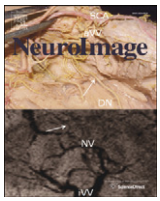
NeuroImage 60 (2012) 1832–1842

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Full Length Article

Brain 'talks over' boring quotes: Top-down activation of voice-selective areas while listening to monotonous direct speech quotations

Bo Yao [a,*], Pascal Belin [a,b], Christoph Scheepers [a]

[a] Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, G12 8QB, UK
[b] Centre for Cognitive Neuroimaging (CCNi), University of Glasgow, Glasgow, G12 8QB, UK

ARTICLE INFO

Article history:
Received 26 September 2011
Revised 31 December 2011
Accepted 22 January 2012
Available online 28 January 2012

Keywords:
Direct speech
Indirect speech
fMRI
Speech perception
Mental simulation
Language comprehension
Emotional prosody

ABSTRACT

In human communication, direct speech (e.g., *Mary said, "I'm hungry"*) is perceived as more vivid than indirect speech (e.g., *Mary said that she was hungry*). This vividness distinction has previously been found to underlie silent reading of quotations: Using functional magnetic resonance imaging (fMRI), we found that direct speech elicited higher brain activity in the temporal voice areas (TVA) of the auditory cortex than indirect speech, consistent with an "inner voice" experience in reading direct speech. Here we show that listening to monotonously spoken direct versus indirect speech quotations also engenders differential TVA activity. This suggests that individuals engage in top-down simulations or imagery of enriched supra-segmental acoustic representations while listening to monotonous direct speech. The findings shed new light on the acoustic nature of the "inner voice" in understanding direct speech.

Introduction

Direct speech (e.g., *Mary said, "I'm hungry"*) and indirect speech (e.g., *Mary said that she was hungry*) are two important reporting styles in everyday communication. The use of direct speech usually coincides with vivid demonstrations of reported speech acts whereas indirect speech provides mere descriptions of what was said (Clark and Gerrig, 1990). For instance, the former usually contains vivid depictions of the reported speaker's voice while the latter does not. Recently, this "vividness" distinction has been shown to underlie language comprehension of the two reporting styles in written text. Combining event-related fMRI and eye-tracking, Yao et al. (2011) found that silent reading of direct speech elicited higher brain activity in the voice-selective areas or temporal voice areas (TVA; Belin et al., 2000) of the right auditory cortex than silent reading of indirect speech. Moreover, Yao and Scheepers (2011) observed that such an "inner voice" experience in reading direct rather than indirect speech was also reflected in behavioural articulation (oral reading) and eye-movement patterns (silent reading). The findings are in line with embodied cognition in language processing (Barsalou, 1999; Zwaan, 2004), suggesting that individuals are more likely to mentally simulate or imagine the reported speaker's voice in understanding direct speech as opposed to indirect speech.

However, with no acoustic stimulation as a reference, it is still unclear what constitutes such "inner voice" experiences during *silent reading* of direct as opposed to indirect speech. In Yao et al. (2011), we speculated that the mentally simulated voice representations entail supra-segmental acoustic information of the quoted speaker's voice (e.g., speech melodies, intonation and emotional prosody), given that a right-lateralised activation pattern was observed. Indeed, "bottom-up" auditory stimulation studies have shown the same lateralisation by contrasting speech or music with acoustically matched noise bursts (Zatorre et al., 1992, 1994), by contrasting speech signals (irrespective of intelligibility) with noise-vocoded signals (Scott et al., 2000), and by contrasting nonverbal sounds comprising extended frequency transitions (supra-segmental) with those comprising rapid frequency transitions (sub-segmental) (Johnsrude et al., 2000). Hence, it seems likely that the right superior temporal gyrus/sulcus (STG/STS) areas are involved in processing dynamic pitch variations which are an important property of supra-segmental vocal information. One type of such information, namely emotional prosody and intonation, is also found to activate similar, right-lateralised activation patterns in various forms including sentences (Mitchell et al., 2003; Wildgruber et al., 2005), words (Wiethoff et al., 2008) and word-like vocalisations (Grandjean et al., 2005). Most importantly, mental simulations of supra-segmental acoustic information in language comprehension would fit well with the notion of direct speech as vivid demonstration — in which vivid depictions of the quoted speaker's voice are characterised in terms of enriched supra-segmental acoustic information.

* Corresponding author.
E-mail address: b.yao@psy.gla.ac.uk (B. Yao).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.01.111

In this paper, we attempted to address whether the mentally simulated voices predominantly consist of supra-segmental acoustic information during comprehension of direct as opposed to indirect speech. In order to create acoustic references to verify such supra-segmental information, we prepared audio recordings of short stories in which direct and indirect speech utterances were spoken monotonously. This manipulation preserved (sub)-segmental acoustic information such as the phonological information of the uttered words in the recordings, but minimized supra-segmental acoustic information such as the global intonation patterns over the utterances. Thus, if direct speech is represented in enriched supra-segmental acoustic representations of voices during language comprehension, individuals would have to mentally simulate such representations, since they are minimized in the stimuli, to supplement what they hear while listening to monotonously spoken direct speech utterances. By contrast, indirect speech is not represented in vivid voice representations and hence individuals need not simulate supra-segmental acoustic information when listening to monotonous indirect speech utterances. Thus, we predict that listening to monotonous direct speech quotations would elicit higher "top-down" brain activity in temporal voice areas of the right auditory cortex (i.e., similar to the brain areas identified in Yao et al., 2011) than listening to equally monotonous, meaning-equivalent indirect speech quotations.

To test this hypothesis, we employed functional magnetic resonance imaging (fMRI) to measure participants' brain activity while they were listening to short stories which contained monotonously spoken direct or indirect speech utterances. Between the two conditions, we compared the evoked BOLD signal changes within participants' temporal voice areas. Moreover, we performed multiple parametric modulation analyses to verify the underlying source of any differential brain activity we observed. To assess consistency of results across studies, we also compared the observed activation patterns with our previous silent reading data (Yao et al., 2011).

Materials and methods

Participants

Twenty-one adult participants were recruited and scanned. They were native English speakers with normal hearing and language abilities, and with no history of neurological or psychiatric disorders. Three participants had to be excluded from analysis due to either (a) no clear response in the voice localiser task and/or excessive head-movements during scanning (2 subjects), or (b) scanning abortion following claustrophobic symptoms (1 subject). Data from the remaining 18 participants (age 18–32 years, 9 males and 9 females) were valid for the final analyses. All of them were right-handed except for one female subject. They signed a consent form and were paid at £6/h for their participation.

Stimuli

Main stimuli

Ninety short stories with different protagonists (indicated by different names) were recorded as stimuli. The stories were exactly the same as in Yao et al. (2011). Transcriptions are available at www.psy.gla.ac.uk/~christop/JOCN_2011/Stimuli.pdf. Each story started with two declarative sentences to set up a scenario (e.g., *Luke and his friends were watching a movie at the cinema. Luke wasn't particularly keen on romantic comedies, and he was complaining a lot after the film.*), followed by either a direct speech or an indirect speech quotation sentence (e.g., *He said: "God, that movie was terrible! I've never been so bored in my life."* or *He said that the movie was terrible and that he had never been so bored in his life.*). The reported clauses in both conditions (underscored in the above examples) were equivalent in terms of linguistic content. Additional comprehension questions were also recorded for 23 stories (ca. 25%) to assess participants' overall comprehension accuracy and to ensure that they read the stories attentively.

The stories and questions were spoken by a professional actress. Critically, in one condition, the direct speech utterances were deliberately spoken as monotonously (*Direct-monotonous* condition) as the indirect speech utterances (*Indirect-monotonous* condition), i.e., without providing vivid depictions of the reported speaker's voice. We also recorded "normal" (i.e., vivid) versions of the direct speech utterances which were used as a control condition (*Direct-vivid* condition). Example recordings are available at: www.psy.gla.ac.uk/~boy/fMRI/samplerecordings/.

Three lists of stimuli with counterbalanced item-condition combinations (i.e., 30 Direct-monotonous trials, 30 Indirect-monotonous trials, and 30 Direct-vivid trials per list) were constructed using a Latin square. Each item appeared once per list, but in a different condition across lists. Each list was assigned to one third of our participants. The presentation order of the items per list was randomised for each participant.

