NeuroImage 141 (2016) 31–39

Contents lists available at ScienceDirect

NeuroImage

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

Exploring the anatomical encoding of voice with a mathematical model of the vocal system

M. Florencia Assaneo [a,b], Jacobo Sitt [c,d,e,f,g], Gael Varoquaux [c,h], Mariano Sigman [i,j], Laurent Cohen [e,f,g,k], Marcos A. Trevisan [a,*]

[a] Department of Physics, University of Buenos Aires-IFIBA CONICET, Ciudad Universitaria, Pab. 1, 1428EGA, Buenos Aires, Argentina
[b] Department of Psychology, New York University, New York, NY 10003, USA
[c] INSERM, Cognitive Neuroimaging Unit, Gif sur Yvette, France
[d] Commisariat à l'Energie Atomique, Direction des Sciences du Vivant, I2BM, NeuroSpin Center, Gif sur Yvette, France
[e] INSERM U1127, Institut du Cerveau et de la Moelle Épinière, Paris, France
[f] CNRS UMR 7225, Institut du Cerveau et de la Moelle Épinière, Paris, France
[g] Sorbonne Universités, UPMC Univ Paris 06, Paris, France
[h] INRIA Parietal, Neurospin, bât 145, CEA Saclay, France
[i] Integrative Neuroscience Lab, Physics dept. UBA-IFIBA CONICET, Pab. 1, 1428EGA Buenos Aires, Argentina
[j] University Torcuato Di Tella, Alm. Juan Saenz Valiente 1010, C1428BIJ Buenos Aires, Argentina
[k] AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Departament of Neurology, Paris, France

ARTICLE INFO

Article history:
Received 16 February 2016
Accepted 14 July 2016
Available online 17 July 2016

Keywords:
Biomechanical model of the vocal system
Neural coding of voice
Voice identity
Auditory cortex

ABSTRACT

The faculty of language depends on the interplay between the production and perception of speech sounds. A relevant open question is whether the dimensions that organize voice perception in the brain are acoustical or depend on properties of the vocal system that produced it. One of the main empirical difficulties in answering this question is to generate sounds that vary along a continuum according to the anatomical properties the vocal apparatus that produced them. Here we use a mathematical model that offers the unique possibility of synthesizing vocal sounds by controlling a small set of anatomically based parameters.

In a first stage the quality of the synthetic voice was evaluated. Using specific time traces for sub-glottal pressure and tension of the vocal folds, the synthetic voices generated perceptual responses, which are indistinguishable from those of real speech.

The synthesizer was then used to investigate how the auditory cortex responds to the perception of voice depending on the anatomy of the vocal apparatus. Our fMRI results show that sounds are perceived as human vocalizations when produced by a vocal system that follows a simple relationship between the size of the vocal folds and the vocal tract. We found that these anatomical parameters encode the perceptual vocal identity (male, female, child) and show that the brain areas that respond to human speech also encode vocal identity.

On the basis of these results, we propose that this low-dimensional model of the vocal system is capable of generating realistic voices and represents a novel tool to explore the voice perception with a precise control of the anatomical variables that generate speech. Furthermore, the model provides an explanation of how auditory cortices encode voices in terms of the anatomical parameters of the vocal system.

© 2016 Elsevier Inc. All rights reserved.

Introduction

Speech perception builds on a cortical structure, extended broadly across the auditory cortex and localized close to the superior temporal sulcus, which is sensitive to voices. This cortical region responds to speech but also to other utterances like laughing, coughing and sighing, suggesting that it is more generally tuned to a specific human vocal system (Belin et al., 2000, Mesgarani et al., 2014).

Voice perception has been previously investigated by the analysis of brain responses to various manipulations of vocal stimuli, including the comparison of forward and reversed speech (Binder et al., 2000; Dehaene-Lambertz et al., 2002) or manipulation of parameters of real speech such as duration, pitch and formants transitions between consonants and vowels (Kühnis et al., 2013, Chang et al., 2010). These studies have demonstrated that the continuum of acoustically varying sounds of speech is represented in the brain as perceptual categories. Such parsing of the acoustical continuum allows for the recognition of phonemes (Lee et al., 2012, Chang et al., 2010) and speaker's identity (Latinus et al., 2013).

[*] Corresponding author.
E-mail address: marcos@df.uba.ar (M.A. Trevisan).

http://dx.doi.org/10.1016/j.neuroimage.2016.07.033
1053-8119/© 2016 Elsevier Inc. All rights reserved.

Thus inferences about how human vocal sounds are processed in the brain rely mostly on theories of auditory perception. However, the faculty of language depends on the interplay between the production and perception of speech sounds. According to the motor theories of speech perception, articulatory gestures are the actual basis of the representation and perception of speech sounds, consisting either in abstract 'intended gestures' specific to the speech domain (Liberman and Mattingly, 1985), or in the actual set of articulatory movements (Fowler, 2010). Recently, important findings led to the conclusion that the brain processes complex information such as the speaker's identity and the articulatory features of the vocal system even at the level of auditory cortex (Bonte et al., 2014; Correia et al., 2015). A relevant question is whether the dimensions that organize voice perception at low levels of processing consist of acoustical, motor, articulatory or anatomical properties of the vocal system that produced it.

One of the main empirical difficulties in addressing this question is to generate sounds that vary along a continuum according to the physical properties the vocal apparatus. Rather than stretching the duration of sounds, or increasing their pitch, one should be able to generate synthetic voice stimuli by controlling anatomical and physiological parameters of the vocal system.

Although the vocal anatomy and physiology are inherently complex, mathematical models capture a wide range of acoustic features of the human voice, and they can be tuned to synthesize sounds that reproduce its main spectral and temporal properties (Story and Titze, 1998; Story, 2013, 2005). These synthetic sounds can effectively convey a recognizable phonetic content (Bunton and Story, 2009; Story and Bunton, 2010); nevertheless, whether these sounds could be perceived as "human" or elicit brain responses comparable to real speech, remain unknown. One encouraging example of this comes from the field of birdsong, where by tuning the parameters of a low-dimensional model it was possible to produce synthesized songs that activated highly selective neurons to the bird's own song, neurons that barely respond to any other sounds, including conspecific songs or slight perturbations of the own song (Amador et al., 2013).

Here, the parameters of a low-dimensional model of the vocal folds (Assaneo and Trevisan, 2013; Lucero and Koenig, 2005) and the vocal tract (Story, 2013, 2005) are controlled to generate utterances with phonological content. These synthetic sounds are compared with real human voices showing that they are perceptually indistinguishable. The synthesizer is then used to test the hypothesis that brain responses to voices in the auditory cortex are tuned to specific anatomical parameters of the vocal system.

Methods

Articulatory voice synthesizer

The human vocal system consists of two main anatomical blocks: the vocal folds and the vocal tract. The vocal folds are a pair of membranes located at the glottis. During the production of vowels, the air coming from the lungs transfers energy to the vocal folds, giving rise to oscillations. Sound is produced by the pressure perturbations generated by these oscillations, determining acoustical properties of the vowel such as its pitch, jitter and shimmer. The vocal tract acts as a wave guide for the sound, emphasizing specific resonant frequencies (formants) that depend on its shape and length, which defines the identity of each vowel. In other cases, the vocal tract itself acts as the sound source. For instance, a turbulent sound source is created as the air is forced to pass through a constriction of the tract, giving rise to the fricative consonants such as /s/ or /f/. Other consonants such as the stops /p/ or /t/ are created when the vocal tract rapidly passes from a completely occluded to an open configuration.

