Auditory–motor integration during fast repetition: The neuronal correlates of shadowing

C. Peschke a, W. Ziegler b, J. Kappes b, A. Baumgaertner c,d,⁎

a Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Germany
b Department of Neuropsychology, City Hospital Bogenhausen, Germany
c Department of Neurology, Schleswig-Holstein University Hospital, Germany
d Neuroimage Nord, University Medical Center Hamburg-Eppendorf, Germany

A R T I C L E   I N F O

Article history:
Received 22 October 2008
Revised 30 January 2009
Accepted 12 March 2009
Available online 1 April 2009

A B S T R A C T

This fMRI study examined which structures of a proposed dorsal stream system are involved in the auditory–motor integration during fast overt repetition. We used a shadowing task which requires immediate repetition of an auditory–verbal input and is supposed to elicit unconscious imitation effects of phonologically irrelevant speech parameters. Subjects’ responses were recorded in the scanner. To examine automated auditory–motor mapping of speech gestures of others onto one’s own speech production system we contrasted the shadowing of pseudowords produced by multiple speakers (men, women, and children) with the shadowing of pseudowords produced by a single speaker. Furthermore, we asked whether behavioral variables predicted changes in functional activation during shadowing.

Shadowing multiple speakers compared to a single speaker elicited increased bilateral activation predominantly in the superior temporal sulci. These regions may mediate acoustic–phonetic speaker normalization in preparation of a translation of perceptual into motor information. Additional activation in Broca’s area and the thalamus may reflect motor effects of the adaptation to multiple speaker models. Itemwise correlational analyses of response latencies with BOLD signal changes indicated that longer latencies were associated with increased activation in the left parietal operculum, suggesting that this area plays a central role in the actual transfer of auditory–verbal information to speech motor representations. A multiple regression of behavioral with imaging data showed activation in a right inferior parietal area near the temporo-parietal boundary which correlated positively with the degree of speech rate imitation and negatively with response latency. This activation may be attributable to attentional and/or paralinguistic processes.

© 2009 Elsevier Inc. All rights reserved.

Introduction

The ultimate target of any speech motor activity is to generate sound sequences which are perceived as intelligible and acceptable linguistic utterances. The production of linguistically meaningful sounds involves a number of complex physical transformations elicited by appropriate vocal tract movements. Implicit knowledge about how articulation creates sound is acquired during speech acquisition in childhood (e.g. Kuhl, 2000), or, in adulthood, when a language with an unfamiliar phonological inventory is learned. The learning of these complex movement-to-sound mappings presumes the existence of a close link between perception and production. The processes of speech perception (e.g. Cutler and Clifton, 1999) and production of speech (e.g. Levelt, 1989) have been investigated extensively. In contrast, relatively few models have specifically addressed the cross-talk of these two functions in auditory–motor mapping.

Among the theoretical accounts, the ‘Motor Theory of Speech Perception’ (Liberman and Mattingly, 1985; based on Liberman et al., 1967) constitutes a first approach towards describing the close connection between speech perception and production. The proponents of this theory postulated that in speech perception speech is perceived as phonetic gestures, not as acoustic features. The ‘phonetic module’ is a structure which is assumed to be shared by the speech perception and production systems, suggesting that the two systems depend on a common code and may have co-evolved (Fowler and Galantucci, 2005). A more recent model addressing an auditory–motor link was advanced by Hickok and Poeppel (2000, 2004; see also Warren et al., 2005). Their functional–anatomical model postulates the existence of a dorsal stream system subserving the auditory–motor integration of speech. The dorsal stream is assumed to connect superior temporal with a posterior perisylvian area at the temporo-parietal boundary (‘area Spt’, for ‘Sylvian–parietal–temporal’) and posterior inferior frontal areas, and is thought to be involved in the
mapping of acoustic speech signals onto articulatory representations. It may be particularly engaged in tasks requiring explicit manipulations of phonological representations, including the segmentation of linguistic units and the retrieval of sublexical elements.

Clinical evidence consistent with a role of a dorsal stream in the auditory-to-motor mapping in speech production can be found in patients with conduction aphasia, who, following damage including left posterior superior temporal gyrus (STG) and perisylvian temporo-parietal cortex, often show fair comprehension and fluent speech production capability while repetition typically is prominently disrupted (Bartha and Benke, 2003; Caramazza et al., 1981; Damasio and Damasio, 1980). Thus, structures associated with the dorsal stream appear to be central to the capability to repeat, a function which requires auditory-to-motor mapping.

Additional behavioral evidence for a direct link between auditory and motor systems is provided by shadowing experiments which require immediate repetition of an auditory–verbal stimulus. In a shadowing condition, participants tend to automatically and consciously imitate phonologically irrelevant acoustic details of the model (Goldinger, 1998). For example, subjects have been found to imitate fundamental frequency when shadowing a short text (Bailly, 2003). In other studies, subjects imitated phonologically irrelevant variations of voice-onset-time (VOT) without an explicit instruction to do so (Shockey et al., 2004; Fowler et al., 2003).

There is also structural evidence supporting a temporal–parietal–frontal link as the central pathway for the auditory–motor integration of speech. Between childhood and adolescence, structural maturation of the white matter pathway connecting temporal and fronto areas (arcuate fasciculus) is more pronounced in the left as compared to the right hemisphere (Paus et al., 1999). This may lead to an increased interaction between temporal and frontal language areas and facilitate a fast bidirectional cross-talk between auditory and motor regions. In adulthood, the dominant left hemisphere shows stronger connectivity than the right hemisphere between posterior superior temporal and ventrolateral prefrontal areas via the arcuate fasciculus (Parker et al., 2005). Catani et al. (2005) used diffusion tensor magnetic resonance imaging to examine the connection between posterior superior temporal regions and the posterior ventrolateral frontal lobe. The authors reported two parallel pathways between these regions: the “classical” arcuate fasciculus, which provides a direct link between temporal and frontal regions, and a second, indirect pathway running from temporal regions to the inferior parietal cortex and from the parietal region to frontal areas. The authors suggested that the function of the direct route is the fast and automated preparation of a motor copy of the perceived speech input, whereas the indirect pathway is used when an intervening stage, such as phonologic transcoding, lies between auditory input and articulatory output. Two recent studies using diffusion MRI tractography demonstrated the relevance of the dorsal language pathway to phonologic processing indirectly and directly, respectively (Glasser and Rilling, 2008; Saur et al., 2008).

