SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PHYSICS
BIOLOGICAL PHYSICS
AUDITORY SYSTEM

Received
21 August 2013

Accepted
12 November 2013

Published
3 December 2013

Correspondence and requests for materials should be addressed to M.A.T. (marcos@df.uba.ar)

Vocal caricatures reveal signatures of speaker identity

Sabrina López[1], Pablo Riera[2], María Florencia Assaneo[1], Manuel Eguía[2], Mariano Sigman[3,4] & Marcos A. Trevisan[1]

[1]Dynamical Systems Lab, IFIBA-Physics dept, University of Buenos Aires, Pabellón 1, Ciudad Universitaria, CABA 1428EGA, Argentina, [2]Acoustics and Sound Perception Lab, Universidad of Quilmes, Roque Sáenz Peña 352, Bernal, Buenos Aires B1876BXD, Argentina, [3]Integrative Neuroscience Lab, IFIBA-Physics dept, University of Buenos Aires, Pabellón 1, Ciudad Universitaria, CABA 1428EGA, Argentina, [4]Torcuato Di Tella University, Almirante Juan Saenz Valiente 1010, C1428BIJ Buenos Aires, Argentina.

What are the features that impersonators select to elicit a speaker's identity? We built a voice database of public figures (*targets*) and imitations produced by professional impersonators. They produced one imitation based on their memory of the target (*caricature*) and another one after listening to the target audio (*replica*). A set of naive participants then judged identity and similarity of pairs of voices. Identity was better evoked by the caricatures and replicas were perceived to be closer to the targets in terms of voice similarity. We used this data to map relevant acoustic dimensions for each task. Our results indicate that speaker identity is mainly associated with vocal tract features, while perception of voice similarity is related to vocal folds parameters. We therefore show the way in which acoustic caricatures emphasize identity features at the cost of loosing similarity, which allows drawing an analogy with caricatures in the visual space.

Speech contains a great deal of information that goes above and beyond its semantic content. Gender, approximate age and affective state of the speaker can be easily and reliably extracted even from small speech samples[1]. Speakers' identity can also be recognized, selecting robust properties from acoustically flexible voices[2]. The non-linguistic information used for these kind of tasks depends on two different classes of factors that shape the human voice: extrinsic factors, determined by culture and speaking habits, as the speaker's accent, and intrinsic factors which depend on the anatomy and physiology of the vocal system[3]. Here we concentrate on intrinsic factors, which are more difficult to imitate than extrinsic ones.

Humans are natural vocal imitators. We copy aspects of other human voices, which is an essential process to acquire language[4,5] and also incorporate to our lexicon sounds of nature in the form of onomatopoeias. This imitation process by which arbitrary sounds (for instance, a knock sound) become vocalized is strongly constrained by the physiology and anatomy of the vocal system[6]. Similarly, although there is flexibility and versatility in vocal impersonation[7,8], this process is constrained by the individual voice production system. The investigation of impersonation, its success and failure, is an empiric manner to address the problem of what determines vocal identity.

During normal speech, voiced sounds are produced by the combined action of the vocal folds and the vocal tract[9]. The vocal folds are a pair of elastic membranes located at the glottis that can be set into oscillatory motion by the transfer of energy from the air expelled from the lungs. The perturbed airflow produced by these oscillations is then injected into the vocal tract, formed by the set of cavities extending from the glottal exit to the lips, whose shape is actively controlled by different articulators as the tongue and jaw[10,11]. The sound wave propagates back and forth along the tract, that acts as a waveguide for the sound. From a spectral point of view, the oscillations of the vocal folds provide a rich sound source characterized by a fundamental frequency f_0 (pitch) and decaying harmonics, and the vocal tract is defined by its resonant frequencies F_i (formants).

Although voiced sounds result of the combined action of vocal tract and vocal folds, both blocks act rather independently during normal speech, because the folds are not appreciably affected by the re-injection of sound from the tract, which is known as source-filter theory[9]. This has consequences on the uttered sounds: from the spectrum of a voiced sound we can extract parameters related specifically to the dynamics of vocal folds or to the anatomy of the vocal tract.

Different vocal anatomies produce different voices. For instance, the female and male typical vocal folds vary in size, producing female voices with higher pitch and formants than male voices. However, although we are good at

recognizing speakers' identities, we can be deceived by vocal impersonators. What are then the vocal features that they select to recreate the identity of a speaker? In this work we address this question by identifying acoustic parameters relevant to evaluative tasks of voice and speaker perception.

Results

We investigate whether voices and identities could be represented in low-dimensional spaces, identifying the acoustical parameters that are perceptually important across subjects. Building on previous efforts that analyzed the cases of one impersonator imitating different targets[8] and different impersonators imitating a single target[12], we created a database of 3 different sentences, each one pronounced by a different public figure (targets T_1 to T_3) along with the corresponding imitations produced by 5 professional impersonators (I_1 to I_5). Using written versions of the sentences, the impersonators first recorded them with their own normal voice (n). Then, again from the written sentences, the impersonators used their memory to imitate the corresponding targets. These imitations rely on internal voiceprints or caricatures (c) of the public figures' voices that are used by the impersonators to build their imitations. Finally, the impersonators replicated the sentences (r) just after listening to the target audio files.

Henceforth, we refer to the caricature and replica produced by impersonator a of the target b as $I_a c_b$ and $I_a r_b$ respectively. Similarly we refer to the natural voices produced by impersonator a of the sentence produced by target b as $I_a n_b$. The complete voice database consists of 48 audio files: targets T_j, impersonators' natural voices $I_i n_j$, caricatures $I_i c_j$ and replicas $I_i r_j$ for each public figure j ($1 \leq j \leq 3$) and impersonator i ($1 \leq i \leq 5$) (see Methods: *voice database* for details. The targets T_1 and T_2 along with their replicas and caricatures can be found as Supplementary Audio files S1 to S22).