Voice localizer stimuli

For the voice localizer session (see Procedure), we presented blocks of vocal sounds and non-vocal sounds provided by the Voice Neurocognition Laboratory (vnl.psy.gla.ac.uk), University of Glasgow. These stimuli were the same as those employed in Belin et al. (2000), and comprised both speech (e.g., spoken vowels) and non-speech (e.g., laughing and coughing) vocal sound clips, as well as non-vocal sound clips (e.g., telephone ringing and dog barking). The contrast in brain activity elicited by vocal versus non-vocal sounds reliably localizes temporal voice areas of the auditory cortex.

Procedure

Participants were positioned in the scanner, wearing MRI-compatible, electrostatic headphones (NordicNeuroLab, Norway) for (1) auditory presentation during both the story listening session and voice localizer session and (2) noise attenuation during fMRI scanning. For the story listening session, participants were instructed to keep their eyes closed, to listen to the stories carefully and to answer comprehension questions which would follow 25% of the short stories they had heard. The stimuli were presented using E-Prime 2.0 (Psychology Software Tools, Inc., USA); each trial started with a 4-second silence period, followed by the presentation of the story and then (in 25% of the trials) a comprehension question regarding the content of the preceding story. Each such question appeared 1 s after termination of the preceding story presentation and prompted a "yes" or "no" response which participants could provide by pressing buttons on a response box with their index or middle fingers, respectively. The 90 listening trials were evenly interspersed with five 30-second "baseline" trials during which no experimental stimulation was present.

After the story listening session, an anatomical scan of the participant's brain was performed, followed by a brief (ca. 10-min) voice localizer scanning session. During the latter, participants were instructed to close their eyes while listening to 20 8-sec blocks of vocal and 20 8-sec blocks of non-vocal auditory stimuli presented in an efficiency optimised, pseudo random order along with 20 8-sec blocks without stimulation, acting as a baseline (cf. Belin et al., 2000).

MRI acquisition

Scanning was performed on a 3-T Siemens Tim Trio MRI scanner using a 12-channel head coil (Erlangen, Germany). Functional scans (for both the story listening session and voice localizer session) were acquired using a T2*-weighted echoplanar imaging (EPI) sequence (32 slices acquired in orientation of the Sylvian fissure;

TR = 2 s; TE = 30 ms; matrix size: 70 × 70; voxel size: 3 × 3 × 3 mm; FOV = 210). T1 whole-brain anatomical scans were obtained using 3D T1-weighted magnetization prepared rapid acquisition gradient echo (MP-RAGE) sequence (192 axial slices; matrix size: 256 × 256; voxel size: 1 × 1 × 1 mm; FOV = 256). The average scanning time for the whole experiment was around 55 min per participant.

Data analysis

Whole brain and ROI analyses

All MRI data were analysed using SPM8 (www.fil.ion.ucl.ac.uk/spm/, University College London). Pre-processing of functional scans included (a) head motion corrections (tri-linear interpolation) whereby scans were realigned to the first volume; (b) co-registration of functional scans to their corresponding individual anatomical scans; (c) segmentation of the co-registered scans; (d) normalisation of functional (3 mm isotropic voxels) and anatomical (1 mm isotropic voxels) data to Montreal Neurological Institute (MNI) space; and (e) smoothing of normalised data (10-mm Gaussian kernel).

fMRI data from the anatomical and voice localizer scanning sessions were used to determine the temporal voice areas in the auditory cortex. The individual voice localizers for most participants (13 subjects) were obtained at $p<0.05$ (FWE-corrected for the whole brain at the peak level). The voice localizers for the other 5 subjects were obtained at $p<0.001$ (uncorrected, to increase sensitivity). The group voice localizer was obtained at $p<0.05$ (FWE-corrected for the whole brain at the peak level). Based on the previous findings of a right-lateralised activation pattern in silent reading of direct vs. indirect speech (Yao et al., 2011), the temporal voice areas of the *right* auditory cortex (i.e., the right TVA) were defined as the main Region of Interest (ROI). The ROI analyses were also run for the left TVA, results of which are presented together with our parametric modulation analysis results in Section Parametric modulation analysis.

For the story listening session, the temporal onset of a critical fMRI event was defined as the temporal onset of the first word within the quotation marks (direct speech) or of the complementizer *that* (indirect speech); its offset was defined as the temporal offset of the last word in the direct or indirect speech quotations. Uncritical events (e.g., listening to background sentences, comprehension questions and instructions, as well as button pressing) were specified as events of no interest in the design matrix. The rest consisted of all "silence" events (including five 30-second baseline trials and all 4-second pre-trial silence periods) and was regarded as baseline. The fMRI data were mapped to the human Colin atlas surface (Van Essen, 2002) in CARET (Van Essen et al., 2001). The mean beta estimates within ROIs were calculated by SPM toolbox easyROI (www.sbirc.ed.ac.uk/cyril/cp_download.html), and submitted to 2-tailed paired-sample *t*-tests.

Parametric modulation analyses

To verify the underlying source of the observed brain activations, we performed parametric modulation analyses with (1) the acoustic parameters, (2) the perceived vividness and (3) the perceived contextual congruency of the critical direct and indirect speech utterances. The acoustic parameters were intended to *objectively* capture the acoustic characteristics of the critical audio recordings independent of linguistic content. The parametric modulations with these measures would unveil whether the observed brain activations were simply engendered by the acoustic differences between conditions. Comparably, the vividness ratings (see below) were intended to provide a more *subjective* measurement of the acoustic characteristics of the critical direct and indirect speech utterances independent of linguistic context. In a way, this measure summarises the joint effect of the acoustic characteristics (i.e., vocal features) and linguistic content (wording etc.) on how vivid the critical utterances would sound to perceivers irrespective of context. Thus, the parametric modulations with the vividness measure would reveal whether between-condition differences reflect evoked brain responses to the differential vocal features that are subjectively perceived "bottom-up" from the stimuli. Finally, the contextual congruency ratings were intended to measure the discrepancy between the actual vocal features of the critical stimuli and the way these stimuli "should have sounded like" in the given contexts. In other words, instead of quantifying the "bottom-up" perceived vocal vividness of the stimuli, the contextual congruency metric was intended to capture the contextually expected vividness (or its discrepancy with the actually perceived vividness) of the critical speech stimuli. In this sense, parametric modulation analyses with the contextual congruency metric would indicate whether observed brain activation patterns reflect 'top-down' mental simulations of enriched vocal depictions while listening to the critical (monotonous) direct and indirect speech utterances.

Acoustic parameters. Using Praat software (Boersma and Weenink, 2010), we characterized the critical audio samples in terms of eight acoustic parameters which are known to be related to speech prosody: (1) the *mean pitch* (fundamental frequency F_0) averaged over the duration of each sample; (2) the *pitch variation*, measured as the standard deviation in pitch over the duration of each sample (pitch *SD*); (3) the *pitch range* (difference between maximum and minimum pitch in Hz over the duration of each sample); (4) the *mean intensity* (in dB) over the duration of each sample; (5) the *intensity variation*, measured as the standard deviation in intensity over the duration of each sample (intensity *SD*); (6) the *intensity range*; (7) the duration of the voiced sample (the recording periods in which the pitch value passed the voicing threshold); (8) the duration of the entire audio sample. These eight parameters were then included simultaneously as modulators in the parametric modulation analyses to partial out their joint contribution to the between-condition differences in the evoked BOLD signals.[1]

Vividness ratings. We recruited twelve native speakers of English with normal hearing and language abilities for this rating study. They were paid at £2 for their participation. A typical session took 10–20 min.

Participants were seated in front of a Dell Duo Core PC, wearing headphones. They were presented with the same auditory stimuli that were used in fMRI scanning; only the critical direct and indirect speech utterances (i.e., without context) were presented. After hearing each utterance, participants had to give a rating by pressing number keys on a keyboard, to indicate how vivid and engaging the utterance they had just heard was. The ratings were given on a 7-point scale in which 7 meant "very vivid and engaging" while 1 meant "extremely monotonous".