In addition, several recent functional neuroimaging studies on speech processing provide evidence that auditory and motor systems are closely intertwined. Passive listening to stories without any overt motor response requirement, in addition to the expected temporal activation, co-activates motor areas which could have generated the speech sounds (Skipper et al., 2005). Similarly, listening to monosyllables activates ventral premotor areas which are also active when producing the same syllables (Wilson et al., 2004). These results are consistent with the assumption that the motor system is automatically engaged when an acoustic speech signal is encoded.

Despite a body of evidence for a strong interconnection between the perception and production of speech, there are still questions concerning the role of a dorsal stream (i.e., a temporal–parietal–frontal circuit) in the auditory–motor integration of speech which remain unresolved. There are now several fMRI studies which involve translating from perception to (covert or overt) production using verbal (Buchsbaum et al., 2001, 2005) and/or tonal stimuli (Pa and Hickok, 2008; Hickok et al., 2003). The tasks in all of these studies (i.e., perceiving and silently rehearsing the stimuli over periods of several seconds) explicitly draw on verbal working memory resources to some degree. It is largely unknown, however, whether the dorsal stream is also involved in the immediate and direct transformation of acoustic input into motor output. A first indication was provided by an fMRI study which showed activation of the parietal aspect of the dorsal stream (left inferior parietal lobule) during covert and overt repetition of long vs. short words (Shuster and Lemieux, 2005). A participation of dorsal stream structures in the implicit processing of word length is also suggested by another study using a covert picture naming task with words which varied in the number of syllables. This study reported the greatest effect of word length in the left posterior Sylvian cortex (Okada et al., 2003).

Thus, the aim of the present study was to examine whether structures associated with the dorsal stream are engaged in immediate auditory–motor integration. We used a shadowing paradigm in order to induce articulatory reproduction of spoken input which, compared to conventional repetition tasks, is relatively free from working memory requirements. Shadowing requires subjects to immediately repeat a verbal auditory input, without any explicit phonological processing requirement. In a reaction time task, subjects respond faster to vowel strings when the response is made by shadowing than when it is made by pressing a button (Porter and Lukber, 1980), suggesting that shadowing involves a direct transfer from speech input to oral articulators (Goldinger, 1998). Experiments based on the shadowing paradigm have shown that speakers, in their shadowing responses, tend to subconsciously echo phonologically irrelevant acoustic details (speaking rate, pitch, intonation, voice-onset time) of the model. The process of shadowing encompasses perceptual as well as production aspects of speech processing, and involves the automated transfer of speech gestures of others, via the perceived acoustic signal, into one’s own speech production system. Specifically, extraction of gestural information from the acoustic signal, in the understanding of various authors, helps to compensate for the acoustic variation between talkers (Studdert-Kennedy and Goldstein, 2003; Liberman and Whalen, 2000; Goldinger, 1998; see also Johnson, 2005). This information is then transferred to the speaker’s own speech production system.

As stimulus materials, we used pseudowords to discourage semantic processing of the stimuli. The items were either spoken by several speakers varying in gender and age (hereafter called ‘multiple speaker condition’), or by a single speaker (hereafter called ‘single speaker condition’). Auditory–motor mapping demands should be higher for the shadowing of utterances of multiple speakers than for the shadowing of a single speaker, because of the higher variability across items in acoustic parameters, such as word duration (as a correlate of speaking rate) or fundamental frequency (F0).

To examine neuronal activation we used event-related fMRI. An event-related design is more suitable for experiments which involve movements of the participants than a block design (Haller et al., 2005; Birn et al., 1999, 2004; Preibisch et al., 2003; Dogil et al., 2002). We decided to test overt speech production mainly because we wanted to record the verbal responses. These recordings enabled us to monitor and analyze subjects’ productions with respect to accuracy and response latency, and allowed for a qualitative analysis of potential imitative behavior with respect to fundamental frequency and speech rate.

Even though sparse temporal sampling is recommended for overt speech production to avoid movement artifacts (Gracco et al., 2005; Tanaka et al., 2000; Eden et al., 1999; Elliott et al., 1999), we used continuous sampling in order to keep the duration of the scanning sessions within a tolerable time frame. Furthermore, speech-related hemodynamic signal change is delayed and reaches its maximum 5–6 s after event onset, while movement-induced signal change occurs...
We hypothesized that the shadowing paradigm activates a shadowing network. To check this, we calculated the behavioral variables (degree of imitation, response latency, and accuracy) to determine their effect on the overall functional activity. We expected that the behavioral variables are associated with the auditory-to-motor transfer process per se, and whether the behavioral variables affect executive aspects of speaking, we expected the behavioral variables to be correlated with signal changes in frontal areas, most likely in the left hemisphere.

### Materials and methods

#### Participants

20 healthy right-handed subjects participated in the study. All subjects were native speakers of German with no history of serious medical, neurological or psychiatric illness, or hearing loss (mean age 26.1 years, range 20–36, ten females). Hand preference was tested with the 10-item version of the Edinburgh Handedness Inventory (Oldfield, 1971). Subjects had an average laterality quotient of 0.87 (range 0.4–1.0).

This study was performed according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the University Hospital Hamburg-Eppendorf. All subjects gave written informed consent prior to the experiment.

### Stimuli

The speech stimuli consisted of 96 bisyllabic pseudowords (plus 16 practice items) with the prefixed German plural article “die” (e.g. /di:ke:pon/). The article was included as a fixed element to facilitate the beginning of the shadowing as each trial began with the same word. The pseudowords had a trochaic stress pattern and a schwa syllable in the unstressed position, thus resembling plural forms of German nouns. Stimuli were controlled for phonological neighborhood to avoid the possibility that changing the first phoneme would produce a real word.

Speech stimuli were recorded in a sound studio. For the multiple speaker condition, all of the 96 pseudowords were read twice by eight speakers of different gender and age (mean age 29.8 years, range 9–77, four females). The recordings were examined for correct pronunciation, intonation, and comprehensibility by the first author (CP). For every speaker, the more intelligible token of each word was selected. Each stimulus was edited using the sound editing software Audacity (http://audacity.sourceforge.net/). Twelve items from each of the eight speakers were pseudo-randomly selected from the material, in a way that none of the pseudowords appeared twice. Sixteen of the recorded items that were not used as experimental items were chosen to serve as practice items for this condition.

For the single speaker condition the stimuli were spoken by an experienced male speaker (age 44 years). He also read the words twice, and one of the two tokens was selected according to the above quality criteria. Additionally, 16 practice items were recorded by the speaker. These materials were adopted from a parallel (behavioral) study, for which it was necessary to manipulate fundamental frequency ($f_0$). That is, $f_0$ was set to 105 Hz in half of the items and to 120 Hz in the other half, with the use of the phonetic analysis software Praat (http://www.praat.org). This variation was significantly lower than the $f_0$ variability in the multiple speaker condition and therefore less, if at all, consciously noticeable to the subjects. Word duration was measured for all experimental items, and median $f_0$ was computed for the items in the multiple speaker condition (see Table 1 for demographic speaker data and a summary of stimulus characteristics). Intensity of stimuli was set to 80dB throughout.

