Motor Control, 2011, 15, 68-84

Masked Auditory Feedback Affects Speech Motor Learning of a Plosive Duration Contrast

Silvia C. Lipski, Stefanie Unger, Martine Grice, Ingo G. Meister

Adult speakers have developed precise forward models of articulation for their native language and seem to rely less on auditory sensory feedback. However, for learning of the production of new speech sounds, auditory perception provides a corrective signal for motor control. We assessed adult German speakers' speech motor learning capacity in the absence of auditory feedback but with clear somatosensory information. Learners were presented with a nonnative singleton-geminate duration contrast of voiceless, unaspirated bilabial plosives /p/ vs. /pp/ which is present in Italian. We found that the lack of auditory feedback had no immediate effect but that deviating productions emerged during the course of learning. By the end of training, speakers with masked feedback produced strong lengthening of segments and showed more variation on their production than speakers with normal auditory feedback. Our findings indicate that auditory feedback is necessary for the learning of precise coordination of articulation even if somatosensory feedback is salient.

Keywords: speech motor learning, articulation, auditory feedback

Auditory and somatosensory feedback are both important for the acquisition and control of speech. It has been hypothesized that, especially when learning new speech sounds, speakers use auditory perceptual categories as a reference for articulation, for example during the acquisition of a second language (Flege, 1995; Flege et al., 1997). The relationship between perception and production is formulated in the neurolinguistic model of speech production by Guenther (2006). By repeated comparison of the acoustic result to stored auditory goals, a speaker refines and stabilizes speech motor commands. The feedback-based learning process results in an internal model of speech movements representing the relation between movement and auditory as well as somatosensory effects (Guenther, 2006; Guenther, Hampson, & Johnson, 1998). The importance of accurate auditory representations as a guidance for motor control has been demonstrated in learning studies which found that perceptual training before or during the production of new speech sounds had a positive effect on articulation (Bradlow, Pisoni, Akahane-Yamada,

Lipski is with the Max Planck Institute for Neurological Research Cologne, Germany. Unger and Grice are with IfL-Phonetik, University of Cologne, Germany. Meister is with the Department of Neurology, University Hospital, Cologne, Germany.

& Tohkura, 1997; Lipski, Grice, & Meister, 2008; Wang, Jongman, & Sereno, 2003). Furthermore, Perkell and colleagues (2004) provided evidence that precise auditory perceptual representations and fine discrimination in articulation are also correlated for native speech sounds.

Adult speakers who have acquired a stable internal model of speech production are not as dependent on auditory feedback as early learners. Since a strong reliance on auditory feedback would probably slow down articulation, and since auditory feedback is not always available, for example, while speaking in a noisy environment, reliable feedforward programs are necessary. These representations may enable speakers who lose hearing as adults to sustain intelligible speech (Menard et al., 2007; Perkell et al., 2007). In addition, it has been shown that somatosensory goals for speech sounds are used for fine-grained articulatory control (Tremblay, Shiller, & Ostry, 2003). Speech control through somatosensory feedback is present in normally hearing speakers even in cases where disturbing the feedback has no effect on the acoustic result (Tremblay et al., 2003). Nasir and Ostry (2008) showed that deaf speakers with their cochlear implants switched off compensated for somatosensory perturbance of the jaw during the production of words of their native language. Moreover, these speakers retained the acquired compensatory movements after the mechanical perturbation ceased. Thus, Nasir and Ostry (2008) demonstrated that somatosensory feedback could be used in learning to modify an existing articulatory program for native speech sounds.

Studies which investigated perturbations in auditory feedback showed that in addition to the contributions of feedforward motor control and somatosensory feedback control, auditory feedback seems to be monitored to plan articulatory movements in accordance with auditory perceptual target representations. If speech is perturbed or articulators change, the auditory feedback system provides corrective commands. Compensations for a mismatch between target and acoustic result lead to a correction of articulation as soon as 100–150 ms after the disturbance. Speakers vary strongly in their magnitude of compensation for changes in auditory feedback, however, adaptive articulation for modifications of pitch or vowel quality has been shown by numerous studies (Houde & Jordan, 1998; 2002; Jones & Munhall, 2000; Natke & Kalveram, 2001; Villacorta, Perkell, & Guenther, 2007; Xu, Larson, Bauer, & Hain, 2004) and also for sibilants (Shiller, Sato, Gracco, & Baum, 2009). Recent findings show that this process is automatic and may be difficult to prevent by the speaker (Munhall, MacDonald, Byrne, & Johnsrude, 2009).

Previous studies on feedback based control and adaptation have only tested the effect of altered or missing feedback for native speech sounds for which the model of control and efference is highly developed. The present study investigated the role of auditory feedback in adult normally hearing speakers for the imitation of a new speech sound contrast. Two groups of adult German speakers practiced the articulation of an unfamiliar contrast which is present in a number of languages, including Italian, but not in the speakers' native language: the phonological length distinction between /p/ and /pp/, generally analyzed as singleton and geminate, respectively. Auditory feedback during learning was masked for one group of speakers while it was not altered for the other group.

Geminate consonants in Italian are generally about twice as long as singletons. Furthermore, the vowels preceding geminates are shorter than those before singleton consonants (D'Imperio & Rosenthall, 1999; Esposito & Di Benedetto,

1999; Gili Fivela, Zmarich, Perrier, Savarinaux, & Tisato, 2007). There is a salient relation between the tactile and auditory sensory dimension in the oral constriction, the duration of labial contact being almost equal to the acoustically measurable closure duration. The actual articulatory gesture for the formation of bilabial stop closure and release is not new for German speakers. What is new is the contrast between two different closure durations along with compensatory shortening of the preceding vowel in the case of the longer, geminated sound /pp/.