The collected ratings were summarised by condition for each subject and then submitted to paired-sample *t*-tests to assess the between-condition differences in vividness. They were also summarised by trial and were then included as a parametric modulator to partial out the contribution of perceived vividness to the between-condition differences in the evoked BOLD signals.

Contextual congruency ratings. We recruited another twelve native speakers of English with normal hearing and language abilities for this rating study. They were paid at £4 for their participation. A typical session took 30–40 min.

The procedure of this rating study was the same as in the above vividness rating study except: (1) participants were presented with the whole stories (i.e., with context), and (2) they had to give 7 point-scale ratings on the "contextual congruency" of the critical direct and indirect speech utterances, i.e., on whether these utterances

[1] Since we are solely concerned with the acoustic parameters' *joint contribution* to the brain activation patterns of interest (and not with each parameter's individual importance), multicollinearity is not an issue here.

matched the context in terms of how vivid and engaging they were; 7 meant "fits in the context extremely well" while 1 meant "does not fit in the context at all".

The collected ratings were first summarised by condition for each subject and were then submitted to paired-sample *t*-tests to assess the between-condition differences in contextual congruency. They were also summarised by trial and were then included as a parametric modulator to partial out the contribution of contextual congruency to the between-condition differences in the evoked BOLD signals.

Parametric modulations. We performed three parametric modulation analyses with the speech utterances' acoustics, vividness and contextual congruency as the modulators, respectively. This was to assess each set of modulators' individual contributions to the observed brain-activation differences between the *Direct-monotonous* condition and the *Indirect-monotonous* condition. Our strategy was to examine whether the differential brain activations would be *reduced* (i.e., accounted for) as a consequence of partialling out the effects of the investigated modulators. First, we performed parametric modulations at the individual level. For each participant, we specified in the design matrix a single trial-type for all three conditions (including the control condition Direct-vivid); it was first followed by one of the three sets of modulator(s) – this would ensure that the effects of the investigated modulator(s) are partialled out across all trials – which were then followed by three experimental conditions coded with binary values. After the experimental trials, other event types and participants' head motion parameters were also included in the design matrix for modelling the collected BOLD signals. We then conducted the contrast analyses in the same way as before with the TVAs as the ROIs. Using the same threshold as the main contrast (i.e., $p<0.05$, FWE-corrected for the localizer-defined volume at the peak level), we examined how the observed brain-activation differences between the *Direct-monotonous* condition and the *Indirect-monotonous* condition were affected when the effects of each set of modulators were partialled out, respectively.

Comparing brain activation patterns with Yao et al. (2011)

We also compared the brain activation patterns observed in the current study (i.e. the contrast between the Direct-monotonous and the Indirect-monotonous conditions) with those from the previous silent reading study (Yao et al., 2011). The comparison was observational: The activation patterns were described using the 3D coordinates (in relation to the standard brain space from the Montreal Neurological Institute) of the peak voxel within each activation "blob". We paired the activation "blobs" with their counterparts between the two studies and compared the peak voxels' coordinates within each pair.

Results

Whole brain and ROI analyses

The voice localizer

Consistent with the findings of Belin et al. (2000), we found that the vocal sounds elicited significantly ($ts>7.6$, $ps<.02$, FWE-corrected for the whole brain at the peak level) greater activity than non-vocal sounds bilaterally in the STG/STS areas (Fig. 1A). The maximum of voice-sensitive activation was located along the upper bank of the central part of the right STS (Table 1).

The main contrast

We found that listening to monotonously spoken direct speech utterances elicited greater BOLD signals in the right TVA ($ts>5.5$, $ps<.006$, FWE-corrected for the localizer-defined volume at the peak level) than listening to monotonous indirect speech utterances (see Fig. 1B). Both conditions were active against baseline. The between-condition difference was significant: For individual Regions of Interest (ROIs), 2-tailed paired-sample $t(17)=5.650$, $p<0.001$; for the group ROI, $t(17)=4.979$, $p<0.001$ (Fig. 1C). The differential brain activity between the Direct-monotonous and the Indirect-monotonous conditions was located in temporal voice areas along the posterior, middle and anterior parts of the right STS brain areas.[2] No region showed an opposite pattern of activity (i.e., the Direct-monotonous condition was always associated with a greater BOLD signal than the Indirect-monotonous condition). As a whole, brain activity outside bilateral TVAs did not significantly differ between the Direct-monotonous and the Indirect-monotonous conditions ($ts<3.8$, $ps>.19$, FWE-corrected for the whole brain at cluster level).

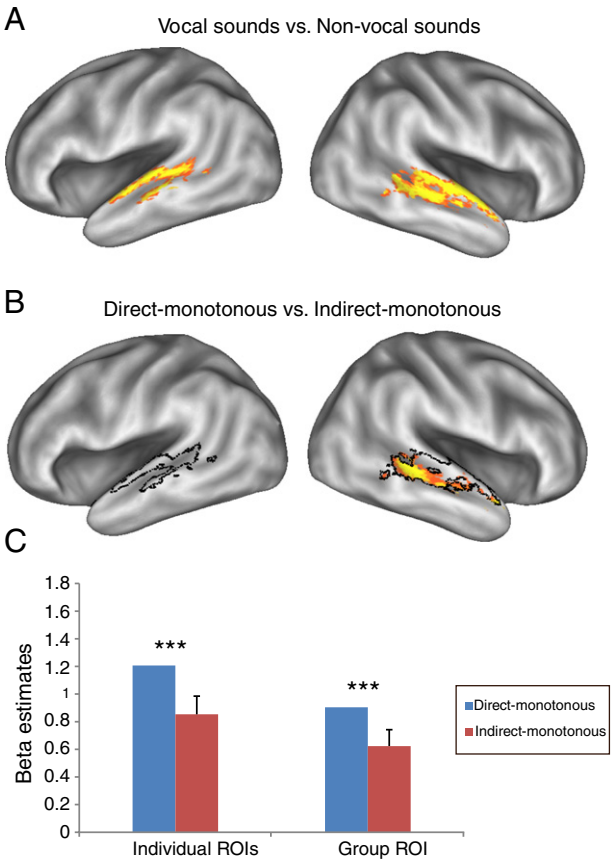


Fig. 1. Illustrations of the localizer and the between-condition differences (Direct-monotonous against Indirect-monotonous), A, brain regions that selectively responded to vocal sounds as opposed to non-vocal sounds (i.e., the TVA) under the threshold of $p<.05$ (FWE-corrected for the whole brain at the peak level), B, Within the TVA (indicated by black lines), brain regions that selectively responded to auditory comprehension of monotonous direct speech utterances as opposed to monotonous indirect speech utterances in the whole brain analysis ($n=18$) under the threshold of $p<.05$ (FWE-corrected for the localizer-defined volume at the peak level), C, Mean signal change (against baseline) regarding the critical Direct-monotonous against Indirect-monotonous contrast in the right TVA Regions of Interest (ROIs), determined individually (left) or using the sample average (right). The single error bar in each panel refers to the 95% CI for the between-condition difference.

In principle, these results could be explained in two very different ways. First, it is possible that the Direct-monotonous condition was somehow acoustically more varied than the Indirect-monotonous condition. In this case, the higher BOLD signals for the Direct-monotonous condition would simply be a reflection of the more varied acoustic information that is carried in those stimuli. However, it is

[2] The loci of the brain activity differences are listed here mainly for the purpose of describing their spatial distribution within the ROI. Since a well-defined functional ROI (i.e., the right TVA) was used for analysis, statistical tests for individual activation peaks are redundant and therefore not reported.

Table 1
Brain regions associated with significant BOLD signal increases for vocal sounds as opposed to non-vocal sounds. Only the first three local maxima (more than 8.00 mm apart) are reported for each cluster.

Regions	MNI coordinates			*t*-Value	FWE-corrected *p*-value
	x	y	z		
Right TVA	60	−28	1	14.66	<.001
	60	−1	−5	12.65	<.001
	66	−19	−2	12.29	<.001
Left TVA	−66	−19	4	12.93	<.001
	−57	−25	4	12.56	<.001
	−45	−31	4	11.35	<.001

Table 3
A summary of the eight acoustic parameters (means, with standard deviations in parentheses) for the spoken stimuli in each experimental condition.