#### Tasks and procedures

The subjects’ task during scanning was to shadow the stimulus upon auditory presentation. Participants were instructed to overtly repeat the target words. They were told to start speaking as quickly as possible, even when the utterance they heard had not yet been terminated, and they were also asked to be maximally accurate. To prepare subjects for the presentation of an auditory stimulus, each trial started with a visual cue (symbol of a loudspeaker) which was presented for a period of 1500 ms, 500 ms before the onset of each auditory stimulus. During the rest of the trial the screen remained dark. The SOA was jittered in an interval of between 4 and 9 s (see Fig. 1).

---

#### Table 1

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Age</th>
<th>Sex</th>
<th>Mean duration (in ms)</th>
<th>Range duration (in ms)</th>
<th>Mean $f_0$ (in Hz)</th>
<th>Range $f_0$ (in Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>9</td>
<td>f</td>
<td>1093</td>
<td>242</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>9</td>
<td>m</td>
<td>792</td>
<td>222</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>15</td>
<td>f</td>
<td>941</td>
<td>228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>16</td>
<td>m</td>
<td>669</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>26</td>
<td>m</td>
<td>945</td>
<td>143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>26</td>
<td>m</td>
<td>1058</td>
<td>91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>60</td>
<td>f</td>
<td>1036</td>
<td>126</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>77</td>
<td>f</td>
<td>912</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (multiple speaker)</td>
<td>–</td>
<td>–</td>
<td>931</td>
<td>167</td>
<td>87-261</td>
<td></td>
</tr>
<tr>
<td>Control (single speaker)</td>
<td>44</td>
<td>m</td>
<td>962</td>
<td>105</td>
<td>105, 120</td>
<td></td>
</tr>
</tbody>
</table>

* Marginally significant difference in word duration between multiple and single speaker condition (two-sample t-test: $t = -3.95$; DF = 129.6; $p = 0.0528$ (for unequal variances)).

---

**Fig. 1.** Example of an experimental trial (SOA 4000–9000 ms).
Subjects were first familiarized with the task before scanning, using the 16 practice items of each condition. None of the subjects required more than one run to become acquainted with the task.

Stimulus presentation and response recording

The visual cue was presented in a grey hue centrally on a dark background. It was projected onto a screen located at the top of the head coil and was viewed by the subjects via a mirror. The auditory stimuli were presented via an MR-compatible electrodynamic headphone with a built-in dual-channel microphone (MR Confon GmbH, Magdeburg, Germany, http://www.mr-confon.de) for the combined presentation and recording of speech. Stimuli were recorded with the sound recording software PhonOr implemented in the Confon system which automatically preprocesses the dual-channel recordings by reducing the whole frequency spectrum of the scanner noise by 20dB in relation to the speech signal. Vocal response recordings had to be customized in order to synchronize them with the recordings of scanner pulses and stimulus onsets. To this end, scanner pulses were recorded in two different channels: once they were routed to the standard presentation software which also recorded stimulus onsets, and once they were fed into the vocal response recording channel as an additional signal. This procedure ensured perfect synchronization of the general presentation parameters with the recording of vocal responses. Vocal responses were recorded by Confon running on a Samsung P30 XTM 1400 laptop situated outside of the scanner room, and saved as wav-files. The volume for the auditory presentation was set to a comfortable level individually for each subject in a preceding test scan. The task sequence was controlled by a PC running “Presentation” software (Neurobehavioral Systems, www.neurobehavioral.systems).

Experimental design

There were six runs, three for each condition (the multiple and the single speaker condition, respectively), with a length of about 4 min each. The sequence of conditions was pseudo-randomized, with the restriction that maximally two runs of one condition could occur in a row and that each run appeared equally often in all positions across all subjects. Each run began with five introductory trials which were followed, without further instruction, by 32 experimental trials. Introductory trials served as warm-up trials. The order of the different speakers in the multiple speaker condition was pseudo-randomized, in a way that the speaker model changed from trial to trial. The trial sequence was fixed within runs for all subjects. The experiment took about 30 min to complete.

Data acquisition and image preprocessing

Imaging was conducted using a 3 T Siemens magnetic resonance imaging system, acquiring around 1320 volumes in total. For each subject, functional T2-weighted gradient-echo echo-planar images were obtained from 20 axial slices (3 mm thickness, no gap, TR 1170 ms, TE 25 ms, flip angle 70°, field of view 210 × 210 mm², matrix 64 × 64) oriented parallel to the anterior and posterior commissure. Furthermore, a 3D high-resolution (1 × 1 × 1 mm voxel size) T1-weighted structural MRI (MPRAGE) was acquired for each subject. Head movement was restrained within the head coil by circumaural headphones fitting tightly into the head coil and by tight foam padding. Additionally, subjects were instructed to minimize head movement. The microphone was placed centrally to the mouth as closely as possible without touching the lips.

The processing and analysis of imaging data were performed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm/). Preprocessing included slice timing, realignment to the first volume, coregistration between the individual structural T1-weighted image and the EPI-images, and segmentation of the structural images. The resulting estimated spatial normalization parameters were then applied to the series of functional images, which where subsequently resampled to a voxel size of 3 × 3 × 3 mm and finally smoothed using a 10 mm full-width at half-maximum isotropic Gaussian kernel.

Statistical analyses

Behavioral data

The first step in vocal response analysis was to determine the accuracy of responses, using the sound editing software ‘Audacity’. Two raters (CP and a second rater who was not familiar with the goal of the study) decided independently whether responses were qualitatively acceptable (against the backdrop of residual scanner noise or interfering scanner pulses; 93 of 3840 items not analyzable, or 2.4%). Next, analyzable verbal responses were evaluated independently by the two raters for accuracy, using the criterion that only words without any audible phonetic or phonological distortions were rated as accurate. Mean concordance between raters was 0.69 (Cohen’s Kappa), which constitutes substantial interrater agreement by convention (Landis and Koch, 1977). Whenever the raters differed (i.e. in 9.7% of the analyzable responses), the respective tokens were rated by a third person. The inclusion criterion for subjects was set to a minimum of 75% correct responses.