The model of the vocal system consists of the differential equations describing the dynamics of the vocal folds and a wave-reflection vocal tract model.

A two-mass model was used to approximate the dynamics of the vocal folds: the cover of each membrane is modeled as two masses m_1 and m_2, one on top of the other, connected with each other and with the glottal tissue. The following are the equations of motion for the displacements x_1 and x_2, that measure the distance of each of the two masses of one of the membranes to the sagittal plane (see Fig. 1A):

$$\begin{aligned} x_i' &= y_i \\ y_i' &= \mathbf{Q}/m_i\left[f_i\left(l_g, d_i, \boldsymbol{P_s}\right) - K_i(x_i) - B_i(x_i, y_i) - \mathbf{Q}k_c(x_i - x_j)\right] \end{aligned} \tag{1}$$

The dynamics of the opposite membrane are assumed to be symmetrical with respect to the sagittal plane. The indices $i, j = 1$ or 2 indicate the lower and upper masses m respectively. Elastic and dissipative forces acting on the folds' tissue are modeled through the non-linear functions K, k_c and B respectively. The parameter Q controls the tension of the folds, and f is the force exerted by the airflow passing through the folds, which depends on their dimensions (l_g and d_i in the sagittal and transverse planes respectively) and on the subglottal pressure P_s. The explicit functional forms of these functions can be found elsewhere (Assaneo and Trevisan, 2013; Lucero and Koenig, 2005).

This simple system captures some of the main features of the vocal folds' dynamics, reproducing speech data as the oscillations' onset and hysteresis (Lucero and Koenig, 2005) and the transversal wave propagating along the surface of the folds (Boessenecker et al., 2007).

Perturbations in glottal airflow produced by these oscillations are injected into the vocal tract, whose shape can be approximated by a series of N concatenated tubes of cross-sectional areas $A(i)$ and lengths $l(i)$, $1 \le i \le N$, for a total vocal tract length $L = \Sigma\, l(i)$, $1 \le i \le N$ (Fig. 1A). Propagation of these perturbations through the tubes is solved by splitting the incoming sound wave into reflected and transmitted waves at each interface, with reflection and transmission coefficients depending on the adjacent areas $A(i)$ and $A(i+1)$. This approximation is called a wave-reflection model, with a long tradition in the literature of voice synthesis (Liljencrants, 1985; Meyer et al., 2010; Murphy et al., 2007; Smith, 2007; Story, 1995; Strube, 1982; Titze and Alipour, 2006). Although the vocal tract can be configured in virtually infinite different shapes, restrictions are imposed by the articulators (jaw, tongue and lips). Taking advantage of this, Story and Titze (Story and Titze, 1998; Story, 2005; Story et al., 1996) developed a representation in which the cross sectional area A of tube i can be described as:

$$A(i) = \pi/4\left[\Omega(i) + \mathbf{q_1}\varphi_1(i) + \mathbf{q_2}\varphi_2(i)\right]^2\ \mathbf{c_k}(i) \tag{2}$$

where Ω is a fixed shape called neutral vocal tract, and $\{\varphi_1, \varphi_2\}$ are the first two spatial modes of an orthogonal decomposition calculated over a corpus of MRI anatomic data. This first squared factor in Eq. (2) represents the vowel substrate. The factor c_k represents a constriction, i.e. a uniform tube of cross section 1 except for a small interval around the k-th tube, where the section smoothly reduces to 0, representing the stop consonant substrate. In this way, the dimensionality of the vocal tract, which can virtually reconfigure into infinite different shapes, is drastically collapsed to a small number of parameters noted in bold type in Eq. (2).

The system of Eqs. (1) and (2) therefore constitute a basic mathematical model capable of reproducing the physics of the vocal system during the production of vowels and plosive consonants.

Stimuli and tasks

Three types of stimuli were used in our experiments:

1. *Non-speech sounds*. The audio samples were downloaded from (Font et al., 2013). These recordings included sounds of nature, animal vocalizations, machine sounds and musical instruments in equal proportions. The duration of the stimuli varied between 0.2 and 0.9 s.

