Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2914942, 9 pages
https://doi.org/10.1155/2020/2914942

Research Article

Forensic Speaker Comparison Using Evidence Interval in Full Bayesian Significance Test

Adelino P. Silva,[1,2,3] **Maurílio N. Vieira**,[4] **and Adriano V. Barbosa**[4]

[1] *Graduate Program in Electrical Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil*
[2] *Institute of Criminalistics of Minas Gerais, Av. Augusto de Lima 1833, 30110-017, Belo Horizonte, MG, Brazil*
[3] *Centro Universitário Newton Paiva, Rua José Cláudio Resende 420, 30494-230, Belo Horizonte, MG, Brazil*
[4] *Department of Electronic Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil*

Correspondence should be addressed to Adriano V. Barbosa; adriano.vilela@cefala.org

Received 10 July 2019; Accepted 7 January 2020; Published 24 March 2020

Academic Editor: Arturo J. Fernández

This paper describes the application of a full Bayesian significance test (FBST) to compute evidence intervals in forensic speaker comparison (FSC). In the FBST approach, the challenge is to apply the test to a large number of observations and to formulate an equation to solve the test quickly. The contribution of the present work is that it proposes an application of the FBST to FSC and develops a method to calculate the FBST for the distribution of expected values (mean) with unknown variance without using Monte Carlo Markov chains (MCMC). Comparisons with other interval inference methodologies indicate that the evidence interval size is 49% greater than that computed with the Gosset approach. The evidence interval presented 71% fewer classification errors than the punctual inference did for the signal-to-noise ratio (SNR) of 17 dB.

1. Introduction

The main task in forensic speaker comparison (FSC) is to analyze two or more voice records to infer whether they come from the same speaker. FSC differs from biometric voice recognition in the hypothesis test approach and in the nature of the voice samples. In the FSC scenario, a *questioned-voice* is compared to a *known-voice*, whereas in biometric recognition, the comparison is made among multiple speakers [1, 2].

The questioned-voice (or voice evidence) is an audio recording accepted as a vestige or evidence in a criminal investigation. The questioned-voice may be recorded in different situations, such as lawful phone interception (wiretapping), recordings of face-to-face conversation, or audio broadcasting.

In FSC, the hypothesis H_0 considers that both the questioned- and known-voices come from different speakers, whereas H_1 assumes that the questioned- and known-voices come from the same speaker.

However, the "individualization" that the hypotheses above propose has been considered a fallacy. This individualization assumes that the result of the confrontation between the questioned and standard voice is unique, without *a priori* probability and without repeating the test for the entire population [3, 4]. According to Saks and Koehler [3], the most reasonable hypotheses would be

$$H\colon \begin{cases} H_0\colon \text{the features of the questioned} - \text{voice are not compatible} \\ \text{with the features of the known} - \text{voice,} \\ H_1\colon \text{the features of the questioned} - \text{voice are compatible with} \\ \text{the features of the known} - \text{voice.} \end{cases} \tag{1}$$

Punctual inference in FSC is based on a score of (dis) similarity [5–7]. Interval inference is a tradeoff between precision and confidence because it sacrifices some precision of the estimate by moving from a point to a range, but results in greater confidence that the statement is correct (inside interval) [8] (pp. 418).

Reports on interval inference in automatic speaker recognition (ASR) began with Bisani and Ney [9], who used bootstrap [10] to compute confidence intervals. Subsequently, Campbell et al. [11] computed confidence intervals using multilayer perceptron (MLP) based on statistical entropy. Later, Koval and Lokhanova [12] used a sigmoid function to approximate the *a posteriori* probability $P(H_0 \mid \vec{x})$, where $\vec{x}$is the voice data and H_0 is the null hypothesis, using Platt scaling [13] and estimated credibility intervals. The credibility interval can also be computed by empirical methods (Morrison et al. [14]).

The present work proposes the application of the full Bayesian significance test (FBST) to compute *evidence intervals* of FSC. This proposal aims to obtain the same confidence of capturing the parameter of interest in FSC and to reduce type I errors, reinforcing the legal aphorism of *Absolvere nocentem satius est, quam condemnare innocentem*. One of the motivations of this work, among others, is to establish a confidence limit of the automatic speaker comparison techniques, primarily when used as a support to quantify an FSC [15].

Applications of the FBST to FSC were not found during the bibliographic survey in the development of this research. Thus, the main contribution of this work is that it proposes an application of the FBST to FSC and develops a method to calculate the FBST for the distribution of the expected value (mean) with unknown variance without using Monte Carlo Markov chains (MCMC).

The results indicate that the application of the FBST to FSC can improve the evaluation of results by the LR framework, reducing the occurrence of type I errors. The FBST also supports decisions on multispeaker comparisons.

The paper is organized as follows. Section 1 presents the FBST and our proposed improvements and proposes adaptations for FSC (the GMM-UBM method was chosen because it presented more satisfactory results in previous experiments than the *i*-vector- and *x*-vector-based methods with deep neural networks (DNN). These experiments were performed with database in Portuguese, quoted in this article, and with voices provided by the Civil Police of Minas Gerais (Brazil) forensic sector. The result of this experiment is in the process of being published). Section 1 compares the evidence interval to other methods. Section 1 presents the conclusion and future research directions.

2. Evidence FSC Interval with the FBST

2.1. Interval Inference in FSC. In classical FSC, the comparison is performed between features of the questioned-voice, $\vec{x}_Q$, and features of the known-voice, $\vec{x}_K$. The features of the universal background model (UBM), $\vec{x}_{\text{UBM}}$, represent the average speaker [5].

The LR can be computed using a GMM-UBM. In this case, the LR equivalent score $\text{LR}(\vec{x}_Q)$ is computed as follows:

$$\text{LR}\left(\vec{x}_Q\right) = \frac{p\left(\vec{x}_Q \mid \lambda_K\right)}{p\left(\vec{x}_Q \mid \lambda_{\text{UBM}}\right)} \begin{cases} \leq \zeta_0, & H_0 \text{ is not rejected,} \\ > \zeta_0, & H_0 \text{ is rejected.} \end{cases} \tag{2}$$

Furthermore, $p(\vec{x}_Q \mid \lambda_K)$ and $p(\vec{x}_Q \mid \lambda_{\text{UBM}})$ are, respectively, the evaluation of the data $\vec{x}_Q$ of the GMM of the known-voice, λ_K, and of the UBM λ_{UBM}.

