Multiparametric Analysis of Speaking Fundamental Frequency in Genetically Related Speakers Using Different Speech Materials: Some Forensic Implications

Julio Cesar Cavalcanti, Anders Eriksson, and Plinio A. Barbosa

Abstract: Objective: To assess the speaker-discriminatory potential of a set of fundamental frequency estimates in intraidentical twin pair comparisons and cross-pair comparisons (i.e., among all speakers). Participants: A total of 20 Brazilian Portuguese speakers of the same dialect, namely 10 male identical twin pairs aged between 19 and 35, were recruited. Method: The participants were recorded directly through professional microphones while taking part in a spontaneous dialogue over mobile phones. Acoustic measurements were performed in connected speech samples, and in lengthened vowels, at least 160 ms long produced during spontaneous speech. Results: f0 baseline, central tendency, and extreme values were found mostly discriminatory in intra-twin pair and cross-pair comparisons. These estimates also displayed the largest effect sizes. Overall, any three identical twins were found statistically different regarding their f0 patterns in connected speech, but not for lengthened vowel-based f0 metrics. Estimates of f0 variation and modulation were found the least discriminatory across speakers, which may signal the control of speaking style and dialect on dynamic patterns of f0. Concerning performance, the base value of f0 (f0 baseline) was found the most reliable metric, displaying the lowest equal error rate (EER). Conclusions: The outcomes suggest that, although identical twins were very closely related regarding their f0 patterns, some pairs could still be differentiated acoustically, only in connected speech. Such findings reinforce the relevance of analyzing long-term f0 metrics for speaker comparison purposes, with particular consideration to f0 baseline. Furthermore, f0 differences across subjects were suggested as more expressive in connected speech than in lengthened vowels.

Keywords: Speaking fundamental frequency—Acoustic phonetics—Forensic phonetics—Identical twins.

INTRODUCTION

This study aimed at analyzing the speaker-discriminatory potential of a set of fundamental frequency descriptors in comparisons performed between genetically related (i.e., identical twin pairs) and cross-pair comparisons (i.e., among all speakers). The analysis performed on the basis of spontaneous speech set out to verify which metrics are considered remarkably “individual” even across very similar speakers, as is the case with identical twins who grow up together. Such individuals represent the highest possible level of intersubject physical similarity and are, in most cases, exposed to very similar environmental and linguistic influences, which characterizes them as an optimal group for exploring idiosyncrasy or individuality in speech production.

Two different speech material types were analyzed: intervals of spontaneous connected speech and lengthened vowels extracted from spontaneous dialogues. The purpose of the study is two-fold: to contribute to the existing body of knowledge of speech-directed twin studies and to examine its usefulness for forensic phonetic studies.

Differences in voice fundamental frequency (hereafter, f0) average values and variability have been reported as a function of several components, ranging from physiological, psychological, linguistic, stylistic to socio-cultural factors. The different levels of contribution of these components on the voice/speech fundamental frequency reveal the multimodal nature of such physical parameters, bearing repercussions on how individuals are perceived by listeners.

Regarding the acoustic dimension, divergences in f0 range as a function of age and sex is perhaps one of the most commonly acknowledged. As pointed out by Zhang, children tend to have higher mean fundamental frequencies when compared to male and female adults, whereas females tend to have higher mean fundamental frequencies than males. The reason for such difference resides in the vocal fold geometry, including length, depth, and thickness of the vocal folds, with the tendency for larger vocal folds to display a lower frequency of vibration and vice-versa. Consequently, it can be said that an individual’s fundamental frequency has an important genetic/organic motivation.

From a physiological viewpoint, the components underlying the determination of the voice f0 (i.e., the acoustic correlate of the vibration frequency of the vocal folds) have been decomposed and explored in detail by the classical myoelastic-aerodynamic theory of voice production proposed by Van den Berg. As described by the author, this frequency depends on a number of five interdependent factors related to the vibrating part of the vocal folds, namely the effective mass; the effective tension; the effective area of the glottis during the cycle; the subglottal pressure; and the
damping of the vocal folds. As pointed out by the author, when these factors are known, the frequency can be estimated.

In terms of forensic-phonetic applications, the analysis of f0 may be regarded as one of the most frequently assessed parameters in speaker comparisons worldwide, applied by the majority of forensic experts. One of the main justifications for such a high applicability regards the easiness with which the parameter can be assessed, being readily available even in small stretches of speech. Another crucial aspect regards the relative resistance of f0 measures to external variables, such as the microphone, the recording device, audio compression, and external noise levels. As pointed out by Jessen, in the forensic phonetic domain, f0 is mostly assessed by its “global” aspects, in which the average f0 and its standard deviation are among the most commonly explored measures. As the author mentions, while the average f0 in speech is regarded as displaying an important anatomical motivation, its variability is less dependent on structural factors and possibly better explained by other elements, such as individuals’ manner of speaking.

In the past few years, a particular f0 metric, the “base value of f0” or “f0 baseline” has been given attention in a number of forensic phonetic studies, due to its higher resistance to many different sources of variability. Initially proposed by Traummüller and Eriksson and revisited in Lindh and Eriksson, the base value of f0 is grounded on a well-known observation in various types of motor activity, namely, the point of departure, a resting position (i.e., baseline). In that instance, as described by Lindh and Eriksson, the f0 base value can be regarded as a neutral mode and frequency of vibration to which the vocal folds return after prosodic or other types of frequency excursions. For that reason, such a measure is regarded as relatively more robust since it better represents the neutral articulation of a given individual. The speaker-discriminatory potential of such measure has also been corroborated by previous research.

In Traummüller and Eriksson, where the concept was first derived, the authors referred to it as the base value which is still the preferred concept name. The reader should be aware, however, that the same concept is also referred to as the baseline. For the sake of consistency and because the concept is referred to as the baseline in Lindh and Eriksson, which we will return to later, we will use the word baseline to signify this concept throughout the present paper.

More recently, in the experiment conducted by Arantes and Eriksson with the purpose of assessing the stabilization time of f0 long-term mean, median, and baseline employing a change point analysis performed in recordings of the same text spoken in 26 languages, average stabilization points of an order of 5 s for the baseline and 10 s for mean and median estimates were observed. Furthermore, the variance after the stabilization point was considerably reduced, with a reduction in variability of approximately 40 times for mean and median and more than 100 times for the baseline. As remarked by the authors, these results show that stabilization points in long-term measures of f0 tend to occur earlier than what has been previously suggested.

It is noteworthy that most studies on f0 have been carried out with semi-spontaneous or read speech which justifies the need for a deeper understanding concerning the robustness levels of the metrics under unscripted speech conditions. Furthermore, the pertinence of assessing f0 extreme values and modulation metrics should also be examined, along with the implications of “speech material” on the discriminatory potential of f0 long-term estimates (e.g., directly comparing the discriminatory patterns of f0 long-term metrics extracted from lengthened vowels vs connected speech).

Twin studies on voice fundamental frequency
Regarding twin studies, within the speech and voice analysis domains, the voice fundamental frequency is probably one of the most commonly explored dimensions. Several studies have systematically reported a high correlation between monozygotic (i.e., identical) twins concerning the physical parameter of voice fundamental frequency. The majority of the studies have focused mainly on analyzing average and standard deviation values of the parameter, disregarding the possible speaker-contrasting potential of other fundamental frequency-related estimates, such as measures of f0 modulation. Some of these studies will be commented on in the following.

With the purpose of assessing the application of f0 as a potential phenotype, conducted a large-scale study with 122 twin pairs in American English, 50 pairs of female twins and 12 pairs of male twins aged between 15 and 75, from which only nine pairs were dizygotic (i.e., nonidentical) twins. As is the case for most twin studies, there was also a higher number of female than male speakers, i.e., 50 pairs of females with a mean age of 41 years and 12 pairs of male twins with a mean age of 40. The analyzed speech samples consisted of readings, and the f0 (in Hz) was assessed following an automatic approach. Overall, when comparing monozygotic (MZ) and dizygotic (DZ) twins solely based on their f0 mean and standard deviation values, no significant intergroup differences were found. Furthermore, when calculating a matrix of correlations for f0 and age, weight, and stature, statistically significant associations of f0 with age and weight were found. When an adjustment for age and weight was applied, coefficients of correlations were separated according to zygosity, revealing a larger discrepancy for f0 measures in DZ twins. According to the authors, such a finding suggests a genetic component underlying the variation of the parameter. The results also appear to suggest the influence of organic covariates on f0, such as age and weight.

