Trabajo Final de Grado

Artículo científico de producción empírica

Dynamic trajectories in the vocal space: Mapping basic acoustic features of the voice with listeners' subjective perception in a public figure identification task.

Franco Miguel Caballero Rosas – 5.154.876-0

Tutores: Dr. Francisco Cervantes y la Dra. Victoria Gradín

Revisora: Dra. Emilia Fló

Facultad de Psicología

Universidad de la República

Montevideo, Uruguay

Septiembre, 2024

Dynamic trajectories in the vocal space: mapping basic acoustic features of the voice with listeners' subjective perception in a public figure identification task.

Abstract:

Using the source-filter model of voice production, previous research has utilized multidimensional maps to quantify differences between speaker voices and explain them in relation to various perceptual dimensions. These approaches may be extended by integrating dynamic aspects of the voice physical signal. It is unclear whether the informational similarity between different voices' acoustic distributions leads to subjective perception of such signals as being from the same person. In this work, we examine a method for representing voices in their important acoustic dynamic dimensions and evaluate these representations in terms of subjective judgment. A database of speech stimuli from famous public figures (targets) was created, along with a collection of verbatim imitations produced by professional impersonators. Imitation pair sets were generated before and after listening to the reference target stimulus. We quantified acoustic similarities between targets and all imitations using two-dimensional probability maps of joint pitch and the average inter-formant distance dimensions, which represent the source and filter stages of vocal production, respectively. Subjective identification data were collected in a large-scale online experiment, where participants evaluated the correspondence of each voice stimulus with the target person. The results showed that the degree of acoustic similarity an imitation has with the original target voice is related to the likelihood of identifying such imitation as a genuine voice. This demonstrates that acoustic similarity mappings are relevant for subjective voice perception and provide a valid way to represent voices. This may open new research directions in voice perception by integrating the dynamic trajectories of the voice's spectral features during production.

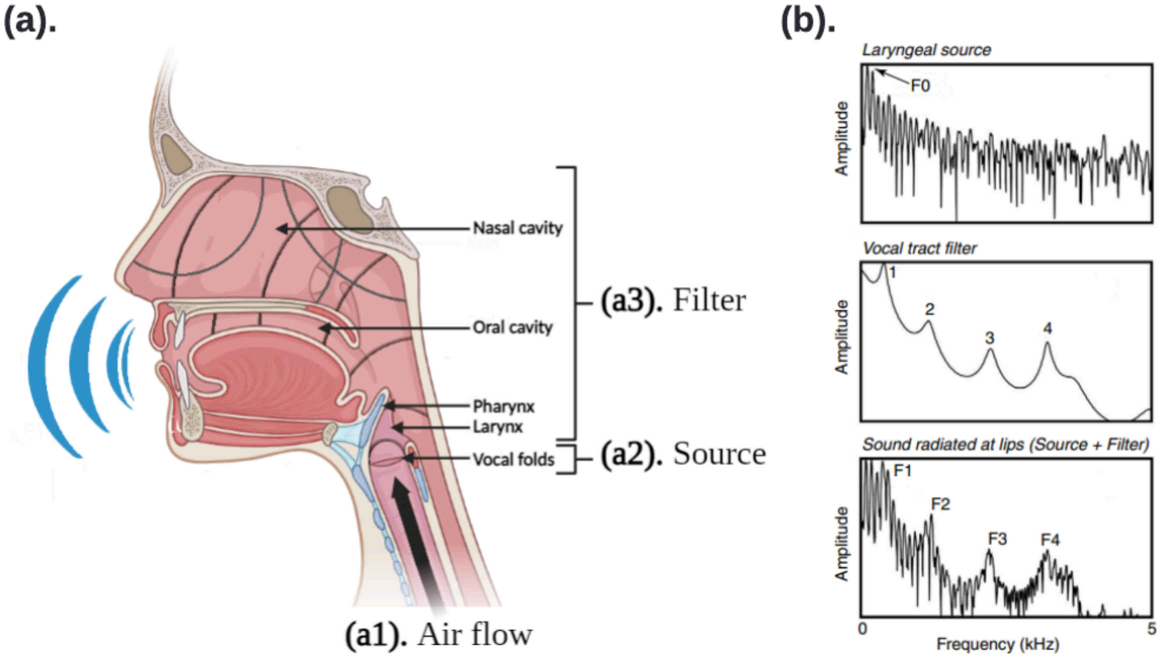
Keywords: *source-filter model, voice imitation, dynamic vocal space, multidimensional mapping, public figure identification, acoustic features*

Introduction

Voice production is a complex phenomenon that has been studied for decades (Simonyan et al., 2016). This process involves a series of physiological and acoustic mechanisms that enable the generation of the sounds that we use to communicate. Speech contains a wealth of information beyond semantic content (Titze, 2000), for example, information crucial for social communication such as identifying the speaker's body size and identity (Ghazanfar & Rendall, 2008). Many characteristics of a person, such as their gender, age, and emotional state, can be discerned even from small samples of their voiced speech (Latinus & Belin, 2011). These signals may allow us to form an instant impression of the speaker and infer some of their traits even without knowing them (Lavan & McGettigan, 2023).

The source-filter model has been a cornerstone in the study of speech production. This model separates the process into a sound production (source) stage first, followed by a modulation stage (filter) (Fant, 1960). The source stage output is the sound generated by the larynx, where the vocal cords are located. When air is expelled from the lungs and passes through the vocal cords, the latter vibrate and generate the sound wave propagating through the air medium within the vocal tract. Its primary characteristics are amplitude (volume), timbre (a quality that allows distinguishing between different sound sources), duration, and most importantly, frequency. Voice sounds are typically produced by numerous frequencies, measured in Hertz (Hz) units, which represent the number of cycles per second that the different vibrations of a sound wave have. Importantly, at this stage, the sounds produced by our vocal cords contain a fundamental frequency (f_0) and its harmonics, where f_0 is the base oscillation frequency of a sound that determines its basic pitch. Harmonics are integer multiples of the fundamental frequency that affect its timbre quality (Figure 1a2).

At the second stage of production, filtering, the sound produced by the cords is modified as it passes through the vocal tract due to cavity resonance effects. The vocal tract may change its shape and size from moment to moment, further altering the acoustic properties of the originally generated sound over time. At any time, the shape and configuration of the vocal tract will determine which harmonic frequency groups that may be amplified and those which are attenuated. These important resonances are known as formants (F_1, F_2, …, F_5) and represent peaks in the frequency spectrum of the human voice. The formants are essential in identifying vowels and some consonant sounds (Peterson & Barney, 1952) (Figure 1a3).



Current Biology

Figure 1 | Key mechanisms and features of voice production. (a) Sagittal section of the vocal tract, showing the vocal cords and orofacial cavities relevant in voice production. The lungs, diaphragm and trachea together supply, control and conduct the necessary airflow (a1) for phonation in the larynx vocal cords (a2). The vocal cords vibrate with the incoming airflow producing sound, which subsequently propagates at supraglottic structures (a3). These structures include the pharynx, the oral cavity, and the nasal cavity, and their configuration jointly which modify the sound produced by the vocal cords. (b) Signatures of the source-filter model in the speech frequency spectrum. Top: frequency spectrum of voiced sound at the larynx, where the vibration of the vocal cords produces regularly spaced peaks. The fundamental frequency (f_0) is the lowest peak. Middle: first four resonances of the vocal tract (peaks numbered 1-4) for an example vowel spoken by a man. Bottom: spectrum of speech at the lips, which combines the laryngeal spectrum (top) with the vocal tract transfer function (middle). The resulting sound contains the harmonic peaks of the f_0 and a frequency envelope shaped by the vocal tract resonances, known as formants (F_1-F_4). Figure 1a was created using BioRender (*https://www.biorender.com/*); Figure 1b was reproduced from Ghazanfar and Rendall (2008).

