FORUM ACUSTICUM EURONOISE 2025

PERCEPTUAL EVALUATION OF SYNTHETIC VOICE DETECTION WITH DYSPHONIC SPEAKERS

Juan José Romero[1*] **Aurora López-Jareño**[1]
Eugenia San Segundo[1] **Jonathan Delgado**[2]
[1] Phonetics Laboratory, Spanish National Research Council (CSIC), Madrid, Spain
[2] Department of Developmental & Educational Psychology, La Laguna University, Spain

ABSTRACT

In the present work, we have designed a perceptual experiment comprising 80 stimuli: 40 samples of natural voices and 40 samples of their corresponding deepfakes. As for natural samples: 20 are from dysphonic patients and 20 are from a control group (half English and half Spanish for both groups). In the former group, we have 5 patients classified as mild-moderate and 5 as severe according to the CAPE-V scale for each language. The experiment involves listeners indicating, for each recording, whether it is a synthetic or human voice. Although some perceptual experiments have tested human performance in detecting synthetic voices, studies involving dysphonic voices are far less common. Our hypothesis is that dysphonic voices are more likely to be perceived as human voices than as deepfakes. In the same way that human faces are characterized by imperfections (e.g. wrinkles) and this allows distinguishing real images from visual deepfakes, human voices are often characterized by dysprosodic and dysphonic phenomena. The aim of this paper is therefore to shed light on new possible predictors of listener performance in perceptual experiments involving audio deepfake detection.

[1]***Corresponding author***: juan.romero@cchs.csic.es

Keywords: *audio deepfake, forensic phonetics, dysphonia, perceptual experiment.*

1. INTRODUCTION

In recent years, advances in voice synthesis technologies have enabled the creation of artificial voices that are practically indistinguishable from human voices. Examples of progress in this field can be seen in the works [1-2] related to Text-To-Speech (TTS) models. We define deepfake voices as those generated by deep neural networks models. This technological development has opened new possibilities in various applications but has also raised concerns regarding their malicious use [3-4].

Several studies have been conducted on human detection of deepfake voices. Several studies concluded that human ability to detect deepfakes is unreliable [5-8]. In [6] they used English stimuli, in [5] they conducted experiments in both English and Mandarin, and the work in [7] was focused on Spanish.

In addition, machine learning and artificial neural networks-based methods have been developed for deepfake detection. Several reviews of these methods can be found in the works in [9-11].

None of the aforementioned studies have examined the perception of deepfake voices with pathological speech, specifically dysphonic voices. Dysphonia is defined by the presence of perceptual and acoustic features related to unstable or asymmetric phonation, such as roughness, breathiness, weak voice or instability in fundamental frequency and intensity [12]. The level of dysphonia

severity for the speakers used in this study had been previously graded according to the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) scale (See Methodology), which is the method proposed by the American Speech-Language-Hearing Association to establish a standardized clinical protocol for auditory-perceptual judgments of vocal quality [13]. This scale has also been adapted into Spanish [14].

To the best of our knowledge, this preliminary study is the first perceptual deepfakes experiment to include dysphonic voices in the sample and to explore the level of dysphonia as a potential biometric marker for distinguishing a natural (bona fide) voice from a deepfake voice. The interest in studying deepfake detection in pathological voices lies in the possibility that voices with "wrinkles" (i.e. imperfections) may be perceived as more natural. Studies such as [12] suggest that phonation-related parameters could help in speech recognition tasks (considering intra-speaker variability), as normophonia is not necessarily the general norm in the population. A stable fundamental frequency in sustained vowels, an absence of breathiness, stable intensity, or absence of roughness in voice cannot always be ensured, even in individuals without organic issues.

This study is also inspired by the results in [15] on deepfake face perception. They assume that the model StyleGAN2 (for generating faces) tends to create white faces close to normality in the "face-space", making them appear familiar, attractive and within the average range, but less memorable than some real human faces.