Next, we performed an error analysis. The number of phoneme substitutions was examined separately for each item. When the target item had been produced with a phonemic substitution, the number of substituted phonetic features of the respective phoneme was counted and described in terms of type of error. Word duration (in ms) and median F0 (in Hz) were then measured for each item with PRAAT. In order to measure vocal response duration, the on- and offset of each individual response was determined by visual and auditory inspection of the speech wave by a trained research assistant. The obtained measurements were randomly checked for accuracy by one of the investigators (CP). The duration of individual responses was then calculated by subtracting the onset from the offset of the response.

Initially, median F0 was analyzed with PRAAT in a fixed frequency range (75–200 Hz for men, 140–300 Hz for women). Because of outliers among the obtained values, it became apparent that the results were affected by remaining scanner noise. To limit the interference of residual concomitant scanner noise in the microphone recordings, the spectral analysis window was narrowed by calculating individual standard deviations of median F0 for each subject and then re-analyzing the data with individual F0-analysis windows of three standard deviations above and below the estimated mean.

To measure the response latency of each item, we calculated the time interval from stimulus onset to vocal response onset. To analyze effects of imitation individually for each participant, response durations and F0 values of responses were correlated with the respective response durations and F0 values of the model stimuli. For analyzing the imitation of F0 in the single speaker condition it was not possible to use correlation coefficients, because model stimuli varied only between two F0-values (120 and 105 Hz). Potential F0-differences between subjects’ responses to the 120 Hz and 105 Hz stimuli were tested with a two-sample t-test. In a last step, all of the behavioral parameters were correlated with each other in order to identify possible dependencies. Only those parameters which were statistically independent were entered into a multiple regression analysis, using the functional data as the outcome variable (see next section).

Functional imaging data

All responses were included in the analyses. Statistical analyses of the functional imaging data were performed in two steps. In a first-
level analysis, a statistical model was computed for each subject. First, the introductory warm-up trials were modeled as a separate regressor of no interest. Then, the primary experimental variable of interest, namely the process of ‘shadowing’, was defined as a single event beginning with stimulus onset, and was entered as the second regressor into the model. This regressor was convolved with a canonical hemodynamic response function (HRF) as implemented in SPM5. In addition, the six realignment parameters (three each for translation and rotation) were inserted as covariates of no interest. Because of possible movement artifacts as a result of overt speech production, we included an additional user-defined regressor of no interest into the model as an on/off function, which was independent of the HRF and indicated the respective speech on- and offset, using TR resolution. This specific ‘speech movement’ regressor labeled those image volumes which were entirely or partly affected by speech production. Voxel-wise regression coefficients for the variable of interest were estimated using the least-squares method within SPM5, and statistical parametric maps of the t statistic (SPM(t)) were generated. Data were high-pass filtered at a cut-off period of 128s and corrected for serial correlation (AR(1) model).

We then computed the main effect of shadowing, both across multiple and single speaker conditions as well as for each of the two conditions separately. Next, we computed the differential contrast of the multiple vs. the single speaker condition. By nature of the task, shadowing encompasses perception and production processes as well as a transfer of auditory information into speech motor representations. Thus, the contrast of the multiple with the single speaker condition may show differences at any or all of these processing levels, even in auditory perceptual areas. In order to more specifically examine the process of transferring speech gestures produced by others into one’s own speech production system, we computed the correlation between item-by-item response latencies and individual BOLD signal changes due to shadowing across speaker conditions. This within-subject analysis is based on the assumption that longer response latencies among the total of a subject’s responses reflect additional computational demands on the transfer process. Vice versa, shorter response latencies are assumed to reflect higher transfer efficiency. To examine the transfer process, we added a parametric regressor containing the individual response latencies for each item. A contrast image representing the effect of transfer efficiency was then computed at the first level for all subjects.

In the second-level analyses, the contrast images of the first-level analyses for each subject were used to perform the group analyses. To identify areas of common activation due to shadowing of pseudo-words across speaker conditions, the contrast images of the main effects of multiple and single speaker conditions were entered into a within-subject ANOVA model including a correction for non-sphericity. These contrast images were submitted to a conjunction analysis under the conservative conjunction null hypothesis, using a threshold of p<0.05 (FWE corrected). Next, areas sensitive to speaker condition were identified by performing a one-sample t-test on the individual contrast images comparing the multiple with the single speaker condition. Likewise, we tested for neuronal correlates of transfer efficiency by computing a one-sample t-test on the images of the correlational analyses between item-by-item response latencies and individual BOLD signal change due to shadowing. Finally, in order to address the question of a correlation of the behavioral variables with overall functional activation due to shadowing across subjects, we conducted a multiple regression analysis with the predictor variables (a) imitation of F0, expressed as one correlation coefficient for each subject, (b) imitation of word duration, expressed by an individual correlation coefficient for each subject, (c) individual mean latency of responses (in ms) and (d) individual accuracy (in %), using the individual main effect of ‘shadowing’ as outcome variable.

In our experimental design the multiple speaker condition contained male, female and children’s voices whereas the single speaker condition contained only a male voice. This raises the question of a potential effect of the gender of the speaker in processing the vocal stimuli. Previously, female voices have been shown to elicit higher activation than male voice in posterior superior temporal lobe due to differences in frequency characteristics (Lattner et al., 2005). To test for potential (speaker) gender effects on the activation during shadowing, we added a parametric ‘gender’ regressor to our first-level model which assigned one of three possible values (male voice = 1, female voice = 2, child’s voice = 3) to each stimulus event. To arrive at independent classifications of gender, three raters independently classified the stimuli according to whether they were spoken by a man, a woman or a child. This rating resulted in perfect agreement among the raters. Then, the main effect of ‘gender’ was computed at the first level for all subjects. The resulting contrast images were submitted to a one-sample t-test at the second level and then used as an exclusive mask (at a threshold of p<0.05 uncorrected) for the differential contrast between multiple and single speaker conditions.

The statistical threshold was set to p<0.05 (FWE corrected) for the conjunction, p<0.05 (FDR corrected) for the comparison of the multiple compared to the single speaker condition, p<0.001 (uncorrected) for the within-subject analysis of the effect of transfer efficiency, and p<0.001 (uncorrected) for the multiple regression. All of the above analyses were computed for the entire volume of 20 slices. Anatomical localization of activation peaks was done by using the SPM Anatomy Toolbox Version 1.3b (Eickhoff et al., 2005). Spatial references are given in Montreal Neurological Institute (MNI) coordinates.