All the results reported here belong to three classes: 1) psychophysical measures of voice similarity and speaker identity (to determine whether an impersonation is successful or not), 2) auditory properties of speech, and 3) the relation between auditory properties and psychophysical measures of similarity to identify auditory signature of vocal identity.

Experiment 1: psychophysical measures of identity. Each subject heard a single target and all its imitations in random order, and gave a rating indicating how likely the voice they had just heard belonged to the public figure in question. They used a scale ranging from 1 (I am sure the voice does not belong to the public figure) to 5 (I am sure the voice belongs to the public figure) (see Methods: *experiment 1* for details).

The 3 targets showed high average ratings of belongingness ($T_1 = 3.9 \pm 0.4$, $T_2 = 3.5 \pm 0.5$ and $T_3 = 4.8 \pm 0.3$) which testifies that the voices of the public figures were easily recognizable for the population used in this study. Next we investigated the effect of three independent factors, 1) Impersonator (I_1 to I_5), 2) Type of impersonation (caricature or replica) and 3) Impersonated character (T_1, T_2 or T_3) by submitting the rating data to an ANOVA with these three factors as fixed variables. The ANOVA (Table 1) revealed that the three factors had significant effects. We followed these main dependence with post-hoc t-test to identify how these factors affected the rating.

First, by pooling together all impersonators and impersonation type, we investigated whether some targets where easier to imitate. Results showed average imitation ratings of 2.58 ± 0.14, 2.81 ± 0.16 and 1.72 ± 0.12 for T_1, T_2 and T_3 respectively. The distribution of ratings can be found in the upper panels of Figure 1. Comparisons of these distributions Bonferroni corrected for multiple comparisons showed a significant difference ratings for T_3 compared to the other two targets (both comparisons $p_{corr} < 0.001$). T_3 is Diego Maradona, a world-wide public figure for around 25 years. In fact, 62% of the impersonations of T_3 ranked as 1 ("*I am sure the voice does not belong to the public figure*"), which shows that psychophysical thresholds of acceptance of vocal identity depend -as could be expected- on the degree of knowledge of the impersonated target. Given that T_1 and T_2 had similar and broad distributions of ratings (and also similar periods of public activity, around 5 years) and that T_3 distribution was different and saturated towards a strong recognition of dissimilarity with the target, we restrict our subsequent analyses mainly to the targets T_1 and T_2 and their imitations.

Next, we submitted the data to independent ANOVAS for each target with impersonator and imitation type as independent factors

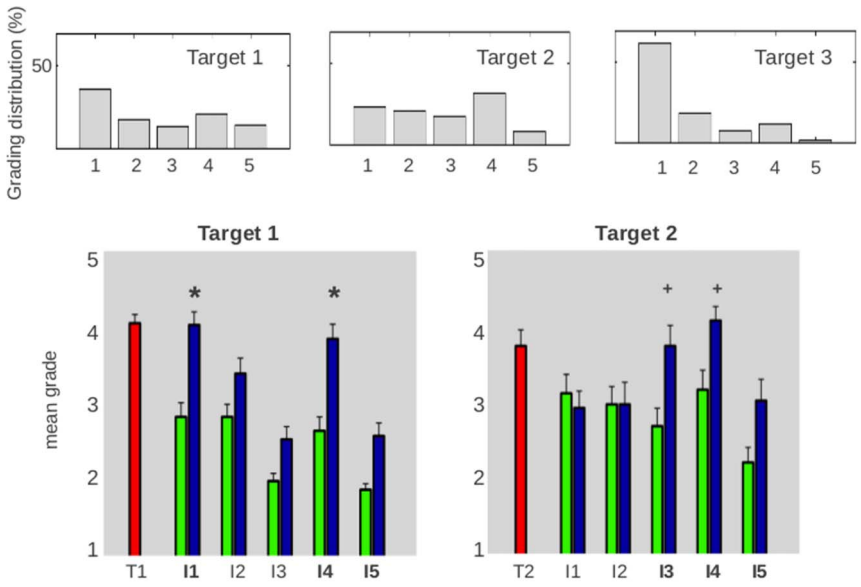

Figure 1 | Experiment 1: at the behavioural level, speaker identity is better elicited by caricatures (blue) than replicas (green). Each participant listened to the set of audio files containing a single target and its imitations (caricatures and replicas) and associated them with the identity of the corresponding public figure using a scale from 1 (the voice does not belong to the public figure) to 5 (the voice definitely belongs to the public figure). In the upper panels we show the distributions of gradings for the 3 targets T_1, T_2 and T_3 across imitation types. In the lower panels, we show the grades (mean $\pm$ sd) for the targets (red), replicas (green) and caricatures (blue) for T_1 and T_2.

(Table 1). ANOVAS revealed for both targets an effect of impersonator (which merely reveals that certain impersonators produce higher quality imitations and caricatures) and, more interestingly, a significant effect of impersonation type. A follow up of this ANOVA revealed that for both targets, the effect of type was accounted by an increase in rating for caricatures than for replicas with average imitation ratings of 2.93 ± 0.09 and 2.03 ± 0.07 respectively (T_1: t = 8.84, df = 179, $p < 10^{-4}$; T_2: t = 3.51, df = 84, $p < 7.2\ 10^{-4}$).

These results indicate that caricatures (produced to recreate individual traits of the speaker) result in voices which are more typically accepted as belonging to the targets than replicas (attempts to produce faithful copies of the target). In other words, vocal caricatures serve the purpose of speaker recognition better than auditory replicas.