We hypothesize that mild-moderate dysphonic voices will tend to be judged as natural because they are perceived as more familiar compared to severe dysphonic voices, regardless of whether they are actually deepfake or bona fide voices.

Furthermore, since this study includes both English and Spanish stimuli, we expected to replicate the findings of [6] regarding the advantage of native speakers in detecting deepfakes in their own language compared to non-native speakers. However, while Müller focused exclusively on native and non-native English speakers, our study has compared native Spanish speakers' performance when detecting audios in both Spanish and English stimuli.

2. METHODOLOGY

2.1 Source and processing of the voices

2.1.1 Source of the voices

The voices used in this study were sourced from two distinct corpora. The English voices were obtained from the Voice Foundation database [16]. All utterances were recorded in a controlled, quiet environment using a condenser microphone placed 6 cm away from the speaker's mouth, with a sampling rate of 48 kHz [16]. The Spanish voices were sourced from [17].

2.1.2 Voice selection and segmentation

For both languages, the voices were categorized into three groups according to the CAPE-V (Consensus Auditory-Perceptual Evaluation of Voice) severity scale: Non-pathological, Mild-Moderate, and Severe. The classification was based on the severity scale and the criteria summarized in Table 1:

Table 1. Summary about the sample of the speakers

Group	Severity (CAPE-V Scale)	Criteria	N. Speakers (women/men)
Non pathological	0-50	Not diagnosed	10 (5/5)
Moderate	50-77	Diagnosed	5 (3/2)
Severe	77-100	Diagnosed	5 (2/3)

The severity ratings were performed by three speech professionals, and voices were selected based on an inter-rater standard deviation lower than 15 in the CAPE-V scale.

The audio files that would constitute the stimuli of the perceptual experiment were cut using Praat [18]."We were away a year ago" was the CAPE-V English sentence selected and "Teresa hace siete regalos pequeños" the CAPE-V Spanish sentence chosen. In total, 20 natural voices were selected for each language (Spanish and English), totaling 40 natural voice recordings.

2.1.3 Voice cloning

To create artificial versions of the natural voices, we used the TTS (Text-to-Speech) software from ElevenLabs [19]. We applied the "Eleven Multilingual v2" model with default settings for Stability, Similarity, Style Exaggeration, and Speaker Boost, following the "Best Practices" recommendations from ElevenLabs and the ElevenLabs Prompt Guide [20] for inputting both the audio and text. This process resulted in 40 cloned voices, with an artificial counterpart for each natural voice. In total, this gave us 80 voices (40 natural and 40 artificial) for use in the experiment.

2.2 Experimental design and procedure

The experiment was designed using PsychoPy [21] and hosted on Pavlovia. Participants were required to complete a demographic questionnaire, listen to the audio files, and classify each audio as either a natural voice or a deepfake (artificial voice). Additionally, participants were asked to rate their confidence in their classification and provide a justification for their response. The experiment followed a unary design, in which the task was repeated for all audio files.
The presentation order of the voices was randomized with a listening break after 40 stimuli. So participants first listened to the Spanish voices, followed by the English voices.
A total of 29 participants were recruited for the experiment, although three were excluded because they reported hearing problems or because they were not native Spanish speakers. The final sample consisted of 17 male and 9 female participants, with an average age of 32.58 years (SD = 12.63 years). Each one of them gave a response to 80 stimuli (half natural, half artificial), so in total there are 2080 responses. For each language, we have 520 responses to non pathological voices and 280 responses to both mild-moderate and severe voices.

2.3 Data analysis

We analyzed the data with the Python data analysis library Pandas [22], the calculus and algebra library Numpy [23] and the data visualization libraries Matplotlib [24] and Seaborn [25]. We decided to attribute the positive value to responding as “artificial” (deepfake) and the negative value to responding as “natural” (bona fide).
To analyze the data, we constructed the normalized confusion matrices based on the classification responses, distinguishing between language (Spanish and English) and severity (Non pathological, Mild-Moderate, Severe). These confusion matrices have in their elements the True Positive Rate (TPR), also called “Sensibility”; the False Negative Rate (FNR), also called “Specificity”; the False Positive Rate (FPR) and the True Negative Rate (TNR) and have the following form in Eqn. (1).

