Hemispheric contributions for processing pitch and speech rate cues to emotion: fMRI data

Chinar Dara & Marc D. Pell

School of Communication Sciences & Disorders, McGill University, Montreal, Canada
chinar.dara@mail.mcgill.ca; marc.pell@mcgill.ca

Abstract

To determine the neural mechanisms involved in vocal emotion processing, the current study employed functional magnetic resonance imaging (fMRI) to investigate the neural structures engaged in processing acoustic cues to infer emotional meaning. Two critical acoustic cues – pitch and speech rate – were systematically manipulated and presented in a discrimination task. Results confirmed that a bilateral network constituting frontal and temporal regions is engaged when discriminating vocal emotion expressions; however, we observed greater sensitivity to pitch cues in the right mid superior temporal gyrus/sulcus (STG/STS), whereas activation in both left and right mid STG/STS was observed for speech rate processing.

Index terms: prosody, fMRI, pitch processing, speech rate processing, emotion comprehension

1. Introduction

Communication of emotions from *speech prosody* has been a topic of research for last three decades and yet a comprehensive understanding of the underlying neural mechanisms is still pending. Although numerous studies have emphasized the importance of the right hemisphere for processing emotional prosody, the right hemisphere structures are increasingly acknowledged to be part of a bilaterally distributed neural network engaged at different stages of processing emotional prosody [1,2,3]. These interactive and possibly complementary processes of the left and right hemisphere remain a matter of debate.

Emotionally salient speech is expressed via modulation of multiple physical parameters, such as, level and range of pitch, speech rate, intensity, and voice quality [8]. A number of investigations using functional magnetic resonance imaging (fMRI) as well as behavioural studies of brain-damaged patients have associated the right superior temporal gyrus/sulcus (STG/STS) with the processing of acoustic cues like pitch, duration and intensity [2,3,4,5]. Most studies on vocal emotion perception show that the STG/STS are sensitive to acoustic cue processes; however, do not allow strong conclusions about how each cue is processed, nor whether there are laterality differences in how STG/STS responds to different acoustic cues conveying emotion.

Studies on auditory processing and speech perception for non-emotional stimuli imply that acoustic processing derive hemispheric differences [6,7]. Zatorre (2002) proposed that the right hemisphere has low temporal resolution and thus seems to be specialized with pitch processing, and in contrast, the left hemisphere has the ability to process rapidly changing information [7]. A more recent description has proposed that the two hemispheres are engaged in acoustic processing in different time domains [6,7]. Poppel (2003) proposed the 'asymmetric sampling in time' hypothesis, according to which the auditory processing occurs on short (20-50 ms) and long (150-250 ms) time scales. This model hypothesizes that the left hemisphere extracts information from short temporal windows (20-50 ms scale) and the right hemisphere from long temporal window (150-250 ms scale).

Further investigations of neural substrates at the stage of acoustic cue processing for vocal emotion comprehension are clearly warranted. Here, two critical cues to decode emotions from the voice – pitch and speech rate – were systematically perturbed and presented in a discrimination task in a mixed block/event-related fMRI paradigm. Participants listened to two utterances that differed only in pitch or speech rate (or not at all) and then made an emotion judgment. This approach can allow us to directly compare the neural activity observed in processing pitch and speech rate to make emotion inferences. In addition, based on the evidence from previous studies, region of interest analysis was conducted in the voice sensitive areas in the temporal regions of both hemispheres to compare the involvement of STG/STS in processing of pitch and speech rate cues.

2. Methods

2.1. Participants

A group of 12 English speaking university students were recruited for this fMRI study. The participant group was balanced for gender with a mean age 22.8 years and education 15.8 years. All participants were right handed, had normal hearing and reported no history of speech, language or neurological disorders.

2.2. Stimulus Material

The *base* stimuli comprised of language-like pseudo utterances (e.g., Someone migged the pazing) taken from an inventory of recordings following a comprehensive perceptual validation study [9,10]. Based on evidence that both pitch and speech rate changes influence how vocal expressions of happiness, sadness, fear and neutral are recognized in speech [9], this study focused on these four emotion types. For each utterance, both pitch and speech rate were increased (+25%) or decreased (-25%) using Praat software by manipulating the mean pitch and duration of the whole utterance by a factor of 1.25 and 0.75 independently and in combination.

