Primers

Human voice perception

Marianne Latinus[1] and Pascal Belin[1,2]

We are all voice experts. First and foremost, we can produce and understand speech, and this makes us a unique species. But in addition to speech perception, we routinely extract from voices a wealth of socially-relevant information in what constitutes a more primitive, and probably more universal, non-linguistic mode of communication. Consider the following example: you are sitting in a plane, and you can hear a conversation in a foreign language in the row behind you. You do not see the speakers' faces, and you cannot understand the speech content because you do not know the language. Yet, an amazing amount of information is available to you. You can evaluate the physical characteristics of the different protagonists, including their gender, approximate age and size, and associate an identity to the different voices. You can form a good idea of the different speaker's mood and affective state, as well as more subtle cues as the perceived attractiveness or dominance of the protagonists. In brief, you can form a fairly detailed picture of the type of social interaction unfolding, which a brief glance backwards can on the occasion help refine — sometimes surprisingly so. What are the acoustical cues that carry these different types of vocal information? How does our brain process and analyse this information? Here we briefly review an emerging field and the main tools used in voice perception research.

Acoustical features of voice

Voice perception is grounded in voice production. To help make clear how vocal information is analysed, we shall briefly consider the particular acoustical characteristics of this sound category. As summarized by Ghazanfar and Rendall (2008), vocal sounds are generated by the interplay of a source (the vocal folds in the larynx) and a filter (the vocal tract above the larynx). The most common vocal sounds ('voiced sounds') correspond to a periodic oscillation of the vocal folds with a well-defined fundamental frequency (f0). The range of f0 values a given individual can achieve during normal phonation or singing is fairly extended, but the average f0 of an individual is largely a function of the size of the vocal folds: men have much larger vocal folds than women or children, resulting in generally lower f0 values. The vocal tract above the larynx acts as a filter reinforcing certain frequencies of the source, called 'formants'. Formant frequencies depend on the particular configuration of the articulators during speech, but also on the individual's vocal tract size. Thus, when pronouncing a same vowel, men have lower formant frequencies than women or children.

An important characteristic of vocal sounds is that they are generally highly harmonic — they are more spectro-temporally regular than the majority of sound categories (apart from many instrumental sounds). This regularity can be captured by indices such as the harmonic-to-noise ratio, or jitter and shimmer, which are measures of short-term perturbation of fundamental frequency and amplitude, respectively. Note that in addition to the 'normal' mode of phonation, the larynx can be used in other modes such as the 'falsetto' register or the 'fry' register, contributing to a greater diversity of possible vocal sounds.

Linguistic information is essentially conveyed by changes in formant frequencies (with the notable exception of tone languages such as Mandarin in which the f0 pattern can discriminate different speech sounds). F0 variations tend to carry information on the linguistic and affective prosody. Other acoustical features of voice are loosely regrouped under the general appellation 'timbre' (the auditory equivalent of visual 'shape') and include widely different aspects of phonation such as an individual's particular repartition of acoustical energy across frequency (long-term average spectrum) or the amount of phonation noise.

Identity information in voices

It is a routine observation that we can recognize voices, and sometimes remember them even after a long time. Although a speaker never utters twice exactly the same sound, listeners extract invariant features in the vocal signal to build representations of a speaker's identity that can be used to recognize that person from novel utterances. This ability is present very early in infants, and is shared by many animal species. Yet we are quite poor at voice recognition compared to face recognition. For instance, ear-witness testimony is notably unreliable and is not routinely accepted as evidence