	Predicted Positive	Predicted Negative
Actual Positive	$\text{TPR} = \frac{TP}{TP+FN}$	$\text{FNR} = \frac{FN}{TP+FN}$
Actual Negative	$\text{FPR} = \frac{FP}{FP+TN}$	$\text{TNR} = \frac{TN}{FP+TN}$

(1)

Given the binary nature of the classification (Natural vs. Artificial), we first computed the proportion of natural and deepfake responses for each group. Instead of using absolute values, we used proportions since the number of non-pathological voices differs from that of mild-moderate and severe voices. Proportions provide more information about both errors and correct responses.
The main hypothesis of the study was that voices with higher degrees of dysphonia would be more likely to be classified as natural compared to those without any pathology. To test this hypothesis, we plotted stacked bar charts to show the proportion of voices classified as natural across the severity groups. Proportions were compared using a Z-test [26] to assess significant differences between the groups.
The last plot we computed was the Receiver Operating Characteristic (ROC) [27] curves across the severity groups and languages, which helped us to have an overall measure of the performance classifying the voices with some degree of severity. These curves require the responses of the participants in form of probability of being a deepfake using the transformation in Eqn (2):

$$p = \begin{cases} \frac{\text{confidence}}{5} & \text{if response = 'a'} \\ 1 - \frac{\text{confidence}}{5} & \text{if response = 'n'} \end{cases} \quad (2)$$

This is the same approach as used in [5]. The confidence is divided by 5 as the confidence was rated in a 5-points Likert scale. Where "a" value corresponds to "artificial" and "n" to natural. We also computed the Area Under the Curve (AUC) and the Equal Error Rate (EER) [28].

3. RESULTS AND DISCUSSION

The first step in our analysis involved computing the normalized confusion matrices for the stimuli, grouped by both speaker language and severity level. In these matrices, the columns represent the participants' responses, while the rows correspond to the ground truth. See Fig 1.

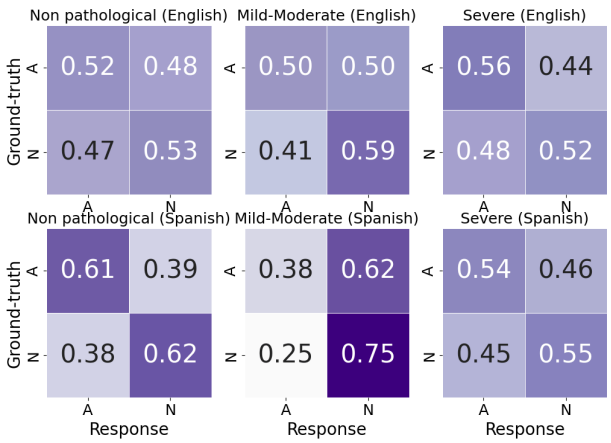

Figure 1. Confusion matrices grouped in languages (English up; Spanish down) and severity (Non pathological, Mild-Moderate and Severe, from left to right) Where labels "A", "N" correspond to "Artificial", "Natural" respectively.

Several remarkable observations can be drawn from these matrices:

First, there is a clear distinction between the Spanish and English stimuli. For the English stimuli, the matrix values range between 0.43 and 0.57, suggesting no strong tendency for participants to correctly identify natural or artificial voices. Moreover, the confusion matrices across the different severity groups show minimal variation.