Anatomical, physiological, and functional factors impose differences in the vocal tract that result in vocal differences among individuals (Titze, 2000; Stevens, 2000). One of the most evident variations is the sexual dimorphism between men and women, but voice differences may exist across different languages or dialects spoken by individuals (Johnson, 2005). To unify these differences, Johnson (2020) proposed a method to represent speech irrespective of vocal tract characteristics. In his normalization approach, the average spacing between formants (ΔF) is used to re-scale vowel formant frequencies, effectively removing acoustic differences that may be due to differences in the vocal tract length. This method has proven effective in standardizing vocal measures and obtaining the average separation of formants regardless of speaker, facilitating comparisons of the acoustic properties of speech among individuals with different dialects or languages. This method explicitly associates ΔF with anatomical correlates of vocal tract length, and demonstrates usefulness in vowel classification and separation, even with few vowel tokens.

The study of voice identification and recognition may be similar to processes involved in face recognition. In the visual domain, Valentine (1991) proposed the hypothesis that recognizable faces may be located in a multidimensional space, representable as points, with axes representing the dimensions in which the faces are mainly encoded. At the center of the coordinate system, an "average face" as obtained from experience is hypothesized where and the farther a face is from this point, the more unusual or singular it is. Under this model, the position of each face as a point is determined by its visual characteristics, with the distance between points representing the similarity between pairs of faces. Following this analogy, Belin et al. (2004) referred to the voice as the "auditory face", hypothesizing that voices may be similarly represented in a common space. One key dimension of this space could be vocal timbre, which is often used to distinguish one voice from another, even if they are producing the same pitch and volume (Cleveland, 1977). What is the most appropriate system of acoustic features to represent voices remains an open question. The source-filter model has been previously employed to identify psychoacoustically relevant parameters to define this space (López et al., 2013; Baumann & Belin, 2008; Chhabra et al., 2012). In voice recognition, the perceptual processes of discrimination may depend on whether the voices are familiar to us (Stevenage, 2018). For example, there is evidence that the recognition of voices to which a listener has been exposed for a considerable amount of time, e.g., family members or famous speakers, is notably accurate even when speakers try to disguise their voices (Hollien et al., 1982; Lavner et al., 2000; Kriegstein & Giraud, 2004; Van Lancker et al., 1985).

In the present study, we hypothesized a vocal space whereby at each time, each utterance from a familiar voice may be positioned on the map according to specific acoustic properties of the dynamic speech stimulus (Figure 2). Due to familiarity, associated with the location is a hypothesized "recognition space," within which, if another vocal stimulus shares similar physical characteristics in this space (i.e., approaches closely enough), there is a chance that both voices will be identified as from the same person.

To address this hypothesis, we invited professional impersonators to imitate well-known voices from people famous in the Uruguayan public scene, to the best of their ability. Imitations were produced under two different styles: first, based on the impersonator's knowledge of a target, referred to as “caricatures”; second, as “replicas” produced upon exposure to an original stimulus from the target with the intent to achieve close resemblance acoustically (López et al., 2013). We subsequently presented the original target stimuli and imitations to listeners in an online study, for them to rate these stimuli as belonging to the target figure or not.

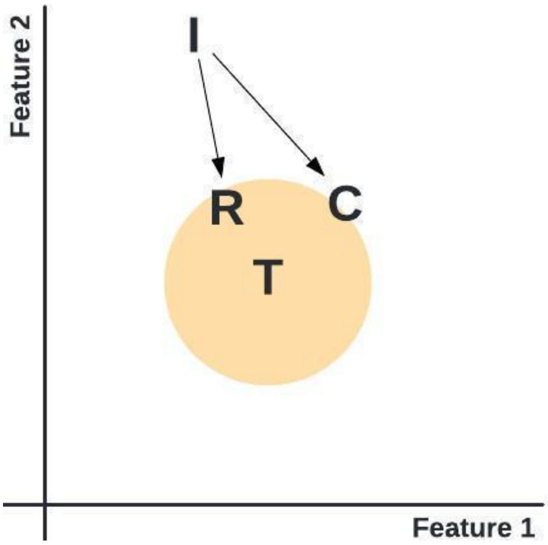


Figure 2 | Representation of a generic two-dimensional voice space: In a hypothesized two-dimensional field, each position represents a voice utterance. "T" indicates the position of a target's voice, while "I" is that occupied by an impersonator with his natural speech. In this model, imitations made either through the replica ("R") or caricature ("C") methods may effectively approach the physical vicinity that defines the perceptual area associated with the target (orange). Shorter distances (e.g., "R" to "T") increase the probability that an imitation will be identified with the target. The model predicts that when "R" and "C" are at the same distance from "T," they may attain equal recognition chances, even if they occupy distinct positions in the field.

To extend this model, we propose to consider a voice space whose locations also inform the dynamic and variable nature of the speech signal as it unfolds over time, via temporal histogram distributions. Here, we address the possibility that imitation-target distances in such a low-dimensional representation may predict the likelihood that imitations are perceived as genuine by listeners in a voice identification. By comparing between the two imitation types, impersonator-independent measures provided an indication of acoustic similarity and whether it predicted which of the two more closely resembled a target voice in participants' responses.

Methods

Subjects. Participants that were 18 years of age or older and residing in Uruguay for at least 10 years were invited to the study. The initial pool of valid respondents included a total of 883 participants, of whom 564 (63.87%) were female, 314 (35.56%) were male, and 7 (0.79%) did not respond about their gender. The average age of the participants was 45.6 years (± 14.6 SD). The participants had an average residence in the country of 45.15 years (± 14.9 SD). 764 (86.5%) of participants were right-handed, 92 (10.4%) were left-handed, and 27 (3.1%) were ambidextrous. The native language of the participants was predominantly Spanish with 876 (99.21%) speakers, while 7 (0.79%) participants spoke natively Portuguese.

From the total of initial participants, 558 completed the minimum number of experiments required (one). Of these, 362 (64.87%) were female, 194 (34.76%) were male, and 2 (0.35%) did not respond about their gender. The average age was 46.91 years (± 14.07 SD), and the average residence in the country was 46.36 years (± 14.36 SD). Regarding handedness, 482 (86.37%) of the participants were right-handed, 54 (9.96%) were left-handed, and 22 (3.94%) were ambidextrous. The native language of the participants was predominantly Spanish, with 554 (99.28%) speakers, while 4 (0.71%) participants spoke natively Portuguese.

The experimental procedures were approved by the Human Research Ethics Committee (CEIH) of the Instituto de Investigaciones Biológicas "Clemente Estable" on June 29, 2022, under registration number 013-3, and the proposal was registered with the Ministry of Public Health on June 6, 2022 (registration number 7221843).

Stimuli. Auditory stimuli were constructed based on public domain sentence samples obtained from 10 public figures ("targets"). These male figures are mainly from the fields of politics, sports, and entertainment in Uruguay, namely, Alberto Kesman (keyword "KE"), sports commentator; Guido Manini ("MA"), politician; Jaime Roos ("RO"), musician; José Mujica ("MU"), politician; Julio María Sanguinetti ("SA"), politician; Luis Alberto Lacalle ("LA"), politician; Luis Suárez ("SU"), football player; Óscar Tabárez ("TA"), football coach; Pablo Bengoechea ("BE"), football coach; and Sergio Puglia ("PU"), chef. Audio recordings of the target speakers were extracted from publicly accessible audiovisual materials, representing normal speaking conditions without clear emotional or stress cues during production.

In addition, imitations of these targets were recorded from five different male professional impersonators. These recordings were obtained in person prior to this study on separate days. Each impersonator first freely selected the individual target figures that they were prepared to imitate. After reading the transcript of the chosen target's sentence, the impersonator was asked to produce a verbatim imitation whereby they read aloud the sentence and performed an ad lib imitation based on their skill with the target speaker (caricature condition). Subsequent to recording and once they were reading to continue, the impersonator was presented with the audio sample of the target's original recording. The impersonator was then asked to reproduce the spoken sentence as faithfully as possible (replica condition). In both cases, impersonators were asked to retry if they did not produce the exact phrase and were allowed to retry if they chose to do so. For replica conditions, target audios could be re-played at the imitator's request. Imitations were never played back to the impersonators.