Furthermore, Debruyne et al. conducted an analysis on f0 and its intra-speaker variability during a reading task. A group of 30 female MZ twins and 30 DZ twins Dutch speakers, aged 15–29 years, and a control group consisting of 30 nongenetically related individuals of equal age were
assessed. In the referred study, f0 was found considerably more similar in MZ twins when compared to DZ twin pairs, while no significant correlation was observed for unrelated peers, which according to the authors, is compatible with a genetic basis on the determination of the parameter. Notwithstanding, when intra-speaker variability of f0 was assessed, highly similar results were observed between MZ and DZ twin pairs, suggesting that an individual’s voice is determined by much more than genetic constitution alone.

The preliminary study carried out by Loakes with Australian English speakers also focused on assessing long-term fundamental frequency in a group composed of eight pairs of twins aged between 18 and 20. Three identical twin pairs and one nonidentical twin pair were analyzed in reading and spontaneous speech across noncontemporaneous samples. The extraction of f0 mean, median, mode, and standard deviation values was performed in the midpoint of all labeled vowel tokens produced by the speakers. For that reason, the number of tokens sampled in spontaneous speech varied considerably across individuals. The outcomes revealed that speakers tend to fall within a specific f0 range and that twins tend to have a more similar mean long-term f0 than unrelated pairs. Notwithstanding, when individual f0 values were compared in a sequence/list, it was noted that twin pairs do not necessarily have the closest mean f0 values, with two pairs being closer to other speakers than to their twin brothers regarding their mean values. Such a difference turned out statistically significant. Evidence was also found suggesting that long-term f0 is relatively stable within-speakers in most cases, even when noncontemporaneous data sets were compared.

A comprehensive study on the voice quality characteristics of a group of 45 monozygotic twin pairs (MT), consisting of 19 male and 26 female pairs, was carried out by Van Lierde et al. considering the effects of sex and age (i.e., male vs. female, and under vs. above 17 years old). One remarkable advantage of the study was the extensive age range. Speakers ranged from 8- to 61-year-old. The analysis consisted of both auditory perceptual evaluation using the GRBAS scale, cf., widely applied within the clinical voice setting, and acoustic measurements of a set of voice quality parameters, including f0 measurements in the middle points of sustained /a/ vowels. Overall, there was no significant influence of sex and age on the levels of vocal similarities in MT, and high correlation scores were observed concerning their f0 estimates, which was also corroborated in the auditory perceptual evaluations. However, it is worth noting that, as it appears, the subjects were compared as a group (male vs. female, younger vs. older than 17) and not individually; hence, intra-pairs specificities were not the focus of the study.

A clinical case study on identical twins’ voice characteristics was performed by Cielo et al. with a small sample consisting of one male and one female MZ pairs of Brazilian Portuguese aged 20 and 28 years, respectively. The voice quality analysis was carried out through auditory-perceptual and acoustic analysis, including automatic measurements of f0 mean, median, standard deviation, maximum and minimum. A descriptive intra-twin pair comparison was made based on the twins’ f0 values and their respective average differences. Moreover, a statistical analysis was performed to verify the extent to which the values deviated from standard normality scores, possibly helping the understanding of genetic and environmental influences on voice patterns. Among the voice quality features assessed, the maximum phonation time was found to deviate from normality for the female twin pair and one male twin, along with other dysphonic marks. According to the authors, environmental factors may be the basis for the male twin pair difference, such as the practice of physical activity and the professional use of the voice by the twin with the highest maximum phonation time.

Another case study was conducted by Whiteside and Rixon on the f0 and temporal patterns of male Southern Irish monozygotic twins (T1 and T2) and an age- and sex-matched sibling (S) recorded 2 years later. The twin pair was aged 21 years and their nontwin sibling 20 years by the time of the recordings. Mean and standard deviation f0 values were automatically extracted from sentences in reading speech, and comparisons among the subjects were made. The results indicated significant differences regarding f0 mean values for all three genetically related speakers. Conversely, no significant between-sibling differences were found for the f0 standard deviation regarding the comparison of the identical twin pair and between one of the twins (T2) and the nontwin sibling (S). Furthermore, when assessing the magnitude of the dissimilarities between sibling pairs through Euclidean distance measures, the smallest distances were observed between the identical twins (T1 and T2), mainly for f0 mean.

It is noteworthy that the commonly reported intridentical twin pair similarities are not solely restricted to the f0 domain, as other voice quality parameters are also regarded as relatively similar for such individuals. In this instance, the experiment conducted by San Segundo and Gómez-Vilda adopted a comprehensive approach on the analysis of twins’ and nontwins’ voice quality. Estimates of glottal source biomechanical parameters derived from vowel fillers, including f0, were used to assess phonation characteristics’ similarities and differences in twins’ voices. The participants consisted of 40 male native speakers of European Spanish, seven MZ pairs, five DZ pairs, four pairs of nontwin siblings, and four pairs of unrelated speakers (i.e., a control group) recorded during spontaneous conversation carried out in noncontemporaneous sessions. The main outcomes suggested a great influence of both genetic and environmental factors on the parameters assessed, as indicated by the relatively similar scores obtained for MZ and DZ twins, whereas nonrelative subjects showed scores well around the background baseline. Furthermore, one MZ twin pair out of seven displayed low matching scores in an intra-twin comparison. According to the authors, such findings may suggest that phonation characteristics may be due to learned styles as much as to imprinted biological patterns. Later on,
in San Segundo and Gómez-Vilda,34 the contribution of both “nature” and “nurture” in the determination of one’s voice was corroborated by the researchers when analyzing a considerably larger number of twin pairs while following the same approach.

RESEARCH QUESTIONS AND HYPOTHESES

1. What set of $f_0$ parameters are regarded as the most speaker-discriminatory in comparisons involving genetically related individuals (i.e., identical twins) and in cross-pair comparisons?
2. Is it possible to differentiate identical twins by assessing their $f_0$ estimates in spontaneous speech even when engaging in a conversation?
3. Are isolated sentence-based and vowel-based $f_0$ estimates equally representative of $f_0$ individual patterns?

Based on previous findings from the specialized literature, some hypotheses have been suggested: Firstly, given the impact of speaking style on $f_0$ modulation, it is assumed that $f_0$ central tendency estimates (e.g., median, mean) and $f_0$ baseline may be more speaker-discriminatory compared to parameters related to $f_0$ variation. Secondly, if divergences can still be observed regarding identical twin pairs, these might be better explained on account of external factors or as the influence of the variable “choice” (i.e., the selection or adoption of phonetic-acoustic patterns). Finally, if $f_0$ metrics in connected speech and lengthened vowels are equally speaker-discriminatory, similar proportions of inter-speaker differences and equal discriminatory performances should be observed.

Note that, although lengthened vowels produced in spontaneous speech do not have the same status as sustained vowels commonly elicited within the voice clinical setting, they share some important features (e.g., their relatively stationary patterns, being long enough so glottal source estimates can be computed).

MATERIALS AND METHODS

Participants

The participants were 20 subjects, ten identical male twin pairs, Brazilian Portuguese (BP) speakers from the same dialectal area. The participants’ age ranged between 19 and 35 years, with a mean of 26.4 years. All identical twin pairs were assigned a letter and a number, according to the following pattern: A1, A2, B1, B2, C1, C2, and so on. The same speaker letters were used to indicate identical twin individuals.

The decision to adopt the term “identical twins” over “monozygotic twins” was based on the following considerations: the latter terminology implies assessing the twins’ genetic material. As no laboratory genetic assessment was carried out, the first term will be preferred. However, it is worth noting that the expression “identical” does not imply that speakers are identical to each other, standing solely to the relative high physical similarities displayed by the twin pairs.

Speakers were recruited through a method known as chain sampling or “snowball”, in which subjects are contacted among their acquaintances or by recommendation of other participants of the study. Each twin within the pair lived and resided in the same city/town. The pairs were recruited in five different cities in the state of Alagoas, which is the second smallest state in Brazil.