There are two possible interpretations of these results: a parsimonious interpretation is that impersonators are simply not trained to copy specific utterances, and their replicas are simply bad copies of the targets. Another more interesting possibility is that they are efficiently reproducing features of the speakers' voice, different from the ones that code identity. In order to investigate this last hypothesis, we designed an experiment to confirm that the replicas are indeed good at copying voices which fail in focusing in the relevant auditory dimensions encoding speaker identity.

Experiment 2: psychophysical measures of voice similarity. For each target, we selected the 3 impersonators that produced the higher ranked caricatures and lower ranked replicas (marked in bold type in Fig. 1). The rationale behind this choice was that we wanted to test the hypothesis that replicas are efficient imitations of the target voice even when they may not focus in the most salient dimensions for identity.

Participants listened to all pairs of audio files and ranked their similarity using a scale from 1 (the two voices are very different) to 5 (the two voices are the same) (see Methods: *Experiment* 2). For each target, the perceptual data produced by each subject can be organized in a similarity matrix M in which the element M_{ij} corresponds to the similarity rating between the pair of audio files i and j. We selected the pairs containing the target files and submitted this data to ANOVAs with impersonator and imitation type as independent factors (Figure 2 and Table 2). The analyses revealed an effect of imitation type. A follow up of this analysis showed that for both targets, this effect was accounted by an increase in rating for replicas with respect to caricatures (T_1: t = 6.99, df = 51, $p < 10^{-4}$; T_2: t = 3.13, df = 47, p = 0.03), with average imitation ratings of 3.31 ± 0.15 and 2.52 ± 0.20 for replicas and caricatures of T_1, and 3.19 ± 0.20 and 1.71 ± 0.12 for replicas and caricatures of T_2. This indicates that replicas are quality copies of the targets' voices and are indeed perceived as more faithful reproductions of the original voice than caricatures, although they fail in eliciting the identity of the speaker.

Acoustic spaces of similarity and identity. Our next aim was to find the acoustic figures that govern the perception of similarity and identity. For each file in the original database, we calculated the 12-dimensional acoustical vector $V = \{jitter, shimmer, f_0, F_i\ (1 \leq i \leq 5), disp(F_5 - F_1), disp(F_4 - F_3), disp(F_5 - F_3), disp(F_5 - F_4)\}$, using mean values calculated over the length of the sentence (see Methods: *acoustic space* for details). The pitch f_0 and the formants F_i provide the most direct information about the anatomy of a vocal system: the first one is related to the mass and elasticity of the folds, while the formants reflect the vocal tract shape. We included two additional vocal folds' parameters: *jitter* and *shimmer*, that measure the cycle-to-cycle variations of frequency and amplitude, respectively. These two parameters have been historically used for qualitative descriptions of voice pathologies and, more recently, have been shown to be strongly associated with voice perception by naive listeners[13,14]. This is also the case for the formant dispersions $disp(F_j - F_i)$[13–16], calculated as the mean interval between formant frequencies, that we included as well (see Methods: *acoustic space* for details). Hence, each audio file is mapped to a 12-dimensional vector and we can then measure how subjective ratings of speaker identity and voice similarity co-vary with its different dimensions.

In experiment 1 we had only one scalar (subjective rating) associating each voice to the target. The results showed a broad variability, with some voices being systematically associated and others never confounded with the target (respectively high and low subjective ratings in Figure 1). Within this variability, analysis clearly showed that caricatures were closer to the target than replicas. Within the caricatures, two impersonators showed particularly efficient imitations of the target. In fact, the higher ranked caricatures (marked with * and + in Fig. 1) were non distinguishable from the targets in subjective ranking of belongingness for both T_1 and T_2 ($P < 0.05$, Friedman test).

To reveal the relevant acoustical variables for speaker recognition, our experimental approach was to identify in which acoustic variables, the efficient caricatures were proximal to their corresponding targets. To this aim we performed the following analysis: first, for each target, we mapped the 11 audio files (the target along with its 5 replicas and 5 caricatures) to all possible combinations of 2-dimensional planes of the original 12-dimensional acoustical space (a total of 66 spaces). For each plane, we measured classification accuracy as the percentage of imitations that were located farther to the target than the two highest subjectively ranked caricatures (the t-value that the highest ranked caricatures were closer to the target than the other imitations were identical yielded identical results).

The results are shown in the upper panels of Figure 3. The dimension $disp(F_5 - F_4)$ is a strong marker of identity, as it shows a good performance for both targets: for T_1, the spaces that result from the combination of this dimension with *shimmer* or *jitter* or F_5 or $disp(F_5 - F_1)$ are such that the best caricatures are more similar to the target than the rest of the imitations. For T_2, the same holds for the combination of $disp(F_5 - F_4)$ with F_5, F_2 or $disp(F_5 - F_3)$. We summarize these results in the histogram of Figure 3, where we show the number of spaces as a function of the correctly classified imitations for both the targets. Only 3 acoustical spaces perform significantly (>2 sd) at