We tested the influence of auditory feedback on the production of the new sound contrast with almost equal saliency of auditory and somatosensory feedback. Our hypothesis was that in addition to somatosensory feedback, auditory feedback is crucially involved in the acquisition of the precise articulation for the contrast to be learned.

Methods and Materials

The experiment consisted of two parts: a baseline test and a learning experiment. The baseline test was carried out first to test participants' initial ability to imitate a singleton-geminate consonant contrast. The results of the baseline test were used to divide participants into two groups, balanced for their accuracy in the articulation of the contrast applied in the baseline test. Subsequently, in the learning experiment, auditory feedback of one group was masked while the other group received normal auditory feedback.

Participants

Our volunteers were 30 female German native speakers (mean age: 24.0, range = 20–30). None of the participants had impaired hearing or speech production. Participants were students or staff of medical science and psychology at the University of Cologne, none of them had any prior phonetic expertise. German was the first language of all participants, none grew up as an active or passive bilingual. Furthermore, none had any knowledge of languages with geminate consonants. Exclusively female participants were tested to increase the homogeneity of verbal memory (cf. review in Andreano & Cahill, 2009) and the time-course of learning among participants. It has been shown that, at least for auditory learning, females show steeper learning curves at an early stage in comparison with male learners who show more improvement later on (Burns & Rajan, 2008).

Stimuli

For the baseline and the learning experiment, naturally spoken, unedited stimuli were used. Stimuli for each experiment were spoken by different female Italian speakers. For the baseline the speaker was from Calabria, for the learning experiment from Apulia.

Stimuli for the baseline were 7 types of fillers: ['ifi], ['izi], ['ili], ['irri], ['iʎʎi], ['iɲɲi], ['iʃʃi], and 2 target items involving a contrast between a single and geminate consonant: ['ini], ['inni]. The first syllable in each stimulus was stressed. The singleton target stimuli had a mean consonant duration of 80 ms (± 4 ms). The geminates had a mean of 232 ms (± 14 ms).

The stimuli for the learning experiment were the single and geminate bilabial plosives as occurring in Italian ['apa] and ['appa] with stress on the first syllable. Five different exemplars per stimulus category were used. Stimuli with singleton consonants ['apa] had a mean first vowel duration of 191 ms (± 3 ms) and mean closure duration of 128 ms (± 6 ms). Stimuli with geminate plosives ['appa] had a mean duration of first vowel of 129 ms (± 8 ms) and a mean closure duration of 287 ms (± 12 ms).

Procedures and Data Analysis

Tests took place in a sound-attenuated room at the University of Cologne. Speech was digitally recorded using an AKG C420 III headset condenser microphone (sampling rate = 44.1 kHz, 16 bit resolution). Auditory signals were presented over Philips SBC HP195 headphones at approximately 75 dB.

Participants were instructed to imitate the sounds as exactly as possible. They were informed that the experiment was designed as a training session for pronunciation. However, participants were not informed about the nature of the contrast or the number of different sound categories.

Before the learning experiment on the influence of auditory feedback, the baseline test was carried out. For this test, the duration contrast between single and geminate intervocalic nasals in ['ini] and ['inni] was used. This contrast does not occur in the German language, the nasal geminate [nn] is not part of the German sound inventory. The task was the same for every participant and no masking of auditory feedback was applied in the baseline test.

In the baseline experiment, 112 filler (4 per category, repeated 4 times) and 16 target (4 per category, repeated 2 times) stimuli were used. Participants were asked to imitate each trial once. A pictogram of a mouth that appeared 800 ms after the presentation of the stimulus indicated when the subject was required to begin speaking. The intertrial-interval was 1500 ms. To prevent direct comparison of the single and geminate contrasts, stimuli were presented in pseudo-randomized order, ensuring that the two categories ['ini] and ['inni] did not follow each other directly.

For the learning experiment, participants were divided into two groups, balanced for age and their accuracy in the articulation of the contrast between ['ini] and ['inni] in the baseline experiment. One group imitated the stimuli without auditory feedback, while the other did so with no impairment to feedback at all. In the first group the feedback was masked with white noise over headphones with a mean intensity of 74 dB. Participants were instructed to whisper the stimuli to prevent bone-conduction. Headphones were fixed additionally with a sweat band to enhance acoustic shielding. Participants reported that they could not hear themselves when they were whispering. The intensity of participants' whispered utterances in the group with masked auditory feedback was about 40–55 dB. To keep conditions equivalent, the second group was also instructed to whisper their imitations.

Five blocks with 20 stimuli (10 x ['apa], 10 x ['appa]) were applied. The mean duration of stimuli containing singleton consonants was 400 ms, (range: 384–425); geminate stimuli: 489 ms (range: 468–517). After listening to a stimulus, participants repeated it three times in a whisper. The repetition was used because repeating utterances after an auditory target has been presented once makes reliance on auditory feedback correction more likely. A pictogram of a mouth that appeared three

times indicated the point at which the subjects were required to begin articulating: the first repetition was at 100 ms after stimulus offset, the second was at 1200 ms after stimulus offset, the third was at 2300 ms after stimulus offset. The intertrial-interval was 4000 ms. There was a pause of several minutes between blocks. The experiment had a total duration of two hours.

Productions were manually labeled from the waveform with reference to a spectrogram using Praat (Boersma & Weenink, 2008). In the baseline test, only the target stimuli ['ini] and ['inni] were analyzed, yielding a total of 3000 productions for all speakers.