The inclusion criteria were: i. identical twins; ii. male speakers; iii. same dialect; iv. aged between 18 and 45 years; v. with at least elementary school completed. The exclusion criteria were: i. reported hearing loss or speech disorder, ii. identical twins raised apart; iii. identical twins that lived apart from each other for more than 5 years.

All twin pairs in the present study were raised together, studied together, were frequently in contact with each other, and displayed a high affinity level. For that reason, the familiarity effect between twin pairs can be regarded as relatively controlled. Furthermore, the researchers made sure the participants were not going through a temporary cold or allergic reaction prior to the recordings, which would have caused the recording session to be postponed.

Some qualitative information about the participants

The participants were also interrogated regarding some health-related topics, such as the regular consumption of alcohol and the habit of smoking. One subject (E2) reported smoking, mainly on weekends, having started about five years ago. Although such information could represent a reasonable justification for excluding this participant and his twin brother (E1) from the experiment, a decision has been made to keep such subjects in the study, in view of the fact that if any differences would be found, they could represent clear support of external influences on the voice fundamental frequency patterns, beyond the genetic influence itself.

When questioned about the habit of alcohol consumption, nine subjects (45%) reported not having such a habit, eight subjects (40%) referred to consuming alcohol only occasionally, and three subjects (15%) reported to have the habit of consuming alcohol, namely A1, E2, and I1. No further questions were made regarding the frequency of such a habit.

Another piece of information concerning the participants worth mentioning regards the use of voice artistically, as is the case for F1 and F2. Although not professionally, this twin pair has experience with theater, performing as actors in small plays in their local church. This information was brought up while conducting the recordings, in which the planning of a theater performance was discussed by the twins, with references to vocal adjustments needed for constructing a particular character being mentioned.

Recordings

The recordings were carried out in silent rooms located in the cities where the twins resided. The speech material used in the present research consists of spontaneous telephone
conversations between twins, with dialog topics being decided by the pairs, aiming at elucidating a more ecologically valid material. During the recording sessions, twin pairs were placed in different rooms, not directly seeing, hearing, or interacting with each other. The speakers were encouraged to start a conversation using a mobile phone while being simultaneously recorded by high-quality microphones. Such a recording approach aimed at eliciting a telephony speaking style and represents an attempt to approximate the experimental conditions to more realistic forensic circumstances, as in San Segundo and Gómez-Vilda. Finally, the unedited and unfiltered audio signals were then processed and registered in two separate channels, with all its acoustic properties preserved for the analysis.

All recordings were carried out with a sample rate of 44.1 kHz and 16-bit, using an external audio card (Focusrite Scarlett 2i2) and two headset condenser microphones (DPA 4066-B). The unedited recordings had an average duration of about 10 minutes. In all cases, the conversation topics were decided beforehand by the twins during the recording sessions.

Besides controlling the variable “similarity” between individuals, the present study presents the advantage of controlling one very often neglected factor in forensic experiments, namely the “familiarity” between speakers.

### Data segmentation, transcription, and extraction

All speech material was segmented and transcribed in the Praat software, following acoustic (i.e., the analysis of spectrogram traces) and auditory criteria. The transcription took into account two types of speech material:

#### Connected speech

Intervals of continuous speech with an average duration of 3 s were segmented throughout different portions of dialogues while preserving intonational phrases’ structure—note that this value corresponds to the right limit of 95% confidence intervals for stress group duration in Brazilian Portuguese, cf. In most cases, silent or filled pauses were applied as a more objective criterion for setting the limits of speech chunks, resulting in inter-pause intervals, containing an average of 9.9 phonetic syllables, with a minimum of 3 and a maximum of 32 units (mode = 14 syllables). Such a high variability regarding the number of syllables within the intervals is due to the fact that, in spontaneous speech, individuals tend to vary substantially both in type and extent of sentences they produce.

A total of 854 speech chunks were analyzed, an average of 42 chunks per speaker with a standard deviation of 5 chunks, from which f0 estimates in connected speech were computed. As for those intervals for which some estimates could not be computed, e.g., f0 peak rate descriptors, these were disregarded during the statistical testing.

#### Lengthened vowels

All vowel segments produced by the speakers in different selections of the dialogues were also segmented and transcribed manually. Furthermore, the vowel segments most frequently lengthened in the corpus were identified. These were: [a:], [e:], and [i:], in a decreasing occurrence frequency order, being oftentimes perceived as filled pauses. Because these vowels were found more often prolonged, they were selected for the extraction of f0 metrics, given their potential forensic applicability (e.g., the extraction of glottal source parameters). It is noteworthy that, besides being the most commonly lengthened vowel in the corpus, the central prolonged vowel [a:] is also one of the most commonly used for the assessment of voice quality aspects within the clinical setting, e.g., A minimum duration threshold of 160 ms was established for the selection of vowel segments, as in San Segundo and Gómez-Vilda.

A total of 399 lengthened vowels were analyzed, a mean of 20 vowels per speaker and a standard deviation of 7 vowels. The lengthened vowels displayed a mean duration of around 250 ms (median of 212 ms), with a standard deviation of 31.7 ms. All vowels produced with a creaky phonation were excluded from the analysis. As for longer vowels produced with both modal and creaky portions, only the modal portion was considered.

#### Data extraction

Following the manual segmentation and transcription of all speech material, the extraction of the parameters was carried out automatically using a modified version of the Praat script “ProsodyDescriptorExtractor” developed by the third author. The f0 floor and ceiling were defined as 60 Hz –300 Hz. For F0 smoothing, a cut-off frequency filter of 2 Hz was used to compute f0 linguistically-relevant peaks throughout the utterances.

### Fundamental frequency (f0) acoustic descriptors

A set of 15 f0 measures were considered for assessment in connected speech (i.e., at the domain of sentences), including descriptors of f0 dispersion, central tendency, and modulation (f0M), as presented below. As for lengthened vowels, given the more stationary f0 patterns observed, only the first seven parameters were considered for the analysis:

| Metric    | Description                                           |
|-----------|-------------------------------------------------------|
| f0mean    | f0 mean in semitones ref 1 Hz/ and in Hertz           |
| f0med     | f0 median in semitones ref 1 Hz/ and in Hertz         |
| f0min     | f0 minimum in semitones ref 1 Hz/ and in Hertz       |
| f0max     | f0 maximum in semitones ref 1 Hz/ and in Hertz       |
| f0sd      | f0 standard-deviation in semitones ref 1 Hz/ and in Hertz |
| f0base    | Base value of f0 in semitones ref 1 Hz/ and in Hertz (i.e., equivalent to the 7.4th quantile of the f0 sample, cf.260). |

(Continued)
As most of the data in the present study have shown not to fit into a normal distribution, as verified through the Shapiro-Wilk normality test ($P < 0.05$), the statistical testing was performed employing nonparametric methods. The Kruskal-Wallis rank-sum test was applied to verify possible differences in each tested parameter, followed by the post hoc analysis with the Dunn’s Multiple Comparison Test (two-tailed). The Bonferroni correction was automatically performed to adjust the alpha threshold due to multiple comparisons.

Comparisons were made across all speakers, twin pairs (intra-twins) and nontwins (cross-pairs), yielding a total of 190 comparisons. The reason for including cross-pair comparisons is mainly because this kind of comparison may be regarded as more realistic from a forensic viewpoint, characterized by speakers who are, in most cases, interacting with someone they know. This procedure also allowed treating all types of comparisons with the same statistical criterion.

For the estimation of the Kruskal-Wallis Effect Size, the following formula was applied, where $H$ is the value obtained in the Kruskal-Wallis test; $k$ is the number of groups; $n$ is the total number of observations:

$$
\eta^2 = \frac{(H - k + 1)}{(n - k)}
$$

(1)

Furthermore, the magnitude of the differences were attributed automatically by the package “rstatix,” version 0.6.0, in the R software, applying the values commonly reported in the literature for the eta-squared ($\eta^2$): 0.01 ≤ $\eta^2$ ≤ 0.06 (small effect), 0.06 ≤ $\eta^2$ ≤ 0.14 (moderate effect), and $\eta^2$ ≥ 0.14 (large effect). As such, the effect size index assumes values ranging from 0 to 1, which when multiplied by 100% indicates the explained variance in percent of the dependent variable by the independent variable (cf. 40, 41). In the present study, the independent variable was the factor “speaker”.