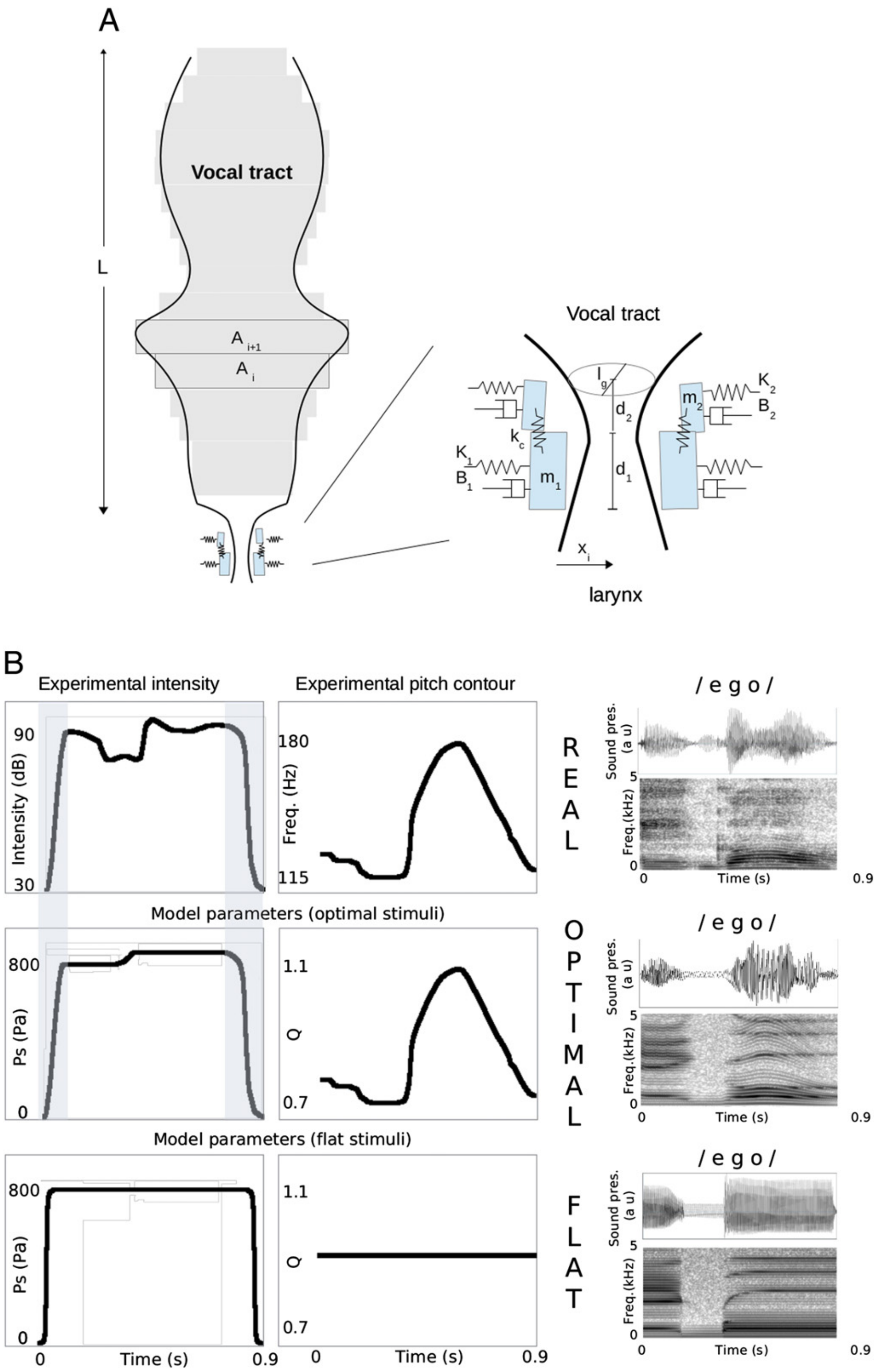


Fig. 1. | Elements of the articulatory synthesizer and voice samples. A. Sketch of the two main blocks that constitute the vocal system: vocal folds and vocal tract. The mean length of the vocal folds is 0.14 cm while the vocal tract is 17 cm for a typical male vocal system (Titze, 1988). Each vocal fold is modeled as a pair of masses m_1 and m_2 coupled by elastic and dissipative forces characterized by the functions K and B respectively. Integrating the equation of motion (Eq. (1)) of the folds leads to oscillations when values of subglottal pressure P_s exceed a critical threshold. The air perturbation caused by these oscillations travels along the vocal tract, approximated by a concatenation of $N = 44$ cylindrical tubes of different cross-sectional areas $A(i)$ and lengths $l(i)$, $1 < i < N$. A sketch of this discretization is show in grey. When the wave reaches an interface between consecutive tubes, one part is reflected and the other is transmitted, with amplitudes depending on the values of the consecutive cross-sections $A(i)$ and $A(i + 1)$. B. Real and synthetic samples of the utterance /ego/. Upper panel: sound intensity, pitch, sound pressure and spectrogram for a real vocalization (R). Middle panel: P_s and Q time traces, sound pressure and spectrogram for an optimal synthetic vocalization (O). P_s is driven to match the mean intensity of the two vowels and the rise and fall times of R (~0.03 s). The tension Q is a scaling of the pitch contour extracted from R. Lower panel: P_s and Q time traces producing intelligible, robot-like speech samples (flat stimuli F). In this case, $Q = 1$ and P_s is a step-like function fixed at 800 Pa with a rapid rise and fall of 0.01 s.

2. *Real speech samples*. Two native Spanish speakers, a female (age 30) and a male (age 40) pronounced, with different intonations and durations, isolated Spanish vowels (/a/, /e/, /i/, /o/ and /u/), diphthongs (/ai/, /ae/, /au/, /ea/, /ei/, /eu/, /ia/, /io/, /oa/, /oi/, /ua/, /ui/ and /uo/) and vowel-consonant-vowel (VCV) structures (/aba/, /abi/, /abo/, / eba/, /ede/, /edi/, /ego/, /igo/, /igu/, /obe/, /odu/, /ogo/ and /uga/). The reasons of this selection of stimuli is that they are simple, short time vocalizations with no meaning, preventing possible brain processing of semantic content. Also, the samples contain vowels and consonants that represent the complex acoustic repertoire of voice.
3. *Synthetic speech samples*. Synthetic sounds were generated by numerical integration of Eqs. (1) and (2) using standard Runge-Kutta functions coded in MATLAB. The parameters of the vocal folds and the vocal tract were set at values described elsewhere (Assaneo and Trevisan, 2013; Lucero and Koenig, 2005) except for the time-dependent parameters of subglottal pressure P_s, vocal fold tension Q and vocal tract shape q_1, q_2 and $c_k\ (i)$ (Eqs. (1) and (2) in bold type) that were varied to construct the different audio samples. Following (Story, 2005), the vocal tract was discretized in $N = 44$ tubes. The phonetic content of the synthetic sounds was the same as the real samples, built as follows: the five Spanish vowels /a/, /e/, /i/, /o/ and /u/ were synthesized using a vocal tract of coefficients q_1 and q_2 of Eq. (2) at constant values (see Assaneo et al., 2013). Diphthongs were synthesized by driving the coefficients $q_1\ (t)$ and $q_2\ (t)$ linearly from the initial to the final vowel for every pair of different vowels. Voiced stop consonants were produced in between vowels by occluding and releasing a small section of the vocal tract (Story, 2013). This was done by multiplying $c_k(i)$ in Eq. (2) by a temporal function that controls the height of the constriction. The consonants /b/, /d/ and /g/ were synthesized for occlusions at tubes $k = 44$, $k = 39$ and $k = 29$ respectively. The final acoustic pressure signals sampled at 44.1 kHz were converted to wav format and used as stimuli for our experiments (audio samples are available at Supplementary materials).

To avoid acoustic clues that could bias the judgments of the participants, the real stimuli were recorded in a sound booth. The three types of audio stimuli were set to the same overall energy (RMS) and presented to the participants with a white noise baseline of 2% of the maximal intensity.

In the fMRI experiments, the auditory stimuli were played binaurally at a mean 90 dB sound pressure, using foam earplugs and noise-attenuated MRI-compatible headphones (Resonance Technology Inc.). In behavioral tasks performed outside the scanner, mono audio files at a sampling rate of 44.1 kHz were presented to the participants via headphones Logitech B530 USB Headset MS Linc Optimi.

The experiments were written in MATLAB, using the Psychophysics Toolbox extensions (Brainard, 1997).

Experiment 1

In order to test the quality of the synthesized speech, an experiment was designed to compare the responses to non-speech sounds (NS), real speech samples (R), and two sets of synthetic speech samples: flat (F) and optimal (O).

The two sets of synthetic samples had the same phonetic content and duration as the R stimuli but differed in the profiles of subglottal pressure $P_s(t)$ and vocal folds' tension $Q(t)$. The flat stimuli (F) were synthesized using constant values for Q and P_s, as shown in Fig. 1B. On the other hand, the optimal stimuli (O) were generated using time traces for $P_s(t)$ and $Q(t)$ that approximate the intensity and pitch contours of real vocalizations (Fig. 1B). For the sound intensity, P_s was set to a phonation value of 800 Pa for the first vowel, and the intensity of the second was adjusted to match R. Experimental recordings show an attack time (i.e. the interval from silence to mean intensity of the first vowel) of about 0.03 s. This was reproduced by linearly increasing P_s from 0 to 800 Pa in 0.03 s. The same procedure was used to approximate the sound offset.

The relationship between vocal folds' tension and fundamental frequency is almost lineal for *Ps* around 800 Pa (Assaneo and Trevisan, 2013). In this way, the time trace of $Q(t)$ was simply a scaling of the experimental pitch contour, extracted with Praat software (Boersma and Weenink, 2013).