Results

Behavioral data

Accuracy

Three of the 23 subjects originally included failed to reach the criterion of at least 75% correct responses and were excluded from further analysis. One subject reached an accuracy rate of 74.4%. Because this was close to the cut-off value, we decided to include the subject. Generally, included participants reached a high level of accuracy in the shadowing task, with a group mean of 86.1% correct responses (range 74.4–93.8%, SD: 5.5%). For the multiple and single speaker condition, the group mean reached 85.1% and 87.0% correct responses, respectively. A paired t-test showed no significant difference in accuracy between the two conditions (t = 1.31, p>0.05).

Among the inaccurate responses, 82.8% contained only one erroneous phoneme. In 77.3% of the responses in which a phoneme had been substituted, only one feature had been changed in the response. The types of feature changes included changes of place (58.7%) and manner of articulation (16.2%) and of voicing (13.5%). Elisions (10.3%) and additions of a phoneme (1.1%) were also observed.

Latency

Average response latency, computed from stimulus onset, was 485 ms (range 273–642 ms; SD: 99 ms). Response latency was higher in the multiple speaker condition (502 ms) than in the single speaker condition (470 ms). This difference between conditions was significant (paired t-test: t = −2.59, p<0.05).

Response duration

The mean duration of responses, computed across both conditions, was 883 ms (range 759–1038 ms; SD: 76 ms). There was a significant difference between the two conditions (paired t-test: t = 2.69, p<0.05). Durations were shorter in the multiple speaker condition (874 ms) than in the single speaker condition (892 ms). This mirrors the difference in stimulus duration between the two conditions (see
Table 1. A group analysis showed significantly more variance in response duration in the multiple speaker condition than in the single speaker condition (Test for equality of variances: $F(1918, 1918) = 1.19$; $p < 0.001$).

Response-$F_0$

Average $F_0$ for the multiple speaker condition was 161 Hz (average $F_0$ for male subjects: 113 Hz; average $F_0$ for female subjects: 208 Hz). $F_0$ in the single speaker condition also was 161 Hz (average $F_0$ for male subjects: 114 Hz; average $F_0$ for female subjects: 208 Hz). The difference between conditions, across subjects, was not significant (paired $t$-test: $t = 0.72$, $p > 0.05$). A test for equality of variances for the group similarly showed no difference in variances between the two conditions ($F(1918, 1918) = 1.01$; $p > 0.05$). Four of the 20 subjects showed significantly higher $F_0$ variance in the multiple speaker condition than in the single speaker condition. In addition, two of the subjects showed a significantly higher variance in the single speaker compared to the multiple speaker condition. Correlations between experimental stimuli and subjects’ responses

Considering the two conditions together, all of the subjects showed significant correlations between the duration of the stimuli and the duration of their vocal responses ($r$ between 0.16 and 0.53; all $p < 0.05$). For the multiple speaker condition, the number of significant correlations was higher than for the single speaker condition. Whereas only 12 of the 20 subjects showed significant imitation of duration in the single speaker condition (correlation coefficients between 0.20 and 0.39), all of the subjects showed significant imitation in the multiple speaker condition ($r$ between 0.25 and 0.67). To examine whether these correlations actually reflect an imitation of speech rate rather than simply being an expression of variation in word length per se, we performed an additional analysis. We used the circumstance that every subject repeated each item twice, once in the single and once in the multiple speaker condition. Word duration for these two instances of each item was different, due to the fact that each of the two speakers (the male speaker in the single speaker condition, and one of the 8 speakers in the multiple speaker condition) used her or his idiosyncratic rate of speech. Itemwise, we then computed the difference in word durations (in ms) between the two speaker models, on one hand, as well as the difference in word durations for the two corresponding responses given by each subject, on the other. These difference scores were then correlated with each other at the single subject level. For all of the subjects, this analysis yielded a significant correlation (Pearson’s product moment coefficients between 0.27 and 0.67 [mean 0.44], all $p < 0.01$). This suggests that significant correlations between durations of model and subject utterances are due to imitation of speech rate rather than to affordances of [structural] word length.

With respect to $F_0$, eight subjects showed a significant positive correlation between response $F_0$ and stimulus $F_0$ across conditions (significant coefficients between 0.15 and 0.35). For one subject, there was a significant negative correlation between stimulus $F_0$ and $F_0$ of response ($r = -0.38$; $p < 0.001$). When considering the multiple speaker condition separately, the number of subjects with significant $F_0$ correlations increased to 13 of 20 ($r$ between 0.20 and 0.45). For seven of the 20 subjects, a comparison of the two $F_0$ levels in the single speaker condition showed a significantly higher $F_0$ for the shadowing of 120 Hz vs. 105 Hz stimuli.

Correlations between behavioral variables

None of the behavioral parameters (accuracy, response latency, imitation of duration and imitation of $F_0$) correlated significantly with any other parameter (all $p > 0.05$).

Functional imaging data

All of the imaging data for the 20 subjects were valid and could be used for functional data analyses. First, we performed a conjunction analysis, examining conjoint BOLD signal changes for multiple and single speaker conditions, compared to rest. This analysis revealed an extensive bilateral pattern of perisylvian activation with maximal changes in posterior and middle aspects of the superior temporal gyrus (STG), along the postcentral gyrus and the inferior frontal gyrus (IFG). There was also activation in the thalamus and the basal ganglia bilaterally. Additionally, there was activation in primary visual areas (Fig. 2). The maximum of functional activation was located in the right STG near Heschl's gyrus (MNI coordinates [69 $-64$ 10]).

We then tested the difference in functional activation between multiple and single speaker conditions. Shadowing of pseudowords spoken by multiple speakers, compared to the single speaker condition, resulted in an increased BOLD signal change along the superior temporal sulcus (STS) bilaterally (Table 2 and Fig. 3) extending into the superior and middle temporal gyri. In the left hemisphere, the temporal activation extended to the most posterior aspect of the STG near the temporal–parietal boundary (‘area Spt’). Furthermore, there was activation in Broca’s area (with a maximal change in pars triangularis bordering on pars opercularis) and in the right thalamus.

Table 2

<table>
<thead>
<tr>
<th>Region (Threshold $p &lt; 0.05$, FDR corrected for the whole brain)</th>
<th>Side</th>
<th>Voxels in cluster</th>
<th>MNI coordinates</th>
<th>$T$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superior temporal gyrus</td>
<td>R</td>
<td>316</td>
<td>63 $-6$ $-9$</td>
<td>5.99</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>L</td>
<td>174</td>
<td>$-60$ $-42$ $9$</td>
<td>5.77</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>L</td>
<td>34</td>
<td>$-57$ $-12$ $-6$</td>
<td>4.10</td>
</tr>
<tr>
<td>Inferior frontal gyrus (pars triangularis bordering on pars opercularis)</td>
<td>L</td>
<td>15</td>
<td>$-54$ $21$ $3$</td>
<td>4.00</td>
</tr>
<tr>
<td>Thalamus</td>
<td>R</td>
<td>29</td>
<td>$3$ $-12$ $6$</td>
<td>3.94</td>
</tr>
</tbody>
</table>
The reverse contrast (single speaker compared to multiple speaker condition) did not yield any significant results.