2.3. Experimental Paradigm

Participants were required to assess emotion information from a pair of nonsense utterances that differed in pitch or speech rate measures in a discrimination task. Each trial consisted of identical utterances spoken in one of the emotion categories (happy, sad, fear, neutral) by the same speaker except that pitch or speech rate of the two utterances was manipulated. This task comprised of two manipulation conditions (*pitch* and *speech rate*) and one control (*no change*) condition. For *pitch* manipulation condition, only mean pitch of the whole utterance was manipulated such that, the mean pitch of one utterance would be higher than the other. Similarly, for *speech rate* condition, speaking rate of one of the utterances in a pair was faster than the other utterance. For the control condition, both the utterances in the pair were same. All trials were blocked by emotion and within each block three cue conditions (pitch, speech rate, no change) were presented in a pseudo-randomized order. Participants were required to listen to the pair of utterances and judge which utterance provides a better exemplar for the pre-defined emotion category from a choice of three alternatives: "1st", "2nd" or "same".

The stimuli were presented to the participants in a mixed event-related/blocked design via Presentation software (Neurobehavioral systems). Every block started by two silent scans followed by displaying the emotion category on the screen to inform the participants about the change in the block. Every trial would start with a cross in the centre of the screen and subsequently, the two utterances were presented serially with an inter-stimulus interval of 100ms. Onset of the stimuli presentation was systematically varied or "jittered" to maximize the likelihood of obtaining the peak of the hemodynamic response for each trial. The participants were instructed to fixate on that cross while listening to the utterances to minimize artifacts due to eye movements. Participants responded using a 3 button response box as soon as the prompt question screen was displayed after the presentation of the utterances. For the baseline "null" condition, the block would start with the screen stating "Please Rest" to inform the patients that they would not be listening to any auditory stimuli. However, for the null trials, in a manner similar to all other blocks, participants were required to fixate their eyes on the cross and press the first button of the response box once the prompt screen appeared. Before the experimental session in the scanner, each participant was provided with a verbal explanation of the task and instructions followed by a short practice session (15 trials) outside the scanner.

2.4. Data Acquisition

Scanning was performed using a Siemens Magnetom 3T syngo magnetic resonance (MR) scanner. A high resolution T1-weighted scan was obtained for anatomical localization. Two out of three functional runs were carried out prior to the T1-weighted scan and one run following the anatomical scan. Structural images were obtained parallel to the plane intersecting anterior and posterior commissures (AC-PC plane) resulting in 160 sagittal slices covering whole brain (1 mm thickness, field of view (FOV) = 256mm^2, TR = 2300ms, TE = 2.98ms). For the functional runs, 42 contiguous T2*-weighted scans were acquired rapidly in a descending manner without any gap between the slices. Acquisition of slices was acquired in a sparse-sampling manner, in which participants listened to the stimuli and gave judgments during the silences between image acquisition to prevent scanner noise from interfering with the auditory stimuli. The design matrix acquired was 64 x 64 with a FOV of 224 mm^2 resulting in the voxel size = 3.5 mm^3. For functional scans, a gradient-echo EPI sequence was used with a TE of 30 ms, TR of 12100 ms, a delay in TR = 9580 ms and a acquisition bandwidth of 2442 Hz/Px.

2.5. Data Analysis

Data processing was performed using Neurolens software and using a series of scripts in MATLAB. The data were corrected for motion artifacts by realigning all blood-oxygen-level-dependent (BOLD) images within each run with the third frame of the first functional scan. Movement parameters (translation and rotation) were inspected manually to exclude frames that showed a movement of 1mm or 1 degree from the previous frame. In addition, the images were spatially smoothed (or low-pass filtered) with a 6-mm full-width at half maximum (fwhm) isotropic Gaussian Kernel. The image analysis was based on a least-squares estimation using the general linear model, $Y=X\beta + \varepsilon$.

To examine the neural activity for pitch and speech rate, data for the three cue conditions were averaged across emotions. Three planned comparisons were entered into the analysis based on our cue conditions: pitch versus null, speech rate versus null and no change versus null. Subsequently, all runs within a subject were combined using a fixed-effects general linear model. Before group statistical maps for each contrast were generated, both anatomical and functional scans of each subject underwent a linear transformation into standard MNI/Talairach stereotaxic coordinate space, using MNI 305 template [11]. Only after the transformations, a mixed-effects linear model was specified in *multistat* to calculate average data across subjects. The data were smoothed with an fwhm Gaussian filter of 8 mm so that the ratio of the random-effects variance divided by the fixed-effects variance results in approximately 100 degrees of freedom. Lastly, the program *stat_summary* generated the minimum *p*-values among those computed with a Bonferroni correction, random field theory and discrete local maxima [12]. Peaks that surpassed an uncorrected *p*-value of .001 (t=3.17) are reported with a cluster size of minimum 693 voxels.