For this experiment we used a fMRI block design. Stimuli were arranged in 9 blocks. Each block contained 8–10 stimuli of each category (NS, R, F and O). Blocks were 9 s long, separated by silences of 4.5 s.

Participants inside the scanner were instructed to listen to the stimuli and to grade them, at the end of each block, by pressing one of four fiber-optic triggers held in the right hand according to the following code: button 0 if "I am sure that the voice is not human"; button 1 if "The voice is likely to be non-human"; button 2 if "The voice is likely to be human"; button 3 if "The voice was definitely human".

Four catch blocks with an extra task were included to encourage subjects to maintain the attention until the end of each block. Those blocks were inhomogeneous, the first 80% of the stimuli consisted of phonetic (O, F or R) and the last 20% to non-speech (NS) stimuli, or vice versa. Subjects were instructed not to respond after catch blocks. Thus, the subjects needed to pay attention until the end of the block to complete the task properly. Each participant performed two scanning runs of Experiment 1 (each 540 s long) in a single session, with a pause of about 2 min between runs. The stimuli were separately randomized for each run.

Experiment 2

Experiment 2 was designed to explore behavioral and brain responses across different dimensions of the vocal system. A scaling factor λ was defined for the vocal tract length L (see Fig. 1A) such as $L = \lambda Lo$, with $Lo = 0.17$ m. Controlling the laryngeal dimensions require a more subtle treatment, as it implies scaling the glottal dimensions and also the vocal fold masses and tissue stiffness. Following (Lucero and Koenig, 2005), a single scaling factor β was adopted. The glottal dimensions were scaled according to the simple rule $d_i = do_i/\beta$ and $lg = lgo/\beta$ (see Fig. 1A); the masses of the vocal folds scale as $m_i = mo_i/\beta^3$, where β^3 compensate the volume variation. Finally, assuming a constant elasticity modulus, the stiffness scale according to the rules $k_i = ko_i/\beta$ and $kc = kco/\beta$. The reference values are $mo_1 = 0.125$ g, $mo_2 = 0.025$ g, $kco = 25$ N/m, $ko_1 = 80$ N/m, $ko_2 = 8$ N/m, $lgo = 1.4$ cm, $do_1 = 0.25$ cm and $do_2 = 0.05$ cm.

Following (Fitch and Giedd, 1999; Lucero and Koenig, 2005), typical vocal tract scaling values are $\lambda = 1$ for males, $\lambda = 0.9$ for females and $\lambda = 0.7$ for 9–10 years old children, and typical laryngeal factors are $\beta = 1$ for males and $\beta = 1.4$ for females (no values reported for children). The vocal tract was scaled in the range $0.6 \leq \lambda \leq 1.3$, and the larynx in the range $0.5 \leq \beta \leq 2.0$, spanning the vocal system dimensions beyond typical values. Each range was divided in seven equally spaced sites, forming a 2-dimensional grid of $7 \times 7 = 49$ sites. Four types of vocalizations were synthesized at each site: the VCV structures /ego/ and /aba/, synthesized for two pitch contours with different durations (0.6 s and 0.9 s). Synthetic samples are available as Supplementary materials.

For this experiment we used an event-related fMRI design. In addition to the above $7 \times 7 \times 4 = 196$ stimuli, we included 38 silence trials of 0.75 s, and 19 randomly selected stimuli were presented twice in immediate succession. The resulting 253 stimuli were presented in random order with an inter-stimulus interval of 2 s.

Participants were instructed to listen to the stimuli while making a 1-back repetition task. In this task, participants should press any of the four fiber-optic triggers held in their right hand when the stimulus they heard was identical to the previous one. This task was included with the sole purpose of maintaining the subjects' attention on the stimuli. Each participant performed four runs (each 695 s long) in one single scanning session with a duration of about 1 h. The stimuli were separately randomized for each run.

In addition, in order to obtain a subjective rating of the stimuli, we also performed a purely behavioral task with the same set of 196 sound stimuli on another set of 30 participants. Stimuli were presented randomly, and subjects were instructed to grade the stimuli according to the same scale of voice quality as described in Experiment 1.

Participants

A total of 47 participants completed the experiments. We scanned 17 adults (7 females) aged 20–40 (mean 30 years), native Spanish speakers. All subjects were free of communication, neurological or medical disorders, passed audiometric screening, and had normal structural MRI scans. An independent second group of 30 native Spanish speakers (18 females, mean age = 34 years, min = 28, max = 40) performed the behavioral tests of experiment 2 outside the scanner.

All subjects were paid for their participation in the study and signed an informed consent form approved by the regional ethical committee.

fMRI data acquisition and processing

Subjects were scanned on a Siemens 3T Verio MRI, 12-channel TIM system. First, high-resolution T1-weighted 3D volumes were acquired for anatomical localization (TR = 2.3 s, TE = 2.98 ms, matrix size 240 × 256 × 176, voxel size 1 × 1 × 1 mm). Second, whole brain functional images were collected using a T2*-weighted EPI sequence, sensitive to BOLD contrast (TR = 2.02 s, TE = 25 ms, matrix size 66 × 66 × 40, voxel size 3 × 3 × 3 mm). The first four scans of all EPI series were not included in the analysis.

Data processing and analyses were conducted using the SPM8 software (Wellcome Trust Centre for Neuroimaging, London, UK) and custom MATLAB code. Functional images were pre-processed as follows: slice timing, motion correction by realignment, co-registration of the T1-weighted image to the mean functional image, normalization of the T1-weighted image to the MNI template, normalization of functional images (resampled voxel size 3 × 3 × 3 mm) by applying the parameters of the anatomical normalization, and Gaussian smoothing (5 mm FWHM).

Each voxel time series was fitted with a linear combination of functions derived by convolving a standard hemodynamic response function with the time series of the stimulus categories. The six movement parameters were entered as regressors of non-interest. The individual contrast images were smoothed (FWHM = 5 mm) and entered in a second level group ANOVA.

