JASA ARTICLE

Aperiodicity cue for fricative voicing contrasts

R. Chottin,[1] D. Demolin,[2,3] X. Pelorson,[1] and A. Van Hirtum[1,a)]

[1]*National Centre for Scientific Research (CNRS), Grenoble INP, Laboratory of Geophysical and Industrial Flows (LEGI), University Grenoble-Alpes, Grenoble, France*

[2]*Université Sorbonne-Nouvelle, National Centre for Scientific Research (CNRS), The Laboratory of Phonetics and Phonology (LPP), Paris, France*

[3]*Rhodes University, Makhanda, South Africa*

ABSTRACT:

This study examines acoustic cues to voicing contrasts in French consonant–vowel–consonant–vowel sequences containing fricatives. We introduce an aperiodicity metric derived from the YIN difference function and evaluate its ability to capture voicing properties in controlled, noise-free speech materials. Results show that the proposed measure enables reliable, objective classification of fricatives as voiced or unvoiced. When combined with a conventional acoustic feature (root mean square amplitude), the metric provides complementary information and supports accurate classification (approximately 83% overall) among unvoiced fricatives, voiced fricatives, and vowels using standard classifiers. These results demonstrate that aperiodicity-based cues offer a valuable addition to established spectral and amplitude-based features for detailed phoneme characterization.

 https://doi.org/10.1121/10.0044096

(Received 9 December 2025; revised 11 May 2026; accepted 20 May 2026; published online 4 June 2026)

[Editor: Benjamin V Tucker] Pages: 4954–4961

I. INTRODUCTION

Fricatives are produced by creating a narrow constriction in the vocal tract that generates turbulent airflow. When this turbulence is accompanied by periodic vocal fold vibration characterized by a fundamental frequency (f_0) typically below 500 Hz, the segment is typically classified as a voiced fricative; in the absence of such vibration, it is voiceless. Voicing contrasts in fricatives present a persistent challenge for acoustic analysis because turbulent noise often masks the low-amplitude periodic energy, and cues to voicing are weaker, more variable, and more context dependent than in sonorants, such as vowels (V).[1–4]

Existing classification methods therefore rely on multiple acoustic cues, including spectral moments, spectral peak locations, amplitude measures, noise duration, and temporal variability.[2–12] These cues are typically interpreted with respect to articulatory and aerodynamic mechanisms. For example, the frequency of the first spectral peak is strongly associated with sibilance and reflects the degree of anterior tongue constriction and resulting aerodynamic jet formation,[1,13,14] whereas spectral moments and other broad-spectrum measures are linked to the place of articulation.[4] Many of these cues therefore serve only as indirect correlates of voicing.

Voicing in fricatives is primarily signaled by fundamental frequency (f_0), segmental duration, and low-frequency energy. Voiceless fricatives are typically tens of milliseconds (or roughly 1.5–2× longer than their voiced counterparts), whereas low-frequency root mean square (RMS) amplitude below 500 Hz is generally 2–6 dB higher (approximately 2–3× greater in linear amplitude) in voiced segments.[4,6,7,11,15–20] In addition to duration and RMS amplitude, low-frequency spectral cues provide further information about fricative voicing.[4,6,17,18] The spectral centroid, or center of gravity (COG), computed over the low-frequency range, reflects the distribution of energy associated with the periodic component and is typically higher in voiced fricatives (VFr) Measures of harmonic structure also contribute to voicing characterization. Cepstral peak prominence (CPP) quantifies the strength of harmonic structure embedded in the broadband frication noise and is consistently higher in VFr. Relatedly, spectral harmonicity measures, such as the harmonics-to-noise ratio (HNR) provide complementary indices of periodic energy, though they are generally less robust than CPP in fricative segments.[18]