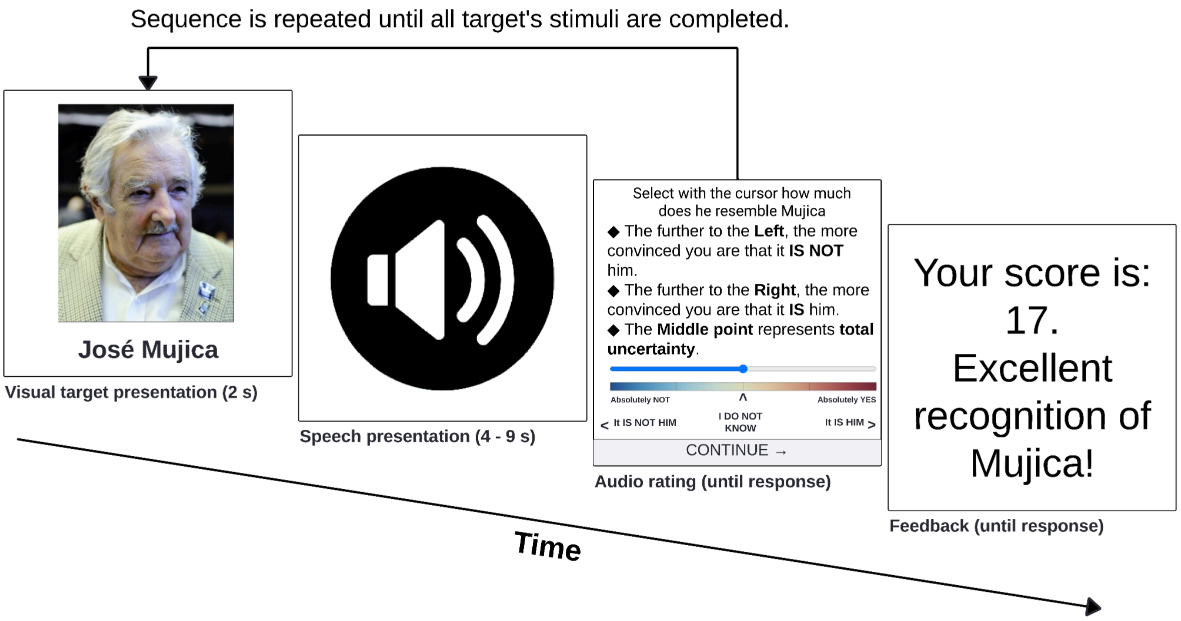
Each sentence and its audio recordings (one target plus two imitation audios) were equally divided into 2 or 3 segments. The procedure resulted in a total of 143 audio recording segments, including 23 target recordings, 60 caricatures, and 60 replicas. Individual segments ranged from 4 to 10 seconds in duration.

Each recording was obtained using an Olympus recorder model VN-541OC (OM Digital Solutions Corporation, Japan) in a low-noise, enclosed space. The audio files were processed in MATLAB at a sampling frequency of 44100 Hz. Cosine ramps with a duration of 5 ms were applied at the beginning and end of all stimuli. Additionally, prolonged silences and repeated filler words were manually removed.

Setup. An online task study was designed to be executed on desktop and mobile devices and was developed using the lab.js platform interface (https://lab.js.org/), a JavaScript library designed for online experiments in the cognitive and behavioral sciences. The study was implemented and administered on the Open Lab platform (https://open-lab.online), which also facilitated data hosting and management. For accessibility and distribution purposes, the presently extinct domain "http://voces.info" was created. The study was disseminated to the local population through single radio and television announcement broadcasts. Subjects freely opted to participate in the experiment using the devices (e.g., headsets) available to them and on the browsers of their choice.

Task. The experiment was divided into an initial introductory practice and instruction phase, followed by the main phase of the experiment. Subjects were then given a third, optional final phase to provide comments and suggestions on the study at the end of their participation. In the introductory phase, participants were first asked to carefully read and, if they agreed and wished to participate, accept the informed consent. They then provided demographic information, including age, years of residence in the country, gender, dominant hand, and native language. Subsequently, task instructions were presented corresponding to the main experimental phase, and participants completed practice trials using three additional public figures that were not included in the main target pool. After the practice trials, participants could begin the main phase of the experiment, which they started by selecting one from multiple target public figures from a main menu screen. Each option was presented with a picture of the target public figure, an indication of task difficulty for that target (based on the available number of imitations for that target), and the expected duration in minutes. Target blocks ranged between 6 and 33 trials. Selection of a target block option initiated the experimental sequence block related to that target's stimuli set. Upon finishing a target block, participants were redirected to the main menu screen, where they could choose to perform another target block, provide optional feedback, or end the session.

Experiment sequence. Upon selection of a target block, participants entered a sequence of trials that presented all individual stimuli from the selected public figure set, including all target, replica, and caricature audio recordings. Each trial began with the name and image of the target figure to be identified, shown for 2 seconds (Figure 3).


Sequence is repeated until all target's stimuli are completed.
José Mujica
Visual target presentation (2 s)
Speech presentation (4 - 9 s)
Audio rating (until response)
Feedback (until response)
Time
Absolutely NOT
Absolutely YES
< It IS NOT HIM
I DO NOT KNOW
It IS HIM >
CONTINUE →

Select with the cursor how much does he resemble Mujica

◆ The further to the **Left**, the more convinced you are that it **IS NOT** him.

◆ The further to the **Right**, the more convinced you are that it **IS** him.

◆ The **Middle point** represents **total uncertainty**.

Your score is: 17. Excellent recognition of Mujica!

Figure 3 | Main experiment sequence: Representation of the main sequence of the experiment illustrating the four steps involved in each trial.

After the visual presentation, the auditory speech stimulus was presented to the participant. Individual speech signals were accompanied by a generic speaker icon to indicate that audio was playing. Following the speech stimulus, a response screen was displayed for the participant to indicate how convinced they were that the voice they heard belonged to the person whose image was shown. Ratings were given on a hundred-point range slider, where a value of 1 indicated the participant was completely sure that the voice did not belong to the public figure, and a value of 100 indicated they were absolutely sure it did. A position of 50 implied that the participant was unsure whether the voice corresponded to the target public figure or not. After marking and validating the trial’s rating, the next trial in the sequence began. All steps were repeated until all stimuli related to the target public figure had been presented and rated by the participant. Upon completion of the target figure’s block sequence, participants were shown with a final feedback slide that scored their performance in the target block. This feedback display consisted of a variable message, randomly selected from a pool of five options (e.g., “Your score is:”; “You achieved”; “You scored”; “Score reached by this public figure:”; “You earned”). The obtained score was displayed below, followed by a qualitative performance indicator (see Table 2).

Scoring system. To assess each participant's accuracy in the identification task, each stimulus rating contributed to the total score computation for the block as follows: Each trial stimulus was assigned a choice value (0 or 1) and a weight value (-1 or 1). Choice assignments depended on whether the participant predominantly believed (rating > 50) that the stimulus was from the target, or that it was not (rating ≤ 50). Weight values indicated the actual authenticity of the voice (1 for a true target voice, -1 for an imitation). See Table 1 for an example. Using the set of ratings, choice values, and weight values for all trials, a d-prime (d’) value was computed for each block. The d’ value measures the participant's ability to distinguish the true voice signal, with greater values indicating a higher likelihood of discriminating the voice signal from noise. Values close to 0 indicate poor ability to distinguish signal from noise, while lower values indicate worse-than-chance performance, which may result from systematic response bias and/or an incorrect decision rule.

Choice	1 (score > 50)	1 (score > 50)	0 (score <= 50)	0 (score <= 50)
Weight	1 (target audio)	-1 (replica or caricature audio)	1 (target audio)	-1 (replica or caricature audio)
Results	The participant correctly identifies a target audio (hit).	The participant incorrectly identifies an imitation as authentic (false positive).	The participant incorrectly identifies a target audio as (miss).	The participant correctly identifies an imitation (correct rejection).