Results

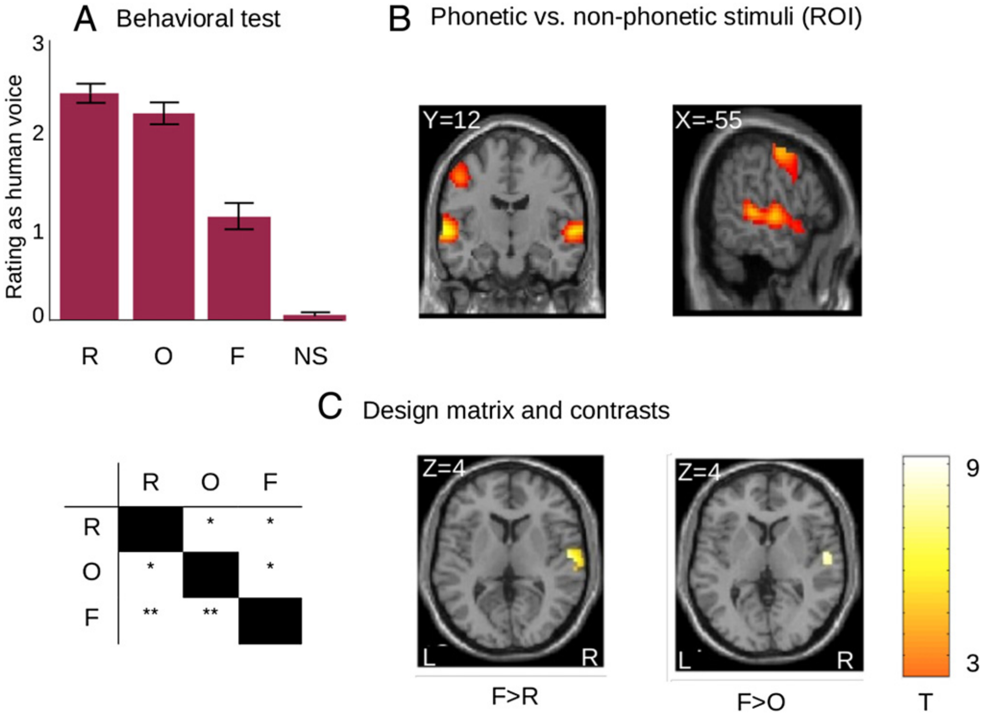
Experiment 1

Behavior

Mean ratings for real voices (R), synthetic voices (F and O) and non-vocal sounds (NS) are shown in Fig. 2A. There was a significant difference across the different types of stimuli (R, O, F, NS), as determined by a Kruskal-Wallis test ($\chi^2(3,1223) = 544$, $p < 0.001$). A post-hoc analysis revealed that the rating for the flat stimuli (1.20 ± 0.07) was lower than optimal and real stimuli (Wilcoxon rank-sum test, $p < 0.05$ Bonferroni corrected), while the optimal (2.44 ± 0.05) and the real (2.52 ± 0.04) stimuli did not differ.

fMRI

To identify brain areas involved in phonetic processing, the condition corresponding to natural sounds (NS) was subtracted from all the conditions corresponding to voices (O,F and R). This contrast showed three brain areas (uncorrected $p < 0.005$, number of



	R	O	F
R	■	*	*
O	*	■	*
F	**	**	■

Fig. 2. | Experiment 1: synthetic vs. real voices. A: ratings of voice naturalness (mean ± SE), for real speech samples (R), synthesized speech using natural time-dependent parameters for pressure and tension (O), synthesized speech using flat time-dependent parameters (F) and non-speech sounds (NS). B: activation by phonetic (O, F and R) minus non-speech stimuli (uncorrected $p < 0.005$, number of vowels > 300). The activated volume was used as a ROI in subsequent analyses. C: design matrix showing all the tested contrasts. The matrix elements marked * correspond to no supra-threshold voxels within the ROI (*p* uncorrected < 0.01). The elements marked ** show active voxels (uncorrected $p < 0.001$). Axial views display the activations for the contrasts F > R (corrected clusterwise p = 0.054) and F > O (corrected clusterwise p = 0.016). Both the behavioral and imaging tests show no differences between R and O synthetic stimuli.

Table 1
Experiment 1. Anatomical location, stereotaxic location, t-value of peak activations and voxel volume of activation clusters (p uncorrected < 0.001).

Anatomical location	Talairach coordinates			t-value	Voxels
	x	y	z		
Phonetic > non-phonetic (ROI)					
Left STG	−66	−25	4	14.6	500
	−69	−7	−2	10.8	
Left precentral G	−51	−7	46	6.5	320
Right STG	60	−19	−2	7	503
	66	−10	1	6.9	
Flat > optimal					
Right STG	54	−10	4	5.9	52
Flat> real					
Right STG	54	−10	4	4	17

voxels > 300) shown in Fig. 2B and Table 1. In consonance with previous results, the superior and middle temporal gyrus and the temporal pole were bilaterally activated (Belin et al., 2000). Another active region was found, extended along left pre and post central gyrus, which showed no overlap with the more superior region corresponding to the button presses (see Supplementary materials: *Estimation of the somatotopic representation of the right hand*). This region has been previously reported as being involved in speech production and also in listening to meaningless monosyllables (Wilson et al., 2004).

The overall active network is in agreement with the dual-stream model of speech processing (Hickok and Poeppel, 2007), were a dorsal stream maps acoustic speech signals into articulatory gestures. These three brain areas involved in phonetic processing were defined as the region of interest (ROI). All further analyses were restricted to, and corrected for multiple comparisons within this activation volume.

Additionally, all pairwise contrasts among phonetic conditions were analyzed (see Fig. 2C). Only the F > R and the F > O contrasts yielded significant activations (uncorrected $p < 0.001$). O and R sounds were indistinguishable even at a lower $p < 0.01$ threshold, for which there were no activated voxels. A cluster in the right STG, including Heschl's gyrus, showed more activation for F than O (clusterwise $p = 0.016$, corrected within the ROI). The same trend emerged in the contrast F > R (clusterwise $p = 0.054$, corrected within the ROI) as shown in Fig. 2 and Table 1. Interestingly, even for synthetic speech that is perceived as 'unnatural' (F stimuli), temporal vocal areas were active and indeed showed higher activity than 'natural' speech.

Analysis of distributed patterns of activity for the different types of auditory stimulations can be found at Supplementary materials: *Multivariate pattern analysis.*

Experiment 2

The results of experiment 1 support the model as a pertinent tool for the study of voice and speech perception at the neural level. The model was then used as an encoding tool to examine our working hypothesis that brain responses to vocal sounds are tuned to a specific anatomical relationship between the vocal tract and vocal folds dimensions. Our procedure serves to identify which specific regions of a broad brain network responding to human vocalizations specifically encode anatomical properties of the vocal tract. We then hypothesize that this region should participate in the recognition of vocal identity, since anatomical properties of the vocal system are a landmark of an individual's voice signature (Lopez et al., 2013). To this end, participants were asked to rate the naturalness of the synthetic voices generated in the 2-dimensional grid of the anatomical parameters λ and β (vocal tract and vocal folds dimensions respectively). We then used the brain responses to these voices in order to identify brain activity that correlates with the naturalness of human vocal sounds.

Behavior

Fig. 3A shows the ratings (averaged over the 30 participants) of 'naturalness' of the synthesized voices in the grid of parameters β and λ (larynx and vocal tract scaling factors respectively). High levels of naturalness (in shades of red) lie in a narrow region that extends roughly from the upper left to the lower right corner of the grid. To characterize this region, a weighted least squares regression $\lambda = m\beta + b$ was fitted using the mean grades as the corresponding weights. The line $\lambda = -0.27\beta + 1.26$ is the best linear approximation ($m\in(-0.2685, -0.04919)$ and $b\in(0.9798, 1.285)$ with a confidence of 95%). We named this the 'voice line'.

fMRI

The identification of a line in parameter space corresponding to human-like utterances allowed to search for brain regions whose activity varies monotonically with the distance to that line. To this end, an *anatomical* model was set up using the Euclidean distance to the voice line as a main regressor. A significant cluster emerged, localized to the right STG. The symmetrical left-hemispheric region showed the same modulation but only in a much smaller cluster (Table 2), for which activation increased with the distance to the voice line (uncorrected $p < 0.001$, clusterwise $p < 0.05$, corrected within the ROI). No cluster was found for which activity decreased for increasing distance to the voice line.