The GMM-UBM is a methodology applied to voice comparison [7, 16, 17]. In the first studies [5, 18], the GMM-UBM methodology was applied using Mel-frequency cepstrum coefficients (MFCC).

The first step in the GMM-UBM procedure is to compute the GMM of the known-voice, λ_K, and of the UBM λ_{UBM}, which can be computed using the expectation-maximization (EM) algorithm [5]. In the second step, the *Score* of the comparison ($\text{LR}(\vec{x}_Q)$) is obtained as a ratio between two likelihoods: the questioned-voice ($\vec{x}_Q$) versus the known-voice (λ_K) and the questioned-voice versus the UBM model (λ_{UBM}).

The *score* proposed by Reynolds et al. [5] is the sample mean of the log-likelihood ratio (LLR) over Tspeech frames:

$$\begin{aligned} \text{LR}\left(\vec{x}_Q\right) &= \frac{p\left(\vec{x}_Q \mid \lambda_K\right)}{p\left(\vec{x}_Q \mid \lambda_{\text{UBM}}\right)} \xrightarrow{\log} \text{LLR}\left(\vec{x}_Q\right) \\ &= \log\left(\frac{p\left(\vec{x}_Q \mid \lambda_K\right)}{p\left(\vec{x}_Q \mid \lambda_{\text{UBM}}\right)}\right). \end{aligned} \tag{3}$$

Because the features $\vec{x}_Q = \{x_Q[0], x_Q[1], \ldots, x_Q [T-1]\}$are not independent and not identically distributed (i.i.d.), the resulting values are not, technically, a likelihood ratio. Normalization by the number of frames, T, also removes the duration effects from the log-likelihood value. However, the $\text{LLR}(\vec{x}_Q)$ of equation (3) allows us to include an interval-based inference.

Calculating the interval inference is possible empirically or analytically over the sample space. The widespread empirical approaches include bootstrap [10], jackknife [19], and the method proposed by Morrison et al. [14]. One possible analytical method uses the t-Student distribution of Gosset [20, 21]:

$$t_{((\alpha/2),T-1)} \frac{\hat{\sigma}}{\sqrt{T}} \leq \text{LLR}\left(\vec{x}_Q\right) - \mu \leq t_{((1-(\alpha/2)),T-1)} \frac{\hat{\sigma}}{\sqrt{T}}, \tag{4}$$

where $\hat{\sigma}$ is the sample standard deviation, μ is the expected value of $\text{LLR}(\vec{x}_Q)$, and $t_{((\alpha/2),T-1)}$ is a t-Student distribution with significance α and $T-1$ degrees of freedom.

In Section 1, we compare our evidence interval computed using the FBST to Morrison's credibility/confidence intervals, the analytical method in equation (4).

Morrison's approach [14, 22] uses two samples of voice per speaker and measures the LLR from the vowel formants. In these works, the credibility intervals were computed from raw data rather than from a statistic such as the mean. We propose a small modification to Morrison's approach such that the computation is based on the sample mean instead of the raw data.

2.2. Full Bayesian Significance Test. The FBST can be used to compute evidence against a precise hypothesis $\mathrm{LLR}(\vec{x}_Q) = \eta$, where η is a value in the parametric space of LLR of equation (2).

The FBST [23, 24] is a coherent Bayesian significance test for sharp hypotheses. The test is based on an evidence concept value, whose original definition was motivated by practical, juridical, and epistemological requirements. Consider the parametric space Θ and a subset $\theta \in \Theta \subseteq \mathbb{R}^n$and a precise (null) hypothesis H_0 that the parameter lies in the null set, defined by the inequality $(g(\theta))$ and equality $(h(\theta))$ constraints given by the vector functions g and h in the parameter space:

$$\Theta_H = \{\theta \in \Theta \mid g(\theta) \leq 0 \wedge h(\theta) = 0\}. \tag{5}$$

For the experimental data $\vec{x}$, the *a posteriori* density of a precise hypothesis is proportional to the product of the likelihood and the *a priori* density [25]:

$$f_n(\theta \mid \vec{x}) \propto f(\theta)\mathscr{L}(\theta \mid \vec{x}), \tag{6}$$

where $f(\theta)$is an *a priori* density and $\mathscr{L}(\theta \mid \vec{x})$ is the likelihood. The points of the parameter space with highest "surprise" in the null set H_0 are

$$\theta^* = \underset{\theta \in \Theta_H}{\arg\max} f_n(\theta \mid \vec{x}), \tag{7}$$

while the highest relative surprise set (HRSS), T^*, is

$$T^* = \{\theta \in \Theta \mid \quad f_n(\theta \mid \vec{x}) > f_n(\theta^* \mid \vec{x})\}. \tag{8}$$

The *Bayesian evidence* value against $H_0$is the *a posteriori* probability of the "tangent" set; that is,

$$\overline{ev} = \Pr(\theta \in T^* \mid \vec{x}) = \int_{T^*} f_n(\theta \mid \vec{x}) \mathrm{d}x, \tag{9}$$

where $\Pr(\theta \in T^* \mid \vec{x})$ is the probability that the parameter θ is inside T^*. The e-value associated with the FBST is

$$e - \text{value} = 1 - \Pr(\theta \in T^* \mid \vec{x}). \tag{10}$$

The e-value is a probability in the parameter space (μ and ρ), whereas the p value is a probability in the sample space [26]. In Section 1, we use the e-value and $\overline{\mathrm{ev}}$ (Bayesian evidence value against H_0) to compute the evidence interval on FSC using hypothesis H: $\mathrm{LLR}(\vec{x}_Q) = \eta$.

2.2.1. Improvement of the FBST over the Mean with an Unknown Variance. This section describes a method to compute the FBST for a distribution of the mean (expected value) with an unknown variance. To lower the computational cost, we focus on a mostly analytical development. This is important in order to limit the computation time of the e-value over the η-space.