### The speaker discriminatory performance of f0 descriptors

Regarding the assessment of the suitability of f0 metrics for potential forensic applications, three estimates will be examined. The first estimate is the Log-likelihood-ratio-cost function (Cllr), an empirical estimate of the precision of likelihood ratios proposed by Brümmel and Du Preez 42 and applied, among others, by da Silva et al. 43 and Morrison 44. According to Morrison et al. 44, such an estimate has the desired properties of being based on likelihood ratios, being continuous, and more heavily penalizing worse results. For computing such estimate, likelihood ratios were calculated through Multivariate Kernel Density analysis - MVKD 45 (i.e., a nonparametric approach), implemented in the package “fvclrr” in R. 46 Multiple pairwise comparisons were performed across individuals in which the background sample consisted of data from all speakers, except those being directly compared (i.e., cross-validation). For comparability, calibrated and raw Cllr estimates are provided. Calibration is a method performed on log-likelihood ratios to reduce the magnitude and incidence of likelihood ratios known to support the incorrect hypothesis, i.e., the contrary-to-fact hypothesis, thereby improving accuracy. Such a procedure is based on a logistic regression model trained with the same set of data (i.e., self-calibration), cf. 44, 47.

The second estimate is the Equal Error Rate (ERR), which represents the point where the false reject rate (type I error) and false accept rate (type II error) are equal, being used to describe the overall accuracy of a biometric system. 48 This estimate is generated along with the Cllr. Both Cllr and ERR values are reported as average values after performing several tests.

Finally, in order to observe the performance of acoustic parameters in terms of their sensitivity and specificity, Receiver Operating Characteristics (ROC) graphs will be plotted. ROC plots are two-dimensional graphs that depict relative tradeoffs between benefits (true positives) and costs (false positives), providing an estimate that allows the comparison across models/metrics, the “Area Under the ROC curve” (AUC) estimate, cf. 49. Moreover, the multi-class ROC function as defined by Hand and Till 50 was applied to compute the multi-class AUC, which basically provides a mean of several AUC estimates.

For the sake of interpretation, it should be observed that an ideal metric for the forensic application should present relatively low Cllr and ERR while displaying relatively high AUC values in relation to the other parameters under comparison.
RESULTS

Average values of \( f_0 \) estimates in connected speech and in lengthened vowels are summarized in Table 1. Furthermore, the statistical analysis outcomes are displayed in Table 2. In Table 2, the numbers of statistical differences for each tested parameter are expressed as a function of all inter-speaker comparisons (i.e., cross-pair differences) and intra-twin pairs comparisons (i.e., twin pairs differences). The parameters are displayed in decreasing order according to the proportion of inter-speaker differences observed. In addition, effect size estimates regarding inter-speaker comparisons and their respective magnitudes are presented. In Table 3, system performance estimates regarding the comparison across all individuals are presented, namely Cllr, EER, and AUC.

Intra-twin pairs and cross-pair differences

Connected speech

Through the inspection of Table 1, it can be observed that central tendency measures of \( f_0 \) (e.g., mean and median)

| Parameter | Interspeaker diff. | Twin pairs diff. | (Interspeaker) Effect size | Magnitude |
|-----------|--------------------|------------------|----------------------------|-----------|
| \( f_0 \) med | 98 (51%) | E1-E2 (2.4 st); F1-F2 (2.6 st) | 50.3% | Large |
| \( f_0 \) mean | 94 (49%) | E1-E2 (1.9 st); F1-F2 (2.7 st) | 49.4% | Large |
| \( f_0 \) base | 91 (47%) | E1-E2 (2.4 st); F1-F2 (1.8 st) | 47.4% | Large |
| \( f_0 \) min | 75 (39%) | E1-E2 (2.6 st); F1-F2 (1.9 st) | 32.1% | Large |
| \( f_0 \) max | 70 (36%) | F1-F2 (3.5 st) | 26.7% | Large |
| \( f_0 \) M8 | 37 (19%) | G1-G2 | 12.7% | Moderate |
| \( f_0 \) M5 | 30 (15%) | F1-F2 | 11.8% | Moderate |
| \( f_0 \) M7 | 21 (11%) | None | 9.0% | Moderate |
| \( f_0 \) sd | 20 (10%) | None | 9.0% | Moderate |
| \( f_0 \) M4 | 18 (9%) | None | 8.8% | Moderate |
| \( f_0 \) M6 | 14 (7%) | None | 5.9% | Small |
| \( f_0 \) SAQ | 13 (6%) | None | 6.1% | Moderate |
| \( f_0 \) M2 | 8 (4%) | None | 5.1% | Small |
| \( f_0 \) M1 | 4 (2%) | None | 3.0% | Small |
| \( f_0 \) M3 | 1 (0.5%) | None | 1.6% | Small |

Lengthened Vowels

Table 2

Number of Significant Differences in Comparisons Among All Speakers and Intra-twin Pairs for \( f_0 \) Descriptors in Connected Speech and Lengthened Vowels, Followed by Effect Size Estimates (\( h^2 \)).
fell within a very specific range, varying from 74 to 93 semitones in connected speech and from 76 to 89 in vowel segments for median values. Nevertheless, some inter-speaker differences could be observed within such intervals.

As can be seen in Table 2, the analysis of the parameters’ discriminatory potential on the basis of the differences observed among all individuals suggested five f0 estimates as being remarkably discriminatory, namely f0 median, f0 mean, f0 baseline, f0 min, and f0 max, in decreasing order of relevance. Notably, although large effect sizes have been observed for all such measures, central tendency measures and f0 baseline were the ones displaying the highest scores, suggesting a higher explanatory potential of the variable “speaker” on the differences observed. Within this group of parameters, two identical twins out of ten (20%) were found different from each other: E1-E2 and F1-F2, except for f0 max that yielded only one difference: F1-F2.

By consulting Table 2, it can be seen that f0 variability and modulation metrics displayed only moderate to small effect sizes, being accompanied by a lower proportion of differences across speakers. Furthermore, all metrics displaying small effect sizes were the ones corresponding to f0 modulation. Regarding such metrics, only one identical twin pair (per parameter) could be differentiated for two out of ten measures, G1-G2 (f0M8) and F1-F2 (f0M5).

In Figures 1 and 2, Kernel density diagrams (i.e., a smoothed version of the histogram) are presented for the six most speaker-contrasting metrics (Figure 1), and the six least speaker-contrasting metrics (Figure 2), according to the proportion of inter-speaker (cross-pair) differences observed. Moreover, individual mean values are displayed at the bottom of the distributions, as expressed by rounded points. Through visual inspection of such Figures, general trends regarding the metrics’ distributions and variability can be observed, revealing convergences or divergences across subjects.

As can be seen, greater variability in mean values and data distributions can be verified across speakers for those parameters displaying higher proportions of inter-speaker differences, such as f0 median, mean, minimum, baseline, and maximum (see Figure 1). Conversely, a higher alignment or congruence regarding mean values and distributions can be noted for those metrics displaying a lower discriminatory potential, as observed for the parameters in Figure 2. Furthermore, it is noteworthy that either a higher or a lower overlap of individuals’ average values directly affects the estimation of effect sizes and is therefore reflected on the estimates. In that regard, greater divergences in the estimates’ average values across speakers tend to yield larger effect sizes and vice-versa, adding valuable information on how variable a metric is.