For the learning experiment, the productions of the first and the last training block were used for analysis. Thus, a total of 120 stimuli were analyzed for each speaker (30 x ['apa], 30 x ['appa]). In total, 1800 tokens were analyzed from each group of speakers.

For the baseline targets the duration of the first vowel and the nasal consonants [n] and [nn] were measured. For the learning experiment, the duration of the first vowel and the closure phase of the plosives, [p] and [pp] were labeled. The end of the closure was identified as the onset of the burst.

Statistical analysis was carried out with R (R Foundation for Statistical Computing, Vienna, Austria, www.R-project.org). For the baseline measures, an ANOVA for nasal duration was calculated with stimulus category and group as the dependent variables. For the results of the learning experiment repeated-measures ANOVAs for vowel and closure duration were conducted separately for each of the three repetitions with group (full feedback, no auditory feedback), stimulus category (singleton, geminate), and block (Block 1, Block 5) as the main factors. Holm adjustment was applied for post hoc comparisons of means. A level of $p < .05$ was considered significant.

Results

Baseline Experiment

Before the learning experiment, the groups of participants were matched according to articulation performance of the target singleton and geminate consonant in the baseline measurement. Thus, no group effects were found for closure duration: two-way interaction between stimulus category and group for nasal duration: $F(1,405) = 0.5$, $p > .5$; and for the duration of the first vowel: $F(1, 405) = 0.0005$, $p > .9$, (Figure 1).

The contrast between [n] and [nn] was realized very clearly with significantly longer nasal duration for the geminate than for the singleton consonant. A strong main effect of stimulus category was found ($F(1,405) = 206.6$, $p < .0001$). The singleton targets (80 ms) were imitated rather closely with a mean of 110.5 ms (*SD*: 25.0), whereas the geminate nasals deviated more strongly from the targets (232 ms) with a mean of 182.2 ms (*SD*: 43.3).

The duration of the first vowel of the ['ini] targets with a mean of 150.0 ms (*SD*: 30.6) was significantly longer than first vowel duration in the ['inni] targets with a mean of 106.6 ms (*SD*: 28.5), ($F(1, 405) = 135.3$, $p < .0001$). This difference was very similar to the contrasts in the vowel duration of the baseline target stimuli (['inni]: 98 ms, ['ini]: 140 ms).

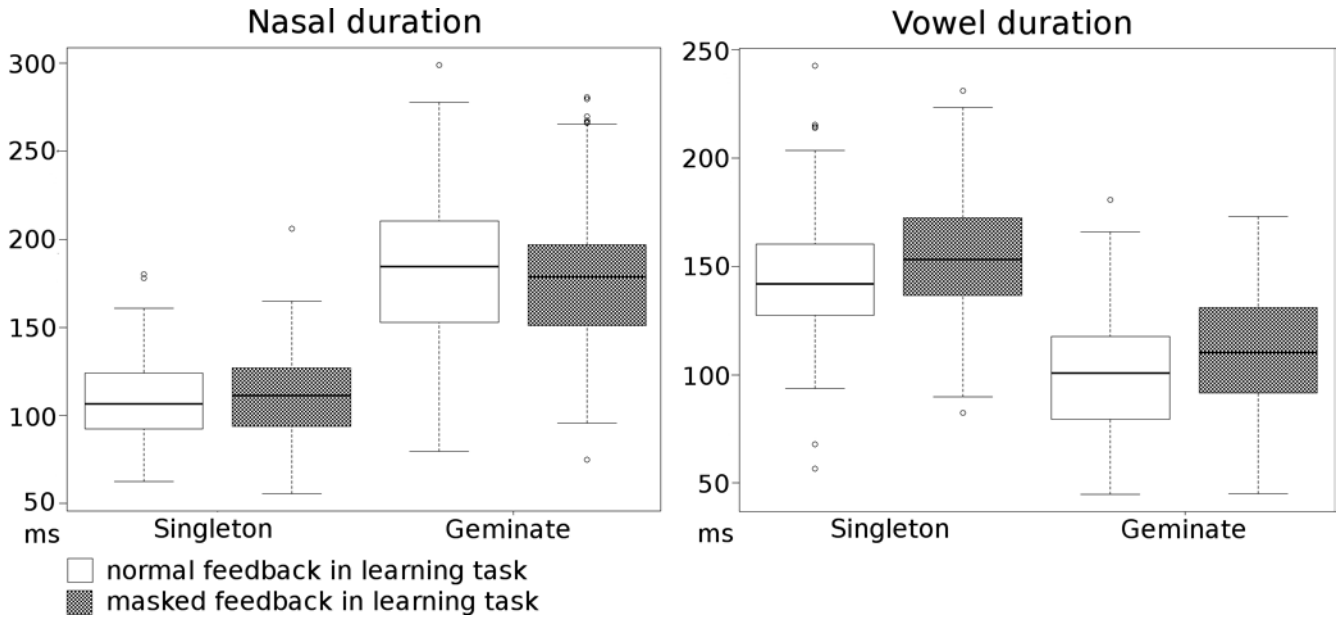

Figure 1 — Baseline test was carried out to ensure equal initial ability to imitate a singleton vs. geminate contrast between groups. Boxplots display the duration of nasal and first vowel in singleton and geminate categories in the baseline test for the participants that were later assigned to the normal auditory feedback group, and for those later assigned to the group with masked auditory feedback. Bottom and top of boxes denote the lower and upper quartile, median is shown as the centreline. Length of whiskers denotes data within lower and upper 1.5 inter-quartile range, the circles show remaining data.