Acoustic parameters	Condition		
	Direct-monotonous	Indirect-monotonous	Direct-vivid
Pitch mean (Hz)	200.43 (12.42)	208.50 (13.39)	265.81 (43.22)
Pitch SD (Hz)	32.62 (6.97)	45.24 (10.51)	66.36 (20.11)
Pitch range (Hz)	143.87 (39.27)	199.97 (52.80)	281.77 (82.48)
Intensity mean (dB)	68.01 (1.17)	66.89 (.93)	70.71 (2.48)
Intensity SD (dB)	7 (.51)	6.87 (.44)	7.51 (.68)
Intensity range (dB)	31.23 (1.62)	29.66 (1.59)	34.32 (3)
Voiced duration (ms)	2044 (534)	2073 (559)	2318 (662)
Audio sample duration (ms)	4028 (984)	4332 (1056)	4702 (1119)

also possible that the Direct-monotonous condition was acoustically *equivalent* to (i.e., not more varied than) the Indirect-monotonous condition. In that case, the higher BOLD signals for the Direct-monotonous condition must have been due to something other than "bottom-up" acoustic characteristics of the stimuli, making an explanation more likely whereby monotonous direct speech utterances were supplemented "top-down" (via mental simulation) with enriched supra-segmental acoustic information.

To distinguish between these contrasting interpretations, we conducted parametric modulation analyses to verify the underlying source of the observed brain activations between Direct-monotonous and Indirect-monotonous conditions. We investigated the signal contribution of the speech utterances' acoustic parameters, their vividness and their contextual congruency to both the critical contrast (Direct-monotonous against Indirect-monotonous) and the control contrast (Direct-vivid against Direct-monotonous), respectively.

Parametric modulation analyses

It was found that the examined parametric modulators (i.e., the acoustic parameters, the perceived vividness and the perceived contextual congruency) were highly correlated with one another (Table 2). This is not surprising since all of them are measured variables that are associated with the vocal vividness of the direct and indirect speech utterances, either objectively, subjectively, or subjectively under consideration of linguistic context (see Section Parametric modulation analyses). However, the correlations were certainly not perfect, and the parametric modulation results below revealed rather distinct parametric contributions to the brain activation patterns of interest.

The effects of acoustics

The descriptives of the eight acoustic parameters are summarised by condition in Table 3.

The parametric modulation analyses showed that using the same threshold of $p<.05$ (FWE-corrected for the localizer-defined volume at the peak level), the Direct-monotonous condition still elicited higher brain activity ($ts>4.85$, $ps<.02$) within the right TVA than the Indirect-monotonous condition after the effects of the acoustic parameters were partialled out (left panel in Fig. 4B). The relevant ROI analyses (for both the left and the right TVAs) are reported in Table 4. These results suggest that the originally observed brain-activation differences between the Direct-monotonous and the Indirect-monotonous conditions (left panel in Fig. 4A) were unlikely to be engendered by the between-condition differences in acoustic characteristics.

In addition, we performed the same parametric modulation analyses on the contrast between the Direct-vivid and Direct-monotonous conditions. The Direct-vivid condition is acoustically more varied and higher in amplitude than the Direct-monotonous condition (Table 3). It was found that the former was associated with significantly increased BOLD signals ($ts>8.28$, $ps<.001$) in bilateral TVAs (right panel in Fig. 4A) as opposed to the latter. After the effects of the acoustic parameters were partialled out, these brain activity differences almost disappeared (right panel in Fig. 4B, with only two voxels surviving under the .05 threshold (FWE-corrected for the localizer-defined volume at the peak level), $ts>4.53$, $ps<.05$). The relevant ROI analyses are reported in Table 4. These results indicate that the brain activity differences between the Direct-vivid and the Direct-monotonous conditions were likely to be engendered by the acoustic differences between the two conditions. In stark contrast, the brain activity differences between the Direct-monotonous and the Indirect-monotonous conditions were unlikely due to differences in acoustics.

Partialling out the effect of vividness

The averaged ratings and corresponding paired-sample *t*-test results are illustrated in Fig. 2. It was found that the control condition Direct-vivid was perceived as significantly more vivid than the two monotonous conditions, $ts(11)>19$, $ps<.001$. However, instead of being "equally monotonous", the Direct-monotonous condition was perceived as significantly *less* vivid than the Indirect-monotonous

Table 2
The cross-correlations between the examined parametric modulators.

Pearson correlation (n = 270)	Pitch mean	Pitch SD	Pitch range	Loudness mean	Loudness SD	Loudness range	Voiced sample duration	Recording duration	Vividness rating	Contextual congruency
Pitch mean	1	.748	.745	.715	.479	.592	.156	.162	.749	.495
Pitch SD	.748	1	.912	.404	.398	.431	.096	.172	.713	.521
Pitch range	.745	.912	1	.402	.359	.385	.201	.270	.693	.509
Loudness mean	.715	.404	.402	1	.528	.782	.267	.161	.600	.297
Loudness SD	.479	.398	.359	.528	1	.699	.031	.138	.432	.261
Loudness range	.592	.431	.385	.782	.699	1	.270	.299	.579	.254
Voiced sample duration	.156	.096	.201	.267	.031	.270	1	.856	.232	.135
Recording duration	.162	.172	.270	.161	.138	.299	.856	1	.288	.224
Vividness rating	.749	.713	.693	.600	.432	.579	.232	.288	1	.632
Contextual congruency rating	.495	.521	.509	.297	.261	.254	.135	.224	.632	1

Table 4
The ROI (group) results for the original contrast analyses and all three parametric modulation analyses. The upper panel reports ROI analyses when contrasting the Direct-monotonous against the Indirect-monotonous conditions. The lower panel reports ROI analyses when contrasting the Direct-vivid against the Direct-monotonous conditions. Standard deviations of the between condition differences are reported in parentheses along with the mean between-condition beta differences.

Direct-monotonous against Indirect-monotonous						
	Left TVA			Right TVA		
Analyses	Mean beta difference	$t(17)$	p	Mean beta difference	$t(17)$	p
Original	.091 (.240)	1.617	.124	.281 (.239)	4.979	<.001
Acoustics partialled out	.098 (.230)	1.809	.088	.261 (.253)	4.389	<.001
Vividness partialled out	.283 (.224)	5.361	<.001	.469 (.279)	7.127	<.001
Contextual congruency partialled out	.015 (.279)	.231	.820	.201 (.301)	2.846	.011
Direct-vivid against Direct-monotonous						
Original	.338 (.186)	7.701	<.001	.419 (.216)	8.218	<.001
Acoustics partialled out	.072 (.221)	1.386	.184	.139 (.259)	2.282	.036
Vividness partialled out	−.211 (.482)	−1.861	.080	−.127 (.561)	−.960	.351
Contextual congruency partialled out	.453 (.241)	7.982	<.001	.535 (.259)	8.777	<.001

condition, paired-sample $t(11) = -7.872$, $p < .001$. The latter suggests that the monotonous direct speech utterances (Direct-monotonous) contained less vivid vocal modulations than the monotonous indirect speech utterances (Indirect-monotonous). This is also inconsistent with a more "bottom-up" explanation of our findings (Fig. 1): Given lower vividness ratings for the Direct-monotonous condition, a "bottom-up" account would predict that this condition would consume less energy (therefore less blood oxygen) to process within the temporal voice areas than the more vivid Indirect-monotonous condition; in other words, the Direct-monotonous condition should have elicited significantly *decreased* BOLD signals within these brain areas compared to the Indirect-monotonous condition. However, exactly the opposite was found (see Section Whole brain and ROI analyses).