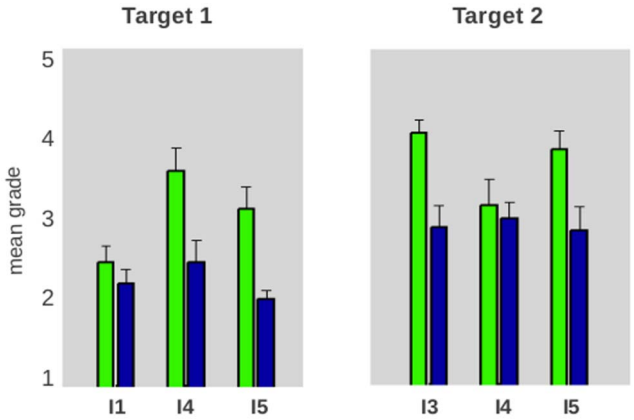

Figure 2 | Experiment 2: at the behavioural level, voice similarity is better elicited by replicas (green) than caricatures (blue). Each participant listened to all pairs from a set of M = 10 audio files ($M(M + 1)/2$ = 55 audio pairs) composed by a target T and the caricatures, replicas and normal voices of the 3 impersonators with the highest ranked caricatures of experiment 1 (bold face in Fig. 1). Participants were asked to rate the voice similarity of each pair using a scale from 1 (the two voices are very different) to 5 (the two voices are the same). For both targets, replicas (green) display higher grades than caricatures (blue) for voice similarity, opposite to the results of experiment 1 for speaker identity shown in Figure 1.

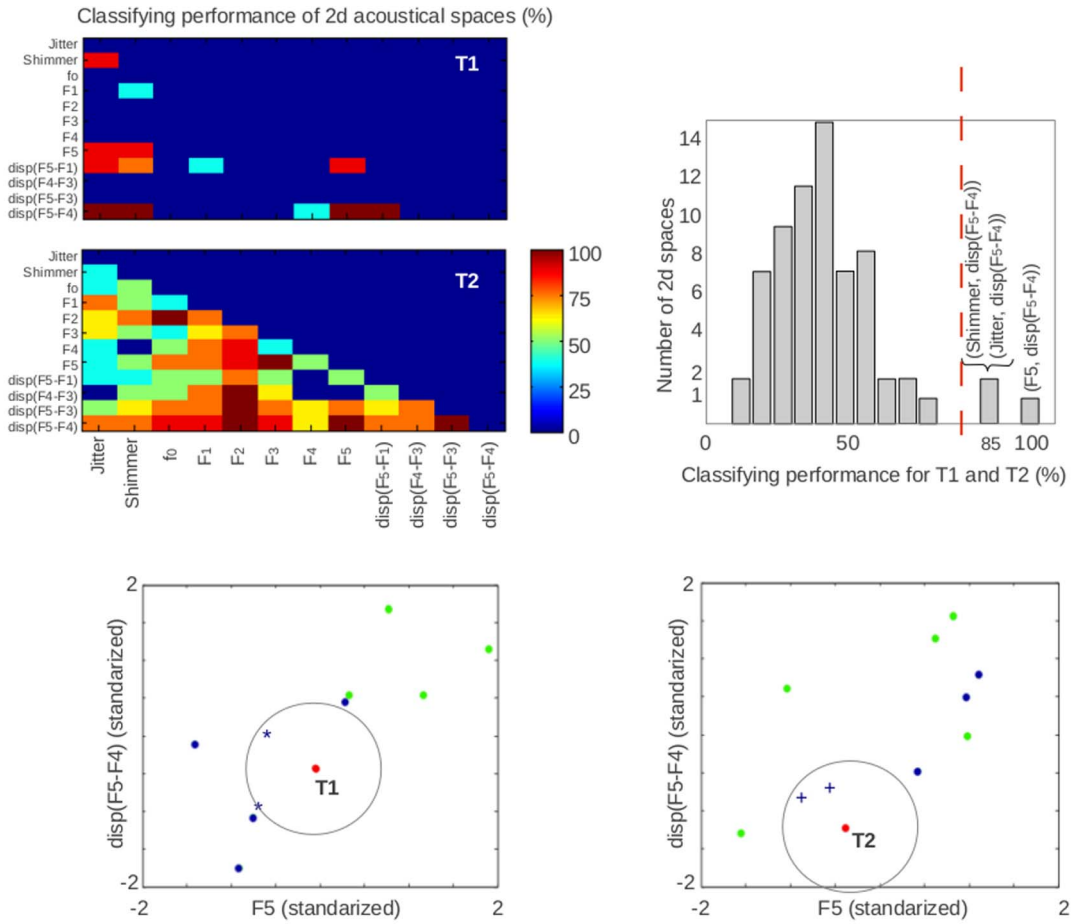

Figure 3 | Experiment 1: at the acoustical level, speaker identity is strongly related to vocal tract features. Upper panels: for every 2-dimensional acoustical space, we measured the classification accuracy as the percentage of imitations that are located farther from the target than the highest ranked caricatures (marked with + and * in Fig. 1) (left). We summarize these results for both targets T_1 and T_2 in a histogram showing the number of acoustical spaces as a function of their classifying performances (right). Lower panels: organization of the targets and imitations in the 2-dimensional space (F_5, $disp(F_5 - F_4)$), where the highest ranked caricatures of experiment 1 are closer to the corresponding target than the rest of the files for both T_1 and T_2.

locating the best caricatures closer to the targets than the rest of the imitations: (*shimmer*, $disp(F_5 - F_4)$), (*jitter*, $disp(F_5 - F_4)$) and (F_5, $disp(F_5 - F_4)$), and only the last one displays a perfect performance. The organization of files for this case is shown in the lower panels of Figure 3.

In experiment 2 the procedure to correspond auditory to psychophysical dimensions is easier because the experiment produced a similarity matrix of dimensionality comparable to the auditory space. Hence, we performed a multidimensional scaling analysis[13,14,17] (see Methods) that allows maximizing the fitting of the dissimilarity matrices to an euclidean space where audio files are organized according to the experimental perceptual distances (upper panels of Figure 4). Note that even if the data is collapsed to such a low dimensional space, this representation allows visualizing the effect of imitation type, with replicas (green) closer to the target (red) than caricatures (blue), and also the rough organization of the two imitation types in two clusters. We also show the natural voices of the impersonators (pink). From their different distributions with respect of the target voices (far from T_1 and relatively close to T_2), we conclude that the relation we observed of replicas being perceived closer to the target than caricatures is not directly related to the proximity between the natural voice of the impersonators and the targets.