One step more was needed to confirm that this vocal encoding is indeed specific to anatomical features. There are in fact two competing models: the *anatomical* model, represented by the distance to the voice line, and the *behavioral* or perceptual model, represented by the raw naturalness ratings. If the two models did not differ, the results could be explained as accurately without any recourse to the anatomy of the vocal system.

To address this point, two distinct analyses were performed. First, the original analysis was repeated using the behavioral ratings (instead of the anatomical model parameters) as main regressors for the fMRI activity. This analysis showed no significant activations.

Even though having significant activations for the anatomical model and non-significant activations for the behavioral model favors the first one, it could still be the case that the difference between the two models would be non-significant. Therefore a second analysis was performed in order to test whether the meta-contrast (computed as the difference between the anatomical and the behavioral model) was significant in the regions activated according to the anatomical model. Indeed, this contrast showed activation areas very similar to the anatomical model, implying that this model represents activations more efficiently than the behavioral model (see Supplementary materials: *Behavioral vs Anatomical model*).

Although 'naturalness' ratings are consistently high for the voices along the voice line, they sound very differently: going from top left to bottom right of the line, voices change from an adult male to an adult female and finally to children's voices (audio samples are provided as Supplementary materials). To investigate how brain responses vary along the line, an *identity* regressor was constructed as follows: first, a coordinate was defined along the voice line with a given fixed origin; for each point on the line, the distance to the origin was assigned to every point perpendicular to the line at that point (see first panel of Fig. 4A). Given that our goal in this case was to evaluate the brain activity associated with vocal identity, the analysis was restricted to the grid elements above the mean of Fig. 3A by using the mask shown in the second panel of Fig. 4A. The resulting map was then used as a regressor (third panel of Fig. 4A).

A regression of brain activity to the position along the voice line (as shown in Fig. 4) revealed two clusters (voxelwise $p < 0.001$, clusterwise $p < 0.05$, corrected within the ROI): one in the right STG and another one

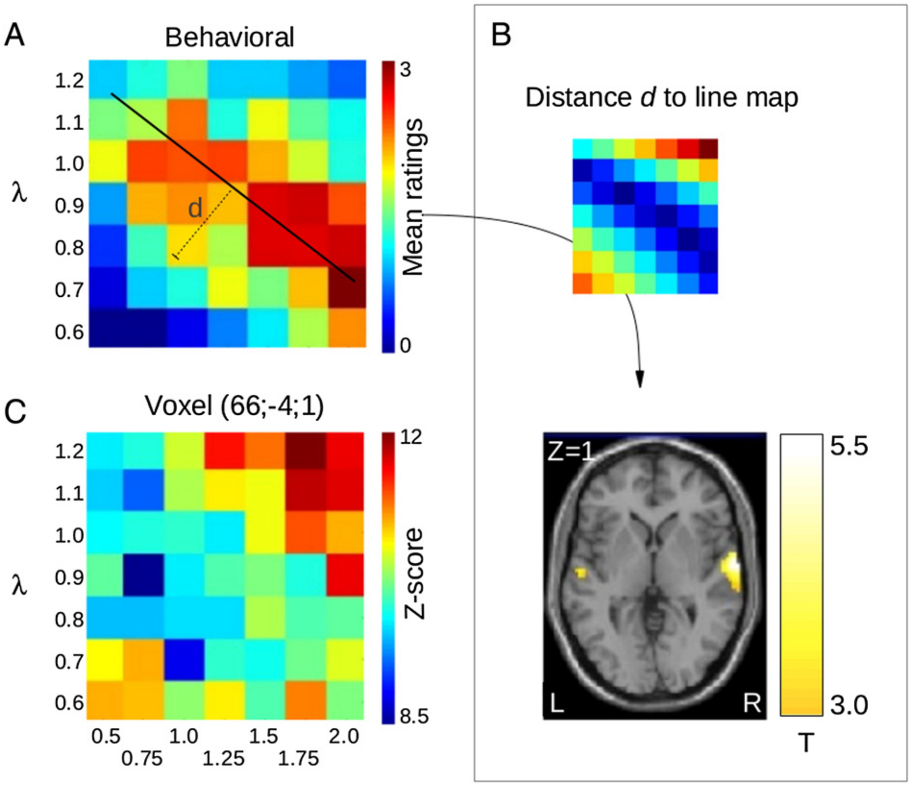


Fig. 3. | Experiment 2: voice identity in anatomical space. A: mean ratings of 'naturalness' for the voices synthesized in the space of anatomical parameters β and λ (representing the scaling factors for the larynx and vocal tract, respectively). The ratings were averaged over the responses of 30 participants. The scale used to rate the voices was 0: the voice is not human, 1: the voice is not likely to be human, 2: the voice is likely to be human and 3: the voice is definitely human. To characterize the region of high level of naturalness (reds in the color map), a weighted least squares regression was applied using the ratings as weights, obtaining the line $\lambda = -0.27\,\beta + 1.26$. B: the map of Euclidean distance *d* to the optimal line λ(β) was used as contrast (top) to obtain the brain regions showing a positive correlation with the distance to the voice line (bottom, uncorrected $p < 0.001$, corrected clusterwise $p < 0.05$). These brain regions were most active for sounds modeled after unlikely combinations of larynx and vocal tract dimensions. C: mean activity in the peak-voxel MNI 66 −4 1, for each point on the grid.

more extended along the left STG (Table 1). The clusters obtained in the regression along the line (Fig. 3 and Table 2) were slightly more posterior and dorsal, but highly overlapping with the ones obtained in the regressions of brain activity to the distance to the voice line in the anatomical space of vocal tract and vocal folds dimensions.

Taken together, these results indicate that the brain region that encodes whether a sound is produced by a human (based on the line in anatomical space) also encodes speaker-specific parameters (based on the position along that line). To check that this topographical overlap is not simply due to a correlation between the regressors, we verified that they are orthogonal (the correlation coefficient between the weights of the two regressors, z-scored and restricted to the elements above the mean was 0.0064 with a p-value of 0.97).

Table 2
Experiment 2. Cf Table 1 approximate anatomical and stereotactic location for maximal activation, t-values and voxel extension for each cluster (uncorrected $p < 0.001$, cluster-corrected $p < 0.05$).

Anatomical location	Talairach coordinates			t-value	Voxels
	x	y	z		
Distance to the typical voice line					
Left STG	−57	13	1	4.2	30
Right STG	66	−4	1	6.1	163
Position along the typical voice line					
Left STG	−60	−22	7	4.9	110
	−66	−31	13	4	
Right STG	66	−19	7	4.4	90
	57	−1	−2	4.2	

Discussion

Voice manipulations and vocal perception

There is wide evidence indicating that speech perception is assisted by internal motor and articulatory models at different stages of brain processing, including the low level auditory cortex (Kuhl et al., 2014; Correia et al., 2015). However, there is still an ongoing debate about the degree of modulation between the processes of production and perception of voice and how exactly this interaction is implemented. This has been difficult to solve in part because brain responses to voice have been probed using acoustically manipulated speech that cannot be mapped directly to specific changes in the vocal system. Here we tried to overcome this difficulty by using a mathematical model of voice production that integrates the physics of the vocal system, generating synthetic voice samples controlled by physiologically inspired parameters. This model offers a unique opportunity to generate synthetic voices for a continuum of anatomical and physiological properties of the vocal system. Operationally, this is performed by changing parameters which can be roughly classified in two families: anatomical parameters and time-dependent parameters. The former specify the sizes of the different parts of the virtual vocal system, while the latter control the instructions that generate the repertoire of sounds over time.