Consider a normally distributed sample $x \in \mathscr{X}$ with n i.i.d. observations, $\mathscr{X}(\mu, (1/\rho))$, where μ is the expected value and $\rho = 1/\sigma^2$ is the precision. The minimal sufficient statistic could be the sample mean $\overline{x}$ and total sum of squares $Q = \sum_{i=1}^{n} (x_i - \overline{x})^2$. The likelihood function for $\mu \in (-\infty, \infty)$and $\rho \in (0, \infty)$ [26] is

$$\mathscr{L}(\mu, \rho \mid n, \overline{x}, Q) \propto \rho^{n/2} e^{-\rho(Q/2)\left(1+(n/Q)(\mu-\overline{x})^2\right)}. \tag{11}$$

Taking the *a priori* noninformative distribution $p(\mu, \rho) = \mathrm{d}\mu\, \mathrm{d}\rho/\rho$ [27], the *a posteriori* probability density function (PDF) is [26]

$$P_n(\mu, \rho \mid n, \overline{x}, Q) = c\rho^{(n/2)-1} e^{-\rho(Q/2)\left(1+(n/Q)(\mu-\overline{x})^2\right)}, \tag{12}$$

where

$$c = \frac{Q^{n-1/2}\sqrt{n}}{2^{n/2}\sqrt{\pi}\Gamma(((n-1)/2))}, \tag{13}$$

and c is calculated such that the integral over equation (12) is 1. The gradient is given by the partial derivatives of $P_n(\mu, \rho)$(henceforth, the we write the PDF $P_n(\mu, \rho| n, \overline{x}, Q)$ as $P_n(\mu, \rho)$) lead to the maximum $P(\mu^*, \rho^*)$:

$$\mu^* = \overline{x},$$
$$\rho^* = \frac{n-2}{Q}. \tag{14}$$

Figure 1 shows an example of the FBST evaluation over H_0: $\mu = 0$. The bell-shaped surface is $P_n(\mu, \rho)$ and the solid black line is the restriction of the null hypothesis $(\mu = 0)$. The maximum value of the black line delimits the "tangent" T^*set, represented as a dash-dot line. The dotted line is the restriction $P_n(\mu, \rho = \rho^*)$.

The evidence against the null hypothesis $(H_0$: $\mu = \eta = 0)$ is evaluated by equation (9). Main works on the FBST over the distribution of a mean with an unknown variance [26, 28, 29] use MCMC to solve the integral of $f_n$in equation (9). However, specifically for equation (12), it shows that the "tangent" set T^* has extreme points ρ_A, ρ_B, $\rho_C e$ and ρ_D(as in Figure 2), where

$$\rho_A = \frac{n-2}{Q\left(1+(n/Q)(\eta-\overline{x})^2\right)},$$
$$\rho_B = \frac{n-2}{Q\left(1+(n/Q)(2\overline{x}-\eta-\overline{x})^2\right)} = \rho_A. \tag{15}$$

Making $P_n(\overline{x}, \rho) = P_n(\eta, \rho_A)$ for equation (12) results in

$$c\rho^{(n/2)-1} e^{-\rho(Q/2)} = c\rho_A^{(n/2)-1} e^{-\rho_A(Q/2)\left(1+((n/Q))(\eta-\overline{x})^2\right)}, \tag{16}$$

and grouping variables and taking the natural logarithm in both sides yields

$$\rho - \left(\frac{n-2}{Q}\right)\log(\rho) + \left(\frac{n-2}{Q}\right)\log(\rho_A) - \rho_A\left(1 + \frac{n}{Q}(\eta-\overline{x})^2\right) = 0, \tag{17}$$

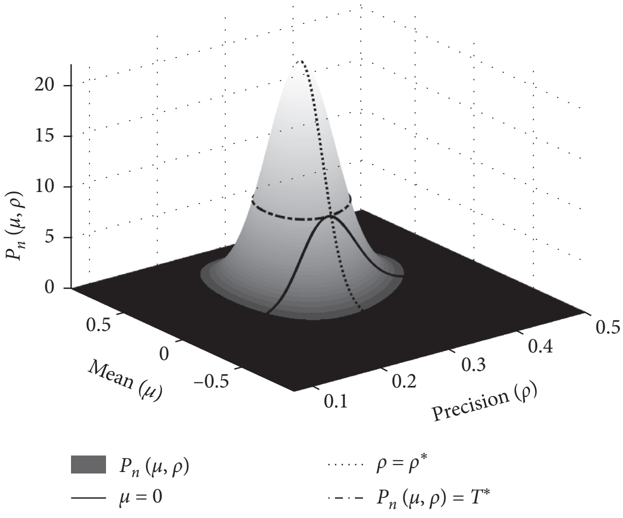


Figure 1: Probability density function $P_n(\mu, \rho \mid n, \overline{x}, Q)$showing the restriction of the null hypothesis and the “tangent” set. The bell shape is $P_n(\mu,\rho)$, while the solid black line is the restriction of the null hypothesis H_0: $\mu = 0$. The maximum value of the black line delimits the “tangent” set T^* (dash-dot line). The dotted line is the restriction $P_n(\mu, \rho = \rho*)$.

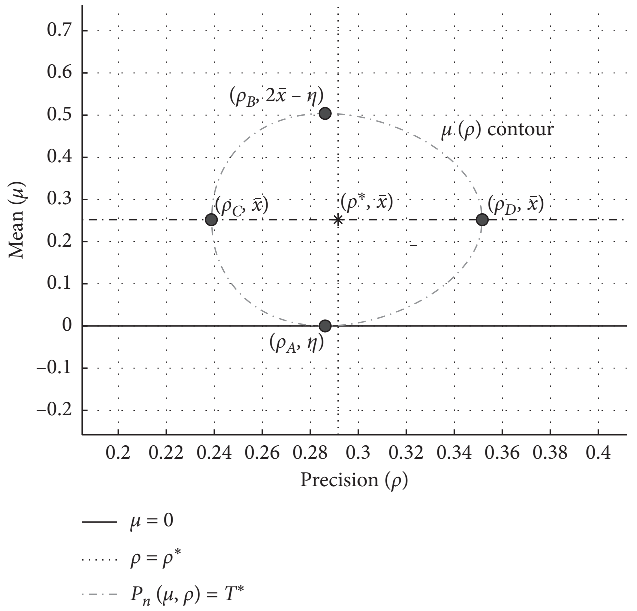


Figure 2: Boundary points of the “tangent” set T^*. The solid black line is the restriction of the null hypothesis H_0: $\mu = 0$. The tangential set T^* is represented by the dash-dot line. The dotted line is the restriction $P_n(\mu, \rho = \rho^*)$.