The fundamental frequency (f_0) is the primary acoustic correlate of voicing. Therefore, accurate voicing analysis depends on reliable f_0 estimation. The YIN algorithm[21] is a nonparametric time-domain f_0 estimator that is computationally efficient and suitable for quasi–real-time applications, with implementations available, for example, in the PYTHON librosa library.[21–24] Unlike traditional autocorrelation-based methods, YIN estimates the degree of periodicity, directly rather than relying on the autocorrelation maximum, thereby reducing errors, such as octave or subharmonic jumps, and improving robustness in signals with moderate noise or spectral complexity.[21,25,26] Nevertheless, like other pitch estimation methods, YIN requires quasi-periodic structure and may exhibit reduced accuracy in signals with irregular phonation or low signal-to-noise ratio.[26,27] Empirical evaluations further show that YIN maintains competitive accuracy at moderate-to-high signal-to-noise ratios (SNRs)

a)Email: annemie.vanhirtum@univ-grenoble-alpes.fr

https://doi.org/10.1121/10.0044096

(≥ 10 dB) across a range of noise conditions (e.g., babble, street, car, factory, restaurant, and white noise), but performance degrades below approximately 0–5 dB SNRs, consistent with its reliance on quasi-periodic structure.[25,26]

Because YIN provides an explicit estimate of signal periodicity in the time domain, it can be interpreted as a time-domain counterpart to spectral or cepstral measures of harmonic organization, such as cepstral peak prominence (CPP). Motivated by this parallel, the present study examines a direct acoustic measure of aperiodicity derived from the difference function underlying the YIN fundamental frequency (f_0) estimator.[21] This measure estimates the relative contributions of periodic and noise-like components in the temporal signal, making it potentially well suited to detecting voicing in fricatives. An additional practical advantage is that the measure is already computed within YIN-based estimators[21] and therefore requires no additional computational cost.

Section II introduces the YIN-based aperiodicity cue. Its application to voicing-contrast detection is then explored in Sec. III, first under clean French speech conditions (SNR ≥ 10 dB), where the noise-like component in fricatives primarily arises from aperiodic excitation, and then under additive white noise at progressively lower SNR levels down to approximately 0–5 dB. Classification is performed using standard approaches, namely, a Mahalanobis distance classifier (MD) and a linear support vector machine. In Sec. IV, the resulting performance is compared with that obtained using other voicing cues shown to be suitable for voicing-contrast detection. The conclusion is formulated in Sec. V.

II. APERIODICITY CUE

The aperiodicity cue κ_k is defined from the cumulative mean normalized difference function computed from a discrete time-signal s_k, $1 \leq k \leq N$, sampled at frequency f_s at time instants $t_k = k/f_s$, using YIN-based methods, such as β-YIN. β-YIN extends the original autocorrelation-based YIN algorithm[21] by providing reliable detection of stable oscillations, as well as their onset, offset, and absence. Following steps 1–3, κ_k is evaluated on each of N_w overlapping signal segments of length W, where $N_{max} \leq W \leq N$. Segments are indexed by $i = 1, \ldots, N_w$ and shifted by ΔW, such that $N_w = N/\Delta W$. The parameter T_{max} denotes the maximum expected oscillation period and N_{max} is the corresponding number of samples. Unless specified otherwise, $T_{max} = 12.5$ ms, $W = N_{max}$, and $\Delta W = W/4$.

In step 1, the autocorrelation function (ACF) at each time index $k = (i-1)\Delta W + 1$ is computed as

$$r_k(\tau) = \sum_{j=k-W/2}^{k+W/2} s_j s_{j+\tau}, \tag{1}$$

with time lag $\tau \leq 3 \times N_{max}$. For an aperiodic or random signal, the ACF attains its maximum at $\tau = 0$ and gradually decays toward zero as the lag increases, without exhibiting any significant local maxima. In contrast, for a periodic signal with period $T_{0,k}$, the ACF reaches its global maximum at $\tau = 0$ and displays a secondary maximum at $\tau = T_{0,k}$. In the presence of additive noise, the secondary maximum at $\tau = T_{0,k}$ is reduced, allowing the magnitude of nonzero-lag peaks to be used as a measure of signal periodicity, which has been explored for tasks, such as fricative classification. However, this approach is found to be limited:[21,28] (i) the amplitude of local maxima varies widely across signals, preventing meaningful comparisons between signals or speakers, and (ii) the strong peak at zero lag can dominate the ACF, leading to frequent f_0-tracking errors, particularly in noisy or weakly periodic signals.