Additionally, region of interest (ROI) analysis were carried out on a set of anatomically constrained 5 mm radius regions that displayed peak activity in group-contrasted BOLD images (as shown in Table 1). The ROI analysis was restricted to the activations in the temporal cortex as that has been implicated in previous studies of emotional prosody processing along with other reports on speech perception [2,3,5,6,7]. Percentage change in the BOLD signal for each ROI was computed for the pitch and speech rate manipulation condition (against null) of all participants individually using in-house software. These values within each ROI were then submitted to repeated measures ANOVA to compare activation between cue conditions (pitch and speech rate);

and hemisphere (left and right). Cue conditions and hemisphere were treated as fixed within-subjects variables.

3. Results

In this section, first the fMRI data from the 'no change' condition (comparison: no change vs. null) will be explored to outline the basic functional network engaged in a simple discrimination task. Subsequently, two planned comparisons were entered into the analysis for the current study to reveal the pattern of activation for processing pitch and speech rate; contrasts against silence baseline: (i) pitch vs. null; (ii) speech rate vs. null.

3.1. Basic functional network in a discrimination task

Simple discrimination in the *no change* condition (when both utterances in a pair were same) compared to null recruited a number of regions in young healthy listeners, which included bilateral inferior frontal gyrus (IFG), bilateral middle frontal gyrus (MFG), right superior frontal gyrus (SFG), bilateral insular gyrus, bilateral STG (with clusters extending towards planum temporal, planum polare (PP) and STS), bilateral anterior transverse temporal gyrus (TTG), left postcentral, left precentral, and left supramarginal gyrus.

3.2. Pitch and speech rate manipulation effects

Conditions in which participants discriminated utterances with either pitch or speech rate differences; both engaged bilateral brain areas in temporal and frontal cortex. Whereas both temporal and frontal regions get activated, the peaks in the superior temporal region form part of large clusters with greater signal strength. The pitch manipulation condition (compared to null) exhibited extensive activation in bilateral temporal regions, especially in mid STG which appeared stronger in the right. Other temporal areas activated for the pitch condition were anterior and posterior STG (with clusters extending to planum temporal, postcentral gyrus and STS), right middle temporal gyrus (MTG) and left anterior and middle TTG. In addition, frontal regions were activated although with smaller cluster size and less strength; bilateral inferior frontal areas, right MFG, right precentral and insular gyrus; and right cerebellum (culmen and declive).

Similarly, speech rate manipulation condition (compared to null) revealed much larger clusters in the temporal regions, especially in the mid STG which also appeared to be stronger in the right hemisphere. In addition, bilateral anterior STG (with cluster extending to PP) and posterior STG (with clusters extending to planum temporal, postcentral gyrus, and STS), bilateral inferior and superior frontal gyrus, right MFG, left insular gyrus and left anterior and middle TTG.

3.3. Region of interest analysis

Analyzing each participant's percentage change in the BOLD signal of the pitch and speech rate manipulation extracted from the whole brain analysis from the temporal regions of both hemispheres provided a series of interesting effects. A 2 x 2 ANOVA evaluated the difference in the percentage change in signal in either of the Hemispheres' (left, right) due to cue manipulation (pitch and speech rate) for each of the peaks observed in the temporal cortex. Most importantly, a significant interaction of hemisphere by cue was observed for the mid STG, $F(1,11) = 5.6$, $p<.05$. Percentage signal change differences for pitch and speech rate conditions were most pronounced in the mid STG/STS region. Signal change for the pitch condition was significantly greater in right hemisphere when compared to the left hemisphere. Also, the percentage change in the signal in mid STG/STS was greater for the pitch condition when compared to the speech rate condition. On the other hand, there was no difference in the signal change in left or right mid STG for speech rate condition.

4. Discussion

The current study was undertaken to explore if the bilateral activations in vocal emotion comprehension is a result of differential pitch and speech rate processing. For the purposes of the present study, pitch and speech rate cues were independently manipulated of pseudo-language like utterances and presented in a discrimination task. As shown previously, an extensive bilateral network including frontal and temporal regions is recruited in discriminating emotion information from voice. The main finding of our study is the differences in the sensitivity of right and left mid STG/STS to pitch and speech rate manipulations.

Both pitch and speech rate manipulations were effective in deriving an increase in the hemodynamic BOLD signal

Table 1. Overview of significant activation peaks in random-effects analysis for which ROI analysis was conducted

	Pitch Discrimination								Speech Rate Discrimination							
	Right				Left				Right				Left			
Site	*x*	*y*	*z*	*t*	*x*	*y*	*z*	*t*	*x*	*y*	*z*	*t*	*x*	*y*	*z*	*t*
ant. STG/STS	62	2	-5	5.4	-55	10	-7	6.2	62	2	-5	5.4				
mid. STG/STS	62	-18	5	8.0	-59	-11	4	6.2	62	-17	4	7.3	-63	-14	2	6.4
post. STG/STS	53	-33	7	7.3	-62	-27	6	7.6	59	-34	7	7.0	-65	-27	7	7.9
PP/ant. STG	56	17	-10	4.3	-50	8	-7	4.3	60	12	-6	4.3	-61	3	-4	4.8