with the roots being

$$\rho_D = \exp\left[-W_{-1}\left(\frac{e^{-(\beta/\alpha)}}{\alpha}\right) + \frac{\beta}{\alpha}\right],$$

$$\rho_C = \exp\left[-W_0\left(\frac{e^{-(\beta/\alpha)}}{\alpha}\right) + \frac{\beta}{\alpha}\right], \tag{18}$$

where $W_n(\cdot)$ is the Lambert-W function [30]. By the symmetry of T^* over the μ- axis, we can compute the evidence $\overline{\text{ev}}$ by

$$\overline{\text{ev}} = 2\int_{\rho_C}^{\rho_D}\int_{\overline{x}}^{\mu(\rho)} c\rho^{(n/2)-1} e^{-\rho(Q/2)\left(1+(n/Q)(\mu-\overline{x})^2\right)} \mathrm{d}\mu\, \mathrm{d}\rho, \tag{19}$$

where $\mu(\rho)$ is the contour function (from any boundary) on the μ-axis of the “tangent” set T^*(Figure 2). The contour of T^* can be defined as

$$P_n(\mu,\rho) = c\rho^{(n/2)-1} e^{-\rho(Q/2)\left(1+(n/Q)(\mu-\overline{x})^2\right)} = P_n^*$$

$$\downarrow \log()$$

$$\mu^2 - 2\overline{x}\mu + \overline{x}^2 + \frac{Q}{n} - \frac{2}{\rho n}\left[\frac{n-2}{2}\log(\rho) - \log\left(\frac{P_n^*}{c}\right)\right] = 0, \tag{20}$$

where $P_n^* = c\rho_A^{(n/2)-1} e^{-\rho_A(Q/2)(1+(n/Q)(\eta-\overline{x})^2)}$. The roots of equation (20) in μ define the left and right sides of the contour (see Figure 2):

$$\mu(\rho) = \overline{x} \pm \sqrt{\frac{2}{\rho n}\left[\frac{n-2}{2}\log(\rho) - \frac{P_n^*}{c}\right] - \frac{Q}{n}}. \tag{21}$$

Note that $\mu(\rho)$ is a contour for values greater and less than $\overline{x}$. By symmetry, we compute equation (19) as

$$\overline{\text{ev}} = 2c\int_{\rho_C}^{\rho_D} \rho^{(n/2)-1} e^{-\rho(Q/2)}\left[\int_{\overline{x}}^{\mu(\rho)} e^{-(n\rho/2)(\mu-\overline{x})^2} \mathrm{d}\mu\right]\mathrm{d}\rho \longrightarrow \overline{\text{ev}}$$

$$= 2c\int_{\rho_C}^{\rho_D} \rho^{(n/2)-1} e^{-\rho(Q/2)} \sqrt{\frac{2\pi}{n\rho}}\left[\operatorname{erf}\left(\sqrt{\frac{n\rho}{2}}(\mu(\rho)-\overline{x})\right)\right.$$

$$\left. - \cancelto{0}{\operatorname{erf}\left(\sqrt{\frac{n\rho}{2}}(\overline{x}-\overline{x})\right)}\right]\mathrm{d}\rho, \tag{22}$$

where $\operatorname{erf}(\cdot)$ is the error function. We can simplify the argument of this function as

$$\nu(\rho,\eta) = \frac{n-2}{2}\log\left(\frac{\rho}{\rho_A}\right) - \frac{Q}{2}(\rho-\rho_A) + \frac{\rho_A n}{2}(\eta-\overline{x})^2, \tag{23}$$

where ρ_A is the inferior limit of ρ, and η is the hypothesis test H_0: $\mu = \eta$. Thus, we can rewrite equation (19) as the one-dimensional integral:

$$\overline{\text{ev}} = \frac{1}{\Gamma((n-1/2))}\left(\frac{Q}{2}\right)^{(n-1/2)} \times \int_{\rho_C}^{\rho_D} \rho^{(n-3/2)} e^{-\rho(Q/2)}$$

$$\cdot \operatorname{erf}\left(\sqrt{\nu(\rho,\eta)}\right)\mathrm{d}\rho. \tag{24}$$

The integral in equation (24) does not need MCMC techniques, thus demanding less computational effort than equation (9) does.

2.3. Proposed Method. This section proposes a method to compute the evidence interval with a Bayesian evidence level α, which can be computed using equation (24). The result in the GMM-UBM scenario is the sample mean $\mathrm{LLR}(\overrightarrow{x}_Q)$ of the time series $\mathrm{LLR}(x[t]_Q)$, as equation (3) shows, on the parametric space η.

Consider the time series $\mathrm{LLR}(x[t]_Q)$ with a parametric mean (expected value) of μ, precision ρ, and sample mean $\mathrm{LLR}(\overrightarrow{x}_Q)$. From this, it is possible to define the evidence interval of μ as the subspace $\eta_L \leq \mu \leq \eta_H$, where η_L and η_H are values above and below $\mathrm{LLR}(\overrightarrow{x}_Q)$, respectively. The Bayesian evidence $\overline{\mathrm{ev}}$ against the precise hypotheses H: $\mu = \eta_L$ and H: $\mu = \eta_H$ is $1 - \alpha$ (see equation (24)).

Outside this range of the LLR, $\eta_L \leq \mathrm{LLR}(\overrightarrow{x}_Q) \leq \eta_H$, the evidence ($e$-value computed by the FBST) that the parametric mean (μ) is higher than η_H or lower than $\eta_L$is less than α.

We are aware that the definition above does not fit the traditional confidence (or credibility) interval as defined in [31]. However, it is an analytical method based on the parameter space and represents the limits of evidence that the sample can provide Bayesian evidence ("significance") of $1 - \alpha$.