Table 1 | Combinations of conditions and their results in audio identification. The table illustrates the combinations of the 'choices' and 'weight' conditions and their associated outcomes.

For each block, feedback was based on the average value of two d’ measures: one obtained based on the participant's rating values and another from the participant's binarized choices. To calculate the rating-based d’ measure, the block’s hit rate (‘hitRate’) was determined by summing the weighted ratings of the authentic target audio instances and normalizing this sum by the maximum possible points from target instances in the block. The false alarm rate (‘faRate’) was similarly calculated by summing the weighted ratings of the imitation audio instances.
To obtain a choice-based d’, a hit rate ‘hitRateCh’ was calculated based on the number of binarized positive choices divided by the number of authentic target trials. The choice-based false alarm ‘faRateCh’ was similarly computed from the number of positive choices of imitations divided by the number of imitation trials.
In either rating- or choice-based cases, the d’ was calculated by z-transforming each rate (using the inverse error function) and subtracting them from each other according to the formula:

$$d' = z(hitrate) - z(falsealarm) \quad (1)$$

A correction was applied to constrain this value, which for scoring presentation purposes was then multiplied by 10 and rounded. This quantitative score was accompanied by a qualitative evaluation according to the level of achieved discrimination (Table 2). This feedback was intended to create an engaging, customized, and interactive experience for the participant, potentially encouraging them to continue with another target block.

Score obtained by participants	Message shown to participants
> 21	*Perfect recognition of *public figure*! Congratulations!*
18 – 21	*Excellent recognition of *public figure*!*
15 – 18	*Very good recognition of *public figure*!*
12 – 15	*Good recognition of *public figure* among their impersonators.*
9 – 12	*Moderate recognition of *public figure* among their impersonators.*
6 – 9	*Poor recognition of *public figure* among their impersonators.*
3 – 6	*Doubtful recognition of *public figure* among their impersonators.*
-3 – 3	*No recognition of *public figure* among their impersonators.*
-3 – -15	*You were fooled by the impersonators!*
≤ -15	*Review the instructions and try again.*

Table 2 | Scoring chart for a target block with qualitative assessments of performance for participant feedback.

Data processing

Speech spectral feature analyses. The raw stimulus audios, including the targets, replicas, and caricatures, were processed and filtered using MATLAB as in Cervantes Constantino and Caputi (2024). For each audio signal, a fixed 6-band gammatone filterbank was applied to assess spectral variations corresponding to male f_0 and formants' F_1-F_5 frequencies (Figure 4). Filterbank coefficients were based on six centre frequencies fc_i set at 146, 671, 1870, 2870, 3946, and 4869 Hz that we defined from a previous analysis of a local public voice database (López et al., 2013). Coefficients were estimated with the Large Time/Frequency Analysis toolbox (Průša et al., 2014; Søndergaard et al., 2012). Sub-bandwidths equalled thrice the critical bandwidth of the auditory filter at the corresponding fc_i, and were defined in equivalent rectangular bandwidths (Glasberg & Moore, 1990). The real part of the filter coefficients was applied to each audio signal. Each filter output was analyzed to obtain instantaneous frequency $f_i(t)$ and, separately, instantaneous amplitude $a_i(t)$ content using the Hilbert method. To perform a multiband demodulation analysis, a mean-amplitude weighted short-time estimate $F_i(t)$ of the instantaneous frequency $f_i(t)$ was computed per sub-band (Grimaldi & Cummins, 2008). Integration of the weighing function was approximated by using a centered moving average with a t=25 ms window, according to the formula:

$$F_i[t=t_0]=\frac{\sum_{t=t_0}^{t=t_0+\tau}[f_i(t)\cdot a_i^2(t)]}{\sum_{t=t_0}^{t=t_0+\tau}[a_i^2(t)]} \qquad (2)$$

The resulting set of discrete short-time instantaneous frequency estimates, sometimes referred to as the pyknogram of the speech signal (Grimaldi & Cummins, 2008), was used to represent the dynamic waveform timeseries for f_0 and formants F_1-F_5.

The dynamic waveform for the formant dispersion ΔF measure was constructed from k=[1,…,K] successive independent linear regression models, with K the total of time samples in the stimulus. At the k-th sample, the formant model is:

$$F_i[t=t_k]=\beta_{0,k}+\beta_{1,k}n_i+\varepsilon_k \qquad (3)$$

with formant frequencies F_i taken as observations at sample k, and scaled by their formant numbers n={1,2,3,4,5} which are taken as predictors. Model coefficient estimation was performed via an iteratively reweighted least squares algorithm with the bisquare weighting function (MATLAB robustfit). The regression slope served as the measure of formant dispersion:

$$\Delta F[t=t_k]=\beta_{1,k} \qquad (4)$$

The seven resulting spectral contour timeseries waveforms (f_0, F_1-F_5, ΔF) were downsampled to 1024 Hz.

To discount the spurious spectral contributions of silent periods and weak transients in natural speech, all periods that corresponded to the signal's envelope below a certain threshold were automatically discarded. The threshold was set per speech sample at the bottom quintile of the full envelope histogram for that speech signal. To obtain the full envelope, the raw speech signal was submitted to a bank of 28 gammatone filters with ERB-spaced centre frequencies corresponding to the 52 – 4956 Hz range. From each sub-band, the envelope was extracted by taking the magnitude value of each sample and raising it to the power of 0.6 (Vanthornhout et al., 2018). The resulting sub-band envelopes were averaged, and this total envelope was resampled to 1024 Hz. Below-threshold epochs were jointly discarded in corresponding periods of the envelope and the seven spectral features.

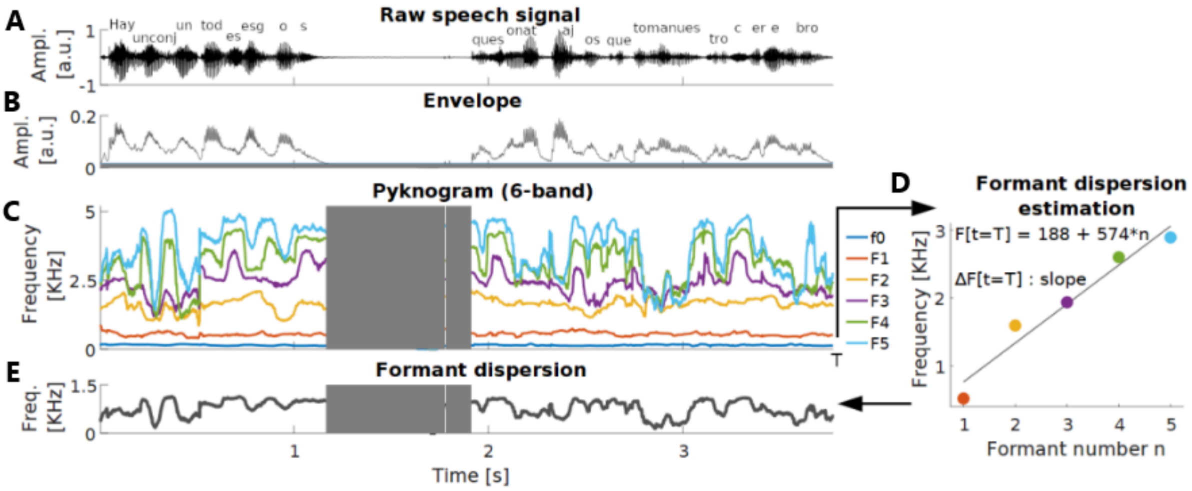


Figure 4 | Setup for extraction of spectral features of voiced speech. (A) Example of a single-speaker speech phrase as presented in a trial. (B) Envelope of the speech example at the original sampling rate. (C) Short-time estimates (mean amplitude-weighted) of instantaneous frequency from the speech example correspond to six spectral bands representing the fundamental frequency f_0 and formants F_1-F_5. Silent regions (grey areas) are not further analyzed. (D) Method of formant dispersion dynamics ΔF estimation, based on sample-by-sample linear regression of formant frequencies. For this example, ΔF at time T is 574 Hz. Figure adapted from Cervantes and Caputi (2024).