In contrast, the Spanish stimuli reveal more pronounced differences between severity groups, indicating potential trends in participants' performance based on voice imperfections. In particular, when the voice has a mid grade of severity, the natural voices are easier to distinguish, as we can see in the TNR (or specificity) = 0.75 in mild-moderate voices against the TNR (or specificity) = 0.62 in non pathological voices. On the other hand, the TPR (or sensibility) for the mild-moderate group is lower than the TPR (or sensibility) in the non-pathological group, showing the increase in the number of type I as II errors (there are more false negatives). Although we cannot say the performance distinguishing mild-moderate voices is greater than distinguishing non pathological voices, we suspect that there is a tendency to judge the voices as natural when these have a moderate grade of severity. This is because, as listeners, we are used to hearing voices of this sort.. For instance, in our English corpus, moderate dysphonic speakers are people with benign vocal cord lesions, such as nodules and polyps, which are noncancerous growths that may form on one or both vocal cords. Most of these lesions are due to vocal abuse or misuse that many of us can have at some point in life. In contrast, severe dysphonia is found in patients with Reinke's Edema, ulcerative laryngitis or cordectomy, to name a few. The prevalence of such conditions is lower in the population [29]. Listeners are less used to hearing this type of voices, so they might have classified them as artificial, simply because they cannot map them to their typical patterns of what a "normal human voice" sounds like.

Secondly, there are more differences between languages for the same severity group. For the non pathological voices, the Spanish ones present a lower proportion of type I and II errors. This is explained by the fact that all the participants are Spanish native speakers, so it is easier to discriminate between bona fide and deepfakes in their mother tongue than those in English [6, 8].

Lastly, we can say that the severe Spanish voices are just as difficult to distinguish as the English ones. This fact points to the lack of familiarity of the participants to voices with a high grade of dysphonia regardless of the language spoken. This phenomenon is observed in Fig 3 again.

In order to study in more detail how severity influences this classification task, we plotted in a stacked bar chart the proportion of audios judged as natural in relation to the total number of audios in each severity group in both languages. The result is in Fig 2.

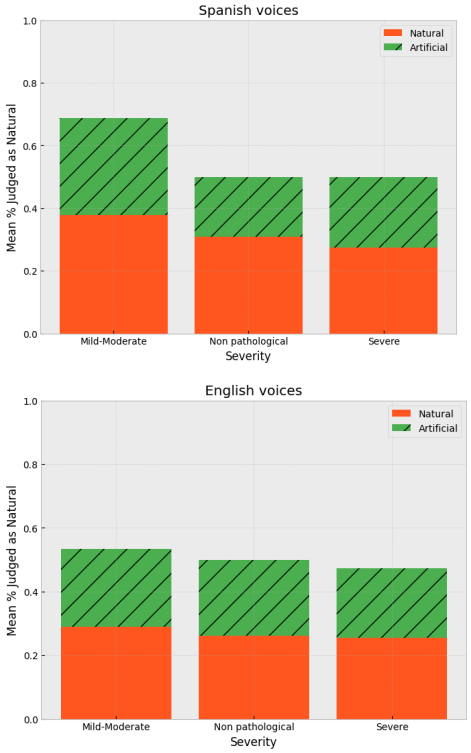

Figure 2. Proportion of stimuli judged as natural grouped by severity (Spanish voices at the top, English voices at the bottom). Orange: true natural voices; green: artificial voices.

Notably, the Spanish stimuli exhibit a clear trend, with a higher percentage of voices judged as natural in the mild-moderate group compared to the non-pathological group. This pattern was further supported by a Z-test for proportion comparison, which revealed statistically significant differences between the severity groups (see Table 2). This is not as clear for the proportion in severe vs non pathological voices, so we applied this Z-test to compare these two groups too. See Table 2.

Table 2. Proportions of voices judged as natural compared with the Z-test across different severity groups in both languages. (***) → p-value < 0.000

	Proportions in Spanish voices	Proportions in English voices
	Mild > Non-P. (***)	Sev. > Non-P.	Mild > Non-P.	Sev. < Non-P.
z	5.04	0.34	0.98	-0.98
p	$2.29*10^{-7}$	0.37	0.16	0.16

In the English stimuli we cannot see a significant difference between severity groups, so this result supports what we have been pointing out about the voices in this language.