In step 2, a difference function is defined as

$$d_k(\tau) = \sum_{j=k-W/2}^{k+W/2} (s_j - s_{j+\tau})^2, \tag{2}$$

which, using Eq. (1), can be expressed in terms of autocorrelations: $d_k(\tau) = r_k(0) + r_{k+\tau}(0) - 2r_k(\tau)$. The difference function estimates the period of a periodic signal by exhibiting a local minimum at $\tau = T_{0,k}$. Unlike the ACF, which is sensitive to amplitude variations that can artificially amplify peaks at higher lags ($\tau = n T_{0,k}$ with $n = 1, 2, \ldots$) and can lead to false detections or overestimations of $T_{0,k}$, the difference function reduces this bias. Consequently, it improves the accuracy and robustness of $f_{0,k}$ estimation by mitigating underestimation errors.

In step 3, the difference function [Eq. (2)] is normalized by its cumulative mean. The resulting cumulative mean normalized difference function, $0 \leq d'_k(\tau) \leq 1$, is proportional to the aperiodic/total power ratio and is calculated as

$$d'_k(\tau) = \begin{cases} 1 & \text{if } \tau = 0, \\ d_k(\tau) \Big/ \left[\frac{1}{\tau}\sum_{j=1}^{\tau} d_k(j)\right] & \text{otherwise.} \end{cases} \tag{3}$$

Nonzero time lags that minimize $d'_k(\tau)$ are used to estimate the period $T_{0,k}$, preventing trivial minima near zero lag from being misinterpreted as the period. This improves the robustness of $f_{0,k}$ estimation by reducing overestimation due to imperfect periodicity or strong first-harmonic resonance.

The aperiodicity measure is defined as

$$\kappa_k = \min_{\tau}(d'_k(\tau)), \tag{4}$$

such that $\kappa_k = 1$ for aperiodic signals, while $\kappa_k = 0$ for perfectly periodic ones.

Figure 1 illustrates the aperiodicity cue $\kappa(t)$ for a signal $s(t)$ consisting of a sinusoid ($f_0 = 200$ Hz) with additive standard white Gaussian noise scaled by an exponentially increasing amplitude $|\alpha(t)|$. The corresponding theoretical κ-curve is derived as outlined in the Appendix. The results indicate that $\kappa(t)$ effectively captures the relative degree of aperiodicity: the instantaneous signal-to-noise ratio (SNR_k)

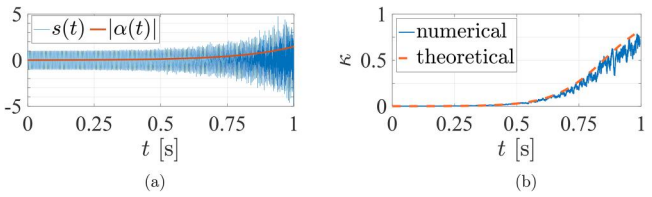

FIG. 1. (a) Signal $s(t)$ consisting of a sinusoid ($f_0 = 200$ Hz) combined with white noise of exponentially increasing amplitude $|\alpha(t)|$, (b) corresponding aperiodicity cue $\kappa(t)$.

decreases from 37 dB at $t = 0$ s ($\kappa = 0$) to 11 dB at $t = 0.8$ s ($\kappa(t) \approx \sigma = 0.4$).