We then submitted these perceptual embeddings to a redundancy analysis in order to explain as much variance as possible using the acoustical parameters (see Methods). In the lower panels of Figure 4 we show the biplots for T_1 and T_2. Although there are contributions from a variety of acoustic parameters, in both cases the most salient ones are f_0 and *jitter*, that correlate with *dim*1 and *dim*2 respectively. In the case of T_1, f_0 correlates with *dim*2 ($r = 0{,}88$) and *jitter* with *dim*1 ($r = 0{,}78$). In the case of T_2, f_0 correlates with *dim*1 ($r = 0{,}91$) and *jitter* with *dim*2 ($r = -0{,}89$).

These acoustical analyses allow drawing a rough correspondence between the different types of imitation and the main building blocks of the vocal system: for the construction of replicas, impersonators focus mainly in vocal folds' properties, producing quality copies of the original voice. Caricatures, on the other hand, are constructed using robust vocal tract features, and the voices produced elicit the identity of the impersonated public figure.

Discussion

In this work we investigated the auditory features that strongly relate to speaker identity and voice similarity. We capitalized on the ability of professional impersonators to generate voices which can simulate another person's identity.

Our two most important findings are: 1) at the behavioural level, replicas are more likely to be perceived as identical to the acoustic target than caricatures. Instead, when listeners are focused on the identity of the speakers whose voices they have been long exposed to, caricatures are more likely to be associated with the speaker than replicas and 2) at the acoustical level we identified different dimensions which are relevant for each task; the information used by listeners to judge voice similarity is related to the vocal folds (f_0 and

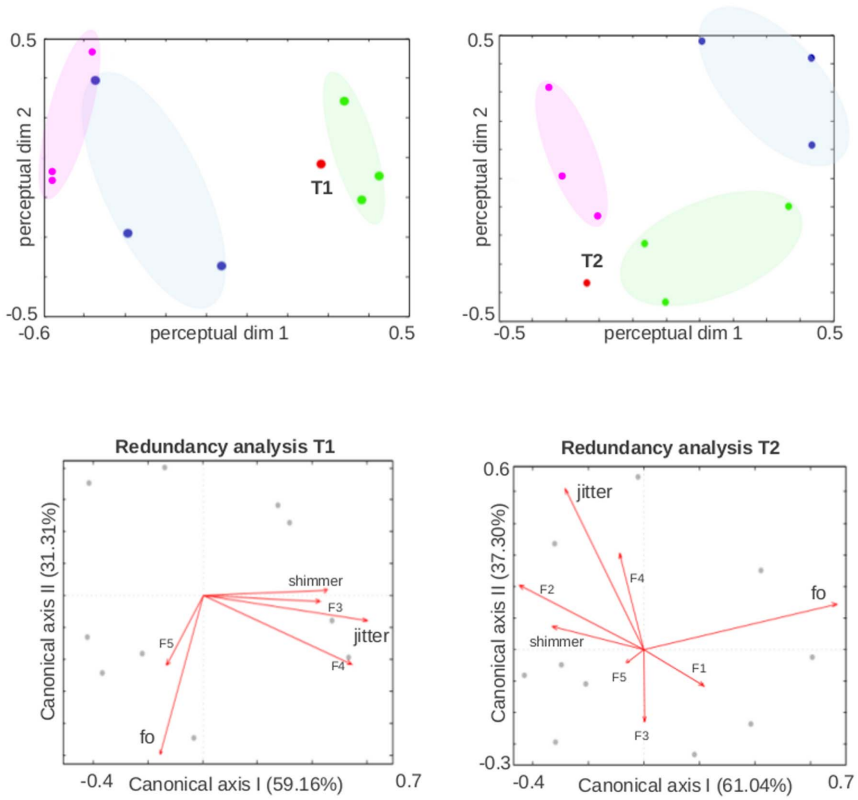

Figure 4 | Experiment 2: at the acoustical level, voice similarity is mainly associated with vocal folds' features. Upper panels: 2-dimensional spaces resulting from the INDSCAL analysis, summarizing the perceptual organization of files from experiment 2 on voice similarity. For each target (red), we show the impersonators' caricatures (blue), replicas (green) and normal voices (pink). We further submitted this data to a redundancy analysis to find the combination of acoustic parameters explaining the variance of the perceptual data. The resulting biplots are shown in the lower panels with the vectors indicating the correlation between the main acoustic parameters and the axes of ordering space.

jitter) and speaker identity is mostly associated with the vocal tract feature $disp(F_5 - F_4)$. Maximal categorization is observed when it is combined with another vocal tract feature (the formant F_5), but high classifying performance is also observed when $disp(F_5 - F_4)$ is combined with the vocal fold parameters *shimmer* and *jitter*.

The higher formants (F_4 and F_5) were good candidates to index identity because they are more stable than the lower formants (F_1, F_2 and F_3) along the utterances[10,15]. This can be seen as two independent channels to communicate semantic content and identity: the lower formants vary to produce utterances that encode linguistic information, typically vowels indexed in the space (F_1, F_2)[9]. While these formants vary along the sentences, higher formants remain largely unchanged, indexing stable properties of the discourse such as speaker identity. A prediction of this model, which can be examined in future studies, is that identity recognition should not be impaired for dysphonic voices, generated using turbulent noise as the sound source, without any vocal folds activity.