Subsequently, visual representations of each stimulus were created in the form of bivariate joint probability density distributions (two-dimensional histograms) with f_0 and ΔF as their dimensions (see Figure 5 for an example of a stimulus segment set). This graphical representation allows visualizing data matrices where the values are represented by a color scale, facilitating the identification of recurrent acoustic patterns in the speech sample.

Here, the X-axis represents the distances between formants (representing the filter contribution), and the Y-axis represents the fundamental frequencies (representing the source contribution). Histograms were created with instantaneous formant distance estimates ranging from 25 Hz to 2000 Hz in increments of 25 Hz bins, and instantaneous fundamental frequency estimates ranging from 0 Hz to 500 Hz in increments of 10 Hz bins. In all cases, classes of audios were compared where the lexical content is equal, i.e., with the same vowels and consonants, but the way they are expressed, and possibly the speakers themselves, are different.

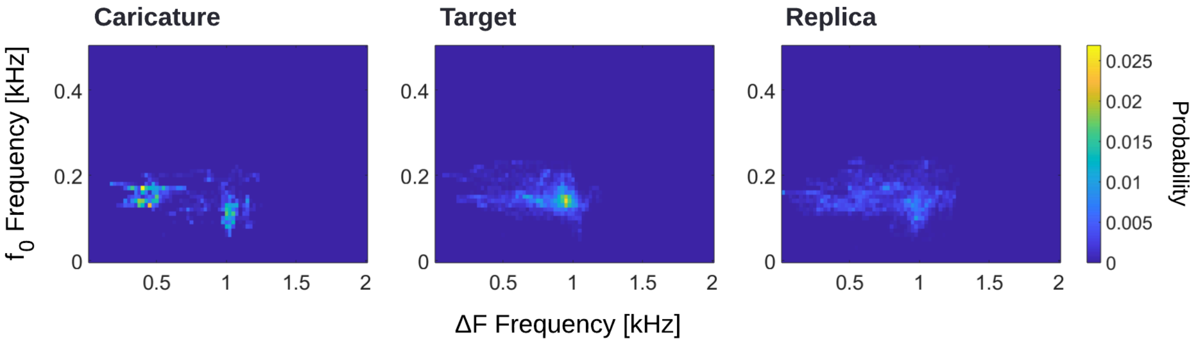


Figure 5 | Stimulus comparison of bivariate probability distributions between target,

caricature, and replica versions of the same sentence: 2D histograms show the joint density of the fundamental frequencies (f_0) and the formant distances (ΔF) for three different audio versions (caricature, replica, and target). Each bivariate probability distribution illustrates the relationship between a speaker's f_0 variations during an utterance and corresponding changes in ΔF. The color intensity in each cell represents the number of occurrences within the defined bins, with more intense colors indicating a higher occurrence rate. The sentence used for creating the 2D histograms (target: "SA") reads: "*Globalization presents us with a scientific and technological explosion in constant change... political systems that are overwhelmed....*"

Representation of acoustic similarity

Acoustic differences between target and imitation stimuli (i.e., caricature versus target, "C-T", and replica versus target, "R-T") were computed using the Jensen-Shannon divergence (JSD) between the corresponding 2D probability distribution pairs. The JSD is a symmetric and bounded measure parameter used to compare informational similarity between probability distributions and is defined even when both distributions have disjoint supports (as occurs, for example, when certain (f_0, ΔF) combinations are present in one stimulus but not the other). A JSD of 0 indicates that the compared distributions are identical. Increasing JSD values reflect greater dissimilarity or separation between the two distributions. After calculating the JSD values for both R-T and C-T pairs per stimulus triplet set, we defined that set's acoustic difference parameter Δφ as the difference between the R-T and C-T divergences:

$$\Delta\varphi = \mathrm{JSD}(\,R \,\|\, T\,) - \mathrm{JSD}(\,C \,\|\, T\,) \quad (5)$$

This acoustic difference parameter was used to estimate which of the two imitation modes, for a given speech segment, lies closer to the target in the (f_0, ΔF) informational space. If $\Delta\varphi > 0$, it means that the replica imitation is acoustically farther from the target compared to the caricature version, with the caricature audio being more similar to the target. Conversely, when $\Delta\varphi < 0$, it indicates the opposite. A Δφ of 0 indicates that both imitation modes are equally distant from the target in this space.

Representation of perceptual similarity

Next, we defined the subjective perceptual similarity of the imitation stimuli by subtracting the average rating attained by each imitation audio from replica and caricature modes. As in the case of acoustic similarity, we evaluated which of the two types of imitation modes, replica or caricature, was perceptually more convincing or better rated by the participants. This approach also ensured that across-imitator perceptual differences due to skill were removed. We defined the perceptual difference parameter Δψ as:

$$\Delta\psi = \overline{\psi}_{replica} - \overline{\psi}_{caricature} \quad (6)$$

The result of this difference for each pair of audios can be interpreted as follows: when Δψ is positive, the replica imitation was better rated than the caricature. Participants were relatively more likely to attribute replica imitation to the target than the caricature. Conversely, a negative value indicates that for a particular speech segment, the caricature imitation was better rated than the replica. A zero value indicates no difference in the subjective evaluation between the replica and the caricature, with neither outstanding in terms of perceived authenticity.

Results

In the online study, target samples received the highest ratings by participants, with an average of 79.947 ± 17.409 SD and a median of 84.794, suggesting a typical association of these stimuli to celebrities' actual voices, as expected. In contrast, the caricature and replica audios received lower ratings on average (caricature 10.952 ± 8.191 SD; replica 12.335 ± 8.334). The imitation median ratings (7.501 and 8.419, respectively) similarly indicate a consistent tendency to recognize the imitation voice samples as belonging to other people (Figure 6).

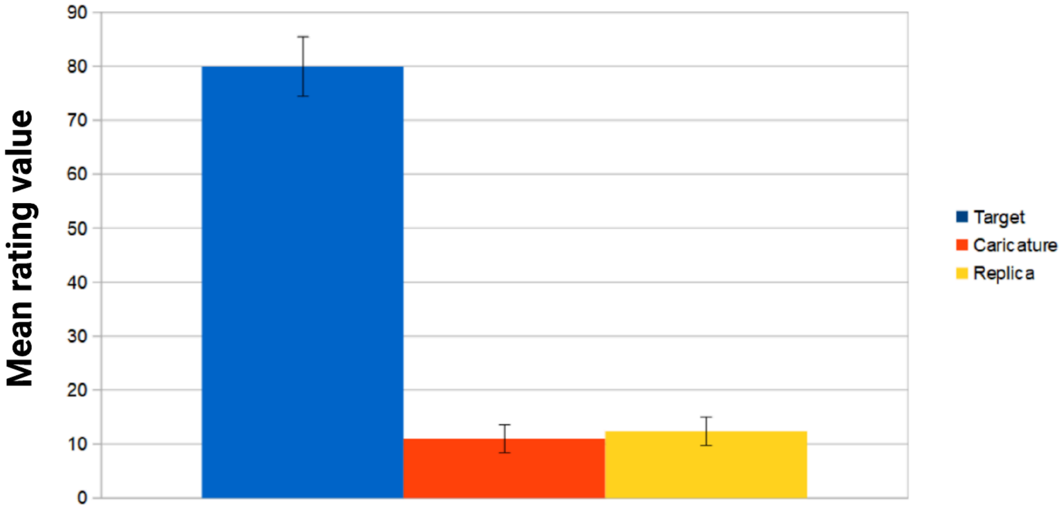


Figure 6 | Average rating results for target and imitation stimuli: Across all target and impersonator voices, participants were more likely to correctly identify authentic audios from public figures' actual voices.

The overall pattern of better recognition of genuine target audios was further analyzed according to the public figures (Figure 7). When examined by public figures, target audios were again perceived as the most authentic. However, for a few public figures (e.g., MU), target identification yielded even lower average scores, with notably higher ratings for their imitations. This may suggest that, for these public figures, impersonations were more successfully credible in relative terms.