One speaker that was assigned to the group with full feedback during training had 8 deviating productions with [nj] instead of [n] (4 times) and [nn] (4 times), 5 similarly deviant productions were found for 3 speakers in the group who had no auditory feedback during training (5 times [nj] instead of [nn]).

Closure Duration

The main finding was a strong lengthening of closure duration in the last block of training by the group without feedback, both for the singleton and even more strongly, for the unfamiliar geminate consonant. This effect was significant during the second repetition of the target stimulus after a delay of 1200 ms (cf. Figure 2).

The first repetition immediately after participants had heard the stimulus showed a significant interaction between block and stimulus category ($F(1, 1042) = 10.9$, $p < .01$). The three-way interaction with group was not significant ($F(2, 1042) = 1.3$, $p = .3$). A comparison of means showed that the stimulus categories were realized with different closure durations in both blocks (all $p < .001$). No change for the singleton consonant was found across blocks ($p = .05$), but the geminate sound was lengthened in Block 5 ($p < .001$).

In the second repetition, a difference between the two groups emerged in the last block of training. The interaction between block, stimulus category, and group was significant ($F = (2, 1048) = 4.2$; $p < .05$). Post hoc comparisons revealed again that both categories were produced differently in both blocks by both groups (all $p < .001$). The groups did not differ for the singleton plosive in the first block ($p > .1$). However, the group without auditory feedback strongly lengthened closure duration for both stimulus categories in Block 5, much stronger than the group with auditory feedback, as revealed by group comparison for the singleton in Block 5 with $p < .01$; and for the geminate with $p < .0001$.

Although a tendency for longer closures by the group without auditory feedback was found for the third repetition, as well, the group effect for the three-way interaction was not significant ($F(2, 1050) = 1.7$; $p = .2$). Again, the stimulus category by block interaction was significant ($F(1, 1050) = 8.9$, $p < .01$). Comparisons of means revealed that the singleton consonant did not change ($p = .4$), but the geminate sound was significantly longer in the last block ($p < .0001$).

The question arises as to whether the whole stimulus was lengthened in the productions of the group with auditory feedback. This will be clarified in the next section in which the vowel duration is discussed.

A striking difference across the groups was found in the variability of their productions. As Table 1 shows speakers without auditory self-perception had much more variability in their productions than speakers with unimpeded feedback.

Vowel Duration

The two groups did not differ in their realization of vowel duration. Generally, vowel duration was realized much longer than the target stimuli in the beginning and at the end of training. Although, as shown in Table 2 and Figure 3, the results for the group without auditory feedback showed a tendency toward more deviation from the target values especially at the end of training, there was no effect of auditory feedback masking.

Table 1 Mean (SD) closure duration for the two plosive categories in ms in first and last block of training for each of the three repetitions.

	Stimulus category	Repetition	Normal feedback	Masked feedback
First Block	[p]	1	151 (46)	154 (44)
		2	153 (46)	161 (59)
		3	152 (47)	165 (58)
	[pp]	1	244 (89)	252 (63)
		2	239 (68)	251 (76)
		3	224 (70)	238 (71)
Fifth Block	[p]	1	138 (34)	152 (60)
		2	150 (40)	172 (76)
		3	152 (42)	173 (76)
	[pp]	1	252 (68)	294 (99)
		2	251 (71)	320 (125)
		3	247 (79)	308 (131)

Target closure duration: [p]: 128 (6) ms; [pp]: 287 (12) ms

Although speakers' vowels were generally much longer than those of the target stimuli, the general patterns of vowel duration for the singleton and geminate stimuli were reproduced by the speakers. In all repetitions, vowel duration before the singleton was significantly longer than before the geminate sound and both groups increased the duration of the vowel before the singleton consonant from Block 1 to Block 5 for all three repetitions.

For the first repetition, the three-way interaction was not significant ($F(2, 1042) = 1$, $p = .3$), but an effect for stimulus category by block comparison was seen ($F(1, 1042) = 19.2$, $p < .001$). The vowel before the singleton plosive was lengthened during training ($p < .0001$) but not for the geminate ($p = .9$).

Similar results for the second repetition were found with no group effect in the three-way interaction ($F(2, 1048) = 2.2$, $p = .2$), but a significant interaction of stimulus category by block ($F(1, 1048) = 19.8$, $p < .001$), again with a longer vowel before the singleton consonant in Block 5 ($p < .001$) but no change from Block 1 to Block 5 was seen for the geminate ($p = .3$).

The third repetition showed the same result. The two-way interaction ($F(1, 1050) = 8.9$, $p < .01$) but not the three-way interaction ($F(1, 1050) = 1.7$, $p = .2$) was significant and a lengthening of the vowel for the singleton category ($p < .0001$) but no difference between blocks was observed for the geminate stimulus ($p = .5$).

Table 2 shows that the group without auditory feedback varies in vowel duration more strongly than the group with normal feedback. However, the absence of a group effect for vowel duration indicates that speakers without auditory feedback did not lengthen the whole logatome.