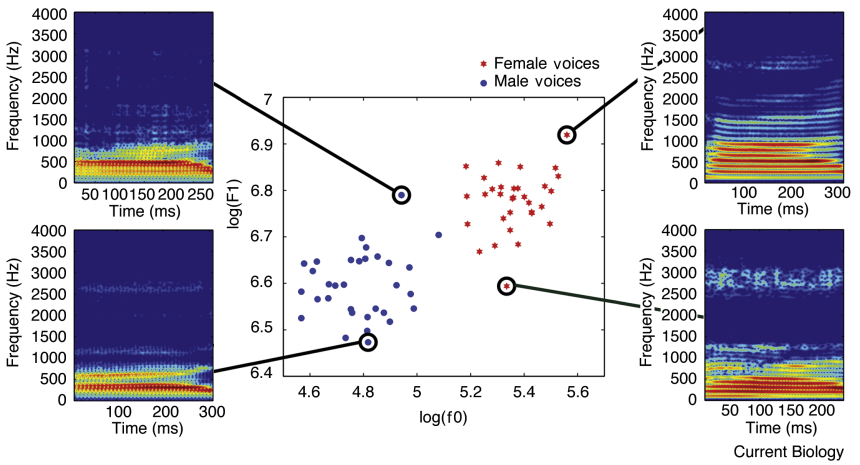

Figure 1. An acoustical voice space.
Individual voices are plotted as points in a two-dimensional acoustical space defined by the average fundamental frequency of phonation (f0) and the average first formant frequency (F1). In this space, the distance between points is a good approximation of how different the two voice identities are perceived. Note the clear distinction between male and female voices along the f0, but not the F1, dimension. Colour panels represent spectrograms (time-frequency decomposition of acoustical energy) of four individual voice examples. Note the large inter-individual variation in f0 (spacing between successive horizontal stripes, or harmonics) and formant frequencies (thicker bands of concentrated energy).

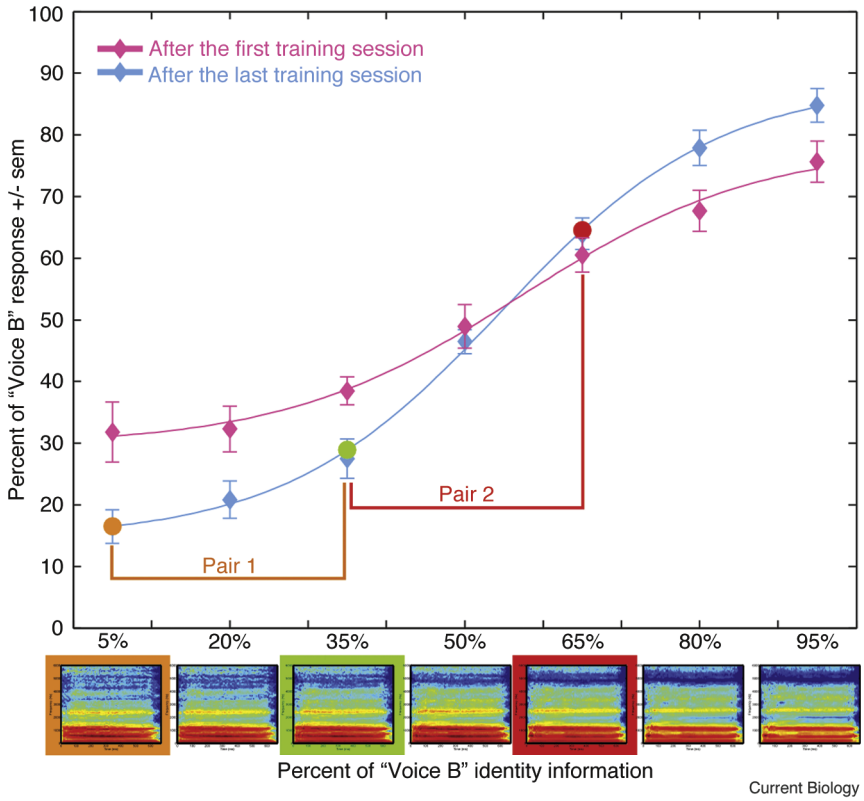

Figure 2. Morphing voice identity.
Lower panels: spectrograms of stimuli generated by morphing between voice A and B pronouncing the same vowel, with the relative weight of voice B changing from 5 to 95 in 15% steps. Upper panel: average categorization curve of the morphed continuum. Pink curve: classification after a single voice learning session. Blue curve: after several voice learning sessions. Note the steeper categorization curve after learning, indicating the formation of distinct identity categories of voices A and B. Pairs 1 and 2 have similar physical difference (30% morph step) but the perceived identity difference is much larger for pair 2 than pair 1.

in a court of law. The phenomenon is made even more complex by some voices being easier to remember (more distinctive) than others.

It has been thought for some time that one might use technology to generate 'voiceprints' that would allow automated voice recognition with a reliability comparable to that of fingerprint identification — but this proved a false hope. Modern voice recognition systems using probabilistic models and huge databases achieve fairly high accuracy when dealing with a small number of possible identities and quiet speaking conditions. But their performance rapidly deteriorates below human levels for many possible identities and natural (noisy) conditions.