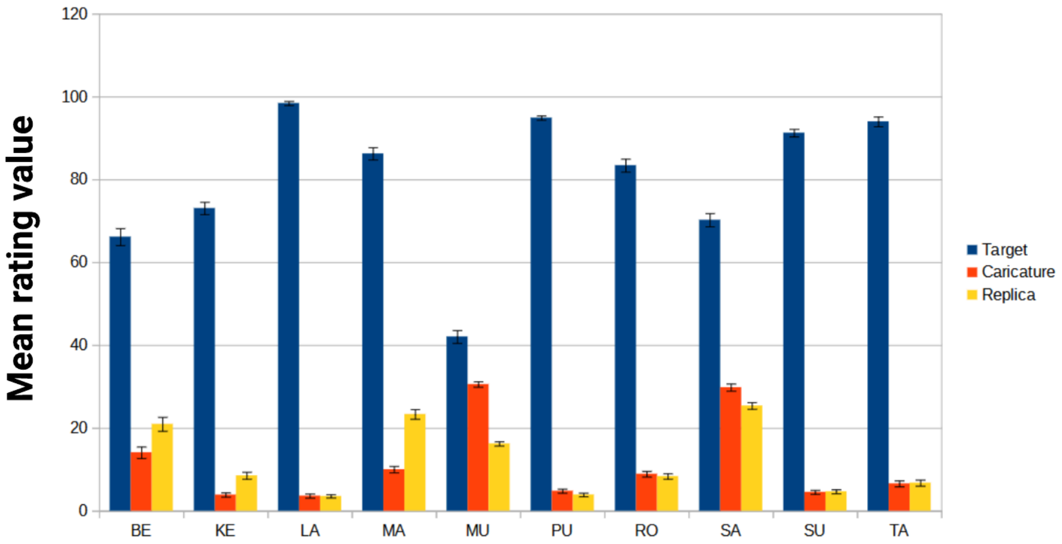


Figure 7 | Results by public figure and types of audio: Average ratings for each public figure's authentic and imitation audio samples.

Next we addressed the replica and caricature imitation modes in terms of the acoustic variables relevant to the study. For each, we estimated the Jensen-Shannon distance in a hypothesized source-filter (f_0, ΔF) space with respect to their target counterparts. Globally, the average acoustic distance between the caricature and the target was 0.469 bits, while the average acoustic distance between the replica and the target was 0.462 bits (Figures 8 and 9).

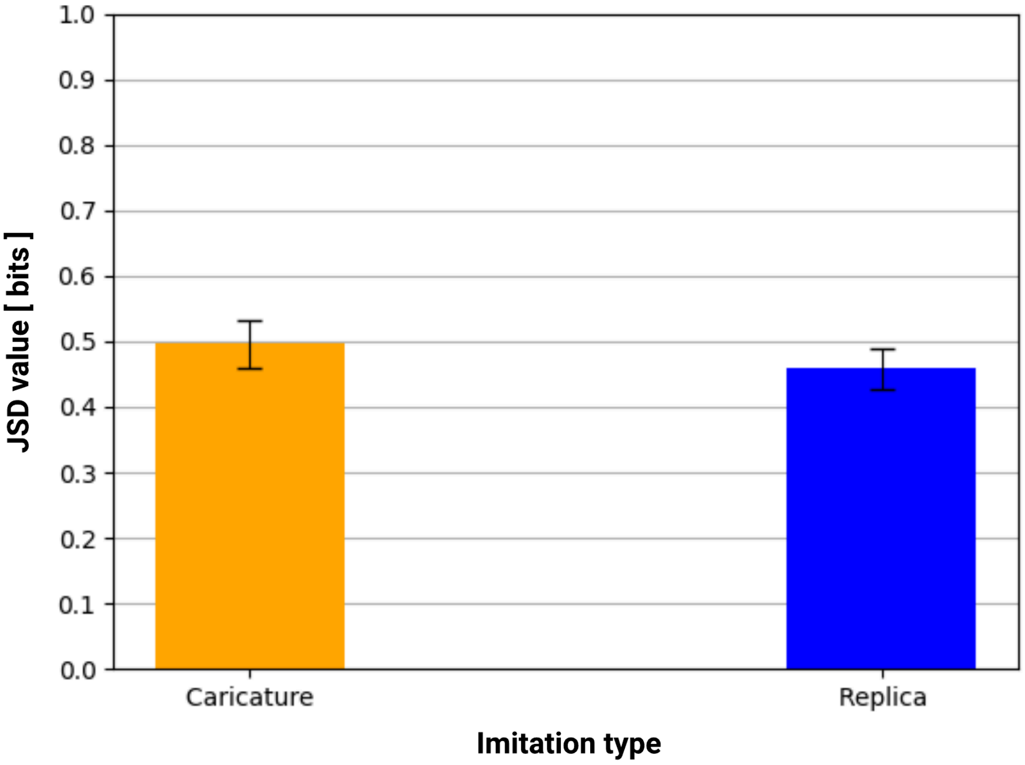


Figure 8 | Average JSD value for replica and caricature imitations with respect to

target voice stimuli. The bars illustrate the average divergence values calculated for each type of imitation in relation to the original voice. Replica imitations appear to show lower divergence compared to caricature imitations.

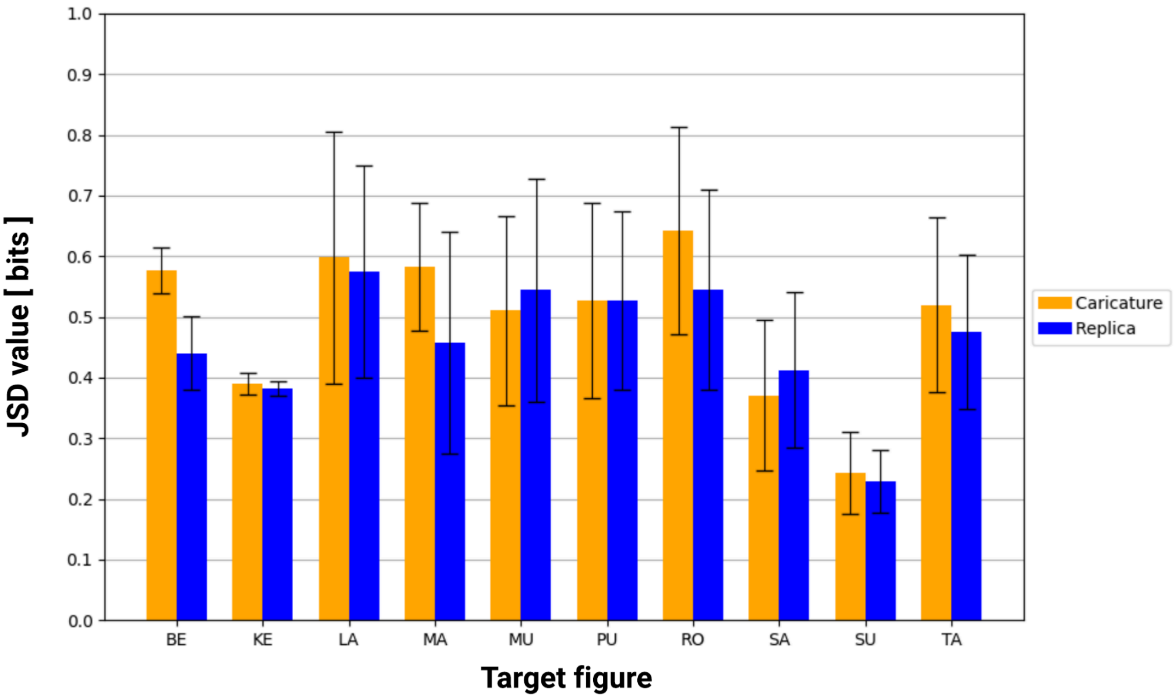


Figure 9 | Average JSD value for replica and caricature imitations per public figure. Acoustic closeness of imitations produced appears to vary according to the type of imitation and the public figure, as measured by the JSD.