The parametric modulation analysis on vividness revealed that using the same threshold of $p < 0.05$ (FWE-corrected for the localizer-defined volume at the peak level), the Direct-monotonous condition still elicited higher brain activity ($ts > 7.14$, $ps < .002$, FWE-corrected for the localizer-defined volume at the peak level) within the right TVA than the Indirect-monotonous condition after the effects of vividness were partialled out (left panel in Fig. 4C). The relevant ROI analyses are presented in Table 4. The results suggest that the originally observed brain-activation difference between the Direct-monotonous and the Indirect-monotonous conditions (left panel in Fig. 4A) is not "explained" by the vividness of the speech utterances (for such a conclusion to be justified, there should have been a *reduction* in the original difference after partialling out the effect of vividness). Instead, it indicates that the original between-condition difference was partially "masked" by the vividness contrast between the two conditions: the vividness ratings indicate that the Direct-monotonous condition was perceived as *less* vivid than the Indirect-monotonous condition (see Fig. 3); when the two conditions were brought to the same vividness level (by partialling out the effect of the vividness modulator), the original brain-activation difference was enhanced. In other words, the influence of factors other than vividness became more pronounced when the negative contribution of vividness (Direct-monotonous < Indirect-monotonous) was eliminated. We conclude that while perceived vividness clearly played a role in the originally reported brain activation patterns, its contributions actually went *contrary to* an actual explanation of those brain activation patterns.

Moreover, we performed the same parametric modulation analysis to assess the signal contribution of vividness to the brain activity differences between the Direct-vivid and the Direct-monotonous conditions. We found that similar to the parametric modulation analysis on acoustics, the brain activity differences between the Direct-vivid and the Direct-monotonous conditions in bilateral TVAs (right panel in Fig. 4A) disappeared after the effects of vividness were partialled out (right panel in Fig. 4C; no supra-threshold voxel was found). The corresponding ROI analysis results are reported in Table 4. The results indicate that the brain activity differences between the Direct-vivid and the Direct-monotonous conditions were likely to be engendered by differences in vocal vividness which, in turn, are likely to be carried by differences in acoustics (see Section The effects of acoustics). Importantly, however, the critical brain activity differences between the Direct-monotonous and the Indirect-monotonous conditions were not explainable in terms of vividness or acoustics.

Partialling out the effects of contextual congruency

The averaged ratings and corresponding paired-sample t-test results are illustrated in Fig. 3. It was found that while the Indirect-monotonous

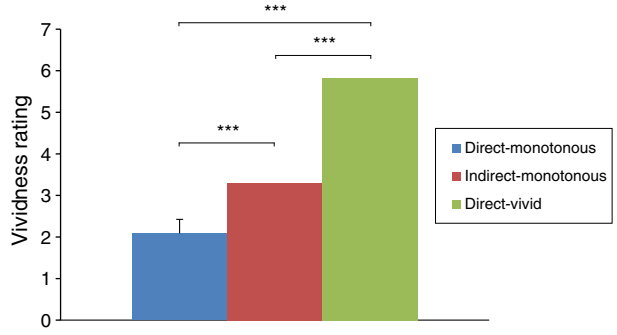


Fig. 2. Illustrations of the between-condition differences in vividness. The significance of the pairwise contrasts (referred to with square brackets) is indicated with asterisks (*** indicates $p < .001$). The single error bar represents the 95% CI for the between-condition difference between Direct-monotonous and Indirect-monotonous conditions.

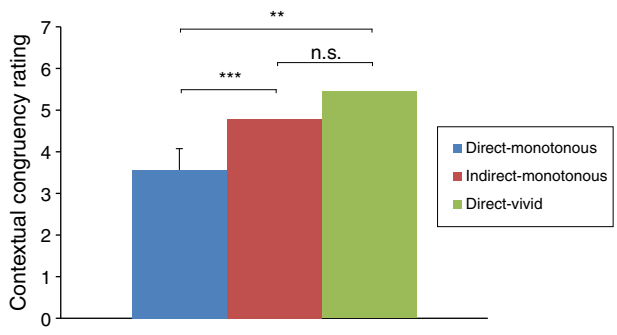


Fig. 3. Illustrations of the between-condition differences in contextual congruency. The significance of the pairwise contrasts (referred to with square brackets) is indicated with asterisks and abbreviations (*** indicates $p < .001$, ** — $p < .01$, n.s. — not significant). The single error bar represents the 95% CI of the between-condition difference between Direct-monotonous and Indirect-monotonous conditions.

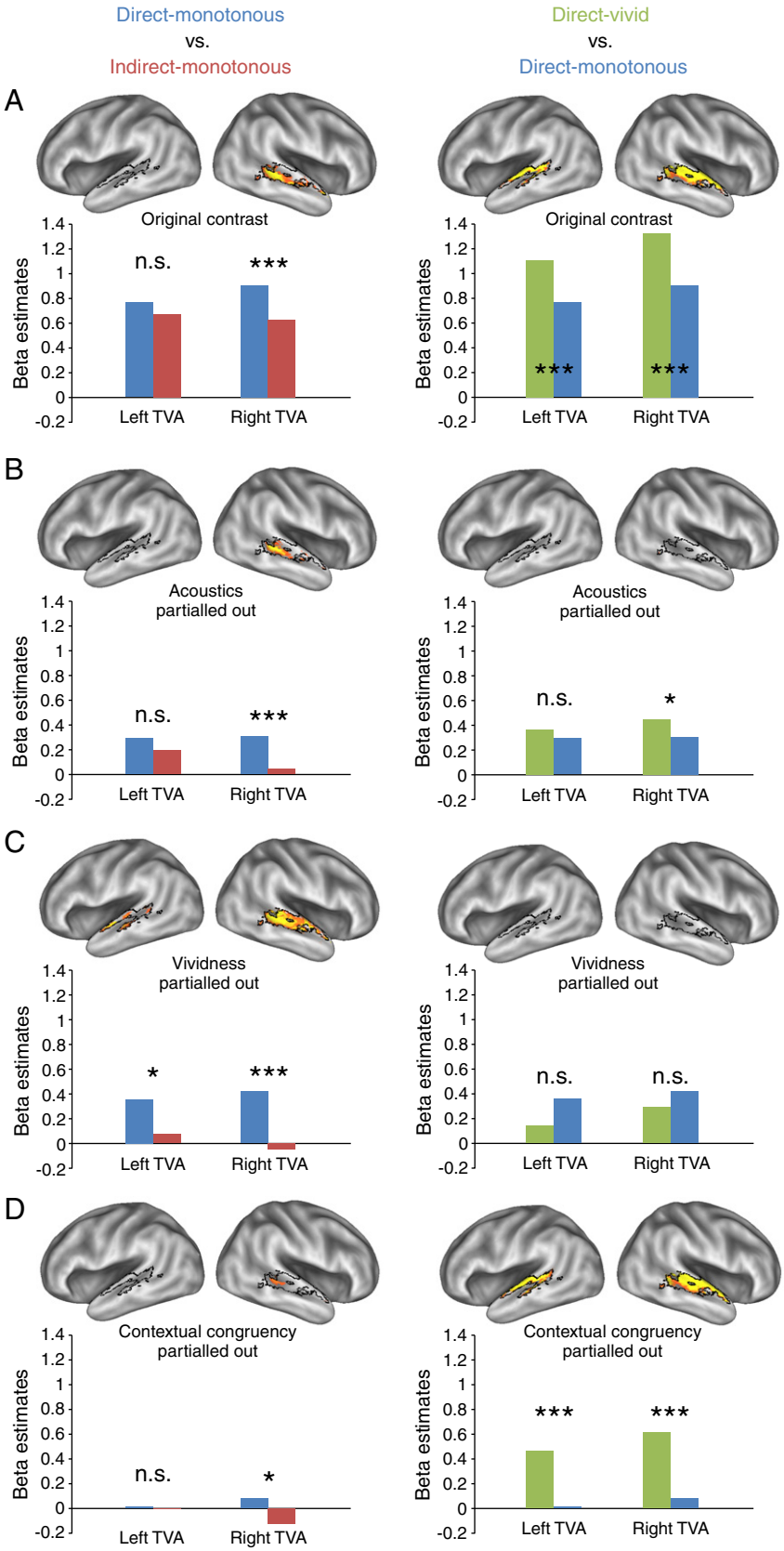


Fig. 4. Illustrations of between-condition differences (Direct-monotonous against Indirect-monotonous [left] and against Direct-vivid [right]) in different parametric modulation models (the TVA is indicated by black lines), A, the original contrasts, B, the same contrasts with effects of acoustics partialled out, C, the same contrasts with effects of vividness partialled out, D, the same contrasts with effects of contextual congruency partialled out.

and the Direct-vivid conditions were perceived as equally congruent with the preceding linguistic contexts, paired-sample $t(11) = 1.754$, $p > .1$, the Direct-monotonous condition was perceived as significantly less congruent with context than either of them, paired-sample $ts(11) > 4$, $ps < .003$ (Fig. 3). The results suggest that listeners routinely expect vivid vocal depictions for direct speech but not for indirect speech; they have to mentally simulate or imagine vivid depictions of the reported speaker's voice to supplement the monotonously spoken direct speech utterances (contextually incongruent) but not the monotonous indirect speech utterances or the vivid direct speech utterances (contextually congruent). This voice simulation process in the Direct-monotonous condition would have required additional energy consumption within the temporal voice areas of the auditory cortex; this could explain why the less vivid Direct-monotonous condition (see Section Partialling out the effect of vividness), which should have elicited decreased BOLD signals in the temporal voice areas of the auditory cortex under a more "bottom-up" interpretation, actually elicited significantly higher BOLD signals within these brain areas as compared to the Indirect-monotonous condition.