We then examined potential influences of the gender of the speaker models on the activation in the multiple speaker condition. At a liberal threshold of $p < 0.05$ (uncorrected) there were gender-related activation changes in posterior STG bilaterally (MNI coordinates $[-60 - 27 12]$ and $[66 - 24 9]$), posteriorly and distally to Heschl’s gyrus (Morosan et al., 2001, http://www.bic.mni.mcgill.ca/cytoarchitectonics). Activation increases were strongest for the children’s voices, in a medium range for female voices, and lowest for male voices. This result was used to exclusively mask the contrast of the multiple as compared to the single speaker condition. The exclusive masking had little effect on the pattern of activation observed without the mask. Fig. 3 depicts the activation unique to the effect of speaker condition, with the gender effect factored out.

Next, the within-subjects analysis of effect of transfer efficiency revealed a single cluster of activation in the left parietal operculum (MNI coordinates $[-39 - 27 18]$; cluster level extent $> 10$; Fig. 4). In this region, functional activation increased with increasing response latency. There were no areas which showed a significant negative correlation between BOLD signal change and response latency.

Finally, we examined whether any or all of the behavioral variables (imitation of $F_0$ and imitation of word duration, accuracy and response latency) were correlated with overall functional activation related to shadowing (corrected for the entire volume) across subjects. The result of the multiple regression analysis indicated significant effects for the imitation of duration and for response latency. The more subjects imitated duration, the more they activated an inferior parietal area near the tempo-parietal boundary (MNI coordinates $[51 - 30 30]$). In addition, activation in nearly the same inferior parietal area (MNI coordinates $[54 - 30 30]$) increased with decreasing average response latency (Fig. 5).

Discussion

Recently a dorsal stream connecting superior temporal with inferior parietal and posterior inferior frontal areas has been suggested to subserve the cortical auditory–motor integration of speech (Hickok and Poeppel, 2000, 2004; Hickok et al., 2003). Previous studies have shown that the dorsal stream is involved in perceptual tasks requiring explicit manipulations of phonological representations by translating acoustic speech signals into articulatory representations (Gandour et al., 2003; Jacquemot et al., 2003; Burton et al., 2000). The aim of the present fMRI study was to investigate whether structures associated with the dorsal stream support auditory-to-motor mapping processes not including explicit phonological processing or verbal working memory requirements. Auditory–motor mapping was induced by a shadowing task in which subjects immediately and overtly repeated (i.e. ‘shadowed’) auditorily presented pseudowords. This type of task is assumed to induce a fast and highly automated transfer of perceived speech gestures of others into a speaker’s own articulations. Even though close shadowing, as a whole, may impose high attentional demands, it is considered free from explicit phonological processing, auditory–verbal short-term storing, or other downstream linguistic processing mechanisms. For the multiple as compared to the single speaker models, we predicted significantly higher neural activation in temporal and/or frontal regions, reflecting a higher demand on speaker model adaptation in the multiple speaker condition.

We recorded subjects’ overt productions during scanning for several reasons: first, to ensure that shadowing responses were actually achieved, despite the adverse conditions of the scanner environment; second, to examine subjects’ individual performance on the task; and third, to gauge the degree of imitation of phonologically irrelevant acoustic details contained in the speaker models. The acceptable degree of accuracy and the relatively short overall latency showed that the participants were able to perform the task well, despite the ambient scanner noise. A subset of the subjects (i.e., 8 out of 20) showed imitation of fundamental frequency of the model stimuli, while all of the subjects imitated their duration. These findings confirm previous behavioral observations of automatic imitation of phonologically irrelevant acoustic parameters. Without being instructed to imitate, all of the participants showed a high correlation of the duration of their own productions with the duration of the model utterances. That is, subjects apparently ‘copied’ the speaking rate of the acoustic stimuli. With respect to $F_0$ imitation, one third of the subjects showed a significant correlation with the speaker models. When only the multiple speaker condition was considered, the number of participants who showed $F_0$ imitation increased to more than half of the subjects. This was likely due to the higher variability in $F_0$ between the eight model speakers in the multiple speaker condition. There were no significant correlations between the behavioral parameters, indicating that subjects who were fast responders did not make more errors when repeating the stimuli.
Furthermore, this finding indicates that subjects who shadowed speech rate did not necessarily also shadow pitch, suggesting that these two parameters are dissociable in imitation behavior. Overall, the behavioral data show satisfactory task performance as well as the expected effect of imitation during shadowing.

As expected, the bilateral cortical activation related to shadowing was in the perisylvian region, encompassing posterior and middle aspects of the STG, the postcentral gyrus, and the inferior frontal and precentral gyrus, suggesting that the incoming auditory information was processed in superior temporal regions, phonologically encoded, and then conveyed to frontal regions. The overall activation maximum in the superior temporal gyrus bilaterally is consistent with functional imaging studies associating the superior temporal gyrus and sulcus with the perception of speech (Price et al., 2005; Davis and Johnsrude, 2003; Binder et al., 2000; Scott et al., 2000; for a review see Scott and Wise, 2004), while activation in the inferior frontal and precentral gyrus likely reflects the motor aspects of shadowing.

The strong postcentral activation was somewhat unexpected. However, a role of a somatosensory component in speech motor control was postulated by Guenther (2006; see also Guenther and Perkell, 2004). The author proposed two feedback control subsystems in speech production, i.e., auditory and somatosensory. In the auditory feedback subsystem, auditory information about one’s own speech production projects to auditory cortical areas in the STG which encode the incoming auditory information and compare it with the expected auditory patterns. Information is then sent back to motor cortex, to translate auditory off-target deviations into corrective motor commands. Similarly, within the tactile-sensory feedback system, somatosensory information derived from the speech movements is encoded via projections to somatosensory cortex, and adaptive corrections are made in the motor cortex. In fact, primary sensory cortex has been shown to be part of a speech motor network in tasks requiring overt articulation (Riecker et al., 2005; Dogil et al., 2002; Lotze et al., 2000). Articulatory movements without phonation, hence without external auditory feedback, have been shown to activate the postcentral gyrus. This does not occur during passive listening to the same stimuli (Pulvermüller et al., 2006). In the present study, auditory feedback was severely restricted by scanner noise and tight-fitting headphones. Additionally, the task required subjects to speak while listening to auditory stimuli. As a result, the participants were not able to use auditory feedback as a control system in a normal fashion. Instead, they may have relied more on the tactile-sensory feedback control system, to compensate for the impoverished auditory feedback. Alternatively, some subjects may have used a specific shadowing strategy to compensate for lacking auditory feedback by directly linking auditory input to somatosensory information before projecting the information to motor cortex.