### TABLE 3

| Metric           | Cllr (raw) | Cllr (cal) | EER | AUC |
|------------------|------------|------------|-----|-----|
| (f0med) / 0 median | 0.56       | 0.55       | 0.24 | 0.78 |
| (f0mean) / 0 mean | 0.54       | 0.54       | 0.20 | 0.77 |
| (f0base) / 0 baseline | 0.52   | 0.50       | 0.13 | 0.77 |
| (f0min) / 0 minimum | 0.68     | 0.68       | 0.23 | 0.72 |
| (f0max) / 0 maximum | 0.80     | 0.76       | 0.30 | 0.69 |
| (f0sd) / 0 standard deviation | 1.03 | 0.95       | 0.45 | 0.61 |
| (f0SAQ) / 0 dispersion | 0.95 | 0.94       | 0.46 | 0.59 |
| (f0M1) smoothed / 0 peak rate | 3.76 | 1.00       | 0.46 | 0.57 |
| (f0M2) sd of / 0 maxima | 1.10 | 0.98       | 0.40 | 0.58 |
| (f0M3) sd of the / 0 maxima positions in seconds | 1.25 | 1.00       | 0.45 | 0.57 |
| (f0M4) / 0 rising rate in the peaks | 1.00 | 0.95       | 0.45 | 0.61 |
| (f0M5) / 0 falling rate in the peaks | 0.98 | 0.91       | 0.40 | 0.62 |
| (f0M6) sd of / 0 rising rate | 1.02 | 0.98       | 0.44 | 0.59 |
| (f0M7) sd of / 0 falling rate | 0.96 | 0.93       | 0.39 | 0.61 |
| (f0M8) / 0 peak width | 1.01 | 0.94       | 0.40 | 0.63 |

### Lengthened Vowels

| Metric           | Cllr (raw) | Cllr (cal) | EER | AUC |
|------------------|------------|------------|-----|-----|
| (f0med) / 0 median | 0.88       | 0.83       | 0.34 | 0.73 |
| (f0mean) / 0 mean | 0.87       | 0.83       | 0.32 | 0.73 |
| (f0base) / 0 baseline | 0.84   | 0.81       | 0.29 | 0.74 |
| (f0min) / 0 minimum | 0.84     | 0.81       | 0.30 | 0.74 |
| (f0max) maximum | 0.91       | 0.85       | 0.38 | 0.72 |
| (f0sd) / 0 standard deviation | 1.09 | 1.00       | 0.49 | 0.57 |
| (f0SAQ) / 0 dispersion | 1.13 | 1.00       | 0.54 | 0.56 |

**Lengthened vowels**

Regarding the analysis of f0 metrics extracted from lengthened vowels, as summarized in Table 2, a relatively smaller proportion of inter-speaker differences could be observed compared to f0 differences in connected speech for the same estimates. Also, relatively similar effect sizes were observed for central tendency and f0 extreme measures in lengthened vowels. Regarding the comparison of the most speaker-contrasting measures, namely f0med, f0mean, and f0base, a reduction in effect size of approximately 7%-10% was verified between connected speech and lengthened vowels, suggesting a higher explanatory potential of the variable “speaker” on f0 patterns in the former.

Furthermore, no intra-twin pair differences could be observed when twins were compared for such linguistic material, which is compatible with a higher congruence of their f0 patterns in the production of lengthened vowels, in most cases, perceived as filled pauses.
As shown in Table 1, \( f_0 \) estimates in lengthened vowels tended to be less variable than in connected speech, as represented by smaller standard deviation and smaller range values for all metrics assessed. Also, the smaller discrepancy between global average \( f_0 \) minimum, median, mean, and maximum values in the production of such phenomena suggest its relative “stationary” status, in contrast to what was verified in connected speech.

FIG. 1. Kernel density diagrams and mean individual values of \( f_0 \)med, \( f_0 \)mean, \( f_0 \)min, \( f_0 \)baseline, \( f_0 \)max (in semitones), and \( f_0 \)M8 (peak width).
The performance of $f_0$ metrics for speaker comparison applications

Through the examination of Table 3, which summarizes the overall performance of $f_0$ metrics in terms of log-likelihood-ratio costs (Cllr raw and Cllr cal), equal error rates (EER), and AUC values, it seems remarkable that, among all metrics, the median, mean and baseline of $f_0$ were the ones presenting the best discriminatory

FIG. 2. Kernel density diagrams and mean individual values of F0M4 ($f_0$ rising rate), F0M6 (sd of $f_0$ rising rate), f0SAQ ($f_0$ dispersion), F0M2 (sd of $f_0$ peaks), F0M1 ($f_0$ peak rate/s), F0M3 (sd of $f_0$ peak’s duration), respectively.
performance. This is expressed by the combined lowest Cllr (raw and calibrated), EER, and highest AUC values observed for such estimates. Furthermore, among all metrics, the $f_0$ baseline was the estimate displaying the lowest Cllr and EER value, outperforming all the others in terms of its accuracy. This observation applies mostly to the analysis of $f_0$ in connected speech.

It is possible to visualize the performance of $f_0$ metrics in terms of their specificity (true positive rate) and sensitivity (false positive rate) for some intra-twin pair comparisons in Figures 3 and 4, which depicts the ROC curves and corresponding the AUC values. By inspecting the graphs, it can be verified that $f_0$ central tendency, floor, and ceiling estimates plus $f_0$ baseline are, in fact, the ones whose curves
most often deviates from the central line towards the upper left point (1.0); hence these are also the estimates presenting the highest AUC values (i.e., the area under the ROC curve). Note that the diagonal line indicates a 50% chance of correct or false acceptance; as such, those metrics closer to this threshold are regarded as less accurate concerning their binary classification performance. Note that, from the plotted estimates, f0 standard deviation (f0sd) and dispersion (f0SAQ) are often the ones closer to the diagonal line, suggesting poorer performances. Notwithstanding, some
exceptions could be noted, such as for the ROC curves of C1-C2 and I1-I2 in Figure 4. It should also be noted that these plotted binary comparisons only consider intra-twin pair comparisons, where the performance of a system is already expected to deteriorate. For visualizing some cross-pair, i.e., nontwin, comparisons in terms of ROC analysis, see Figures 5 and 6, where either better or worse discriminatory performances can be observed, depending on the pairs under comparison, as suggested by the variation in AUC values.

The ROC plots in Figure 5 are not in alphabetic order to draw attention to E1-E2 and F1-F2 in Figure 3, i.e., the twin pairs with the highest proportions of intra-twin pair...
differences. As can be seen, in terms of ROC analysis, these twin pairs were the ones that most differed from each other, as indicated by the relatively higher AUC values for centrality measures, such as $f_0$ baseline, $f_0$ median, and $f_0$ mean.

Finally, regarding the comparison of a system’s performance based on $f_0$ metrics extracted from connected speech vs. lengthened vowels samples, relatively better performance is suggested for the former compared to the latter. In Table 3, it is possible to verify that EER and particularly Cllr values were found relatively increased, whereas a slight decrease of AUC values in lengthened vowels is indicated. Such an increase in Cllr values was observed even when
considering calibrated Clr estimates. Clr cal estimates were found relatively better in connected speech than in lengthened vowels. Regarding the EER of f0 baseline, note that this metric was 16% higher in lengthened vowels (29%) compared to connected speech (13%).

DISCUSSION

The purpose of this study was to evaluate the speaker-discriminatory potential of a set of f0 metrics in comparison carried out with genetically related subjects (i.e., identical twins) and in cross-pair comparisons among all speakers. The data assessed here may be regarded as ecologically valid since it comprises dialogues between very familiar subjects, namely identical twin pairs who have grown up together. The inter-speaker comparisons were performed on the basis of two different kinds of speech material: intervals of continuous speech (i.e., at the level of sentences) vs. lengthened vowels extracted from dialogues which were perceived, in most cases, as filled pauses.

Some general discriminatory patterns were suggested regarding the variability of f0 metrics in intra-twin pair and cross-pair comparisons. Overall, central tendency long-term f0 metrics plus f0 baseline were found the most speaker discriminatory estimates in connected as well as in lengthened vowels. The second most contrasting metrics category regards f0 floor and ceiling estimates. Such f0 descriptors were also the ones displaying the largest effect sizes. Conversely, f0 variation and modulation-dependent estimates, such as f0 standard deviation and dispersion, were the ones displaying the lowest proportions of inter-speaker differences.

The observation of higher discriminatory potential of long-term f0 measurements in more extensive speech intervals is not surprising. As expected, f0 excursions on such intervals are much more expressive than in lengthened vowel tokens, which present a relative stationary f0 contour pattern.

Such an observation may also have critical perceptual implications, since the listener’s discriminatory performance may be potentially better when larger speech stretches are used in auditory discrimination trials, as more acoustic information will be present.

Although the set of f0 parameters assessed in the study conducted by Whiteside and Rixon was considerably smaller than the set of parameters examined in the present study, the main outcomes point in the same direction. When comparing siblings, it was observed that the measurement associated with f0 variability (i.e., f0 standard deviation) was the parameter displaying the greatest level of similarity across the siblings, while mean f0 values presented the highest discrepancies, even between an identical twin pair. According to the authors, the results suggest that more dynamic aspects of f0 may be under a greater influence of environmental variables, which are responsible for shaping intonation patterns, such as dialect and speaking style. Such a higher environment influence on the f0 variability was also previously suggested by Debruyne et al.