The parametric modulation analyses revealed that under the same threshold ($p < 0.05$, FWE-corrected for the localizer-defined volume at the peak level), a much smaller number of voxels (< 50, as opposed to 281 voxels in the original contrast) at the right aSTS/pSTS area survived ($ts > 4.9$, $ps < .013$) when contrasting the Direct-monotonous condition with the Indirect-monotonous condition after the effects of contextual congruency were partialled out (left panel in Fig. 4D). Indeed, the relevant ROI analyses in Table 4 indicate that the signal difference between the Direct-monotonous and the Indirect-monotonous conditions in the right TVA was notably smaller and less significant when compared to the original contrast analysis. The results show that the observed brain activity difference in the right TVA between the Direct-monotonous and the Indirect-monotonous conditions can indeed be partially explained by the discrepancy between the contextually expected vividness and the actually perceived vividness of the corresponding speech stimuli. This suggests that the observed brain activation may indeed reflect 'top-down' mental simulations of enriched vocal depictions while listening to the monotonous direct speech utterances rather than monotonous indirect speech utterances.

We performed the same parametric modulation analysis to examine whether contextual congruency can also 'explain' the brain activity differences between the Direct-vivid and the Direct-monotonous conditions. We found that the Direct-vivid condition still elicited significantly increased BOLD signals ($ts > 5.08$, $ps < .015$, FWE-corrected for the localizer-defined volume at the peak level) in bilateral TVAs as opposed to the Direct-monotonous condition after the effects of contextual congruency were partialled out (right panel in Fig. 4D). The relevant ROI analyses (Table 4) confirmed that the original signal contrast between the Direct-vivid and the Direct-monotonous condition was by no means reduced when effects of contextual congruency were partialled out (if anything, the contrast became slightly stronger). This indicates that the contextual congruency of the speech stimuli cannot explain the brain activity differences between the Direct-vivid and the Direct-monotonous conditions, which were suggested to be engendered by the 'bottom-up' acoustic characteristics (or vocal vividness) of the stimuli (see Sections The effects of acoustics and Partialling out the effect of vividness).

Summary

The parametric modulation results showed that increased brain activity during auditory language comprehension of monotonous direct speech as opposed to monotonous indirect speech can in part be explained by the contextual congruency of the direct or indirect speech utterances, but not by their acoustic characteristics or their perceived vividness out of context. It suggests that listeners routinely expect vivid depictions of the reported speaker's voice for direct speech but not for indirect speech, and that they are more likely to mentally simulate such enriched supra-segmental vocal representations while listening to direct speech utterances which are spoken monotonously as opposed to monotonous, meaning-equivalent indirect speech utterances.

Comparing brain activation patterns between studies

We found that the activation patterns observed when listening to monotonous direct speech against monotonous indirect speech hugely resembled those observed in silent reading of direct speech against indirect speech (cf. Yao et al., 2011). The brain activation patterns in both studies were located at the posterior, the middle and the anterior parts of the right STS areas. Within the MNI space, the peak voxels within each activation cluster were spatially close to their counterparts across the two studies (Fig. 5). The between-study consistency in the activation patterns suggests that the "inner voice" we observed in silent reading of direct as opposed to indirect speech (Yao et al., 2011) is similar in nature to the enrichment of monotonous direct speech (as opposed to monotonous indirect speech) that we found in the current study. Given that the "vocal enrichments" we observed in the present study entailed supra-segmental acoustic information that was hardly available in the actual stimuli, the "inner voice" we observed in silent reading may also have been supra-segmental in nature.

Discussion

The current experiment investigated mental simulations of supra-segmental acoustic representations during auditory language comprehension of direct as opposed to indirect speech. We employed audio recordings in which direct and indirect speech utterances were spoken monotonously. This manipulation preserved the (sub)-segmental acoustic information (e.g., phonological information associated with individual words) but minimized the supra-segmental acoustic information (e.g., the intonation patterns across the speech utterances). Using event-related fMRI, we found that listening to

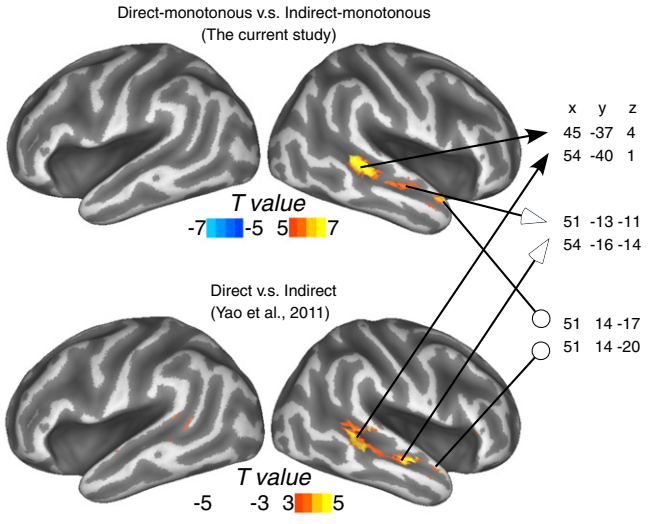


Fig. 5. Illustrations of the critical contrasts between the two studies (masked by TVA). The top panel shows the contrast between Direct-monotonous and Indirect-monotonous conditions in the current study (listening). The bottom panel shows the contrast between Direct speech and Indirect speech in Yao et al. (2011) (silent reading). The arrows point to the peak voxels' 3D coordinates (in MNI space) in the activation clusters. The peak voxels were paired with their anatomical counterparts between the two studies. The thresholds for the two contrasts were adjusted to better illustrate the activation blobs.

monotonously spoken direct speech utterances elicited significantly higher brain activity within temporal voice areas of the right auditory cortex as compared to listening to monotonous, meaning-equivalent indirect speech utterances. Part of the between-condition difference was significantly accounted for by the perceived contextual congruency of the direct and indirect speech utterances but not by the utterances' acoustic characteristics or out-of-context vividness (see Section Parametric modulation analyses for the rationale behind these three measures).