Using acoustic distances for each audio segment, impersonator, and target, we then computed the acoustic difference parameter Δφ between replicas and targets and evaluated its ability to predict perceptual differences Δψ attained between both imitation modes. Δψ is defined as the difference in the average ratings between replica and caricature imitation modes across all participants for a given audio segment, impersonator, and target. A two-dimensional scatter plot was created (Figure 10), with Δφ as the independent variable and Δψ as the dependent variable.

In the scatter plot, a total of 60 observations (audio segment, impersonator, and target) were obtained, each shown with a standard error bar of the mean. Each observation represents a pair of replica and caricature audios, summarizing which of them was more similar to the target voice and which of the pair obtained a better average rating among all subjects who participated in that target block.

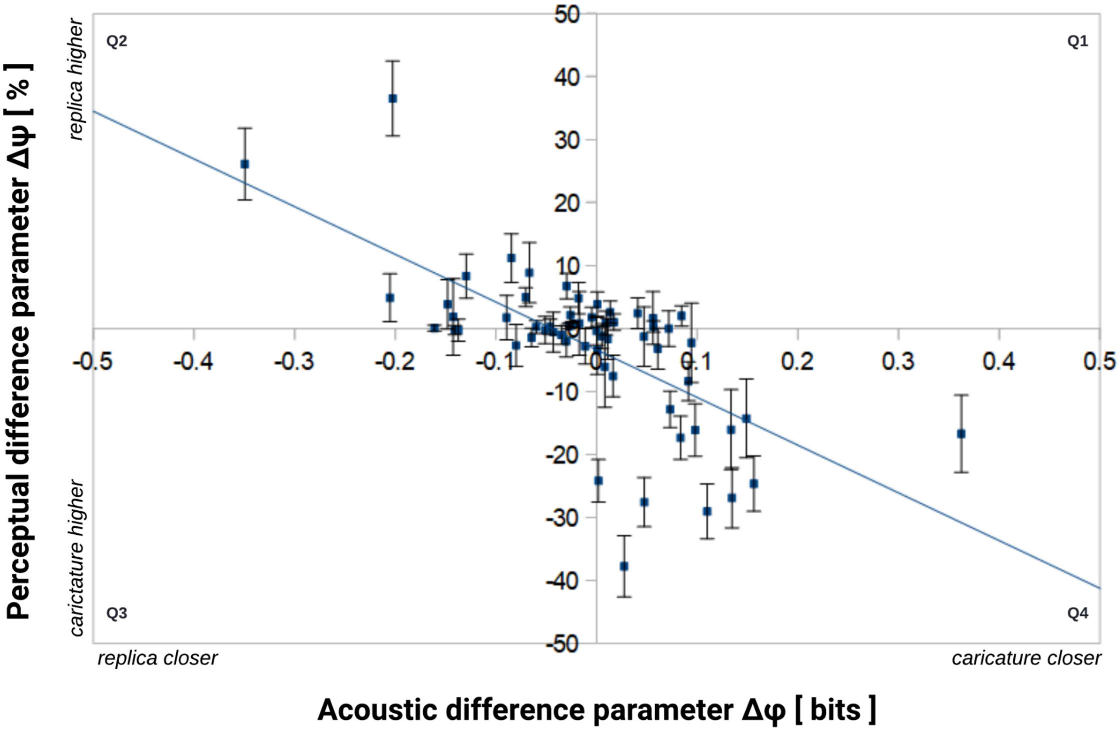


Figure 10 | Scatter plot relating acoustic and perceptual differences between imitation modes. The abscissa represents the relative distance between the replica and the caricature imitation modes with regards to the target voice (Δφ). The ordinate represents the average rating difference between replicas and caricatures across participants (Δψ). Error bars represent the standard error of the mean.

Observing the data distribution across quadrants, we can interpret that when an observation falls predominantly into the fourth quadrant (22 observations) and the second quadrant (20 observations), and less frequently into the third (9 observations) and first quadrants (8 observations). In other words, the pattern of data suggests an associative negative relationship, where as Δφ increases, Δψ tends to decrease. We assessed whether this relationship may be explained by a linear model. A simple linear regression revealed that the overall regression was statistically significant, the linear model was not strong ($R^2 = 0.416$, $F(1, 58) = 41.39$, $p = 2.60 \times 10^{-8}$), explaining approximately 41.6% of the variance in Δψ. Nevertheless, it was found that Δφ significantly predicted Δψ ($\beta = -74.2810$, $p < 0.001$), confirming a negative relationship between the two variables. The significant negative slope indicates that as Δφ increases, Δψ tends to decrease. Additionally, the intercept was statistically significant ($\beta_0 = -3.3643$, $p = 0.008$).

To further explore the relationship between Δφ and Δψ without assuming linearity, a Spearman's rank correlation analysis was performed. This analysis indicated a negative monotonic correlation ($\rho = -0.634$, $p = 5.23 \times 10^{-8}$), suggesting that as Δφ increases, Δψ tends to decrease in a consistent manner. The significant negative correlation confirms an overall negative association between the two variables.

We note that in these results, the trend line crosses the X-axis at approximately -0.0453 bits, representing the point of subjective equality (Δψ = 0), where participants did not favor either the replica or caricature options perceptually. This acoustic difference parameter level may be interpreted in terms of a slight closeness of replicas and distance of caricatures from the target, with participants still equivalently rating both modes. On the other hand, as the trend line crosses the Y-axis at approximately -3.3643%, which corresponds to $\Delta\varphi = 0$ (where the estimated acoustic distances of the replica and caricature are the same with respect to the target), it suggests that participants may tend to perceive the caricature imitation modes as slightly more convincing than replicas given the same level of separation from the target. Whether these biases are significant is a matter that will require further replication and investigation in future work, for example, through the use of psychometric function models, which describe how the probability of a response varies as a function of stimulus parameters.

Discussion

What specific acoustic parameters are most prominent in voice perception and identification remains unclear. Numerous studies have used the source-filter model as a starting point to identify the relevant acoustic properties for voice perception in relation to the organization of voices in perceptual space. This study aimed to construct a two-dimensional representation of a voice space with physical dimensions based on the source-filter model. Acoustic distance measures between voice stimuli represented in this space were defined in terms of information units, and the corresponding perceptual distances determined by listeners were estimated through an online study. The findings indicate a correspondence between the acoustic and perceptual distances, where imitations that are acoustically closer to a familiar target predict higher chances of identification as such. Therefore, the results suggest a familiar voice space where perceptual identification may be based on, at least, acoustic characteristics consistent with the source-filter model. As a result, emitters with an ability to adjust their voice in these parameters to approximate an intended target may achieve an increased likelihood of being perceived as that person.

In the present study, we examined a two-dimensional model of perceptual identification based on familiar male voices. In this regard, speaker familiarity is a factor that may positively influence voice perception (Lavan et al., 2021; Stevenage, 2018). Moreover, recognition increases when the speech is in the same language as the listener's (Wester, 2012). Prior research addressing the typical dimensions for listeners to map perceptually the acoustic characteristics of speaker voices, has been motivated by determining the basic acoustic characteristics of the voice that efficiently represent spatial relationships in perceptual space. It is noteworthy that so far, these mappings, whose dimensions may be typically constructed in frequency units or their derivates, for simplicity usually represent voices as static points, which discards many of the speech dynamic properties. Generally, f_0 has been a consistently used dimension in these studies as it plays a critical role as the lowest frequency of the vocal spectrum that is to be shaped through the vocal tract resonances. By contrast, the complementary dimensions to pitch have been more varied (Baumann y Belin, 2008; Chhabra et al., 2012; López et al., 2013; Stevenage et al., 2023), and it remains unclear which is the most optimal space with the fewest possible dimensions.

In our study, we extended upon the original work by López et al. (2013), where a perceptual space for familiar speaker identification was mapped to acoustic features from the vocal tract, particularly the dispersion between the upper formants. In that study, the perceptual judgments of voice similarity, i.e. discriminating between voice pairs, were found to be by contrast more related to acoustic parameters associated with the vocal cords, such as fundamental frequency and its variability. Using behavioral responses, López et al. obtained a dimension map based only on the acoustic properties they deemed most relevant for constructing that space.