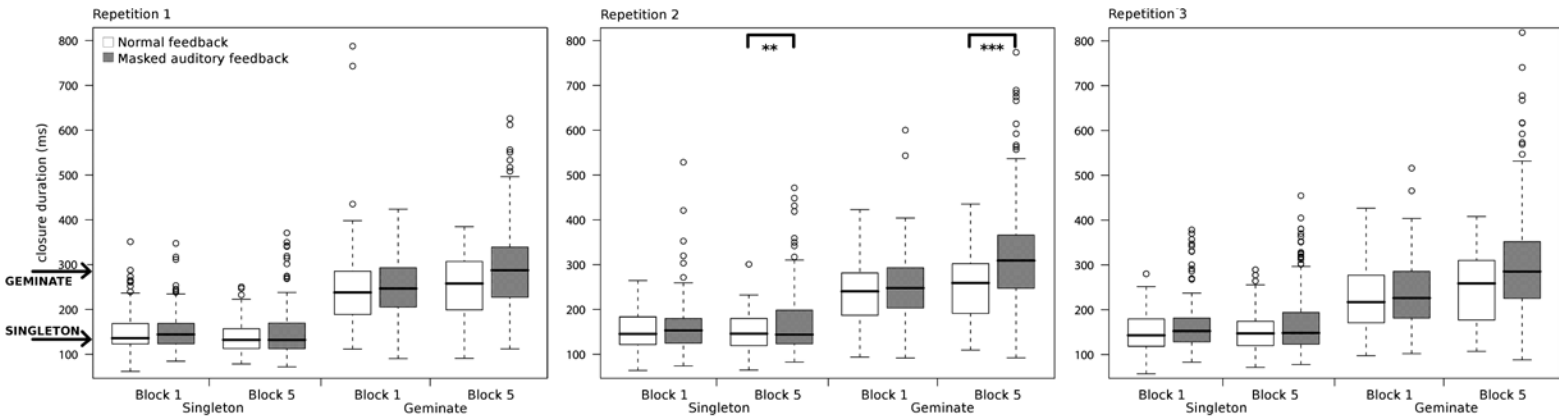

Figure 2 — Closure duration: The lack of auditory feedback affected speech production at later training stages. Speakers without auditory feedback showed higher variability of closure duration in Block 5. The boxplots show closure duration distribution in speakers with (white) and without auditory feedback (dark) in the three repetitions in Block 1 and Block 5. Bottom and top of boxes denote the lower and upper quartile, median is shown as the centreline. Length of whiskers denotes data within lower and upper 1.5 inter-quartile range, the circles show remaining data. Significant group differences are marked by ** for $p < .001$, and *** for $p < .0001$. Arrows on the left indicate mean values of target stimuli.

Table 2 Mean (SD) duration of the vowel preceding the two plosive categories in ms in first and last block of training for each of the three repetitions.

	Stimulus category	Repetition	Normal feedback	Masked feedback
First Block	[p]	1	208 (43)	247 (77)
		2	214 (46)	250 (72)
		3	205 (43)	256 (77)
	[pp]	1	169 (58)	185 (68)
		2	171 (54)	188 (65)
		3	176 (54)	201 (83)
Fifth Block	[p]	1	235 (54)	293 (94)
		2	245 (56)	289 (86)
		3	244 (57)	293 (84)
	[pp]	1	167 (62)	188 (64)
		2	182 (60)	189 (64)
		3	185 (64)	198 (72)

Target vowel duration: [p]: 191 (3) ms; [pp]: 129 (8) ms

Deviant Productions

In both groups, several sounds were produced which clearly deviated from the target categories. Both groups produced the same types of deviations. A group difference can be found in the number of deviant productions, which were more frequent for the group with normal feedback. Furthermore, the availability of auditory feedback may have had a negative effect of strengthening the development of false production strategies in one speaker who produced two consecutive stops with two releases instead of the geminate plosive with a long closure phase and one release. This speaker stabilized this pattern in Block 5 and substituted most geminate sounds with the two-plosive pattern. Thus, she may have adopted and stabilized an inappropriate strategy to realize the unfamiliar sound. No such tendencies were found for the group without auditory feedback where deviations occurred only on occasion.

Six speakers in the group with normal feedback produced 31 deviant productions in Block 1, 14 for the singleton and 17 for the geminate category, consisting of 5 types of errors voiced stop [b], insertion of a second plosive: [pb], [p^{h}ʔ], and [p^{h}p] or substitution with the fricative [f]. In Block 5 three speakers in the group with normal feedback produced 32 deviations, 29 of them were found for one speaker, as mentioned above, for the geminate category, the rest were produced for the singleton consonant.

In the group with masked auditory feedback, in Block 1 four speakers had a total of 14 deviations which consisted of 4 different types. Twelve deviations were