For the English stimuli, the distribution of natural voice classifications appears more uniform across severity groups. The Z-test results confirmed the absence of significant differences between these groups, suggesting that participants' ability to distinguish between natural and artificial voices was less influenced by severity in the English stimuli compared to the Spanish ones.

Last but not least, we computed the ROC curves and their corresponding AUC and EER across the severity groups in the way described in Eqn (2).

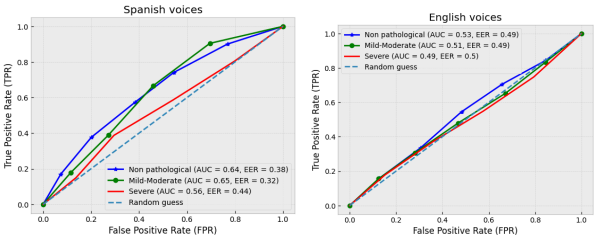

Figure 3. ROC curves of the participants divided by severity groups (Spanish voices at the top, English voices at the bottom).

The presented ROC curves, along with the corresponding AUC and EER values, reveal notable differences in participants' performance across languages and severity groups. For the English stimuli, the performance closely aligns with a random guess, as indicated by AUC values near 0.5 and EER values approaching 0.5 as well. This suggests that participants struggled to reliably distinguish between natural and artificial voices in this language.

A similar pattern can be observed for the severe Spanish voices, where the ROC curve and corresponding metrics also approximate random guessing. This indicates that

participants faced considerable difficulty in correctly identifying the authenticity of severely dysphonic Spanish voices.

The ROC curves indicate that the performance in classifying mild-moderate and non-pathological voices in Spanish is similar (see the curves in Fig 3 and the corresponding AUC and EER results. This suggests that although the TNR is higher in the mild-moderate group (there are more natural voices judged as natural), this advantage is offset by an increase in type II errors (i.e., the FNR also increases, there are more natural voices judged as artificial too). Consequently, participants do not perform better when classifying mild-moderate voices, rather they tend to classify these voices more frequently as natural ones.

The results show that the imperfections of the voices could be an important factor to ensure that a voice judged as natural is actually natural. The stimuli typically used in perceptual experiments aimed to distinguish between natural and artificial voices are usually voices without any pathology. The results of this preliminary work show the need to include pathological voices (in this case dysphonic) in these designs.

Overall, these findings suggest that participants' performance is influenced by their familiarity with the stimuli. Given that the participants were native Spanish speakers with no background in clinical voice assessment, they are likely more used to hearing non-pathological or mildly dysphonic voices in their own language. Consequently, their improved performance with mild-moderate Spanish voices compared to other conditions may reflect this greater exposure and familiarity.

It is also important to list some limitations of our research. First, since the experiment was performed online, and although participants were said to complete it in a quiet environment, using headphones, and paying close attention, we cannot be sure of the actual conditions in which they participated. Additionally, since participants always classified the Spanish voices first, the results for the English voices may have been influenced by increased fatigue, which could have affected their attention.

Last but not least, it is worth exploring the 'other-accent' effect, which has been investigated in numerous perceptual studies on talker recognition [30-32] since all the Spanish speakers were from the Canary Islands, while the English speakers came from several places in the United States such as Baltimore, New York or Los Angeles. Although participants did not know the origin of the speakers, the dialectal heterogeneity of the English voices could have confused them.

4. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In the presented work we have highlighted key differences in participants' performance classifying natural and artificial voices across language and severity degrees of dysphonia.

Firstly, while English stimuli showed a performance similar to tossing a coin, Spanish stimuli revealed notable tendencies relative to voice severity. Specifically, participants showed a tendency to classify mild-moderate Spanish voices as natural more frequently than non pathological and severe voices. This result aligns with the observed increase in the True Negative Rate for mild-moderate voices, indicating that participants were not necessarily better at identifying mild-moderate voices but rather more prone to labeling them as natural.