How does our brain differentiate and recognize voices? One initial step in answering this question is to look at ways to describe the variability in voice characteristics associated with different identities. Voice coaches and professionals such as voice-over artists and radio journalists use a rich and varied terminology to characterize different voices. However, because the vocal apparatus is largely hidden from sight, descriptions are not based on shape as for faces, but rather on sensory or metaphorical analogies, sometimes quite hard to grasp for the non-initiated. Thus, voice artists for radio advertisement can be classified by the 'colour' of their voice and selected based on the match with the product's colour (for example, a 'brown' voice for a beer ad). Voice care professionals routinely use the GRBAS — Grade, Roughness, Breathiness, Aesthenia, Strain — scale to rate the presence of perceived phonation problems, such as pathological breathiness, roughness or harshness. The scale works well to characterize voice pathologies, but it is not very useful for distinguishing normal voices.

Another approach is to use multidimensional scaling of voice similarity judgments. Baumann and Belin (2010) asked a small group of listeners to rate the degree of perceived identity dissimilarity between a large number of pairs of voices pronouncing brief vowels. The pattern of dissimilarity thus obtained was well explained by representing individual voices as points in a two-dimensional space reflecting contributions of the source and filter, respectively. This suggests we represent voices via a 'perceptual voice space' with a small number of dimensions, similar to evidence obtained with faces. In this voice space (Figure 1), voices located close to one another are perceived as from similar identities, while voices located far apart are perceived as with very different identities. One should note, however, that these results obtained with brief vowels (the closest possible approximation to the static face stimuli typically used in face perception research) may not generalize to more natural speaking situations in which many other sources of variation, such as intonation and speaking style, are contributing.

Voice morphing

It would be useful to be able to manipulate the position of a stimulus in the voice space, in order to perform controlled experiments, in a similar way to how finely-controlled synthetic tones or noise have been used in studies of early-level auditory cortex. Recently available voice morphing tools provide such an exciting opportunity. Recent advances in speech analysis/resynthesis tools allow the generation of natural-sounding interpolations between voice recordings. Voice morphing notably allows experiments that examine the effects of perceptual differences in voices while acoustical differences are controlled.

We studied voice identity categorization and learning by generating a voice continuum between recording of two subjects (A and B) speaking the same vowel, and asking listeners to perform a forced-choice identity categorization task (identity A or B) on stimuli from this continuum, before and after learning the two voice identities. The results, illustrated in Figure 2, show the classical sigmoid curve with good categorization of morphed stimuli close to the original identities, and a steeper slope at the

level of the ambiguous, 50%A–50%B stimulus, suggestive of categorical perception. (Note that 'true' categorical perception would also require easier discrimination of pairs of stimuli across the categorical boundary.) Notably, the categorization curve slope becomes steeper after voice learning, showing that the listeners start creating distinct categories for the two identities. These results can then be used to generate pairs of voice stimuli with controlled acoustical difference — number of steps along the continuum — but different perceptual differences (Figure 2). For instance, pair 1 and pair 2 in Figure 2 can be used in psycho-acoustical or neuroimaging experiments dissociating the effects of physical vs. perceptual differences.