For example, consider that the comparison between a questioned-voice and a known-voice generates a time series $\mathrm{LLR}(x_Q[t])$, where the values of frames $\overrightarrow{x}_Q = \{x_Q[0], x_Q[1], \ldots, x_Q[T-1]\}$ in equation (2) are used. Figure 3 shows the statistical distribution of these LLR values on the normalized histogram (Norm. Hist.) in the left panel. In this panel, the solid light gray line is the empirical PDF (emp. PDF) and the small circle over this curve indicates the sample mean ($\mathrm{LLR}(\overrightarrow{x}_Q)$). The dash-dotted rectangle on the left graph is the region on the right graph. The sample mean of the $\mathrm{LLR}(x_Q[t])$ series is $\mathrm{LLR}(\overrightarrow{x}_Q) \approx -0.8$ Np (nepers) (neper is the natural logarithm of ratios, named after John Napier).

The evaluation of the hypothesis H: $\mathrm{LLR}(\overrightarrow{x}_Q) = \eta$ along the variable η in the LLR space with the FBST (equation (24) yields the *e-value curve*. The variation of η values results in the *e-value curve* (ev-curve, solid dark gray) indicated in the right graph of Figure 3. This curve is computed by sampling the η space and solving equation (24) for each sample. On this graph, the horizontal dash-dotted line ($\mathrm{ev} = 0.05$) indicates the Bayesian evidence (significance) $\alpha = 0.05$ (evidence value against hypothesis $\overline{\mathrm{ev}} = 95\%$ or e-value $= 0.05$). The horizontal solid black error bar ($\mathrm{ev} > 0.05$) indicates the evidence interval and the sample mean.

3. Comparison with Other Methods

This section presents an experiment and a case study involving the range of evidence. We conducted training and testing stage with a voice data set CEFALA-1 [32], containing 104 speakers (55 men and 49 women) recorded with five microphones (generating 520 records). The validation step used 50 recordings that do not belong to the corpus CEFALA-1. This validation emulates an open-set database in speaker comparison.

We designed an experiment to compare the proposed interval inference method with other methods used in FSC. The experiment used 104 voices narrowband filtered (4th order butterworth) in the 300–3500 Hz range and resampled to 8 kHz, compatible with the Brazilian mobile phone system.

In order to compare the various interval inference methods, we need to use the speech database to define the known-voice and questioned-voice sets. We do this as follows. For each subject 50% of voice content was used as known-voice and 50% as questioned-voice, both in the CEFALA-1 corpus and in the validation recordings.

In order to emulate forensic conditions, both the known-voice and questioned-voice data are subject to 3 types of degradation. First, the data are contaminated with pink noise at the following SNR levels: 25 dB, 23 dB, 20 dB, 17 dB, 15 dB, and 12 dB. Next, the data are encoded and then decoded by a GSM 06.60 codec [33]. Finally, the data are run through a narrowband filter (300–3500 Hz).

The features were extracted with MFCC ($c[n]$) using 13 critical bands (filters), a frame length of 25 ms, and frame step of 10 ms. The features include delta $\Delta c[n]$ and delta-delta $\Delta^2 c[n]$. We used Sonh's [34] method for voice activity detection (VAD) to identify the voiced frames.

The methods used to compute interval inference (significance $\alpha = 0.05$) were

Gosset: confidence interval computed by equation (4)

Morrison: empirical credibility interval computed by combining the k-nearest neighborhood (KNN) with the linear regression, as described by Morrison [14]

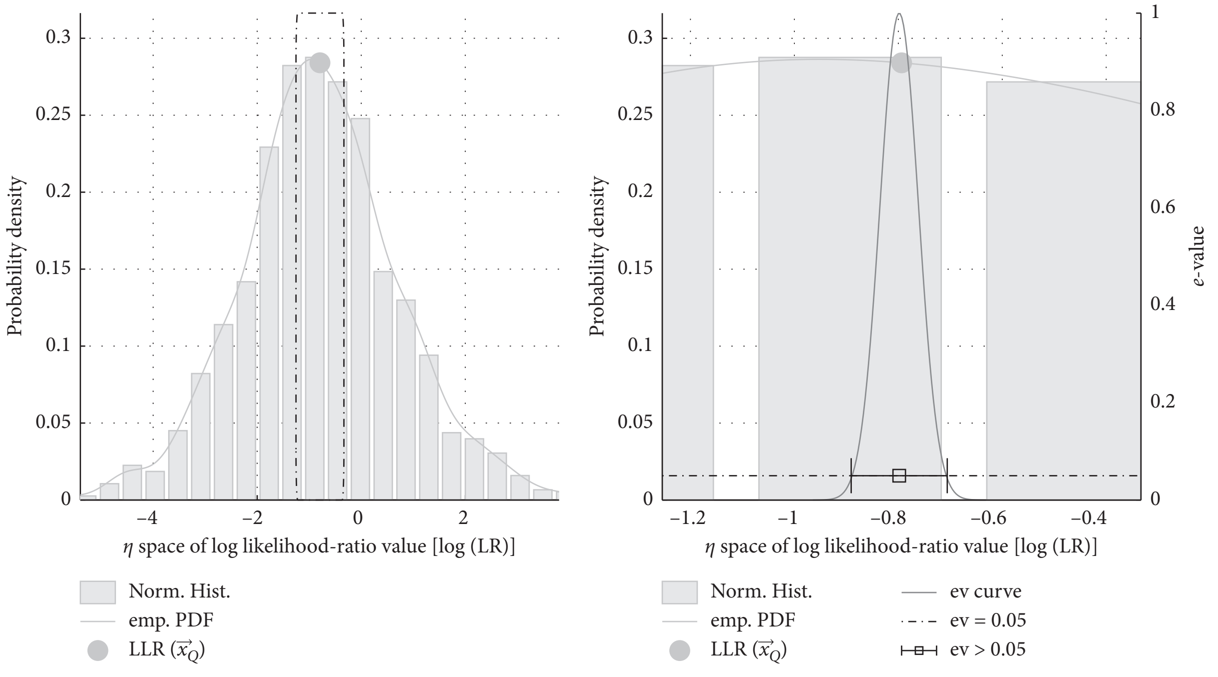
FBST: the proposed method that computes the evidence as a subspace of the parametric space, where the e-value is α

Figure 4 presents examples of the interval inference. In the figure, we show the LLR values along the horizontal axis. The inference intervals are shown as horizontal lines, with the circles indicating the mean values and dot-dashed vertical line indicating the decision threshold.

