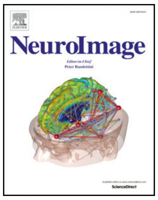
Contents lists available at ScienceDirect

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.

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 Published by Elsevier Inc.

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):

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

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.