The results suggest that the brain keeps track of context-based expectations of vivid acoustic information for direct speech rather than indirect speech utterances on a supra-segmental level. The increased 'top-down' brain activity in the right TVA when listening to monotonous direct speech utterances (contextually incongruent) as opposed to monotonous indirect speech utterances (contextually congruent) may reflect resolution of increased expectation violations for the former. It is noteworthy that this kind of expectation violation in the current experiment is very similar to the detection of discrepancies between semantic content (e.g., *She won the lottery jackpot*) and the emotional prosody of utterances (e.g., a sad emotional prosody) (Mitchell, 2006; Schirmer et al., 2004; Wittfoth et al., 2010). However, the latter have consistently been found to activate the left inferior frontal gyrus rather than the right STS areas. These studies employed explicit judgement tasks (e.g., a judgement task on the emotion of the utterances), which may encourage the processing of the *emotional* content rather than the *acoustic* aspects of the speech utterances. The left inferior frontal gyrus activation observed in these studies therefore seems to be linked more to the processing of the *emotional* discrepancies between speech semantics and prosody (e.g., valence-related emotional discrepancies; Schirmer et al., 2004; Wittfoth et al., 2010) rather than the processing of the *acoustic* discrepancies between the expected and the perceived prosody of the speech utterances. In contrast, the current experiment observed a highly-localized, right-lateralized locus of effect for the Direct-monotonous versus Indirect-monotonous contrast using a normal language comprehension task which does not require explicit metacognitive judgments on emotions. In this experimental setting, the observed right STS activations seem unlikely to reflect 'top-down' processing of emotional discrepancies. Moreover, studies have demonstrated that the decoding of the emotional content of affect prosody would most likely be represented by a network of brain regions other than the right STS. For example, the amygdala may be involved in registering the subjective value or relevance of emotionally charged intonation contours (Ethofer et al., 2009; Schirmer et al., 2008; Wiethoff et al., 2009); the right ventral frontoparietal cortex is perhaps involved in simulating the body states and visceral feelings associated with the emotions expressed by intonation contours (Adolphs et al., 2002; Pitcher et al., 2008; van Rijn et al., 2005); the basal ganglia may be involved in facilitating various aspects of the simulation process guided primarily by the ventral frontoparietal cortex (Pell and Leonard, 2003; Wittfoth et al., 2010); finally, the bilateral orbitofrontal cortex is perhaps involved in the explicit judgement of appraisal of affectively tinged tones of voice (Adolphs et al., 2002; Hornak et al., 2003; Sander et al., 2005; Wildgruber et al., 2004). In contrast, the comparison between the Direct-monotonous and the Indirect-monotonous conditions in the current experiment did not reveal increased BOLD signals in those emotion-related regions beyond the right TVA. Overall, it seems that the brain resolves violations of expectation of vivid acoustic information locally within the auditory cortex without communicating with the frontal cortex or other emotion-related regions. It is likely that the enhanced engagement of the right TVA when listening to monotonous direct speech utterances reflects the 'top-down' processing of the discrepancies between the expected vivid acoustic information and the perceived monotonous acoustic information of the direct speech utterances.

An alternative interpretation of the results adopts the notion of *perceptual simulations* (e.g., Barsalou, 1999): Listeners expect vivid vocal depictions (i.e., enriched supra-segmental acoustic information) for direct speech but not for indirect speech; they are more likely to mentally simulate such information (if it is not available in the stimulus itself), thereby effectively supplementing what they hear when listening to monotonously spoken direct speech utterances (contextually incongruent) but not monotonously spoken indirect speech utterances (contextually congruent). Such a 'voice simulation' mechanism may be one possible way in which the brain resolves the discrepancy between the expected (vivid) and the perceived (monotonous) acoustic information in monotonous direct speech utterances. As such, it is potentially complementary to the above 'expectation violation' account. Theoretically, voice simulations can be automatically implemented within the auditory cortex without necessarily communicating with the frontal cortex or other brain regions (Barsalou, 1999). This fits well with the highly localized activations of the right auditory cortex that were observed in the current study. Furthermore, a 'simulation' account also fits well, both theoretically and empirically, with our previous results on voice simulations in silent reading of direct versus indirect speech. Combining fMRI and eye-tracking, our previous study showed that the right TVA became more active when individuals silently read a direct speech as opposed to an indirect speech quotation, suggesting that readers are more likely to mentally simulate the reported speaker's voice when reading the former (Yao et al., 2011). Behaviourally, such voice simulations were found to be reflected in articulation (oral reading) and eye-movement patterns (silent reading): Readers automatically adjusted their reading rates in accordance with the contextually-implied speaking rate (fast vs. slow) of the reported speaker during oral and silent reading of direct speech but not of indirect speech (Yao and Scheepers, 2011). Both studies suggest that voice simulations are an integral part of language comprehension of direct speech. It is therefore reasonable to assume that voice simulations also take place when listening to direct speech, particularly when presented in an 'incongruent', 'awkward' monotonous manner. Indeed, further support for this assumption stems from the striking similarity between brain activation patterns observed in the current study and in the previous fMRI study on silent reading (Yao et al., 2011; also see Section Comparing brain activation patterns between studies). Taken together, it seems plausible that the increased right STS activations when listening to monotonous direct speech reflect enhanced mental simulations of vivid vocal depictions (i.e., enriched supra-segmental acoustic information), as an integral part of a normal direct speech comprehension process.

The current experiment sheds new light on the nature of the "inner voice" representations that are mentally simulated in language comprehension of direct as opposed to indirect speech. In the previous silent reading experiment (Yao et al., 2011), there was no experimentally manipulated auditory stimulation which could be used as a reference to the representational nature of the observed "inner voice" activations. The current study resolved this issue by using monotonously spoken direct versus indirect speech utterances in which the enriched supra-segmental vocal information is scarcely available while the (sub)-segmental acoustic information is intact. This manipulation suggested that the mentally simulated (as opposed to acoustically perceived) voice representations must be supra-segmental rather than (sub)-segmental in nature. Intriguingly, the "inner voice" activations observed in the current investigation and in Yao et al. (2011) were located in virtually the same brain areas. Reconciling the findings of the two studies, we infer that the "inner voice" we observed in silent reading of direct as opposed to indirect speech may also entail supra-segmental acoustic representations.

Moreover, the notion of mental simulations of supra-segmental acoustic information is consistent with the findings from "bottom-up" auditory stimulation studies; the latter suggest that the right

auditory cortical areas are indeed specialised in processing slower (e.g., spectral) variations of pitch such as speech melody (Scott et al., 2000), musical melody (Grandjean et al., 2005; Patterson et al., 2002; Zatorre et al., 1994, 2002) and emotional prosody (Grandjean et al., 2005; Mitchell et al., 2003; Wiethoff et al., 2008; Wildgruber et al., 2005). Most importantly, mental simulations of supra-segmental acoustic information in language comprehension would fit well with the notion of direct speech as vivid demonstration — in which vivid depictions of the quoted speaker's voice are characterised in terms of enriched supra-segmental acoustic information.

At a broader level, the current findings also provide novel insights into speech perception. Speech perception is not a passive information processing mechanism in that it involves both bottom-up and top-down processes. The top-down influences in speech perception have been documented in various respects: Listeners use prior lexical knowledge (Kolinsky and Morais, 1996; Mattys, 1997; Pitt and Shoaf, 2002) or perceptual experience (Clarke and Garrett, 2004; Davis et al., 2005; Goldinger et al., 1999) in perceptual grouping of speech, segmenting connected speech (Davis et al., 2002; Mattys et al., 2005), perceptual learning of distorted speech (Davis et al., 2005; Hervais-Adelman et al., 2008), and perceiving speech categorically (Norris et al., 2003; Pisoni and Tash, 1974). However, such top-down influences have mostly been documented in terms of low-level "speech recognition" — researchers have mainly focused on how listeners use prior lexical knowledge or perceptual experience to interpret *distorted*, *unintelligible*, or *ambiguous* speech. This neglects the fact that speech perception also involves comprehension of the recognised linguistic information. It is currently less clear whether and how on-line comprehension of auditory linguistic information also influences upcoming *intelligible* and *unambiguous* speech perception. Moreover, many studies revolved around the top-down interactivity at the *(sub)-segmental* level (e.g., word interpretation) whereas the top-down influences at the *supra-segmental* level (e.g., emotional prosody) have received limited attention. The current study demonstrated top-down activations of *supra-segmental* acoustic representations during *intelligible* and *unambiguous* speech perception of direct versus indirect speech utterances. It provides evidence that during natural speech perception, top-down interpretations of incoming acoustic signals routinely take place even at the supra-segmental level. More importantly, such top-down interpretations are modulated as a function of linguistically/pragmatically different reporting styles (direct vs. indirect speech). Our findings emphasise that in addition to prior lexical knowledge and perceptual experience, other linguistic factors such as reporting style should also be considered in modelling the top-down interactivity of natural speech perception.