Secondly, the influence of language familiarity was also evident, with Spanish participants demonstrating better performance in identifying non pathological Spanish voices compared to English voices. This agrees with the results in [6] but with native Spanish speakers. This is likely due to their greater exposure to their mother tongue. Additionally, the difficulty in distinguishing severe voices in both languages suggests that participants struggled with highly dysphonic voices regardless of language.

These results underline the importance of voice imperfections as a factor in the perception of humanity. The tendency to classify mildly damaged voices as natural may reflect participants' familiarity with common voice "wrinkles", which TTS models may fail to reproduce convincingly.

Future works could aim to tackle several aspects:

It may be interesting whether a listener's musical training [33] or linguistic background can help distinguish a human voice from a deepfake.

It would also be valuable to examine the in-domain and out-domain performance of an algorithm trained with the

voices used in this experiment and compare its performance with that of the participants, similar to the approach described in [5].

It is also worth developing a model that allows us to assess how much of the choices variability can be explained by factors such as severity, language, response confidence or sex of the speaker, while also accounting for errors associated with random effects like the participant, the audio sample, or the device used during the experiment. And being aware of the listeners variability [34].

Additionally, it could be interesting exploring whether response confidence is an indicator of response accuracy or whether phenomena like the Dunning-Kruger effect emerge, as observed in [15].

Investigating the influence of reaction time in distinguishing between deepfakes and bona fide voices, following the approach in [7].

Studying the qualitative responses provided by participants to determine whether human perception-based discriminatory criteria can aid in distinguishing between cloned and bona fide voices. This would align with the methodology in [5], although their study focused on English and Mandarin voices with native speakers of these languages.

In addition to potential perceptual parameters, it is important to explore other acoustic parameters of both natural and synthetic stimuli that may also contribute to this classification task as in [35].

5. ACKNOWLEDGMENTS

This paper was done under the Project PID2021-124995OA-I00 funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE. The present work has been also funded by the European Union-Next Generation EU. However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

6. REFERENCES

[1] Y. Wang et. al, “Tacotron: Towards End-to-End Speech Synthesis” in *Proc. of Interspeech,* (Stockholm, Sweden), pp. 40006–4010, 2017.

[2] S. E. Eskimez, et. al: “E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS,” *Preprint from arxiv*, 2024.

[3] T. Brewster, “Fraudsters Cloned Company Director’s Voice In $35 Million Bank Heist, Police Find”. Forbes: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=7dfbccf67559, 2021.

[4] T. Chivers, “What do we do about deepfake video?” The Guardian. https://www.theguardian.com/technology/2019/jun/23/whatdo-we-do-about-deepfake-video-ai-facebook, 2019.

[5] K. T. Mai et. al: “Warning: Humans cannot reliably detect speech deepfakes,” *PLoS One*, vol. A, no. B, pp. 111–222, 2023.

[6] N. Müller, K. Pizzi, and J. Williams, “Human Perception of Audio Deepfake,” in *Proc. of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM ‘22),* (Lisboa, Portugal), pp. 85-91, 2022.

[7] E. San Segundo and M. Gibson, “Reconocimiento perceptivo de hablantes: un experimento con voces clonadas artificialmente y con voces de gemelos idénticos,” in *Proc. of the 13th Congreso Ibérico de Acústica Tecniacústica (DDAM ‘22),* (Faro, Portugal), 2024.

[8] Di Cooke, et. al: “As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli,” *Preprint from arxiv*, 2025.

[9] X. Liu et. al: “The title of the journal paper”ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[10] J. Yi: “Audio Deepfake Detection: A Survey,” *Preprint from arxiv*, 2023.

[11] R. Mubarak et. al: “A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats,” *IEEE Access*, vol. 11, pp. 144497–144529, 2023.

[12] P. Gómez-Vilda et. al: "Using Dysphonic Voice to Characterize Speaker's Biometry," *Language and Law*, vol. 1, no. 2, pp. 42–66, 2014.

[13] G. B. Kempster et. al: "Consensus auditory perceptual evaluation of voice: Development of a standardized clinical protocol," *Am J Speech Lang Pathol.*, vol. 18, no. 2, pp. 124–132, 2009.