Unlike the present study, which achieved the correlation of the existence of a space where the acoustic properties of the vocal cords and the vocal tract coexist with the listeners' subjective perception, López et al. used participants' perceptions to identify the acoustic properties necessary for the construction of the vocal space. Additionally, their study distinguished between replica and caricature imitation mode suggesting that caricatures are more effective in evoking the speaker's identity, while replicas achieve greater acoustic similarity. They further suggested that different acoustic properties may be exploited depending on the type of imitation: while caricatures may emphasize the characteristics of the vocal tract to highlight identity, replicas may focus on the characteristics of the vocal cords to achieve greater similarity. In the present work, we suggest a unified framework where acoustic properties from both the source and filter stages of production may be jointly represented in the perceptual characterization of acoustic distances, regardless of imitation mode.

In a related study, Baumann and Belin (2008) addressed the acoustic features that subjects may use to discriminate speakers, using different vowel sound stimuli. A two-dimensional space, based on f_0 and the separation between the upper formants (F_5 – F_4), was found to account for perceptual difference judgments of unfamiliar voices, highlighting the joint contribution of the larynx and vocal tract in voice identification. Another study (Chhabra et al., 2012) similarly found a two-dimensional vocal space, based on f_0 and formant dispersion (Hz), to be perceptually relevant in a voice discrimination task, and such organization was found to be also effective for listeners with schizophrenia. In another study (Stevenage et al., 2023), the acoustic differences between public figure targets and their impersonators were addressed using a multidimensional combination of acoustic features, including f_0, the standard deviation of f_0, the harmonic-to-noise ratio, jitter, shimmer, and F_4. In this case, however, the feature sets achieved by impersonators were not found to be systematically closer to the targets than control voices in acoustic space. However, despite the lack of significant acoustic differences, listeners were able to correctly discriminate voices across conditions, suggesting other perceptual and cognitive factors may play an important role in the task. In addition, the findings may illustrate the limitations of ‘fixed-point’ static in multidimensional characterizations of voiced speech. In contrast to the previous studies, the present work addresses the dynamic time-varying distributions defined by the trajectory of voiced speech, through the use of joint probability distributions. Our data suggest that by measuring the informational differences between distributions, spatial relationships in acoustic feature space may be expressed to define associated perceptual spaces.

Limitations

In this self-administered online study, participants performed the task in an unsupervised manner using their own interface and audio playback devices. For this reason, the level of the participant's attention is an uncontrolled variable as it would be in an on-site study. The environmental conditions, including distractors or noise, were also not controlled.

In addition, variability on the acoustic features may be expected across instances from the same speaker, but in the current design this source of uncertainty is not accounted for. This relates to the estimated variability within each speaker because, in practice, voices are dynamic and may change according to their context (Lavan et al., 2019; Stevenage et al., 2023).

Conclusions

The present work introduces a novel approach to vocal perceptual spaces through a spatial mapping that is based on source-filter spectral properties of a voice signal, using informational relationships to assess acoustic distance, or similarity. The approach represents dynamic and time-varying acoustic feature distributions of familiar voices and their imitations, to inform perceptual identification judgments in relation to such voice space. We demonstrate that such acoustic similarity mappings are relevant in subjective voice perception, suggesting that these mappings represent a valid means for representing voices. These findings may open new avenues for vocal perception research through the creation of more precisely defined dynamic trajectories in acoustic feature space.

References:

Baumann, O., & Belin, P. (2008). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological research*, *74*, 110-120. https://doi.org/10.1007/s00426-008-0185-z

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129-135. https://doi.org/10.1016/j.tics.2004.01.008

Chhabra, S., Badcock, J. C., Maybery, M. T., & Leung, D. (2012). Voice identity discrimination in schizophrenia. *Neuropsychologia*, *50*(12), 2730-2735. https://doi.org/10.1016/j.neuropsychologia.2012.08.006

Cleveland, T. F. (1977). Acoustic properties of voice timbre types and their influence on voice classification. *The Journal of the Acoustical Society of America*, *61*(6), 1622-1629. https://doi.org/10.1121/1.381438

Constantino, F. C., & Caputi, Á. A. (2024). *Cortical tracking of speakers' formant changes predicts selective listening* (p. 2024.05.23.595545). bioRxiv. https://doi.org/10.1101/2024.05.23.595545

Fant, G. (1960). *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations*. Mouton.

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology: CB*, *18*(11), R457-460. https://doi.org/10.1016/j.cub.2008.03.030

Grimaldi, M., & Cummins, F. (2008). Speaker Identification Using Instantaneous Frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(6), 1097-1111. IEEE Transactions on Audio, Speech, and Language Processing. https://doi.org/10.1109/TASL.2008.2001109

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10*(2), 139-148. https://doi.org/10.1016/S0095-4470(19)30953-2

Johnson, K. (2005). Speaker Normalization in Speech Perception. En *The Handbook of Speech Perception* (pp. 363-389). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470757024.ch15

Kriegstein, K. V., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*(2), 948-955. https://doi.org/10.1016/j.neuroimage.2004.02.020

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology: CB*, *21*(4), R143-145. https://doi.org/10.1016/j.cub.2010.12.033

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology (London, England: 1953)*, *110*(3), 576-593. https://doi.org/10.1111/bjop.12348

Lavan, N., Kreitewolf, J., Obleser, J., & McGettigan, C. (2021). Familiarity and task context shape the use of acoustic information in voice identity perception. *Cognition*, *215*, 104780. https://doi.org/10.1016/j.cognition.2021.104780

Lavan, N., & McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Communications Psychology*, *1*(1), 1-11. https://doi.org/10.1038/s44271-023-00001-4

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9-26. https://doi.org/10.1016/S0167-6393(99)00028-X

López, S., Riera, P., Assaneo, M. F., Eguía, M., Sigman, M., & Trevisan, M. A. (2013). Vocal caricatures reveal signatures of speaker identity. *Scientific*

Reports, *3*(1), 3407. https://doi.org/10.1038/srep03407

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175-184. https://doi.org/10.1121/1.1906875

Průša, Z., Søndergaard, P. L., Holighaus, N., Wiesmeyr, C., & Balazs, P. (2014). The Large Time-Frequency Analysis Toolbox 2.0. En M. Aramaki, O. Derrien, R. Kronland-Martinet, & S. Ystad (Eds.), *Sound, Music, and Motion* (Vol. 8905, pp. 419-442). Springer International Publishing. https://doi.org/10.1007/978-3-319-12976-1_25

Simonyan, K., Ackermann, H., Chang, E. F., & Greenlee, J. D. (2016). New Developments in Understanding the Complexity of Human Speech Production. *The Journal of Neuroscience*, *36*(45), 11440. https://doi.org/10.1523/JNEUROSCI.2424-16.2016

Søndergaard, P. L., Torrésani, B., & Balazs, P. (2012). THE LINEAR TIME FREQUENCY ANALYSIS TOOLBOX. *International Journal of Wavelets, Multiresolution and Information Processing*, *10*(04), 1250032. https://doi.org/10.1142/S0219691312500324

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, *116*(Pt B), 162-178. https://doi.org/10.1016/j.neuropsychologia.2017.07.005

Stevenage, S. V., Singh, L., & Dixey, P. (2023). The Curious Case of Impersonators and Singers: Telling Voices Apart and Telling Voices Together under Naturally Challenging Listening Conditions. *Brain Sciences*, *13*(2), 358. https://doi.org/10.3390/brainsci13020358

Stevens, K. N. (2000). *Acoustic Phonetics*. MIT Press.

Titze, I. R. (2000). *Principles of Voice Production*. National Center for Voice and Speech.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19-38. https://doi.org/10.1016/S0095-4470(19)30723-5

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for Research in Otolaryngology: JARO*, *19*(2), 181-191. https://doi.org/10.1007/s10162-018-0654-z

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781-790. https://doi.org/10.1016/j.specom.2012.01.006