The horizontal light gray line indicates a same-speaker comparison, and dark gray indicates a different-speaker comparison. The scenarios are (a) correct comparison, (b) an intermediate region where the comparison threshold is within the inference interval, and (c) comparison error (Type I or Type II).

We used the method proposed by Morrison et al. [14] to compute the credibility interval over the data themselves, not over the mean (expected value) of the data. Morrison's method was adapted to compute the mean of 50 subsamples with replacement (similar to bootstrap [10]).

We evaluated the performance of each interval inference method based on results presented in Figure 4. We expected that a comparison between the GMM model of a given speaker and a set of features coming from that speaker (same-speaker comparison hereafter) results in a higher LLR value than a comparison between that same GMM model and a set of features coming from a different speaker (different-speaker comparison hereafter). The training and



(a)

(b)

Figure 3: The panel on the left shows the normalized histogram of LLR ($x_Q[t]$) occurrences and the empirical PDF (thick line). On the same panel, the dashed rectangle indicates the region shown in the panel on the right. In this panel, the bell-shaped dark gray solid line indicates the *e-value curve* and the horizontal solid black error bar indicates the evidence interval and the sample mean.

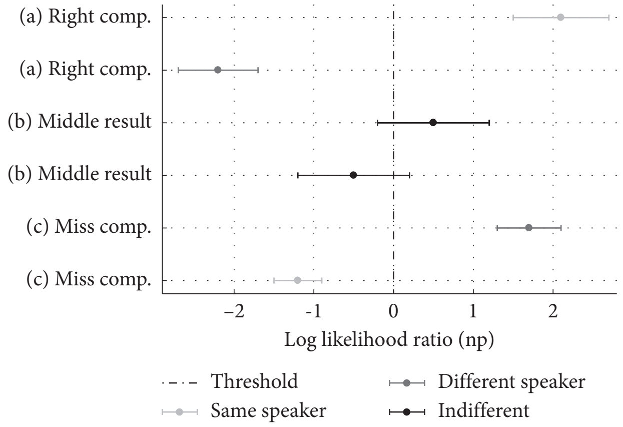


Figure 4: Scenarios of the inference intervals. The dot-dashed vertical line indicates the decision threshold. The light gray line indicates a same-speaker comparison, and dark gray indicates a different-speaker. The scenarios are (a) correct comparison ou “Right comp,” (b) an intermediate or “middle result” region where the comparison threshold is within the inference interval, and (c) comparison error (Type I or Type II) ou “Miss comp.”

testing stage, using only samples from the CEFALA-1 corpus with contaminations between 12 and 25 dB, presented an equal error rate (EER) of 8.1% with threshold at LLR = 0.25 Np. The results presented below cover the test and validation steps.

Figure 5 shows the number of correct classifications in scenario (a). The occurrences of correct classifications for the evidence interval (vertical light gray bar) is smaller than that of other methods (interval and punctual). The comparisons of the best interval methods yield values of 84.0% against 84.4% for SNR 12 dB, 84.4% against 88.6% for 15 dB, and less than 1% for the other SNR values. These values represent a loss of the accuracy of less than 0.5% compared to interval inference. Compared to the punctual inference, the loss in the accuracy is less than 1.6% for the other SNR values.

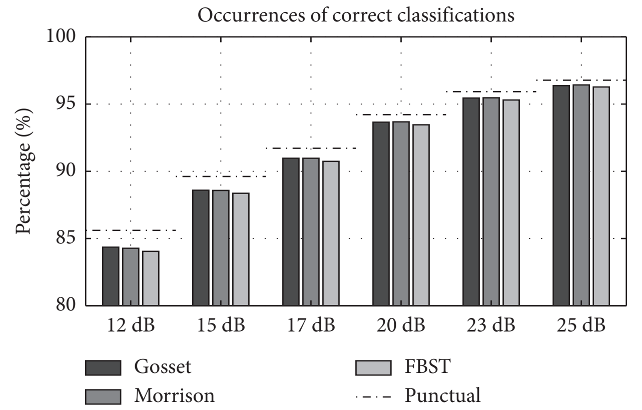


Figure 5: Percentage of correct classifications, scenario (a), for each method at different SNR values.

The intermediate results, in which the intervals overlap, are exemplified in Figure 4 by comparisons (b). These scenarios are deemed inconclusive and represent an *In dubio pro reo* condition, meaning that a defendant should not be convicted when doubts remain about his or her guilt (association between questioned- and known-voices).