The comparison of the multiple with the single speaker condition showed a higher change in activation for the shadowing of multiple speakers which was located along and around the STS bilaterally, in parts of Broca’s area, and in the right thalamus. This activation cannot be attributed to differences between conditions in stimulus intensity or duration, since stimulus intensity was equalized, and the loudness level was kept constant for all runs. Stimulus as well as response duration was shorter in the multiple than in the single speaker condition, but if at all, one would have expected this fact to lead to a decrease, not an increase, in activation. Furthermore, activation related to speaker condition (multiple vs. single) cannot be explained by the gender of the speaker models because this activation was exclusively masked with activation due to the effects of gender in processing the vocal stimuli in this study.

One explanation for the effect of the speaker condition may be that the multiple speaker condition imposed higher demands on speaker normalization processes than the single speaker condition. Stimuli in the multiple speaker condition were produced by speakers of different age and gender, each of whom had her/his own idiosyncratic articulation, speaking rate, and voice fundamental frequency. For instance, it is known that the same vowel has entirely different acoustic properties when it is spoken by a child as compared to an adult, or by a male as compared to a female person (Whiteside, 2001; Zahorian and Jagharghi, 1991). In the shadowing task, speakers were required to compute acoustic goals for their own articulations from these highly variable sources. In the single speaker condition, on the contrary, a shadower’s audio-motor adaptation processes had to cope with only a single speaker model, hence with always the same speaker-specific features and only little rate and pitch variation. The transformation of auditory input into speech motor output therefore remained largely invariant when participants shadowed the single-speaker model. Thus, higher activation in response to the multiple speaker models compared to the single speaker model may reflect a higher number of computations for the adaptation to a continuously changing speaker model as compared to the case in which the speaker model remained the same throughout a number of trials.

Due to the partial temporal overlap of listening and speaking in the shadowing paradigm, the present study is not able to dissociate...
perceptual from motor activations. Conceivably, however, STS activation in this comparison (i.e., multiple vs. single speaker conditions) may reflect the perceptual component of speech processing in the present study. This interpretation matches previous reports from functional imaging studies examining voice perception which shows higher activation along the bilateral STS for speech compared to non-speech stimuli (Belin et al., 2002; Vououmanos et al., 2001; Binder et al., 2000). Furthermore, the STS may be particularly sensitive to a change in voice identity. The anterior part of the right STS has been shown to respond stronger to different voices compared to a single voice (Belin and Zatorre, 2003). Thus, (anterior) aspects of the STS activation in the present study may be associated with voice perception, while posterior aspects may be associated with the recognition of speech. Monitoring for a specific word in a list of spoken words has been shown to result in a cluster of increased activation in superior temporal gyrus near the transverse temporal gyrus bilaterally (apart from activation in superior parietal areas) when the words were spoken by different speakers compared to a single speaker (Wong et al., 2004). Since subjects in their study specifically attended to phonetic information (and not to vocal information), the authors state that this difference in activation in middle/superior temporal regions (as well as in parietal regions) may be due to the computational demands of resolving acoustic–phonetic ambiguities associated with multiple speaker models. Similarly to Wong et al. (2004), we propose that the observed activation in and around the STS in the present study may represent the neural correlate of the extraction and normalization of speech gestures.

An alternative explanation for the higher level of activation in the multiple compared to the single speaker condition in the STS could be that perceptual adaptation to the constant voice occurred in the single speaker condition. Consecutive stimuli of the same type activate the same subpopulation of neurons and therefore cause a reduction in activation (Krekelberg et al., 2006). This adaptation effect should occur at a relatively early stage of perception. In addition, higher activation in the multiple speaker condition could reflect a novelty effect due to the trial-to-trial change in voices, which may have induced a stronger orienting response.

Interestingly, the left temporal activation in the multiple speaker condition extended into the most posterior aspects of superior temporal gyrus (pSTG). Several studies have shown that this area is activated by speech perception and by speech production tasks (e.g. Okada and Hickok, 2006; Hickok et al., 2003; Buchsbaum et al., 2001). In the present study, this activation borders on the tempo-parietal boundary near the Sylvian fissure, the area called Spt, which is considered to be part of the dorsal stream and to map auditory onto motor representations of speech (Hickok and Poeppel, 2004, 2000). Other authors consider this area to be equivalent to the posterior aspect of the planum temporale (Buchsbaum et al., 2005). This area has been proposed to be involved in optimizing the efficiency of auditory–motor mapping by comparing the incoming speech signal with stored phonological and articulatory templates derived from auditory experience (Warren et al., 2005). Thereby, auditory signals are transformed “into a form that constrains motor output” (Warren et al., 2005, p. 638). During fast repetition, as needed for shadowing, subjects may have had to rely on these efficient template-matching mechanisms in order to achieve fast and highly accurate performance. Furthermore, incoming auditory information was more variable in the multiple compared to the single speaker condition, perhaps imposing a higher number of computations on the template-matching algorithms.

In addition, the differential contrast between speaker conditions activated the left posterior IFG. Among other cognitive processes, this area has been argued to play an important role in articulatory processing (Bonilha et al., 2006; Hillis et al., 2004). In the present study, IFG activation may represent the motor component of the multiple speaker adaptation during shadowing. A rather straightforward interpretation could be that this activation reflected the fact that all subjects imitated the duration of the pseudowords produced by the speaker models, consequently their responses had more variable rates in the multiple as compared to the single speaker condition. Hence, the trial-to-trial change in speech rate in the multiple speaker condition required more computation related to the planning and execution of motor acts. Another explanation may be that the audio–motor translation process per se, i.e. the transformation of auditory input patterns into one’s own speech gestures, depends on left IFG activity. The more variable the auditory input, the higher the demands imposed not only on auditory normalization processes in temporal cortical areas, but also on the motor extension of these processes in frontal language areas. The thalamic activation, also observed in the present study, is congruent with previous findings implicating the thalamus as part of a neural network subserving speech motor control during overt speech production (Riecker et al., 2008; Bohland and Guenther, 2006).