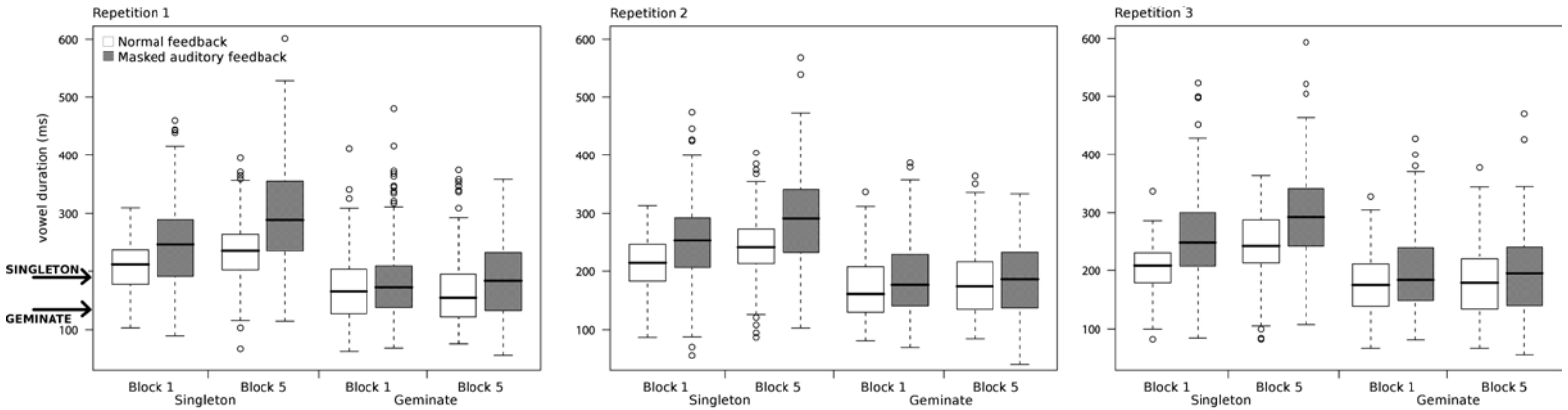

Figure 3 — Vowel duration: No significant group differences were found for vowel duration. Boxplots show vowel duration distribution in speakers with normal feedback (white) and with masked feedback (dark) in the three repetitions in Block 1 and Block 5. Bottom and top of boxes denote the lower and upper quartile, median is shown as the centreline. Length of whiskers denotes data within lower and upper 1.5 inter-quartile range, the circles show remaining data. Arrows on the left indicate mean values of target stimuli.

found for the singleton category. In Block 5, six speakers of this group produced 18 deviations of 5 types, of which 11 were found for the singleton consonant. The deviations were sporadic without indication of regular substitutions.

Discussion

This study demonstrates the importance of auditory feedback for learning of the precise articulation of a new sound contrast in adult speakers. The lack of auditory feedback did not have an immediate effect on the production of target speech sounds. Rather, the effect became evident over the course of the learning process. The sounds that the speakers were trained to produce—voiceless, unaspirated Italian singletons [p] and geminates [pp]—provide clear somatosensory feedback, since they are mainly distinguished by the duration of lip closure.

Both groups exhibited a change in production which can be interpreted as a learning effect. They both increased the contrast between singleton and geminate closure duration from the beginning to the end of training in all repetitions, and increased the duration of the first vowel before singleton consonants.

The main finding of this study is that the absence of acoustic feedback in the group which spoke while white noise masked auditory self-perception resulted in stronger variance of closure duration and longer closures for both stimulus categories. The increase in variance and duration were at their strongest for the new geminate sound at the end of training. This effect was found in the second imitation of the sounds after listening to the target stimulus. In contrast, the speakers with unhindered feedback varied the duration of closure less strongly and did not overshoot the closure durations. Thus, speakers relied more strongly on auditory feedback when imitating after some delay.

The absence of a difference in the first repetition may be explained by the promptness of repetition: only 100 ms after stimulus presentation. At this point, the stimulus would have been available in auditory short-term memory, a factor which appears to improve translation to motor commands (Peschke et al., 2009). With increasing response latency, the role of sensory feedback on production may be stronger. The overshoot in closure durations by the group with masked feedback in the second repetition toward the end of training may result from an attempt to increase the contrast in somatosensory feedback. It may also be an attempt to correct for inappropriate auditory feedback and the lack of a perceived contrast. Studies on songbirds have shown that the deprivation of auditory feedback itself is processed as an auditory error signal (Brainard & Doupe, 2000). Whether this is the case for human speakers is, however, unknown.

The two groups did not differ in the third repetition although there still is a tendency for the group with masked feedback to produce longer closure durations and display stronger variability than the control group. The decrease of duration of the geminate closure for speakers without auditory feedback in the third repetition may result from processing a sensory-feedback error. This correction may be caused by the influence of native category representations since German plosive closures are normally much shorter. Producing such long closures in the second repetition may enhance attempts for self-correction.

Consonant closures are often longer in whispered than in voiced speech, arguably as a strategy for air-conservation (Schwartz, 1972). Since all speakers in both groups imitated the stimuli in a whisper, it can be ruled out that speaking mode caused the group difference found in the current study.

The lack of auditory feedback in our study affected articulation in a gradual way. In the short term, the absence of auditory feedback does not seem to affect vowel and closure duration for adult speakers. We found no differences between speakers with and without auditory feedback in the initial phase of learning, which is considered to be strongly affected by cognitive strategies (Fitts & Peterson, 1964). Our results indicate that articulatory precision and increasing consistency of contrast production are facilitated by a continuous process of auditory feedback correction, consisting of repeated comparison and correction as proposed by Guenther (2006). A slow impact of hearing loss has also been observed in postlingually deafened speakers (Lane et al., 2007; Matthies, Svirsky, & Perkell, 1994; Svirsky, Lane, Perkell, & Wozniak, 1992), whereas for prelingually deafened children, articulation was affected rapidly after auditory feedback was interrupted (Higgins, McCleary, & Schulte, 2001).

In the current study, the speakers with and without auditory feedback initially had comparable control over vowel and closure duration. This finding demonstrates that adults who have acquired internal models of articulation are able to rely on sensorimotor representations during early stages of learning. The finding that all speakers regardless of whether they could hear themselves or not produced the vowel duration pattern with a short vowel before the long and a long vowel before the short consonant may be explained by the fact that it conforms with German syllable structure where a long vowel (after a glottal closure) can be followed by a syllable boundary but a short vowel is followed by a consonant. When German speakers place the syllable boundary after [a] for the singleton stimuli and treat [pp] as an ambisyllabic consonant for the geminate category, the vowel duration pattern is not difficult at all.