[14] F. Núñez-Batalla: "Validation of the Spanish Adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)," *Acta Otorrinolaringología*, vol. 66, no. 5, pp. 249–257, 2015.

[15] E.J. Miller et. al: "AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones?," *Psychological Science*, vol. 34, no. 12, pp. 1390–1403, 2023.

[16] Walden P. R.: "Perceptual Voice Qualities Database (PVQD)," *Mendeley Data, , v2*, https://data.mendeley.com/datasets/9dz247gnyb/1, 2020.

[17] S. Hochmuth et. al: "A Spanish matrix sentence test for assessing speech reception thresholds in noise, " *Int J Audiol*, vol. 51. pp. 536-544, 2012.

[18] P. Boersma: "Praat: doing phonetics by computer," *Glot International*, vol. 5. pp. 341-345, 2001.

[19] ElevenLabs: Free Text to Speech & AI Voice Generator. https://elevenlabs.io/.

[20] ElevenLabsPrompting-ElevenLabs Documentation. https://elevenlabs.io/docs/best-practices/prompting/controls.

[21] J. Peirce et. al: "PsychoPy2: Experiments in behavior made easy," *Behav Res Methods*, vol. 51, no. 1, pp. 195–203, 2019.

[22] W. McKinney, "Data structures for statistical computing in python," in *Proc. of the 9st Python in Science Conference,* (Austin, Texas), pp. 56–61, 2010.

[23] C. R. Harris et. al: "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.

[24] J. D. Hunter: "A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3 pp. 90–95, 2007.

[25] M. L. Waskom: "seaborn: statistical data visualization," *Journal of Open Source*, vol. 6, no. 60 pp. 3021, 2021.

[26] J. Gorgas, N. Cardiel and J. Zamorano: *Estadística básica para estudiantes de ciencias*. Madrid: Departamento de Astrofísica y Ciencias de la Atmósfera, 2009.

[27] N. A. Macmillan and C. D. Creelman: *Detection theory: A User's Guide*. Mahwah: Lawrence Erlbaum Associates, 2005.

[28] D. W. Hosmer and S. Lemeshow: *Applied Logistic Regression*. Toronto: John Wiley & Sons, 2000.

[29] S. M. Cohen et. al: "Prevalence and causes of dysphonia in a large treatment.seeking population," *Laryngoscope*, vol. 122, no. 2, pp. 343–348, 2012.

[30] E. San Segundo and V. Marrero: *El efecto "otro acento" en una rueda de reconocimiento de voces: un estudio perceptivo en español*. In W. Elvira-García and P. Roseano: *Avances metodológicos en fonética y prosodia* (pp. 167-178). Madrid: Editorial UNED, 2024.

[31] A. G. Goldstein et. al: "Recognition memory for accented and unaccented voices," *Bull. Psychon. Soc.*, vol. 17, pp. 217–220, 1981.

[32] D. Van Lancker, J. Kreiman, and K. Emmorey: "Familiar voice recognition: patterns and parameters Part I: Recognition of backward voices," *Journal of Phonetics*, vol. 13, no. 1, pp. 19–38, 1985.

[33] E. San Segundo: "El entrenamiento musical y otros factores que pueden influir en el reconocimiento perceptivo de hablantes," *Fonética experimental, educación superior e investigación*, vol. 1, pp. 571–588, 2014.

[34] T. H. Kinnunen et. al, "Speaker Detection by the Individual Listener and the Crowd: Parametric Models Applicable to Bonafide and Deepfake Speech," in *Proc. of Interspeech,* (Kos, Greece), pp. 3654–3658, 2024.

[35] E. San Segundo and J. Delgado, "Deepfakes frente a voces de gemelos idénticos: un análisis acústico preliminar," in *Proc. of the 13th Congreso Ibérico de Acústica Tecniacústica (DDAM '22),* (Faro, Portugal), 2024.