As mentioned above, shadowing encompasses perception and production processes as well as the translation of auditory into speech motor information. Since the transfer process per se was the most interesting aspect of shadowing to us, we conducted an item-by-item correlation of the individual within-subject response latencies with the BOLD signal change due to shadowing across both speaker conditions. Our assumption was that longer response latencies reflect a higher demand on the transfer process. This correlation showed activation in the left parietal operculum. One question may be whether item-wise response latency may have been modulated by factors other than transfer efficiency, for example by perceptual difficulty. We cannot entirely dismiss the possibility that perceptual difficulties may have contributed to the increase in response latency. In order to explore this alternative explanation, we examined whether longer latencies were associated with a higher error rate (which one might expect if perceptual difficulty were causing the increase in latency). A paired test, contrasting average response latency for correct with average response latency for incorrect responses, demonstrated significantly higher latencies for those items that were repeated accurately ($t = -2.39; DF = 19; p = 0.0276$; mean latency for correct responses $= 488$ ms; mean latency for incorrect responses $= 469$ ms). If indeed perceptual difficulties were contributing to the increase in response latency, then these difficulties did not result in a higher number of errors. Furthermore, the fact that overall subjects took significantly longer to respond to the multiple models than to the single speaker model (while being equally accurate in both conditions) suggests that it may have been factors other than mere perceptual difficulty that contributed to the increase in BOLD signal in the left parietal operculum — for example increased auditory–motor transfer demand. In a study of cortical auditory regions, cell structure in the primary auditory cortex and the parietal operculum were shown to be cytoarchitectonically similar (Galaburda and Sanides, 1980). An involvement of the parietal operculum in the acoustic–phonetic processing in speech perception was proposed by Caplan et al. (1995) who showed that an impairment of phoneme identification and discrimination abilities was associated, together with the left posterior supramarginal gyrus, with damage in this region. However, the parietal operculum is not only associated with the perceptual part of speech processing. In a functional imaging study, this area was shown to participate in phonological aspects of both speech perception and speech production. Listening to as well as covertly repeating pseudowords activated two areas in the left posterior perisylvian region: posterior STG and the parietal operculum (Buchsbaum et al., 2001). The authors proposed this perisylvian region to be part of the dorsal stream associated with auditory–motor mapping (Hickok et al., 2000). Similarly, in our study activation in the left parietal operculum may represent the neuronal correlate of the process of transferring auditory information into speech motor
representations. Notably, we extend Buchsbaum et al.'s results, which were obtained in a task with a relatively heavy verbal working memory load (i.e., the rehearsal of strings of multisyllabic pseudo-words), to fast and relatively automatic repetition, a task with relatively low demands on verbal working memory. Proceeding from our original assumption of a tempororo-frontal network in fast and automated repetition, the localization of this area in the parietal operculum was somewhat unexpected. However, the dorsal pathway has been found to be connected with the inferior parietal lobe in a study using MRI tractography, supporting the notion that inferior parietal cortex may serve as a general interface between perceptual and motor systems (Parker et al., 2005) — an interpretation that is consistent with our finding.

Finally, across participants we found significant effects of two behavioral parameters, the degree of imitation of speech rate and the response latency, on overall functional activation related to shadowing. Specifically, the more subjects imitated rate of speech, the more they associated the right inferior parietal lobule, an area located near the posterior ascending ramus of the Sylvian fissure termed the ‘planum parietale’ (Binder et al., 1996; Jancke et al., 1994). We also found that faster responders showed higher activation in nearly the same parietal region near the tempo-parietal boundary. The planum parietale, which has been demonstrated to be more pronounced in the right compared to the left hemisphere, has been associated with higher order cognitve function such as attention (Jancke et al., 1994). This interpretation squares with other reports showing temporoparietal activation for the processing of competing vs. congruent auditory stimuli in a dichotic listening task (Pugh et al., 1996). Thus, a potential explanation for this finding may be that the shadowing task, as a whole, required auditory attention. Even though we consider the auditory-to-motor mapping process, which is at the core of the shadowing task, a highly automated process, appropriate implementation of such an immediate repetition mode probably requires considerable attention. In order to keep the audio–motor transformation running efficiently, the process must obviously be shielded from other perceptual or cognitive intrusions. Consistent with this is that most of the subjects examined here reported that the task required high effort and focused attention. In the present study, subjects who performed particularly well on the temporal aspects of the shadowing task may have been particularly attentive, which may have resulted in a higher activation of right inferior parietal cortex.

There also may be a more language-specific explanation for the fact that activation in inferior temporoparietal cortex near the Sylvian fissure was correlated with the imitation of speech rate. One might speculate that there is a division of labor between speech-related areas of the left and right hemisphere, with the left-sided areas being specifically involved in translating the linguistic information components in the perceived speech signal, while speech-related areas of the right hemisphere may deal with paralinguistic aspects of the signal, such as, for instance, rate of speech. It is known that cortical regions in the temporal, parietal and frontal lobes of the right hemisphere are activated by non-linguistic auditory information, such as complex non-verbal stimuli (Roland et al., 1981), or by emotional-expressive sounds, e.g., laughter (Meyer et al., 2005). Furthermore, paralinguistic auditory information contained in speech signals, e.g., a speaker’s voice characteristics, are known to activate right perisylvian regions (Lattner et al., 2005). Moreover, processing of the duration of auditory signals – a parameter which was used here as a correlate of speech rate – has been demonstrated in an MEG study to activate right inferior parietal cortex (Levane, et al., 1996). Taken together, such evidence may explain why in the present study imitation of speech rate correlated with right inferior parietal activation.

In sum, our results have demonstrated bilateral involvement of dorsal stream structures in the relatively automated, immediate repetition of speech stimuli. These structures included the superior temporal and the inferior frontal and precentral gyri. Additional postcentral activation may reflect feedback mechanisms of speech motor control. Specific activation in the superior temporal sulcus bilaterally related to the shadowing of multiple speaker models may constitute a neural correlate of the extraction and normalization of articulatory gestures of other speakers in preparation for speech production, while activation in Broca’s area may indicate motor aspects of the adaptation to multiple speaker models. The left parietal operculum may play a central role in efficient shadowing (i.e., an effortless transfer from auditory-to-motor information). Finally, a right inferior parietal area near the tempo-parietal boundary may be associated with attentional and/or paralinguistic processes.

Acknowledgments

This study was funded by a grant from the German Federal Ministry of Education and Research (BMBF-01GW0572) to W.Z. and A.B. and carried out as part of the collaborative BMBF research project “From dynamic sensorimotor interaction to conceptual representation: Deconstructing apraxia”. We thank Christian Buechel for helpful suggestions on the experimental design, and Gregory S. Hickok and an anonymous reviewer for constructive comments on an earlier version of the manuscript.

References