Duration distinctions are generally rather easily perceived by nonnative speakers (Escudero, Benders, & Lipski, 2009; Ylinen, Shestakova, Alku, & Huotilainen, 2005). This might explain why all speakers were able to differentiate between the two plosive categories, both in terms of closure duration as well as the duration of the preceding vowel. Moreover, the German language has a vowel length distinction which is operative on the vowel [a]. This might explain why performance was comparable across the two groups for the vowel duration.

However, even though the pattern of vowel duration with a short vowel before the geminate and a long vowel before the singleton consonant was imitated very clearly by both groups, their vowels were generally much longer than the target vowels. It is likely that this is caused by the speaking mode, since speakers regularly produce longer syllables when whispering as compared with speaking in a normal voice (Tartter, 1989). The increase of duration before the singleton consonant was found for all speakers by the end of the training and may be an attempt to increase the contrast. Trying to speak more clearly leads to increased syllable duration (Ferguson & Kewley-Port, 2002; Ohala, 1994) but probably not to a shortening of short vowels or syllables.

Previous studies of speech motor learning with manipulated auditory or somatosensory feedback have examined the production of native speech sound contrasts or native words (Houde & Jordan, 1998; 2002; Jones & Munhall, 2000; Munhall et al., 2009; Nasir & Ostry, 2008; Natke & Kalveram, 2001; Shiller et al., 2009; Tremblay et al., 2003; Villacorta et al., 2007; Xu et al., 2004). These studies demonstrated the strong influence of sensory feedback on speech production and attest that adult learners have established sensory target representations in their native language.

In contrast, the current study provides a demonstration of speech motor learning without auditory feedback for a nonnative speech sound contrast. Our results show that auditory feedback is necessary to acquire more precise and less variable articulation. The observed drift toward longer closures that exceeded the target values and greater duration variability is consistent with findings for deaf speakers who show a lengthening of vowels and anomalous variation in voice onset time for plosives (Lane, Wozniak, & Perkell, 1994; Waldstein, 1990). The increase of variation in closure duration for the familiar singleton plosive in speakers without auditory feedback in the current study indicates that even aspects of speech which are salient in somatosensory terms are monitored auditorily.

Anatomically, the auditory feedback loop system associated with articulatory learning involves the bilateral superior temporal gyrus, where auditory variances are processed, and premotor and primary motor cortex, which process motor commands (Hickock, Buchsbaum, Humphries, & Muftuler, 2003; Tourville, Reilly, & Guenther, 2008). During speaking, the activity of the auditory cortex is reduced (Houde, Nagarajan, Skihara, & Merzenich, 2002). However, in primates, when a mismatch between the intended and the produced sound is perceived, exactly those neurons with reduced activity during normal vocalization respond most strongly (Eliades & Wang, 2008). Therefore, it is possible that the suppression of the auditory cortex during speaking enhances the sensitivity to detect articulatory errors (Eliades & Wang, 2008).

Of course, the reliance on auditory and somatosensory self-perception by itself does not ensure that new speech sounds are accurately learned by adult speakers, since interference from native sound categories can have a strong impact on learning (Best, 1995). The analysis of deviating productions in the current study indicates that at least one speaker in the group with full feedback adopted an erroneous production strategy of the new geminate sound during training. Perceptual training before or during production of new sounds can help to enhance correct auditory representations and thus enable effective articulatory learning (Bradlow et al., 1997; Lipski et al., 2008; Wang et al., 2003).

In summary, our findings indicate that auditory feedback is important for the acquisition of a new speech sound in adult speakers, even if somatosensory feedback is readily available. The effect of auditory feedback, however, was found at later stages of motor learning. Whereas masking of auditory feedback did not affect early learning, we have shown that the establishing of more precise feedforward models requires both somatosensory and auditory feedback.

Acknowledgments

We thank Thomas Kaleta for excellent support in data analysis. We would also like to thank two anonymous reviewers for their suggestions which have led to an improved manuscript.

References

Andreano, J.M., & Cahill, L. (2009). Sex influences on the neurobiology of learning and memory. *Learning & Memory (Cold Spring Harbor, N.Y.), 16,* 248–266.

Best, C.T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 171–206). Baltimore, MD: York Press.

Boersma, P., & Weenink, D. (2008). Praat: doing phonetics by computer (Version 5.0.20) [Computer program]. Retrieved July 22, 2008, from http://www.praat.org/

Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101,* 2299–2310.

Brainard, M.S., & Doupe, A.J. (2000). Interruption of basal ganglia-forebrain circuits prevents plasticity of learned vocalizations. *Nature, 404,* 762–766.

Burns, O.M., & Rajan, R. (2008). Learning in a task of complex auditory streaming and identification. *Neurobiology of Learning and Memory, 89,* 448–461.

D'Imperio, M., & Rosenthall, S. (1999). Phonetics and phonology of main stress in Italian. *Phonology, 16,* 1–28.

Escudero, P., Benders, T., & Lipski, S.C. (2009). Native, non-native, and L2 perceptual cue weighting for Dutch vowels: the case of Dutch, German, and Spanish listeners. *Journal of Phonetics, 37,* 452–465.

Esposito, A., & Di Benedetto, M.G. (1999). Acoustical and perceptual study of germination in Italian stops. *The Journal of the Acoustical Society of America, 106,* 2051–2062.

Eliades, S.J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature, 453,* 1102–1107.

Ferguson, S., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 112,* 259–271.

Flege, J.E. (1995). *Second language speech learning: Theory, findings, and problems*.

Flege J. E., Bohn O. - S., & Jang S. (1997). Effect of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25,* 437–470.