In a previous study, Baumann and Belin[13] asked participants to listen to pairs of voices and determine whether they belonged to the same person. To make this judgment, participants relied on voice similarity and speaker identity (that was unknown to the participants). In very close coincidence to our findings, they found that parameters f_0 and $disp(F_5 - F_4)$ presented dominant contributions in nearly orthogonal dimensions of the perceptual space. Our work can be seen as a zoom in on this study, by separating identity (as stored in memory of a known voice) and similarity (as auditory proximity of two consecutive voices). We identify the same components f_0 and $disp(F_5 - F_4)$ as being key auditory features with different roles: $disp(F_5 - F_4)$ encodes identity and f_0 similarity. This results from two combined analyses: first, by identifying that similarity and identity are different processes, showing that two different types of imitation (replicas and caricatures) differently affect these tasks. Second, by relating variability in perceptual performance (in similarity and identity) with variability in the auditory features of the voices.

The poor results obtained with the T_3 (Figure 1) suggest that the period of exposure to a speaker's voice is critical to separate different scenarios of speaker recognition. One in which impersonators convince their listeners by copying specific voice features of the original voice, and another where a separate consideration of the acoustic parameters is not enough for eliciting the identity of the speaker[18].

Humans use faces and voices as strong identity carriers. Although the performance is quite poor for speaker recognition compared to face recognition[1], some parallels can be traced between the visual and acoustic perceptual processing, as was recently suggested by the reconstruction of visual and speech objects from neural populations using similar models[2,19].

An overall conclusion of our experiments is that acoustic caricatures seem to emphasize identity features at the cost of loosing similarity, which allows drawing an analogy between acoustic and visual caricatures. However, this requires a note of caution. In our work we did not work on the idea of exaggeration of vocal features, which is the first thing that naturally comes into mind when one uses the metaphor of caricatures. Instead, our focus was on identifying the acoustical dimensions in which caricatures are effectively proximal to the targets.

Methods

A total of 128 native Spanish speakers (82 females, age 32 ± 13) with normal hearing and no vocal training participated in 2 perceptual experiments. A total of 5 native

Spanish speakers (0 females, age 34 ± 7) participated in the construction of the voice database. All the participants signed a written consent form.

All the experiments described in this paper were approved by the ethics committee *Comité de Ética del Centro de Educación Médica e Investigaciones Clínicas 'Norberto Quirno'* (*CEMIC*) qualified by the Department of Health and Human Services (HHS, USA): IRb00001745-IORG 0001315.

Voice database. We constructed an audio database containing 3 sentences pronounced by public figures (*targets* T_1, T_2 and T_3) and the imitations recorded by 5 professional impersonators. The public figures were selected out of the common repertoire of the 5 professional impersonators: a TV entertainer (T_1), a former Argentinian president (T_2) and a world-wide famous former soccer player (T_3). Target audio files were extracted from public access audiovisual media, selecting the best quality audio files (sampling frequency of at least 22.05 kHz and low reverberation effects). The sentences were chosen from the targets in normal speech situations and their content was selected to avoid inducing strong emotions or stress during impersonation.

The impersonators produced imitations of each sentence in 3 different conditions: first, using written versions of the sentences, they recorded them with their normal voices (n) and impersonating the corresponding targets (caricatures c). Finally, they recorded imitations produced right after listening to the targets (replicas r). In this way, we recorded and stored a database of 48 audio files $\{T_j + I_i n_j + I_i c_j + I_i r_j\}$, $1 \leq j \leq 3$, $1 \leq i \leq 5$ (4.9 ± 2.3 s mean duration).

At the moment of the database construction, the professional impersonators worked at the principal radio stations in Argentina. They were recorded in a low-noise room at a sampling frequency of 22.05 kHz, with a Takstar SGC568 microphone on Praat[20]. Audio files in the database were equalized in loudness. A low-level pink noise (power spectral density $S(f) \propto 1/f$) was added to mask low frequency differences between the copies recorded at the laboratory and the original target files taken from audiovisual media. The audio files of targets T_1 and T_2 and their imitations are available as Supplementary Information on line.

Experiments. The perceptual experiments were written in MATLAB, using Psychtoolbox[21]. Mono audio files at a sampling rate of 22.05 kHz were presented to the participants via headphones Logitech B530 USB Headset MS Linc Optimi.

Experiment 1: identity of the speaker. The participants declared to be able to recognize the characters by their voices. To avoid saliences coming from the topics associated with the public figures[22], each participant completed the experiment for a single target and imitations, i.e. each participant listened to one single sentence uttered by the public figure and his different impersonators. $N_1 = 36$ participants listened to the set $\{T_1 + I_i c_1 + I_i r_1\}$, $N_2 = 22$ to the set $\{T_2 + I_i c_2 + I_i r_2\}$ and $N_3 = 40$ to $\{T_3 + I_i c_3 + I_i r_3\}$, for a total of $N = 98$ participants (42 females, age 31 ± 10) with normal hearing and no vocal training. The participants were asked to associate the audio file with the identity of the target using the following scale: 1 (the voice doesn't belong to the public figure), 2 (it is unlikely that the voice belongs to the public figure), 3 (the voice probably belongs to the public figure), 4 (it is very likely that the voice belongs to the public figure) and 5 (I am sure that the voice belongs to the public figure). The results of the experiment are summarized in Figure 1. We explicitly excluded the participants that did not recognize the voice of the corresponding public figure (that graded the target file with 1).