The outcomes of the present study do not disprove the assumption that individuals vary in how they exploit fundamental frequency. However, the observation of lower interspeaker difference proportions for f0 variation/modulation estimates invites the hypothesis of such metrics as varying to a lesser extent in comparisons based on data representing the same speaking style, same speaking condition, same dialect, and with a relatively homogeneous group composed only by young male speakers.

Variation in the fundamental frequency dimension has also been referred to as associated with different speaking styles’ particular characteristics. In the study conducted by Higuchi et al., while analyzing f0 contours of Japanese sentences on the basis of f0 min, f0 amplitude in the phrase domain, and f0 amplitude in the lexical domain, the authors observed clear differences when comparing four speaking styles, e.g., unmarked, hurried, angry, and gentle. Amongst the analyzed speaking styles, more expressive differences were observed for the angry speech, which was characterized by a high f0 min and a minimum change in both phrase and lexical amplitudes, yielding reasonable flat f0 contours.

Furthermore, the higher convergence among individuals for f0 variation metrics in the present study seems to be in broad agreement with Arantes and Eriksson, when observing that speaking style has a significant effect on the shaping of f0 distributions, with a higher or lower variation depending on the speaking style being analyzed. In the study carried out with a multilingual corpus (including the analysis of Brazilian Portuguese, British English, and Swedish) comprised of speech productions in three different speaking styles, the authors of the referred study observed that interview or sentence reading were the styles in which speakers differed the most in terms of f0 distribution shape. Furthermore, encouraging evidence of a remarkable intra-speaker stability of the shape of f0 distributions was also found, in which f0 contour by the same speaker tended to vary less when speaking in different styles than the contours of different speakers that are speaking in the same style, which was especially the case for the spontaneous speaking style. As suggested by the researchers, this could indicate the suitability of this feature as a good parameter in forensic speaker comparison, as expressed by a low within-speaker variation and a high interspeaker variability.

Intratwin pair comparisons

Concerning intratwin pair comparisons, it was possible to observe that from a total of 11 intrapair significant differences, only two of those concerning f0 modulation metrics, whereas nine differences derived from f0 central tendency, baseline, and extreme values, mostly from f0 median, mean, baseline and minimum. Moreover, from the group of 10 identical twin pairs assessed, only three pairs could be differentiated according to the present study’s approach regarding the connected speech analysis, namely E1-E2, F1-F2, and G1-G2; however, only two of these were more systematically and consistently differentiated: E1-E2 and F1-
F2. As for the analysis of lengthened vowel segments, only one twin pair was statistically different: I1-I2.

It is noteworthy that E1-E2, mostly E2, are the only speakers in the present study with a reported smoking habit, whereas both F1 and F2 are the ones who use their voice artistically (i.e., acting). Although, according to E1-E2, smoking is not a shared daily-based habit, such information signals the relevance of further analysis, given the widely reported and well-described effects of smoking on \( f_0 \), characterized mainly by a \( f_0 \) lowering.\(^{51–53}\) Notwithstanding, the fact that such difference could not be observed when considering the analysis of \( f_0 \) metrics in the set of lengthened vowels selected does not allow any inference to be made. In addition, in a previous study conducted with the same twin pairs, which mainly focused on the analysis of vowel formants (i.e., F1, F2, F3, and F4), E1-E2 was also found statistically distinct, particularly regarding their fourth vowel formant (F4),\(^ {54}\) revealing that the phonetic-acoustic dissimilarity between E1-E2 is not solely restricted to the \( f_0 \) dimension.

It should be noted that F1-F2 was the twin pair displaying the highest proportion of significant differences. This twin pair is responsible for more than half of the intratwin pair differences observed. The fact that such speakers have a reported experience in using their voice artistically (which was, in fact, one of their main topics during the dialogues, where vocal quality adjustments for composing a certain character were mentioned) may suggest a higher “self-awareness” regarding how they sound and are perceived by others. In this instance, the factor “choice” cannot be disregarded, as suggested and considered by many other twin studies at different phonetic-acoustic levels.\(^ {20,55–58}\) Nonetheless, future studies must consider the possible influence of other factors, such as recent heavy voice usage or vocal overload, which were not explored here, and the objective evaluation of hearing acuity, for instance. In this regard, relying on the participants’ judgment can sometimes be a source of error.

The observation that no significant differences were observed in the comparison of MZ and DZ twins regarding the \( f_0 \) variation in Debruyne et al.\(^ {28}\) is a rather interesting finding, since it suggests an unbalanced contribution of genetic and environmental factors on the parameter average values and variation. As remarked by the authors, while the \( f_0 \) variability may be highly determined by behavioral and adaptive elements (i.e., under a strong environmental influence), a comparable influence on both DT and MT is presumed. As such, genetic factors may be overwhelmed by external factors, no longer accounting for the difference between MT and DT.

In the present study, the verification of a possibly higher intratwin pair congruence regarding \( f_0 \) variation and modulation metrics solely can not be taken as an indication that such divergences do not occur in the speech of twins. Here, the mere fact that identical twins were taking part in a simultaneous dialogue may induce some sort of prosodic entrainment or synchronicity, which may be on the basis of their higher congruence. Such synchronicity has been explored and corroborated in the experiment conducted by Buder and Eriksson\(^ {69}\) while analyzing the prosodic cycles of mean speech fundamental frequency and mean voice intensity between subjects while engaging in a conversation.

Notwithstanding, a potential “synchronicity effect” can not be generalized to the comparisons performed across all subjects (i.e., cross-pair comparisons), where a higher congruence for \( f_0 \) variation and modulation across speakers has also been suggested. Note that, except for the ten intratwin pair dialogues in the 190 cross-pair comparisons, 180 comparisons were carried among individuals who were not taking part in the same dialogue (e.g., A1 - B1, C2 - A2, B2 - C1).

It should also be remarked that \( f_0 \) patterns, like any other phonetic dimension, are regulated by the interaction of both intrinsic (system-oriented) and extrinsic (output-oriented) factors.\(^ {60}\) Hence, assuming that identical twins were, in all cases, interacting with someone they were very familiar with is also expected to affect the way they exploit \( f_0 \) in their speech. This observation adds to the ecological validity of the speech material analyzed.

Finally, of an unquestionable relevance is whether listeners could potentially perceive the discrepancies in \( f_0 \) central tendency and extreme values in intratwin pair comparisons observed. In this regard, while assessing listeners’ sensitivity to differences in the amount of change in \( f_0 \) in an experiment with a “forced-choice” sample comparison procedure, where stimulus contained a pitch movement of variable size ranging from 1 to 6 semitones, ‘t Hart\(^ {61}\) found evidence suggesting that only differences of more than three semitones seem to play a part in communicative situations. Notwithstanding, there was evidence for good discriminators as being able to perceive \( f_0 \) modulations of about 2.1 - 2.8 semitones, with \( f_0 \) rises being better perceived than \( f_0 \) falls.

As verified in the present study, the magnitude of the differences observed for E1-E2 and F1-F2 concerning \( f_0 \) central tendency and extreme values in speech were, in some cases, on average, just below three semitones, except F1-F2 for \( f_0 \) maximum. Therefore, it is uncertain whether such differences, for \( f_0 \) only, were large enough to be perceived by the individuals, which certainly demands further analysis.

**Connected speech vs. lengthened vowels**

Regarding the comparison of \( f_0 \) metrics in connected speech versus lengthened vowels, a reduction in speaker-discriminatory potential was suggested, as expressed by lower proportions of interspeaker differences, mostly for \( f_0 \) median, mean, and baseline. Besides, smaller effect sizes were observed in lengthened vowels, which is compatible with a higher explanatory potential of the variable “speaker” on global patterns of \( f_0 \) in connected speech. From a statistical viewpoint, such effect size reduction may be interpreted as the consequence of a lower variability of \( f_0 \) in lengthened vowels, which tend to present a relatively stationary pattern, in contrast to more extensive speech intervals.
In this regard, different studies within the clinical voice domain (i.e., vocology studies) support the observation that different acoustic outcomes regarding \( f_0 \) may be obtained depending on the task and consequently on the nature of the material under assessment (e.g., comparisons of sentences and sustained vowels), suggesting that for a more reliable assessment of the parameter, materials collected in more than one task have to be considered.\(^{37,38,62,63}\)

Although the differences observed in the present study on account of the comparison of \( f_0 \) parameters in speech intervals and lengthened vowels can not be attributed to the variable “task” given that the measurements were extracted from the same recordings, they can be potentially assigned to differences in the speech material under assessment. Overall, the results suggest that not all the acoustic complexity present in speech regarding \( f_0 \) is reflected in the production of lengthened vowels as observed in the present study.