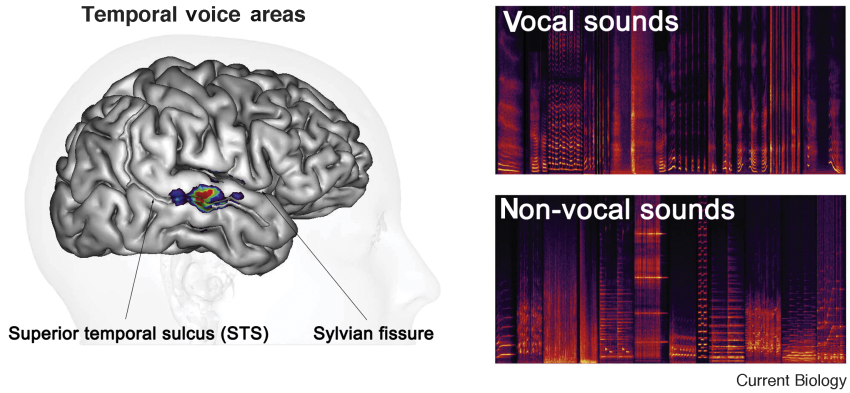

Figure 3. Temporal voice areas (TVAs) in the adult brain.
The contrast of cerebral activity measured in the adult brain by functional magnetic resonance imaging (fMRI) in response to auditory stimulation with vocal versus non-vocal sounds (stimuli available at http://vnl.psy.gla.ac.uk) highlights voice selective TVAs with greater activity in response to the vocal sounds. The TVAs (shown here in an individual young adult subject) are mostly located along the middle and anterior parts of the superior temporal sulcus (STS) bilaterally. Reproduced with permission from Belin and Grosbras (2010).

Voice sensitivity in the brain

How does our brain extract information in voices? Most neuropsychological or neuroimaging studies of cerebral voice processing have focused on speech perception, for understandable reasons. But this has led to an unfortunate situation in which the vocal nature of speech has been ignored, and speech stimuli are compared in neuroimaging studies to much cruder 'control' stimuli such as tones or modulated noise, leading to claims of 'speech-selective' cerebral activations. Yet there are many intermediate levels of stimulus complexity that are worth exploring. For instance Binder and colleagues used functional magnetic resonance imaging (fMRI) to measure an indirect index of cerebral activity when normal subjects listen to speech sounds. They observed extensive activations along the anterior and middle parts of the superior temporal sulcus (STS) that were not observed for simpler sounds such as tones or noise. Yet these activations, which one could think would reflect linguistic processing unique to speech sounds, remained largely unaltered when the speech stimuli were played backwards, thus removing most of the linguistic content (but leaving voice timbre largely unaffected). This suggests that these regions of higher-level auditory cortex might be more interested in the *vocal* nature of the speech stimulus that in its potential linguistic content.

Our group has confirmed a particular sensitivity of the brain to sounds of human voice (Belin *et al.* (2000)). We showed using fMRI that the auditory cortex of normal subjects contains 'temporal voice areas' (TVAs; Figure 3), located along the mid STS bilaterally, which show a particular sensitivity to voices whether they contain speech or not. TVAs respond more strongly to vocal sounds (speech or nonspeech) than to a range of control sound categories, such as amplitude modulated noise of frequency filtered sounds or animal vocalizations, a finding replicated by several groups. TVAs appear within a few months after birth and are already present in the brain of macaques, suggesting a long evolutionary history and an early development of cerebral voice processing (Figure 3). But little is known on the exact functional role of the TVA, or even whether their greater response to voice implies a specific role in cerebral voice processing.

Future directions

Much work remains before we understand human voice perception at least as well as face perception is currently understood. A number of important questions remain unanswered, and finding answers to these questions is not just important for basic research and the general advancement of understanding. Research into voice processing is also likely to impact everyday life in a world in which we increasingly not only speak to computers, but have computers use voice to communicate with us. Engineers need the results of this research to know how to best automatically extract and process the different types of vocal information in voice (identity, affect, ...) in the emerging domain of automated social signal processing. They also need to obtain reliable information on the perceptually-relevant acoustic characteristics of voices and their impact on perceived 'personality' of the synthetic voice. How does one synthesize an attractive, or a trustworthy voice?

Further reading

Belin, P., and Grosbras, M.H. (2010). Before speech: cerebral voice processing in infants. Neuron *65*, 733–735.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and 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., and Possing, E.T. (2000). Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex *10*, 512–528.

Baumann, O., and Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol. Res. *74*, 110–120.

Ghazanfar, A.A., and Rendall, D. (2008). Evolution of human vocal production. Curr. Biol. *18*, R457–R460.

Kreiman, J. (1997). Listening to voices: Theory and practice in voice perception research. In Talker Variability in Speech Research, K. Johnson and J. Mullenix, eds. (New York: Academic Press), pp. 85–108.

Nass, C., and Brave, S. (2005). Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship (Cambridge: MIT Press).

[1]Voice Neurocognition Laboratory, Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [2] International Laboratory for Brain, Music and Sound Research, Université de Montréal, and McGill University, Montreal, Canada.
E-mail: marianne.latinus@glasgow.ac.uk; pascal.belin@glasgow.ac.uk