Experiment 2: voice similarity. We selected the files of the 3 impersonators that produced the higher ranked caricatures and lower ranked replicas of experiment 1. The set $\{T_1 + (I_1 + I_4 + I_5)(c_1 + r_1 + n_1)\}$ was presented to $N_1 = 17$ participants (7 females, age 27 ± 7) and the set $\{T_2 + (I_3 + I_4 + I_5)(c_2 + r_2 + n_2)\}$ to $N_2 = 13$ participants (5 females, age 25 ± 6). Each set consisted on $M = 10$ files, and the participants listened to all pairs $M(M-1)/2 = 55$ in the set in random order. The specific order of appearance of a given pair AB or BA was also randomized. The participants were asked to grade the similarity of the voices of each pair using the following scale: 1 (the two voices are very different), 2 (the two voices are different), 3 (the two voices are similar), 4 (the two voices are very similar) and 5 (the two voices are the same). We excluded the participants whose matrices presented diagonal elements different from 5 (they graded identical files as not been identical).

Data analysis. The audio files and perceptual data were subjected to the following analyses.

Multidimensional scaling (MDS). A standard way to summarize a set of matrices containing dissimilarity measures is to fit them as distances in some kind of perceptual space (usually an euclidean, low-dimensional space) through multidimensional scaling. Several MDS models and techniques have been developed and applied to different musical and vocal spaces. Here we used a standard weighted Euclidean model in which the salience of each dimension is different for each subject (INDSCAL), as provided by Praat[23]. The detailed description of the method can be found elsewhere[13,14,24]. The perceptual spaces for target T_1 and T_2 are shown in the upper panels of Figure 4.

Acoustic space. Each audio file was associated with the 12-dimensional vector of acoustical parameters $V = \{jitter, shimmer, f_0, F_i\ (1 \leq i \leq 5), disp(F_5 - F_1), disp(F_4 - F_3), disp(F_5 - F_3), disp(F_5 - F_4)\}$, using the mean values of the parameters over the length of the sentence. The parameters were calculated using Praat[23] at recommended default values. One important question is if, beyond their mean values, the time evolution of the acoustic parameters is relevant to speaker and voice recognition. Although studies that focused on prosodic aspects were inconclusive[25], some temporal properties as pitch $f_0(t)$, sound intensity $I(t)$ and duration $D(t)$ have been shown to be cues for differentiating voices[26]. In Figure 5 we show these time traces for targets and imitations of T_1 and T_2. With the exception of pitch, the prosodic contours of caricatures (blue) and replicas (green) follow similar patterns for the short sentences used in this work, and were excluded from the analysis. With respect to pitch, we use the mean values to account for the differences between caricatures and replicas.

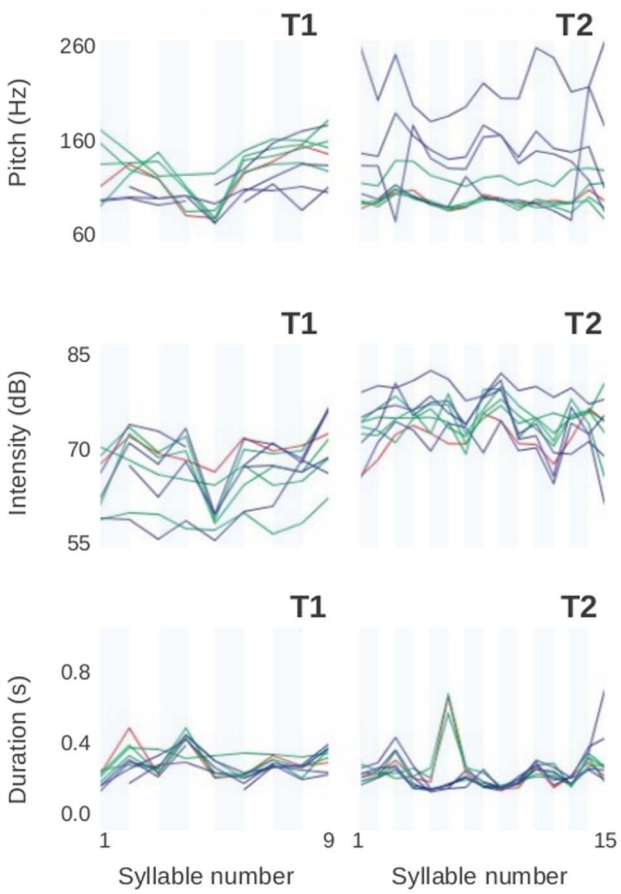

Figure 5 | For the short sentences of our database, prosodic contours do not contribute to the separation of caricatures and replicas. We show the prosodic contours of 3 variables: syllabic duration (upper panels), pitch (middle panels) and sound intensity (lower panels). Targets are shown in red, replicas in green and caricatures in blue. With the exception of pitch, the two types of imitation do not show stereotyped patterns nor clustering into different groups.

Redundancy analysis (RDA). We investigated which acoustic parameters contribute to explain the organization of the perceptual data in the acoustic space V. We submitted the data of the perceptual 2-dimensional spaces calculated with INDSCAL to a redundancy analysis using the statistical toolbox Fathom for MATLAB[27]. The fraction of variance explained for T_1 is 59% and 31% for canonical axes (90% cumulative). For T_2, the fraction explained is 61% and 38% for each canonical axis (98% cumulative). The most salient acoustic parameters are, for both targets, f_0 and *jitter*, that correlate with *dim*1 and *dim*2 respectively. In the case of T_1, f_0 correlates with *dim*2 (r = 0,88) and *jitter* with *dim*1 (r = 0,78). In the case of T_2, f_0 correlates with *dim*1 (r = 0,91) and *jitter* with *dim*2 (r = −0,89). A posterior Monte-Carlo test showed signification ($p = 0.022$ and $p = 0.028$ for T_1 and T_2 respectively), which implies that at least one of the parameters presents an effect in the ordering of the audio files. The ordination distance biplots are shown in the lower panels of Figure 4.