Conclusions

The current study shows that listeners routinely expect vivid depictions for direct speech but rarely for indirect speech; they spontaneously engage in mental simulations of vivid vocal depictions while listening to monotonously spoken direct speech rather than to monotonously spoken indirect speech. The findings replicate our previous findings of an "inner voice" during silent reading of direct as opposed to indirect speech, but within a different modality. This highlights the universality of such voice simulation process in understanding direct speech. Furthermore, it provides evidence that the nature of the mentally simulated "inner voice" entails supra-segmental acoustic representations. It also verifies the neural correlates of such voice simulation process, which include the anterior, the middle and the posterior parts of the right STS brain areas. Future research would be sought to specify the exact function of the involved brain areas during such simulation process. Finally, from a broader perspective, the current findings extend the scope in modelling natural, intelligible speech perception, emphasising that comprehension-driven top-down influences at the supra-segmental level should also be considered.

Acknowledgments

We thank F. Crabbe for support in fMRI scanning. The study was supported by the Institute of Neuroscience and Psychology and the Centre for Cognitive Neuroimaging, University of Glasgow.

References

Adolphs, R., Damasio, H., Tranel, D., 2002. Neural systems for recognition of emotional prosody: a 3-D lesion study. Emotion 2 (1), 23–51.
Barsalou, L.W., 1999. Perceptual symbol systems. Behav. Brain Sci. 22 (4), 577–609.
Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403 (6767), 309–312.
Boersma, P., Weenink, D., 2010. Praat: doing phonetics by computer [computer program] version 5.1.43retrieved 4 August 2010 from http://www.praat.org/2010.
Clark, H.H., Gerrig, R.J., 1990. Quotations as demonstrations. Language 66 (4), 764–805.
Clarke, C.M., Garrett, M.F., 2004. Rapid adaptation to foreign-accented English. J. Acoust. Soc. Am. 116 (6), 3647–3658.
Davis, M.H., Marslen-Wilson, W.D., Gaskell, M.G., 2002. Leading up the lexical garden path: segmentation and ambiguity in spoken word recognition. J. Exp. Psychol. Hum. Percept. Perform. 28 (1), 218–244.
Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., McGettigan, C., 2005. Lexical information drives; perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. Gen. 134 (2), 222–241.
Ethofer, T., Kreifelts, B., Wiethoff, S., Wolf, J., Grodd, W., Vuilleumier, P., et al., 2009. Differential influences of emotion, task, and novelty on brain regions underlying the processing of speech melody. J. Cogn. Neurosci. 21 (7), 1255–1268.
Goldinger, S.D., Kleider, H.M., Shelley, E., 1999. The marriage of perception and memory: creating two-way illusions with words and voices. Mem. Cognit. 27 (2), 328–338.
Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., et al., 2005. The voices of wrath: brain responses to angry prosody in meaningless speech. Nat. Neurosci. 8 (2), 145–146.
Hervais-Adelman, A., Davis, M.H., Johnsrude, I.S., Carlyon, R.P., 2008. Perceptual learning of noise vocoded words: effects of feedback and lexicality. J. Exp. Psychol. Hum. Percept. Perform. 34 (2), 460–474.
Hornak, J., Bramham, J., Rolls, E.T., Morris, R.G., O'Doherty, J., Bullock, P.R., et al., 2003. Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. Brain 126 (Pt 7), 1691–1712.
Johnsrude, I.S., Penhune, V.B., Zatorre, R.J., 2000. Functional specificity in the right human auditory cortex for perceiving pitch direction. Brain 123, 155–163.
Kolinsky, R., Morais, J., 1996. Migrations in speech recognition. Lang. Cognit. Process. 11 (6), 611–619.
Mattys, S.L., 1997. The use of time during lexical processing and segmentation: a review. Psychon. Bull. Rev. 4 (3), 310–329.
Mattys, S.L., White, L., Melhorn, J.F., 2005. Integration of multiple speech segmentation cues: a hierarchical framework. J. Exp. Psychol. Gen. 134 (4), 477–500.
Mitchell, R.L., 2006. How does the brain mediate interpretation of incongruent auditory emotions? The neural response to prosody in the presence of conflicting lexico-semantic cues. Eur. J. Neurosci. 24 (12), 3611–3618.
Mitchell, R.L.C., Elliott, R., Barry, M., Cruttenden, A., Woodruff, P.W.R., 2003. The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. Neuropsychologia 41 (10), 1410–1421.
Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cogn. Psychol. 47 (2), 204–238.
Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., Griffiths, T.D., 2002. The processing of temporal pitch and melody information in auditory cortex. Neuron 36 (4), 767–776.
Pell, M.D., Leonard, C.L., 2003. Processing emotional tone from speech in Parkinson's disease: a role for the basal ganglia. Cogn. Affect. Behav. Neurosci. 3 (4), 275–288.
Pisoni, D.B., Tash, J., 1974. Reaction-times to comparisons within and across phonetic categories. Percept. Psychophys. 15 (2), 285–290.
Pitcher, D., Garrido, L., Walsh, V., Duchaine, B.C., 2008. Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. J. Neurosci. 28 (36), 8929–8933.
Pitt, M.A., Shoaf, L., 2002. Linking verbal transformations to their causes. J. Exp. Psychol. Hum. Percept. Perform. 28 (1), 150–162.
Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., et al., 2005. Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. NeuroImage 28 (4), 848–858.
Schirmer, A., Zysset, S., Kotz, S.A., Yves von Cramon, D., 2004. Gender differences in the activation of inferior frontal cortex during emotional speech perception. NeuroImage 21 (3), 1114–1123.
Schirmer, A., Escoffier, N., Zysset, S., Koester, D., Striano, T., Friederici, A.D., 2008. When vocal processing gets emotional: on the role of social orientation in relevance detection by the human amygdala. NeuroImage 40 (3), 1402–1410.
Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J.S., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.
Van Essen, D.C., 2002. Windows on the brain: the emerging role of atlases and databases in neuroscience. Curr. Opin. Neurobiol. 12 (5), 574–579.
Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H., 2001. An integrated software suite for surface-based analyses of cerebral cortex. J. Am. Med. Inform. Assoc. 8 (5), 443–459.

van Rijn, S., Aleman, A., van Diessen, E., Berckmoes, C., Vingerhoets, G., Kahn, R.S., 2005. What is said or how it is said makes a difference: role of the right fronto-parietal operculum in emotional prosody as revealed by repetitive TMS. Eur. J. Neurosci. 21 (11), 3195–3200.

Wiethoff, S., Wildgruber, D., Kreifelts, B., Becker, H., Herbert, C., Grodd, W., et al., 2008. Cerebral processing of emotional prosody — influence of acoustic parameters and arousal. NeuroImage 39 (2), 885–893.

Wiethoff, S., Wildgruber, D., Grodd, W., Ethofer, T., 2009. Response and habituation of the amygdala during processing of emotional prosody. Neuroreport 20 (15), 1356–1360.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., et al., 2004. Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. Cereb. Cortex 14 (12), 1384–1389.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., et al., 2005. Identification of emotional intonation evaluated by fMRI. NeuroImage 24 (4), 1233–1241.

Wittfoth, M., Schroder, C., Schardt, D.M., Dengler, R., Heinze, H.J., Kotz, S.A., 2010. On emotional conflict: interference resolution of happy and angry prosody reveals valence-specific effects. Cereb. Cortex 20 (2), 383–392.

Yao, B., Scheepers, C., 2011. Contextual modulation of reading rate for direct versus indirect speech quotations. Cognition 121 (3), 447–453.

Yao, B., Belin, P., Scheepers, C., 2011. Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. J. Cogn. Neurosci. 23 (10), 3146–3152.

Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A., 1992. Lateralization of phonetic and pitch discrimination in speech processing. Science 256 (5058), 846–849.

Zatorre, R.J., Evans, A.C., Meyer, E., 1994. Neural mechanisms underlying melodic perception and memory for pitch. J. Neurosci. 14 (4), 1908–1919.

Zatorre, R.J., Belin, P., Penhune, V.B., 2002. Structure and function of auditory cortex: music and speech. Trends Cogn. Sci. 6 (1), 37–46.

Zwaan, R.A., 2004. The immersed experiencer: toward an embodied theory of language comprehension. Psychol. Learn. Motiv. Adv. Res. Theory 44, 35–62.