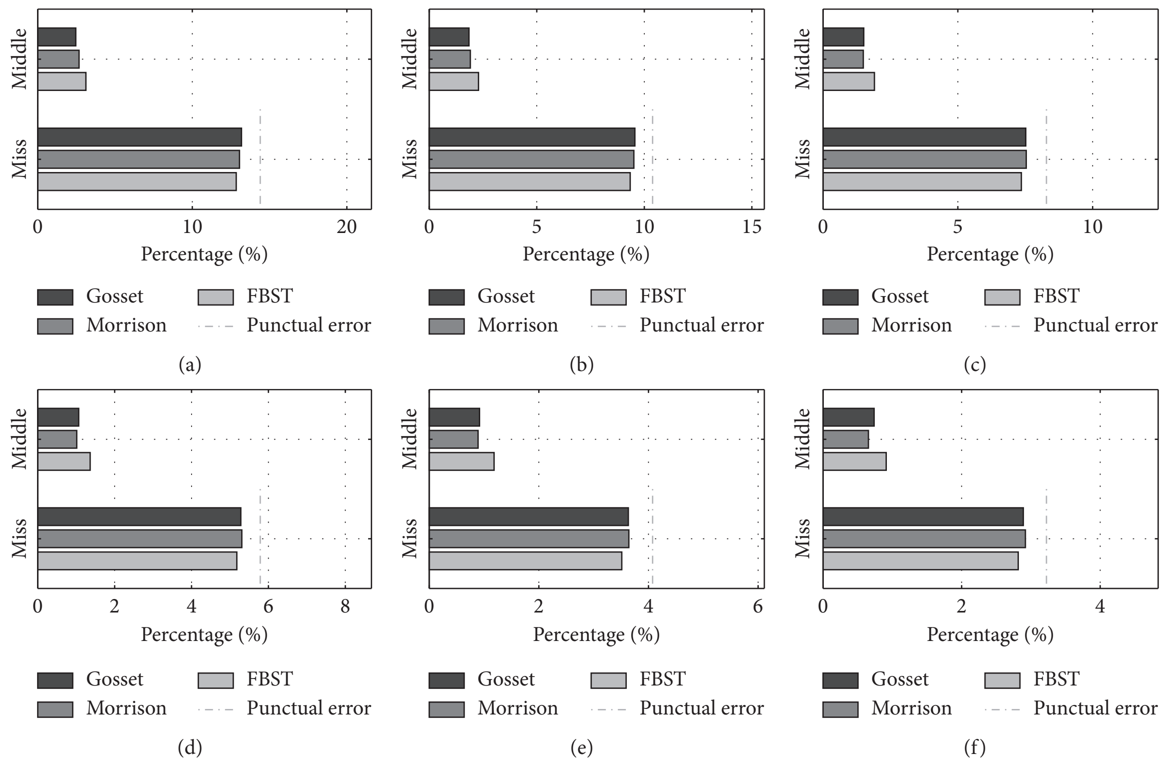


Figure 6: Results from the interval inference method. Each chart—grouped by SNR—shows the percentage of scenarios described in Figure 4. The horizontal bars indicate the percentages by the method, while the dark gray-dashed vertical line indicates the error percentage of the inference without an interval. (a) SNR 12 dB. (b) SNR 15 dB. (c) SNR 17 dB. (d) SNR 20 dB. (e) SNR 23 dB. (f) SNR 25 dB.

In the punctual inference, scenario (b) does not occur, and there is no transition region. Thus, in the interval inference, scenarios (a) and (c) are decisive, and the intermediate scenario, (b), indicates that the results have some equivalence; that is, there is a chance that the comparison between different speakers will be larger (or smaller) than the comparison between the same speakers.

Figure 6 shows the comparison results for various interval inference methods (Gosset, Morrison's method, and FBST). The results are grouped by the SNR level. The panel indicates the percentage of inconclusive interval inferences (b), wrong interval inferences (c), and punctual error inferences (dashed vertical line).

Compared to the punctual inference (dashed vertical line), the evidence interval computed by the FBST (horizontal light gray bar) reduces the number of wrong inferences in 1.6%, 1.1%, 0.9%, 0.7%, 0.6%, and 0.4%, respectively, for SNRs from 12 dB to 25 dB (see Figure 6). Compared to the other methods of the interval inference, the evidence interval (horizontal light gray bar) presents an incorrect number of inferences (c) less than or equal to the other methods (horizontal bars).

These results can be explained by checking the size of the intervals for each method in Figure 7. In this figure, points represent the raw data (jittered horizontally), the horizontal line shows the sample mean, and the lateral lines represent a smoothed density. Table 1 summarizes the values contained in Figures 5, 6, and 7.

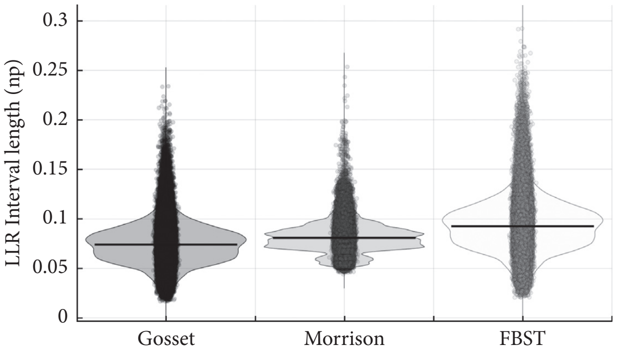


Figure 7: RDI (raw description and inference) of the interval length for different computational methods. Each column points represent the raw data (jittered horizontally), the horizontal line shows the sample mean, and the lateral lines represent a smoothed density.

On an average, the length of the evidence interval (computed by the FBST) is 24% larger than the interval calculated by the Gosset method and 15% larger than the interval calculated by Morrison's method (see Table 1). They also present a higher dispersion than the other methods do.

Another attempt to measure the influence of interval inference is to exclude from the confusion matrix the comparisons that result in scenario (b) of Figure 4. In this way, a fifth category, "In dubio pro reo," may be included. The Table 2 presents a comparison of how the inclusion of

Table 1: Occurrence of the scenarios described in Figure 4 for interval inference methods.

		Percent of occurrences ±0.1(%)			
Scenario	SNR (dB)	Gosset	Morrison	FBST	Punctual
(a) Correct classification	12	84.4	84.3	84.0	85.6
	15	88.6	88.6	88.4	89.6
	17	91.0	91.0	90.7	91.7
	20	93.7	93.7	93.5	94.2
	23	95.4	95.5	95.3	95.9
	25	96.4	96.4	96.3	96.8
(b) Middle interval	12	2.4	2.6	3.2	
	15	1.8	1.9	2.3	
	17	1.5	1.5	1.9	
	20	1.0	1.0	1.4	
	23	0.9	0.9	1.2	
	25	0.7	0.7	0.9	
(c) Miss classification	12	13.2	13.1	12.8	14.4
	15	9.6	9.5	9.3	10.4
	17	7.5	7.5	7.4	8.3
	20	5.3	5.3	5.1	5.8
	23	3.7	3.6	3.5	4.1
	25	2.9	2.9	2.8	3.2
Average interval length (Np)	0.074	0.080	0.092		

Table 2: Percentage of true positives, true negatives, false positives, and false negatives obtained in the test and validation steps.