1. Latinus, M. & Belin, P. Human voice perception. *Curr. Biol.* **21**, R143–5 (2011).
2. Eriksson, E. That voice sounds familiar: factors in speaker recognition (2007) at <http://umu.diva-portal.org/smash/record.jsf?pid=diva2:140217>. Accessed 29 October 2013.
3. *A Figure of Speech. A Festschrift for John Laver* [Hardcastle, W. J. & Mackenzie Beck, J. (*eds.*)] (Lawrence Erlbaum Associates, 2005).
4. Hauser, M. D., Chomsky, N. & Fitch, W. T. The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–79 (2002).

5. Markham, D. *Phonetic imitation, accent, and the learner* (Lund University Press, 1997). Accessed 29 October 2013.
6. Assaneo, M. F., Nichols, J. I. & Trevisan, M. A. The Anatomy of Onomatopoeia. *PLoS One* **6**, e28317 (2011).
7. Kitamura, T. Acoustic analysis of imitated voice produced by a professional impersonator. *INTERSPEECH* 813–816 (2008) at <http://basil.is.konan-u.ac.jp/pub/is2008.pdf>. Accessed 21 July 2013.
8. Zetterholm, E. Same speaker – different voices. A study of one impersonator and some of his different imitations. in *Proc. 11st. Aust. Int. Conf. Speech Sci. Technol.* (Warren, P. & C. I.) 70–75 (2006).
9. Titze, I. R. *Principles of voice production.* 354 (Prentice Hall, 1994).
10. Fant, G. *Acoustic theory of speech production* (Mouton De Gruyter, 1970).
11. Titze, I. The physics of small-amplitude oscillation of the vocal folds. *J. Acoust. Soc. Am.* 1536–1552 (1988) at <http://link.aip.org/link/jasman/v83/i4/p1536/s1>. Accessed 16 May 2013.
12. Zetterholm, E. The same but different-three impersonators imitate the same target voices. in *Proc. 15th Int. Congr. Phonetic Sci.* 2205–2208 (2003).
13. Baumann, O. & Belin, P. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychol. Res.* **74**, 110–20 (2010).
14. Murry, T. & Singh, S. Multidimensional analysis of male and female voices. *J. Acoust. Soc. Am.* **68**, 1294–1300 (1980).
15. Collins, S. Men's voices and women's choices. *Anim. Behav.* **60**, 773–780 (2000).
16. Fitch, W. T. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J. Acoust. Soc. Am.* **102**, 1213–22 (1997).
17. Samson, S., Zatorre, R. J. & Ramsay, J. O. Multidimensional scaling of synthetic musical timbre: perception of spectral and temporal characteristics. *Can. J. Exp. Psychol.* **51**, 307–15 (1997).
18. Zetterholm, E. in *Speak. Classif. II SE - 16* (Müller, C.) **4441**, 192–205 (Springer Berlin Heidelberg, 2007).
19. Pasley, B. N. *et al.* Reconstructing Speech from Human Auditory Cortex. *PLoS Biol.* **10**, e1001251 (2012).
20. Boersma, P. & Weenink, D. Praat: doing phonetics by computer (2013), at <http://www.praat.org/>. Accessed 30 October 2013.
21. Brainard, H. D. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
22. Eriksson, E. J. *et al.* Detection of imitated voices, who are reliable earwitnesses? *Int. J. Speech Lang. Law* **17**, 25–44 (2010).
23. Boersma, P. & Weenink, D. Praat: doing phonetics by computer (2013), at <http://www.praat.org/>. Accessed 30 October 2013.
24. Caclin, A., McAdams, S., Smith, B. K. & Winsberg, S. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *J. Acoust. Soc. Am.* **118**, 471 (2005).
25. Farrús, M., Wagner, M., Erro, D. & Hernando, J. Automatic Speaker Recognition as a Measurement of Voice Imitation and Conversion. *Int. J. Speech Lang. Law* **17**, 119–142 (2010).
26. Farrús Cabeceran, M. *Fusing prosodic and acoustic information for speaker recognition* (Universitat Politecnica de Catalunya 2008). Accessed 30 October 2013.
27. Jones, D. L. FATHOM for Matlab at <http://www.marine.usf.edu/user/djones/matlab/matlab.html>. Accessed 30 October 2013.

Acknowledgments

We want to thank the professional impersonators for their kindness and patience during the recording sessions. This work was partially funded by the University of Buenos Aires (UBA) and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). M.S. is sponsored by the James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition - Scholar Award.

Author contributions

S.L. and M.A.T. designed the experiments; P.R., M.E. and M.A.T. programmed the experiments; S.L. and M.F.A. conducted the experiments; S.L., P.R., M.F.A., M.S. and M.A.T. analyzed the data; M.S. and M.A.T. wrote the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Supplementary audio files are available online. Audio file S1 contains the target T_1. Audio files S3, S5, S7, S9 and S11 are the replicas, and S2, S4, S6, S8 and S10 the caricatures produced by impersonators I_1 to I_5 respectively. Audio file S12 contains the target T_2. S14, S16, S18, S20 and S22 are the replicas, and S13, S15, S17, S19 and S21 the caricatures produced by impersonators I_1 to I_5 respectively.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: López, S. *et al.* Vocal caricatures reveal signatures of speaker identity. *Sci. Rep.* **3**, 3407; DOI:10.1038/srep03407 (2013).