Finally, in the present study, a higher intratwin pair convergence for lengthened vowels comparisons in relation to connected speech comparisons was suggested. According to the present study’s analysis approach, no identical twin pair seemed to diverge significantly in the former condition. Such an observation may indicate that a greater interspeaker variability in long-term \( f_0 \) estimates is present in continuous speech, a condition in which twins can avail of their vocal plasticity and, deliberately or not, diverge from each other.

In summary, the findings of the present experimental research appear to point in the direction that the magnitude of phonetic-acoustic similarity/dissimilarity between identical twins is, in part, dependent on the nature of speech/voice material under analysis, and consequently, on how representative of the speaker’s voice/speech behaviour it is in a broader sense.

**Discriminatory performance**

The verification of the base value of \( f_0 \) (i.e., \( f_0 \) baseline) as the most accurate measure is in broad agreement with previous literature reporting. While conducting a comprehensive study involving different \( f_0 \) metrics with Brazilian Portuguese speakers using a Multivariate Kernel Density (MVKD) analysis, da Silva et al.\(^{23}\) noted that the \( f_0 \) baseline was the estimate displaying the lowest Equal Error Rate (ERR) in relation to other \( f_0 \) measures. Moreover, the authors noted an even better discriminatory performance when combining \( f_0 \) baseline and \( f_0 \) median. Furthermore, \( f_0 \) baseline has also been reported in the literature as the most stable parameters with regard to different sources of variation, such as speaking style, vocal effort, and recording quality.\(^{26}\)

Additional support for the use of baseline and median values of \( f_0 \) instead, especially when conducting forensic analysis, regards the fact that, as remarked by Lindh and Eriksson\(^{26}\) the extraction of \( f_0 \) values from recordings is often done automatically. When manual examinations are carried out after automatic extractions, they may very likely reveal measurement errors which are also dependent on the audio quality, such as octave jumps. It goes without saying that, within a forensic context, audio samples with poor quality are the rule, and good audio recordings the exceptions. Therefore, using median values instead of means would be more reliable as it is less sensitive to possible outliers. In that instance, perhaps the reason why a discrepancy between \( f_0 \) mean and \( f_0 \) median could not be observed in the present study has to do with the fact that only very high-quality recordings were assessed, not allowing such verification.

Future studies should also explore more in-depth how the patterns observed here vary as a function of different speaking styles and different communicative contexts (e.g., spontaneous dialogue vs. interview with an unfamiliar interlocutor), providing relevant information on the levels of stability and variability of the parameters assessed for forensic speaker comparison purposes.

It is important to emphasize that, although in the present experimental study, \( f_0 \) metrics have been assessed individually, different phonetic-acoustic estimates should be acknowledged in a realistic forensic context. The combination of different metrics from different phonetic-acoustic dimensions (e.g., spectral, temporal) tends to yield better explanatory models and more accurate speaker profiles.

Finally, as remarked by Braun\(^{2}\) given the social-cultural influences on how \( f_0 \) is perceived in a language, one should be extremely cautious when considering the use of reference material acquired from one language group to another. This fact justifies the relevance of cross-linguistic studies, with particular consideration to those parameters commonly analyzed in a forensic setting.

**CONCLUSION**

The present study allowed identifying a set of potentially relevant speaker-discriminatory \( f_0 \) estimates for forensic speaker comparison. The findings in the present study add to the body of knowledge regarding how \( f_0 \) parameters are exploited by individuals under different levels of interspeaker variation control. Overall, \( f_0 \) baseline, central tendency, and extreme values were found to display higher proportions of intratwin pair and cross-pair differences while also presenting the largest effect sizes. Conversely, \( f_0 \) variation and modulation estimates were found relatively more stable across different subjects, inviting the hypothesis that such metrics may be partly controlled by speaking style and dialect. Moreover, \( f_0 \) metrics assessed in connected speech tended to present a better discriminatory potential than lengthened vowels. The outcomes also suggest that, although identical twins were found very closely related regarding their \( f_0 \) patterns, some pairs could still be differentiated acoustically, mainly in connected speech. Whether such differences in \( f_0 \) estimates solely were large enough to be perceived by external listeners is uncertain. However, such a finding reinforces the relevance of analyzing long-
term f0 metrics for forensic purposes, particularly of f0 baseline, which displayed the lowest equal error rates.

Acknowledgments
This research, which derives from the first author’s doctorate thesis, was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, in cooperation with the Swedish Foundation for International Cooperation in Research and Higher Education - Sweden (STINT), and in part by the National Council for Scientific and Technological Development - Brasil (CNPq). Special thanks to Justin Lo (University of York) for providing one of the scripts for the analyses.

SUPPLEMENTARY DATA
Supplementary data related to this article can be found online at doi.org/10.6084/m9.figshare.16689433.v1.