The present work dealt with both families, but deliberately excluded the systematic exploration of the time-dependent parameters that control the changes of shape of the vocal tract during the utterances. This *articulatory timing* is a key element to impart naturalness to the voice. It could be used to explore the representation of articulatory features in the brain during voice perception, which would be a natural

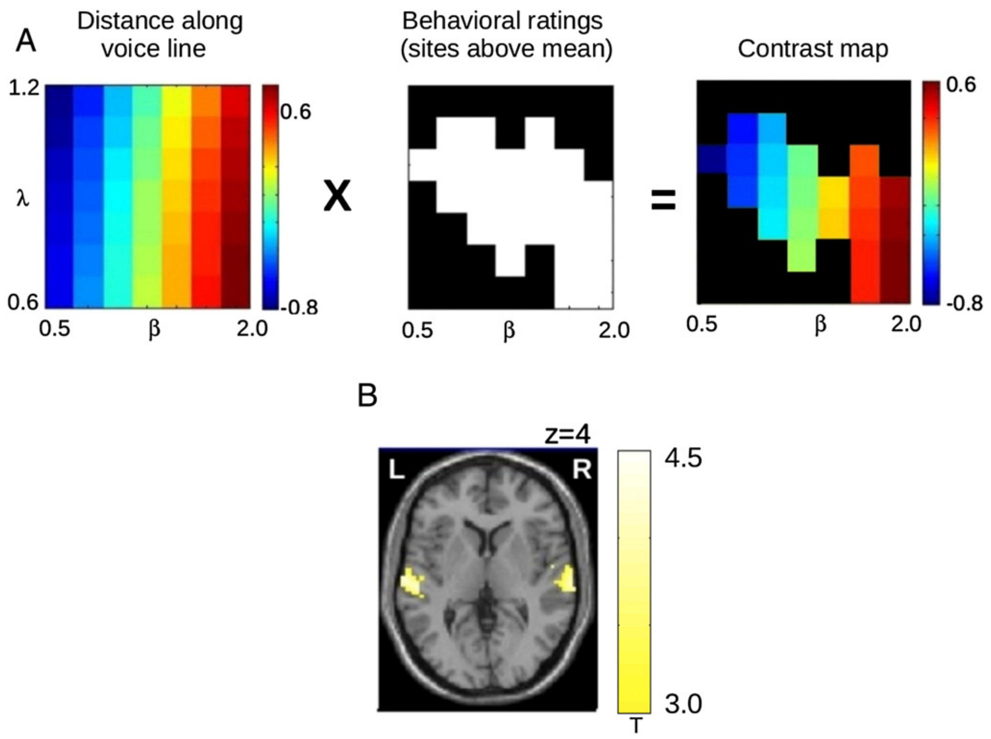


Fig. 4. | Experiment 2: Identity coding in anatomical space. A. The contrast map was obtained by superimposing two maps: one is the map that measures the distance along the voice line from an arbitrary origin, and the other is the mask of the behavioral ratings above the mean. B: transverse view displaying areas showing significant positive activation at group level (p uncorrected < 0.001 p cluster-corrected < 0.05) for this contrast. These areas are most active for sounds at the right-hand extreme of the voice line, which corresponds to child-like vocalizations with small larynges and vocal tracts.

follow-up to this work. In this first stage, the focus was set on the representation of anatomical properties of the vocal system during voice perception.

The first necessary step for the use of mathematical models to explore voice perception is to evaluate the quality of the synthetic voice. All musical instruments (the human vocal system included) are based on a set of physical attributes that provide a 'signature' of the instrument. This signature is known as timbre, a most elusive perceptual attribute that have been associated with a corpus of acoustical dimensions, including temporal features such as attack-time, jitter and shimmer and spectral features such as brightness and formant distributions (Baumann and Belin, 2010; Caclin et al., 2005). Instead of exploring these timbric dimensions in the acoustical space, in Experiment 1 we focused in the temporal variables that are actively driven during speech and vocal production: the sub-glottal pressure and vocal folds' tension profiles.

High ratings of naturalness for "optimal" synthetic stimuli, which participants confounded with actual human vocal sounds, showed that we succeeded in generating high-quality utterances. This goes beyond previous evidence that this class of models produces intelligible speech (Bunton and Story, 2009; Story and Bunton, 2010). By showing that synthetic sounds are confused (both behaviorally and in the fMRI responses) with real human vocalization, we demonstrate that the model faithfully captures basic timbre properties of the human voice.

This first step allowed us to study, in Experiment 2, which combinations of the two core anatomical parameters (length of the vocal tract and dimensions of the vocal folds) result in human-like vocalizations. We showed that for utterances to be perceived as human, these two variables have to show a specific linear relationship. We identified a line in parameter space along which voices are perceived as typically human, while voices distant from the line are not. In the next section we will discuss why different positions along the optimal vocal line map to speakers of different identities, ages or gender. In summary, (1) we showed that the proposed mathematical model is capable of generating synthetic voices within a wide range of individual variability; (2) the construction of this model allows to investigate in a parametric manner how the cortex responds to sounds depending on the vocal apparatus that produced them.

Vocal recognition and identity

Speech carries information through two different channels: one of them communicates semantic content, and the other one information about the identity of the speaker. The identity information depends on extrinsic factors such as the speaker's accent and speaking habits and, more critically, on intrinsic factors such as the anatomy and physiology of the vocal system. The extraction of identity information can be as specific as recognizing a speaker from a database of known voices (Formisano et al., 2008; Lopez et al., 2013), but also includes more general tasks such as the recognition of the gender and approximate age of an unknown speaker (Smith and Patterson, 2005).

This broader sense of identity information has been recently explored (Latinus et al., 2013) using morphing techniques, showing that voices are perceived as more atypical (and elicit more activation in the temporal vocal areas) as they become more distant from a male or a female prototypical voice in a 3-dimensional acoustical space. Although the dimensions of that space were derived from acoustic features of the voices, two of them are related to anatomical properties of the vocal system: the pitch and the formant dispersion, which correlate with the dimensions of the vocal folds β and the vocal tract λ respectively (Fitch, 1997).

Using a generative model of human voice, we found a cluster localized in the STG (mostly in the right hemisphere) that distinctively responded as the model drifts towards anatomically unrealistic parameters (distance to the typical voice line). This is consistent with (Latinus et al., 2013) who showed that atypical voices produce larger responses in the STG, but here, instead of using complex acoustical parameters, we directly manipulated in our model the mathematical representation of anatomical parameters. Moreover, we found that these brain regions overlap with those encoding speaker's identity (regions whose activity varies as stimuli move along the typical vocal line).

Conclusion

In this work we investigated voice perception processes using a low-dimensional mathematical model of the vocal system. The equations of motion for the vocal folds and the propagation of the airflow perturbation along the vocal tract are numerically integrated to generate synthetic voices.