Fitts, P.M., & Peterson, J.R. (1964). Information capability of discrete motor responses. *Journal of Experimental Psychology, 67,* 103–112.

Gili Fivela, B., Zmarich, C., Perrier, P., Savarinaux, C., & Tisato, G. (2007). Acoustic and kinematic correlates of phonological length contrast in Italian. *Proceedings of the 16th International Conference of Phonetic Sciences (ICPhS),* Saarbrücken, Germany, pp. 469-472.

Guenther, F.H. (2006). Cortical interaction underlying the production of speech sounds. *Journal of Communication Disorders, 39,* 350–365.

Guenther, F.H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105,* 611–633.

Hickock, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience, 15,* 673–682.

Higgins, M.B., McCleary, E.A., & Schulte, L. (2001). Articulatory changes with short-term deactivation of the cochlear implants of two prelingually deafened children. *Ear and Hearing, 22,* 29–45.

Houde, J.F., & Jordan, M.I. (1998). Sensorimotor adaptation in speech production. *Science, 279,* 1213–1216.

Houde, J.F., & Jordan, M.I. (2002). Sensorimotor adaptation of speech I: compensation and adaptation. *Journal of Speech, Language, and Hearing Research: JSLHR, 45,* 295–310.

Houde, J.F., Nagarajan, S.S., Sekihara, K., & Merzenich, M.M. (2002). Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience, 14,* 1125–1138.

Jones, J.A., & Munhall, K.G. (2000). Perceptual calibration of F0 production: evidence from feedback perturbation. *The Journal of the Acoustical Society of America, 108,* 1246–1251.

Lane, H., Matthies, M.L., Guenther, F.H., Denny, M., Perkell, J.S., Stockmann, E., et al. (2007). Effects of short- and long-term changes in auditory feedback on vowel and sibilant contrasts. *Journal of Speech, Language, and Hearing Research: JSLHR, 50,* 913–927.

Lane, H., Wozniak, J., & Perkell, J.S. (1994). Changes in voice-onset time in speakers with cochlear implants. *The Journal of the Acoustical Society of America, 96,* 56–64.

Lipski, S.C., Grice, M., & Meister, I.G. (2008). Auditory perception influences speech motor learning. *Proceedings of the 8th International Seminar on Speech Production*, Strasbourg, France, pp. 405-408.

Matthies, M.L., Svirsky, M., & Perkell, J.S. (1994). A preliminary study of the effects of cochlear implants on the production of sibilants. *The Journal of the Acoustical Society of America, 96,* 1367–1373.

Ménard, L., Polak, M., Denny, M., Burton, E., Lane, H., Matthies, M.L., et al. (2007). Interactions of speaking condition and auditory feedback on vowel production in postlingually deaf adults with cochlear implants. *The Journal of the Acoustical Society of America, 121,* 3790–3801.

Munhall, K.G., MacDonald, E.N., Byrne, S.K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America, 125,* 384–390.

Nasir, S.M., & Ostry, D.J. (2008). Speech motor learning in profoundly deaf adults. *Nature Neuroscience, 11,* 1217–1222.

Natke, U., & Kalveram, K.T. (2001). Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *Journal of Speech, Language, and Hearing Research: JSLHR, 44,* 577–584.

Ohala, J.J. (1994). Acoustic study of clear speech: a test of the contrastive hypothesis. *Proceedings of the International Symposium on Prosody, September 18, 1994,* Yokohama. 75-89.

Perkell, J.S., Guenther, F.H., Matthies, M.L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America, 116,* 2338–2344.

Perkell, J.S., Lane, H., Denny, M., Matthies, M.L., Tiede, M., Zandiour, M., et al. (2007). Time course of speech changes in response to unanticipated short-term changes in hearing state. *The Journal of the Acoustical Society of America, 121,* 505–518.

Peschke, C., Ziegler, W., Kappes, J., & Baumgaertner, A. (2009). Auditory-motor integration during fast repetition: The neural correlates of shadowing. *NeuroImage, 47,* 392–402.

Schwartz, M.F. (1972). Bilabial closure durations for /p/, /b/, and /m/ in voiced and whispered vowel environments. *The Journal of the Acoustical Society of America, 51,* 2025–2029.

Shiller, D.M., Sato, M., Gracco, V.L., & Baum, S.R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America, 125,* 1103–1113.

Svirsky, M.A., Lane, H., Perkell, J.S., & Wozniak, J. (1992). Effects of short-term auditory deprivation on speech production in adult cochlear implant users. *The Journal of the Acoustical Society of America, 92,* 1284–1300.

Tartter, V.C. (1989). What's in a whisper? *The Journal of the Acoustical Society of America, 86,* 1678–1683.

Tourville, J.A., Reilly, K.J., & Guenther, F.H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage, 39,* 1429–1443.

Tremblay, S., Shiller, D.M., & Ostry, D.J. (2003). Somatosensory basis of speech production. *Nature, 423,* 866–869.

Villacorta, V.M., Perkell, J.S., & Guenther, F.H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America, 122,* 2306–2319.

Waldstein, R.S. (1990). Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. *The Journal of the Acoustical Society of America, 88,* 2099–2114.

Wang, Y., Jongman, A., & Sereno, J.A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America, 113,* 1033–1043.

Xu, Y., Larson, C.R., Bauer, J.J., & Hain, T.C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America, 116,* 1168–1178.

Ylinen, S., Shestakova, A., Alku, P., & Huotilainen, M. (2005). The perception of phonological quantity based on durational cues by native speakers, second-language users and nonspeakers of Finnish. *Language and Speech, 48,* 313–338.