REFERENCES
1. Williams CE, Stevens KN. Emotions and speech: some acoustical correlates. J Acoust Soc Am. 1972;52:1238–1250.
2. Braun A. Fundamental frequency: how speaker-specific is it? Beiträge zur Phonetik und Linguistik. 64. 1995:1951995:1959:9–23.
3. Traummüller H., Eriksson A. The frequency range of the voice fundamental in the speech of male and female adults [unpublished manuscript]. 1995. https://www2.ling.su.se/staff/hartmut/f0_m&f.pdf.
4. Awan SN, Mueller PB. Speaking fundamental frequency characteristics of white, african american, and hispanic kindergartners. J Speech Lang Hear Res. 1996;39:573–577. https://doi.org/10.1044/jslhr.3903.573.
5. Higuchi N., Hirai T., Sagisaka Y.. Effect of speaking style on parameters of fundamental frequency contour1997;:417–421;10.1007/s0892-1997(03)00031-6.
6. Escudero P, Boersma P, Rauber AS, et al. A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. J Acoust Soc Am. 2009;126:1379–1393. https://doi.org/10.1121/1.3180321.
7. Arantes P, Eriksson A. Quantifying fundamental frequency modulation as a function of language, speaking style and speaker. INTERSPEECH2019:1716–1720.
8. Probst L, Braun A. The effects of emotional state on fundamental frequency. In: Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia2019.
9. Rilliar A, de Moraes JA, Erickson D, et al. Social affect production and perception across languages and cultures—the role of prosody. Leitana. 2013;2:15–41. https://doi.org/10.28998/2317-9945.2013v2n2p15-41.
10. Signorello R, Demolin D, Bernardoni NH, et al. Vocal fundamental frequency and sound pressure level in charismatic speech: a cross-gen-der and-language study. J Voice. 2020;34:808–811. https://doi.org/10.1016/j.jvoice.2019.04.007.
11. Klofstad CA, Anderson RC, Nowicki S. Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. PLoS ONE. 2015;10:e0133779. https://doi.org/10.1371/journal.pone.0133779.
12. Zhang Z. Mechanics of human voice production and control. J Acoust Soc Am. 2016;140:2614–2635. https://doi.org/10.1121/1.4696409.
13. Van den Berg J. Myoelastic-aerodynamic theory of voice production. J Speech Hear Res. 1958;1:227–244.
14. Gold E, French P. International practices in forensic speaker comparision. Int J Speech Lang Law. 2011;18:293–307.
15. Louks D. Variation in long-term fundamental frequency: measurements from vocalic segments in twins’ speech. In: Proceedings of the 11th Australian International Conference on Speech Science & Technology. University of Auckland, New Zealand; 2006:205–210. https://assta.org/proceedings/sst2006/sst2006-107.pdf.
16. Carson CP, Ingrisano DR-S, Eggleston KD. The effect of noise on computer-aided measures of voice: a comparison of eschepshp and the multi-dimensional voice program software using the cal 4300b module and multi-speech for windows. J Voice. 2003;17:12–20. https://doi.org/10.1016/s0892-1997(03)00031-6.
17. Fuchs R, Maxwell O. The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range. In: Proceedings of Speech Prosody, ISCA; 2016:523–527. https://doi.org/10.21437/SpeechProsody.2016-107.
18. Maryn Y, Ysenbaert F, Zarowski A, et al. Mobile communication devices, ambient noise, and acoustic voice measures. J Voice. 2017;31:248–e11. https://doi.org/10.1016/j.jvoice.2016.07.023.
19. Jannetts S, Schaeffler F, Beck J, et al. Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types. Int J Lang Commun Disord. 2019;54:292–305. https://doi.org/10.1111/1460-6984.12457.
20. Cavalcante JC, Englert M, Oliveira Jr M, et al. Microphone and audio compression effects on acoustic voice analysis: a pilot study. J Voice. 2021. https://doi.org/10.1016/j.jvoice.2020.12.005.
21. Jessen M. Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. Formal Linguistics and Law Berlin: Mouton de Grayter. 2009:115–139. https://doi.org/10.1515/9783110218398.2.115.
22. Arantes P, Eriksson A. Spatial and temporal sampling of a long-periodic frequency. Speech Prosody. 20142014:1149–1152. https://www.isca-speech.org/archive/SpeechProsody_2014/pdfs/224.pdf.
23. da Silva RR, da Costa JPC, Miranda RK, et al. Applying base voltage of fundamental frequency via the multivariate kernel-density in forensic speaker comparison. 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE; 2016:1–8. https://doi.org/10.1109/ICSPCS.2016.7843307.
24. Arantes P, do Nascimento Linhares ME. Investigando a robustez de uma metodologia para determinação do valor de base da frequência fundamental. Revista de Estudos da Linguagem. 2018;26:535–570.
25. Passetti R.R., Estudo acústico-perceptual do estilo de fala telefônico com implicações para a verificação de locutor em portugues brasileiro2018.
26. Lindh J, Eriksson A. Robustness of long time measures of fundamental frequency. INTERSPEECH2007:2007–2008.
27. Przybyla BD, Horii Y, Crawford MH. Vocal fundamental frequency in a twin sample: looking for a genetic effect. J Voice. 1992;6:261–266.
28. Debruyne F, Decoster W, Van Gijssel A, et al. Speaking fundamental frequency in monozygotic and dizygotic twins. J Voice. 2002;16:466–471.
29. Van Lierde KM, Vinck B, De Ley S, et al. Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach. J Voice. 2005;19:511–518.
30. Wuyts FL, Bodt MSD, Molenberghs G, et al. The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. J Speech Lang Hear Res. 2000;43:796–809. https://doi.org/10.1044/jslhr.4303.796.
31. Cielo CA, Agustini R, Finger LS. Caracteristicas vocais de gêneros monozigóticos. Revista CEFAFAC. 2012;14:1234–1241.
32. Whiteside SP, Rixon E. Speech tempo and fundamental frequency patterns: a case study of male monozygotic twins and an age-and sex-matched sibling. Logoped Phoniatr Vocat. 2013;38:173–181. https://doi.org/10.1093/jshr.140154399.2012.742562.
33. San Segundo E, Gómez-Vilda P. Matching twin and non-twin siblings from phonation characteristics. VII Jornadas de Reconocimiento Bio-métrico de Personas. 2013:10–17.
34. San Segundo E, Gómez-Vilda P. Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings. Lang Law/Linguagem e Direito. 2014;12.
35. Boersma P, Weenink D. Praat: doing phonetics by computer [computer program], version 6.1.38. URL http://www.praat.org/ Retrieved January. 2021;14.
36. Barbosa PA. Incações Em Torno Do Ritmo Da Fala. Pontes; 2006.
37. Moon KR, Chung SM, Park HS, et al. Materials of acoustic analysis: sustained vowel versus sentence. J Voice. 2012;26:563–565. https://doi.org/10.1016/j.jvoice.2011.09.007.
38. Sotome T, Kanazawa T, Konomi U, et al. Analysis of speech fundamental frequencies for different tasks in Japanese. J Voice. 2021;1–8. https://doi.org/10.1016/j.jvoice.2020.12.021.
39. Barbosa P.A.. Prosodydescriptorextractor [praat script]. 2020.https://github.com/pabarbosa/prosody-scripts/tree/master/ProsodyDescriptorExtractor.
40. Tomczak M, Tomczak E. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. Trends Sport Sci. 2014;21:19–25.
41. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol: General. 2012;141:2. https://doi.org/10.1037/a0024338.
42. Brümmer N, Du Preez J. Application-independent evaluation of speaker detection. Comp Speech Lang. 2006;20:230–275. https://doi.org/10.1016/j.csl.2005.08.001.
43. Morrison GS. Forensic voice comparison and the paradigm shift. Sci Justice. 2009;49:298–308.
44. Morrison GS, Zhang C, Rose P. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic Sci Int. 2011:208:59–65.
45. Aitken CG, Lucy D. Evaluation of trace evidence in the form of multivariate data. J R Stat Soc Ser C (Appl Stat). 2004;53:109–122.
46. Lo J. Fvclrr: likelihood ratio calculation and testing in forensic voice comparison [r package], version 1.1.1. https://rrdio.github.io/justinjhlo/fvclrr. 2020.
47. Morrison GS, Zhang C, Enzinger E. Forensic speech science. Forensic Sci Int. 2019. https://doi.org/10.1016/j.forsciint.2010.11.001.
48. Conrad E, Misner S, Feldman J. CISSP Study Guide. Newnes; 2012.
49. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
50. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45:171–186. https://doi.org/10.1023/A:1010920819831.
51. Murphy CH, Doyle PC. The effects of cigarette smoking on voice-fundamental frequency. Otolaryngol-Head Neck Surg. 1987;97:376–380. 10.1177/2F019459988709700406.
52. Gonzalez J, Carpi A. Early effects of smoking on the voice: a multidimensional study. Med Sci Monit. 2004;10:CR649–CR656.
53. Ayoub MR, Larrouy-Maestri P, Morosmonme D. The effect of smoking on the fundamental frequency of the speaking voice. J Voice. 2019;33:802–e11. https://doi.org/10.1016/j.jvoice.2018.04.001.
54. Cavalcanti JC, Eriksson A, Barbosa PA. Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. PLoS ONE. 2021;16:e0246645. https://doi.org/10.1371/journal.pone.0246645.
55. Nolan F, Oh T. Identical twins, different voices. Int J Speech Lang Law. 1996;3:39–49.
56. Loakes D. A forensic phonetic investigation into the speech patterns of identical and non-identical twins. Int J Speech Lang Law. 2008;15:97–100. https://doi.org/10.1558/ijssl.v15i1.97.
57. Zuo D, Mok PPK. Formant dynamics of bilingual identical twins. J Phon. 2015;52:1–12. https://doi.org/10.1016/j.wocn.2015.03.003.
58. San Segundo E, Yang J. Formant dynamics of Spanish vocalic sequences in related speakers: a forensic-voice-comparison investigation. J Phon. 2019;75:1–26. https://doi.org/10.1016/j.wocn.2019.04.001.
59. Buder EH, Eriksson A. Prosodic cycles and interpersonal synchrony in american english and swedish. Fifth European Conference on Speech Communication and Technology - EUROSPEECH1997 1997:235–238. https://www.isca-speech.org/archive/archive_papers/eurospeech_1997/e97_0235.pdf.
60. Lindblom B. Explaining phonetic variation: a sketch of the H&H theory. Speech Production and Speech Modelling. Springer Netherlands; 1990:403–439. https://doi.org/10.1007/978-94-009-2037-8_16.
61. ’t Hart J. Differential sensitivity to pitch distance, particularly in speech. J Acoust Soc Am. 1981;69(3):811–821. https://doi.org/10.1121/1.385592.
62. Zraick RI, Skaggs SD, Montague JC. The effect of task on determination of habitual pitch. J Voice. 2000;14(4):484–489. https://doi.org/10.1016/S0892-1997(00)80005-3.
63. Viegas F, Viegas D, Guimarães GS, et al. Comparison of fundamental frequency and formants frequency measurements in two speech tasks. Revista CEFAC. 2019;21(6).