		True positive (%)	True negative (%)	False positive (%)	False negative (%)	*In dubio pro reo* (%)
Train	Punctual	91.9	91.9	8.1	8.1	
	Gosset	91.1	90.8	7.0	7.4	1.8
	Morrison	91.2	90.9	7.1	7.5	1.7
	FBST	90.8	90.0	6.79	7.12	2.4
Validation	Punctual	85.3	90.6	9.4	14.7	
	Gosset	85.0	90.0	8.8	14.0	1.1
	Morrison	85.0	90.0	8.8	13.9	1.2
	FBST	85.0	89.9	8.7	13.8	1.3

the interval inference, when including "In dubio pro reo," changes the percentage of true positives, true negatives, false positives, and false negatives. The table shows the EER calibration of 8.1%. However, for open-set validation, the GMM-UBM methodology presents false positive rates of 9.4%, which reduces to 8.4% using the range of evidence calculated from the FBST.

4. Conclusion and Future Work

This paper presented an improvement to the FBST calculation for the distribution of a mean with an unknown variance. These improvements obviate the need for MCMC techniques to calculate the FBST integral. Compared with other methods, the evidence interval was more conservative, reducing incrementally Type I and Type II errors in low-SNR scenarios.

Although the results do not present a significant improvement in the reduction of the false positive rate, for open sets, the present work helps to understand the limits of the GMM-UBM methodology applied to FSC. The contribution of the range of evidence may seem insignificant. However, in the case of sex crimes, especially against children, understanding the limits of each tool in the FSC helps the forensic expert to make more informed decisions.

Possible developments of the present work include improving the FBST for the Behrens–Fisher problem, combining the evidence interval with background database calibration and tests with different features such as Power Normalized Cepstral Coefficients (PNCC), Perceptual Linear Predictive (PLP), and noise. The application of the interval inference in speaker verification techniques, such as *i*-vector and *x*-vector, are under development and should be discussed in future work.

Data Availability

The audio files (corpus) used in the experiments can be found at http://www.cefala.org. It is the intention of the authors to make available the processed data and the algorithms as soon as the work is published. Basically the data are acoustic features (Mel-frequency cepstrum) and Gaussian mixture models.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank N. Phillips for always sharing codes. The authors thank the SPAV team of the Institute of Criminalistics of Minas Gerais and all colleagues and teachers of CEFALA for their practical contributions. This work was carried out with the financial support of the Centro Universitário Newton Paiva.

References

[1] P. Rose, *Forensic Speaker Identification*, CRC Press, Boca Raton, FL, USA, 2003.

[2] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Science & Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[3] M. J. Saks and J. J. Koehler, "The individualization fallacy in forensic science evidence," *Vanderbilt Law Review*, vol. 61, no. 1, pp. 199–219, 2008.

[4] G. S. Morrison, "Forensic voice comparison," in *Expert Evidence*, I. Freckelton and H. Selby, Eds., p. 106, Thomson Reuters, Toronto, Canada, 2010.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[6] G. S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM)," *Speech Communication*, vol. 53, no. 2, pp. 242–256, 2011.

[7] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: a tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[8] G. Casella and R. L. Berger, *Statistical Inference*, Vol. 2, Duxbury Press, Pacific Grove, CA, USA, 2002.

[9] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004.

[10] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, CRC Press, Boca Raton, FL, USA, 1994.

[11] W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. J. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proceedings of the (ICASSP'05) IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, March 2005.

[12] S. Koval and A. Lokhanova, "Confidence bounds curves as a tool for evaluation of automatic speaker recognition results uncertainty," in *Proceedings of the 14th International Conference on Speech and Computer. SPECOM*, pp. 284–289, Athens, Greece, September 2011.

[13] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, vol. 10, pp. 61–74, MIT Press, Cambridge, MA, USA, 1999.

[14] G. S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system," *Forensic Science International*, vol. 208, no. 1–3, pp. 59–65, 2011.

[15] E. Gold and P. French, "International practices in forensic speaker comparison," *International Journal of Speech Language and the Law*, vol. 18, no. 2, pp. 293–307, 2011.

[16] R. Togneri and D. Pullella, "An overview of speaker identification: accuracy and robustness issues," *IEEE Circuits And Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[17] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: a systematic review," *Expert Systems With Applications*, vol. 90, pp. 250–271, 2017.

[18] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[19] M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, 1956.

[20] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, p. 1, 1908.

[21] S. L. Zabell, "On student's 1908 article "the probable error of a mean"," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 1–7, 2008.

[22] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*, vol. 51, no. 3, pp. 91–98, 2011.

[23] C. De Bragança Pereira and J. Stern, "Evidence and credibility: full bayesian significance test for precise hypotheses," *Entropy*, vol. 1, no. 4, pp. 99–110, 1999.

[24] C. A. D. B. Pereira, J. M. Stern, S. Wechsler et al., "Can a significance test be genuinely Bayesian?" *Bayesian Analysis*, vol. 3, no. 1, pp. 79–100, 2008.

[25] C. A. d. B. Pereira, "Full Bayesian significant test (FBST)," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., pp. 551–553, Springer, Berlin, Germany, 2011.

[26] M. R. Madruga, C. A. B. Pereira, and J. M. Stern, "Bayesian evidence test for precise hypotheses," *Journal of Statistical Planning and Inference*, vol. 117, no. 2, pp. 185–198, 2003.

[27] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 186, pp. 453–461, 1946.

[28] N. L. Oliveira, C. A. d. B. Pereira, M. A. Diniz, and A. Polpo, "A discussion on significance indices for contingency tables under small sample sizes," *PLoS One*, vol. 13, no. 8, Article ID e0199102, 2018.

[29] C. C. Assane, B. d. B. Pereira, and C. A. d. B. Pereira, "Model choice in separate families: a comparison between the fbst and the cox test," *Communications in Statistics-Simulation and Computation*, vol. 48, no. 9, pp. 2641–2654, 2019.

[30] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambertw function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[31] W. M. Bolstad, *Introduction to Bayesian Statistics*, John Wiley & Sons, Hoboken, NJ, USA, 2013.

[32] A. F. Neto, A. P. Silva, and H. C. Yehia, "Corpus CEFALA-1: base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia/corpus CEFALA-1: audiovisual database of speakers for biometric, phonetic and phonology studies," *Revista de Estudos da Linguagem*, vol. 27, no. 1, p. 191, 2019.

[33] G. ITU, Gsm full rate speech transcoding, Gsm Rec 6.10, 1991, https://www.etsi.org/deliver/etsi_en/300900_300999/300961/06.00.00_40/en_300961v060000o.pdf.

[34] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.