprikwe
spurlos	without a trace	1.25	spurkos
Steuer	tax/steering wheel	3.54	steuel
Stiefel	boot(s)	2.44	stieser
Stempel	seal	2.06	stengkel
streiken	to strike	2.55	streipen
Stunde	hour	5.75	stungke
stöbern	to browse	0.29	stögeln
stiften	to endow	2.30	stichken
Standard	standard	2.70	stambald
stemmen	to lift	1.67	stennen
stinken	to smell (badly)	1.90	stimpen

Table 5 /ɪk/- and /ɪç/-final control words

/ɪk/-final word	English translation	log (frequency + 1)	/ɪç/-final word	English translation	log (frequency + 1)
Lyrik	poetry	2.92	Anstich	tapping	2.48
Klinik	clinic	2.94	Dietrich	lock pick	2.55
Anblick	sight	3.00	Kranich	crane	0.51
Technik	technique	4.91	Rettich	radish	2.01
Klassik	classical music	1.47	Teppich	carpet	3.18
Grafik	graphics	1.95	plötzlich	suddenly	5.08
Panik	panic	1.99	nämlich	namely	4.73
Chronik	chronicles	2.32	glimpflich	without serious consequences	0.29
Taktik	tactics	2.46	grässlich	horrible	1.50
Plastik	plastic	2.58	ziemlich	quite	4.34

References

Baayen, H. R. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, H. R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (release 2 ed.): Linguistic Data Consortium.

Baron-Cohen, S., & Staunton, R. (1994). Do children with autism acquire the phonology of their peers? An examination of group identification through the window of bilingualism. *First Language, 14,* 241–248. doi:10.1177/014272379401404216

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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

Caspers, J. (1998). Who's next? The melodic marking of question vs. continuation in Dutch. *Language and Speech, 41,* 375–398. doi:10.1177/002383099804100407

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin and Review, 11,* 1084–1089.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82,* 515–535. doi:10.1353/lan.2006.0130

Diehl, R., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55,* 149–179. doi:10.1146/annurev.psych.55.090902.142028

Dixon, P. (2008). Models of accuracy in repeated-measures design. *Journal of Memory and Language, 59,* 447–456.

Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaughnessy, K., & Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review, 16,* 964–969. doi:10.3758/PBR.16.5.964

Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology. Human Perception and Performance, 27,* 229–240. doi:10.1037//0096-1523.27.1.229

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One, 5,* e10068. doi:10.1371/journal.pone.0010068

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99,* 1730–1741. doi:10.1121/1.415237

Fowler, C. A., Brown, J. M., Sabadini, L., & Welhing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49,* 396–413. doi:10.1016/s0749-596x(03)00072-x

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics, 71,* 1138–1149. doi:10.3758/app.71.5.1138

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415,* 755. doi:10.1038/415755a

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequences. In H. Giles, N. Coupland, & J. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1–68). Cambridge: Cambridge University Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279. doi:10.1037//0033-295X.105.2.251

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.

Gregory, S. W. J., & Webster, S. W. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perception. *Journal of Personality and Social Psychology, 70,* 1231–1240.

Harrington, J., Palethorpe, S., & Watson, C. (2000). Does the Queen speak the Queen's English? *Nature, 408,* 927–928. doi:10.1038/35050160

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *The Behavioral and Brain Sciences, 24,* 849–937. doi:10.1017/S0140525X01000103

Honorof, D. N., Weihing, J., & Fowler, C. A. (2011). Articulatory events are imitated under rapid shadowing. *Journal of Phonetics, 39,* 18–38. doi:10.1016/j.wocn.2010.10.007

Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience, 7,* 942–951. doi:10.1038/nrn2024

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical Mechanisms of Human Imitation. *Science, 286,* 2526–2528. doi:10.1126/science.286.5449.2526

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. [Review]. *Cognition, 92,* 101–144. doi:10.1016/j.cognition.2002.06.001

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. [Article]. *Journal of Experimental Psychology. Human Perception and Performance, 26,* 634–647. doi:10.1037//0096-1523.26.2.634

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception 36*, ECVP Abstract Supplement.

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

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

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences, 13,* 110–114. doi:10.1016/j.tics.2008.11.008

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

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30,* 1113–1126. doi:10.1207/s15516709cog0000_79

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75–78. doi:10.1126/science.198.4312.75

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science, 35,* 184–197. doi:10.1111/j.1551-6709.2010.01140.x

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109,* 168–173. doi:10.1016/j.cognition.2008.08.002

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS One, 4*. doi:10.1371/journal.pone.0007785

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*. doi:10.1016/j.wocn.2010.12.007

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204–238. doi:10.1016/S0010-0285(03)00006-9

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 9,* 1718–1725. doi:10.1121/1.414696

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119,* 2382–2393. doi:10.1121/1.2178720

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics, 72,* 2254–2264. doi:10.3758/BF03196699

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. *The emergence of language* (pp. 381–415). Erlbaum: Mahwah, NJ

Quene, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59,* 413–425. doi:10.1016/j.jml.2008.02.002

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*. doi:10.1146/annurev.neuro.27.070203.144230

Sacks, H., Scheglof, E. A., & Jefferson, G. (1974). Simplest Systematics for Organization of Turn-Taking for Conversation. *Language, 50,* 696–735. doi:10.2307/412243

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

Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action: Candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience, 10,* 295–302. doi:10.1038/nrn2603

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 10587–10592. doi:10.1073/pnas.0903616106

Stone, G. O., & Orden, G. C. V. (1993). Strategic control of processing in word recognition. *Journal of Experimental Psychology. Human Perception and Performance, 19,* 744–774. doi:10.1037//0096-1523.19.4.744

Sundara, M., Namasivayam, A. K., & Chen, R. (2001). Observation-execution matching system for speech: A magnetic stimulation study. *Neuroreport, 12,* 1341–1344. doi:10.1097/00001756-200105250-00010

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,* 178–196. doi:10.1037/0096-1523.32.1.178

Van Bezooijen, R. (2005). Approximant /r/ in Dutch: Routes and feelings. *Speech Communication, 47,* 15–31. doi:10.1016/j.specom.2005.04.010

Van der Heijden, A. H. C., Hagenaar, R., & Bloem, W. (1984). Two stages in postcategorial filtering and selection. *Memory and Cognition, 12,* 458–469. doi:10.3758/BF03198307

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. [Article]. *Neuropsychologia, 41,* 989–994. doi:10.1016/s0028-3932(02)00316-0

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12,* 957–968. doi:10.3758/BF03206432

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702.

Yuen, I., Davis, M. H., Brysbaert, M., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 592–597. doi:DOI10.1073/pnas.0904774107