Regions of peak neural activity associated with pitch and speech rate discrimination (compared to null). Coordinates are reported in MNI305/Talairach space. STG – superior temporal gyrus, STS- superior temporal sulcus, ant. – anterior, mid. – middle, post. – posterior, PP – planum polare

from anterior to posterior regions of STG/STS of the left and right hemisphere. Interestingly, the BOLD response for pitch as well as speech rate processing in right mid STG/STS was stronger than anterior and posterior STG/STS activations (see Figure 1). The BOLD response was observed bilaterally for both cue conditions, however, the percentage change in the signal in right mid STG was greater for pitch manipulations when compared to change in signal for speech rate manipulations. Furthermore, right mid STG/STS seemed to be more sensitive than the left mid STG/STS for pitch processing, whereas both left and right mid STG/STS were recruited for speech rate processing.

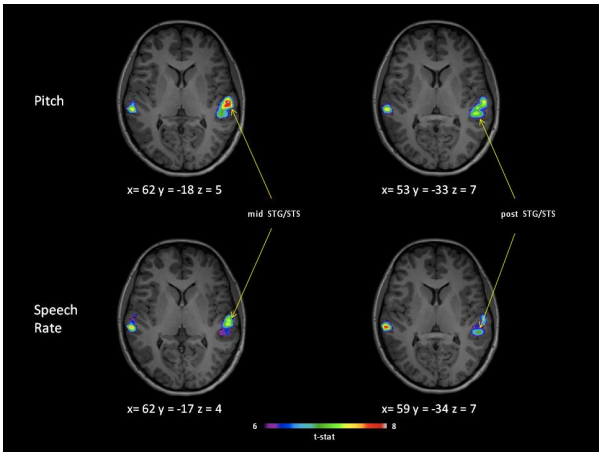

Figure 1. Random effects BOLD activation map obtained from comparison of discrimination judgments of pitch and speech rate conditions relative to null. These axial slices reveal increased activation in right mid STG/STS relative to the posterior STG/STS region for pitch condition as well relative to corresponding activation pattern in mid STG/STS in speech rate condition.

These results are in line with previous findings suggesting acoustic cue processing derives differential hemispheric contributions [6,7]. Many reports have elaborated on the specialized role of the right auditory cortices in processing complex pitch patterns over a longer period of time [2,3,5,13]. Large clusters in the right STG/STS region for pitch condition demonstrated that evaluating emotion information from pitch measures seems to recruit right mid STG to a greater extent. In contrast, speech rate processing seems to derive STG/STS region of both hemispheres. Instead of viewing differential hemispheric contributions for processing speech rate information, it seems that both left and right STG/STS regions are required to extract emotion information from speech rate cues.

Using manipulations in speech we have demonstrated the subtle differences in hemispheric asymmetry in processing pitch and speech rate cues to evaluate emotion information. Our findings provide direct evidence for the notion that both hemispheres engage in interactive and complimentary manner at the stage of acoustic cue processing for vocal emotion comprehension although they are differentially weighted based on the acoustic cues.

5. References

[1] Adolphs, R., Damasio, H., & Tranel, D. (2002). Neural systems for recognition of emotional prosody: a 3-D lesion study. *Emotion, 2*(1), 23-51.

[2] Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., et al. (2004). Hemispheric roles in the perception of speech prosody. *NeuroImage, 23*, 344-357

[3] Kotz, S. A., Meyer, M., Alter, K., Besson, M., von Cramon, D. Y., & Friederici, A. D. (2003). On the lateralization of emotional prosody: An event-related functional MR investigation. *Brain and language, 86*, 366 - 376.

[4] Beaucousin, V., Lacheret, A., Turbelin, M.-R., Morel, M. l., Mazoyer, B., & Tzourio-Mazoyer, N. (2007). FMRI study of emotional speech comprehension. *Cerebral Cortex, 17*(2), 339-352.

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

[6] Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*, 245 - 255.

[7] Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences, 6*(1), 37 - 46.

[8] Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychol. Bull., 99*(2), 143-165.

[9] Dara, C., & Pell, M. D. (In Review). Influence of pitch and speech rate manipulation on vocal emotion recognition.

[10] Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: a comparison of four languages. *Journal of Phonetics, 37,* 417-435.

[11] Talairach, J., & Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain* (3rd ed. ed.). New York: Thieme Medical

[12] Worsley, K. J. (2005). An improved theoretical P value for SPMs based on discrete local maxima. *Neuroimage, 28*, 1056-1062.

[13] Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24 - 30.