Our two most important findings are that (1) using profiles for sub-glottal pressure and tension of the vocal folds (P_s, Q) matching experimental pitch profiles, the synthetic voices generate perceptual responses which are indistinguishable from those of real speech at behavioral and fMRI levels, and (2) the synthetic sounds are perceived as human vocalizations when the scaling parameters of the vocal folds and vocal tract (β, λ) follow a linear relationship. Voices are judged more atypical and neural activation is stronger as the distance from that line increases.

On the basis of these results, we propose that the low-dimensional model of the vocal system analyzed in this work has the necessary ingredients to generate realistic voices, which makes it a pertinent tool in the study of voice perception. The results presented here are consistent with previous works on voice coding. More importantly, in the framework of a unified sensory-motor program for speech (Cogan et al., 2014, Kuhl et al., 2014, Assaneo et al., 2013), our results allow for a straightforward interpretation of the problem of voice identity in terms of simple relations between anatomical scaling factors.

Acknowledgements

This work was supported by the University of Buenos Aires (UBA), the Concejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and ECOS A11S01 scientific collaboration program (France-Argentina). JS and MFA thank Mariam Chammat and Pablo Barttfeld for their kind help with fMRI data analysis, and Aaron Schurger for revising the manuscript. MAT and MFA thank Gabriel Mindlin for inspiring conversations.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2016.07.033.

References

Amador, A., Perl, Y.S., Mindlin, G.B., Margoliash, D., 2013. Elemental gesture dynamics are encoded by song premotor cortical neurons. Nature 495, 59–64.

Assaneo, M.F., Trevisan, M.A., 2013. Revisiting the two-mass model of the vocal folds. Pap. Phys. 5, 1–7.

Assaneo, M.F., Trevisan, M.A., Mindlin, G.B., 2013. Discrete motor coordinates for vowel production. PLoS One 8, e80373.

Baumann, O., Belin, P., 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol. Res. 74, 110–120. http://dx.doi.org/10.1007/s00426-008-0185-z.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech. Cereb. Cortex 10, 512–528.

Boersma, P., Weenink, D., 2013. Praat: Doing Phonetics by Computer.

Boessenecker, A., Berry, D.A., Lohscheller, J., Eysholdt, U., Doellinger, M., 2007. Mucosal wave properties of a human vocal fold. Acta Acust. united with Acust. 93, 815–823.

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., Formisano, E., 2014. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J. Neurosci. 34, 4548–4557.

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.

Bunton, K., Story, B.H., 2009. Identification of synthetic vowels based on selected vocal tract area functions. J. Acoust. Soc. Am. 125, 19–22.

Caclin, A., McAdams, S., Smith, B.K., Winsberg, S., 2005. Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. J. Acoust. Soc. Am. 118, 471. http://dx.doi.org/10.1121/1.1929229.

Chang, E., Rieger, J., Johnson, K., 2010. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. 13, 1428–1432.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory-motor transformations for speech occur bilaterally. Nature 37.

Correia, J.M., Jansma, B.M.B., Bonte, M., 2015. Decoding articulatory features from fMRI responses in dorsal speech regions. J. Neurosci. 35, 15015–15025.

Dehaene-Lambertz, G., Dehaene, S., Hertz-Pannier, L., 2002. Functional neuroimaging of speech perception in infants. Science 298, 2013–2015.

Fitch, W.T., 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J. Acoust. Soc. Am. 102, 1213–1222.

Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Am. 106, 1511–1522.

Font, F., Roma, G., Serra, X., 2013. Freesound technical demo. Proc. 21st ACM Int. Conf. Multimed. - MM '13, pp. 411–412.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. “Who” is saying “what”? Brain-based decoding of human voice and speech. Science 322, 970–973. http://dx.doi.org/10.1126/science.1164318.

Fowler, C.A., 2010. The reality of phonological forms: a reply to Port. Lang. Sci. 32, 56–59.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8 (5), 393–402.

Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lotus, L.J., Imada, T., 2014. Infants' brain responses to speech suggest analysis by synthesis. Proc. Natl. Acad. Sci. 111, 11238–11245.

Kühnis, J., Elmer, S., Meyer, M., Jäncke, L., 2013. The encoding of vowels and temporal speech cues in the auditory cortex of professional musicians: an EEG study. Neuropsychologia 1–11.

Latinus, M., McAleer, P., Bestelmeyer, P.E.G., Belin, P., 2013. Norm-based coding of voice identity in humanauditory cortex. Curr. Biol. 23, 1075–1080.

Lee, Y.-S., Turkeltaub, P., Granger, R., Raizada, R.D.S., 2012. Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. J. Neurosci. 32, 3942–3948.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception reviewed. Cognition 1–36.

Liljencrants, J., 1985. Speech Synthesis With a Reflection-Type Line Analog. Royal Institute of Technology, Stockholm.

Lopez, S., Riera, P., Assaneo, M.F., Eguía, M., Sigman, M., Trevisan, M.A., 2013. Vocal caricatures reveal signatures of speaker identity. Sci. Rep. 3, 3407.

Lucero, J.C., Koenig, L.L., 2005. Simulations of temporal patterns of oral airflow in men and women using a two-mass model of the vocal folds under dynamic control. J. Acoust. Soc. Am. 117, 1362–1372.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 1006.

Meyer, P., Wilhelms, R., Strube, H.W., 2010. A Quasiarticulatory Speech Synthesizer for German Language Running in Real Time. pp. 523–539.

Murphy, D., Kelloniemi, A., Mullen, J., Shelley, S., 2007. Acoustic modeling using the digital waveguide mesh. IEEE Signal Process. Mag. 24 (2), 55–66.

Smith, J.O., 2007. Introduction to Digital Filters: With Audio Applications. Julius Smith.

Smith, D.R.R., Patterson, R.D., 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J. Acoust. Soc. Am. 118, 3177.

Story, B.H., 1995. Physiologically-based Speech Simulation Using an Enhanced Wave-reflection Model of the Vocal Tract. University of Iowa.

Story, B.H., 2005. A parametric model of the vocal tract area function for vowel and consonant simulation. J. Acoust. Soc. Am. 117, 3231.

Story, B.H., 2013. Phrase-level speech simulation with an airway modulation model of speech production. Comput. Speech Lang. 27, 989–1010.

Story, B.H., Bunton, K., 2010. Relation of vocal tract shape, formant transitions, and stop consonant identification. J. Speech. Lang. Hear. Res. 53, 1514–1528.

Story, B.H., Titze, I.R., 1998. Parameterization of vocal tract area functions by empirical orthogonal modes. J. Phon. 26, 223–260.

Story, B.H., Titze, I.R., Hoffman, E.A., 1996. Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am. 100, 537–554.

Strube, H.W., 1982. Time-varying wave digital filters and vocal-tract models. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82, pp. 923–926.

Titze, I., 1988. The physics of small-amplitude oscillation of the vocal folds. J. Acoust. Soc. Am. 1536–1552.

Titze, I.R., Alipour, F., 2006. The Myoelastic Aerodynamic Theory of Phonation. National Center for Voice and Speech.

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