III. APPLICATION AS VOICING-CONTRAST DETECTOR

Because vocal fold vibrations are quasi-periodic under normal conditions, this section considers the use of aperiodicity cue $\kappa(t)$ to detect voicing contrasts in consonant–vowel–consonant–vowel (CVCV) sequences composed of repeated consonant–vowel (CV) patterns (e.g., /sasa/).

A. Data: CVCV sequences

Normalized acoustic signals $s_a(t)$ [Fig. 2(a)] corresponding to 458 CVCV (or C1V1C2V2) sequences, in which consonants (C) and V were repeated ($C1 = C2$ and $V1 = V2$, so that the indices indicate repetition), were analyzed. The data came from five adult native speakers of French (three males and two females[29]) who produced the sequences with fundamental frequencies ranging from 80 to 300 Hz. All participants reported having no speech disorder. Recordings were made in a quiet laboratory environment at $f_s = 44.1$ kHz.

V segments [/a/ (spa), /i/ (see), /u/ (foot)] appeared in all 458 CVCV combinations. C consisted of either unvoiced fricatives (UFr) [239 CVCV; /ʃ/ (shoe), /f/ (foot), /s/ (sun)] or their VFr (219 CVCV) counterparts [/ʒ / (treasure), /v/ (vase), /z/ (zoo)]. Measured mean ± standard deviation SNRs for CVCV segments were 12 ± 5 dB for $C1$, 21 ± 4 dB for $V1$, 13 ± 5 dB for $C2$, and 19 ± 4 dB for $V2$. Thus, mean SNRs were comparable across the consonant ($C1$, $C2$) and vowel ($V1$, $V2$) repetitions, with differences remaining within the respective standard deviations.

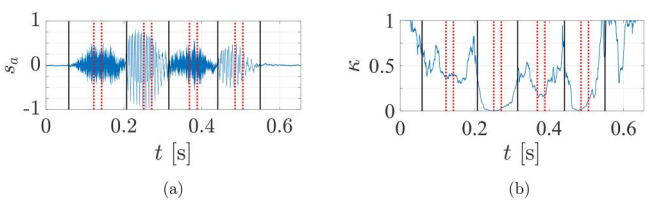

FIG. 2. Illustration of a CVCV sequence /ʃuʃu/. (a) Normalized acoustic signal $s_a(t)$, (b) aperiodicity cue $\kappa(t)$. For each phoneme, vertical lines indicate boundaries (black solid) and central 40 ms analysis interval (red dotted).

B. Phoneme acoustic cues: $\overline{\kappa}$ and A_{rms}

As illustrated in Fig. 2, the aperiodicity cue $\kappa(t)$ was computed for each normalized CVCV signal $s_a(t)$, following the procedure described in Sec. II. Each manually segmented CVCV phoneme was then characterized by its mean aperiodicity $\overline{\kappa}$, obtained by averaging $\kappa(t)$ over a centered 40 ms analysis interval (ΔT_a, with $\Delta T_a > T_{max}$). The analysis window duration (approximately 20%–60% of the segment duration; 20%–40% for UFr), and its midpoint placement were selected to be consistent with established acoustic analysis practices for fricatives.[2,4] Specifically, the window was centered within each phoneme segment to minimize onset and offset boundary effects, thereby capturing the quasi-stationary central portion of the segment.[17] In addition, for each phoneme, the RMS amplitude A_{rms} was computed over the same interval

$$A_{\text{rms}} = \sqrt{\frac{1}{N_a}\sum_{n=1}^{N_a} s_a(n)^2}, \tag{5}$$

where N_a denotes the number of samples corresponding to the interval ΔT_a. Each CVCV phoneme (or $C1$, $V1$, $C2$, $V2$) was thus represented by a two-dimensional acoustic cue vector $\mathbf{c} = [\overline{\kappa}, A_{\text{rms}}]^\top$, comprising mean aperiodicity $\overline{\kappa}$ and RMS amplitude A_{rms}, where the superscript $\top$ denotes the transpose.

The resulting feature space, spanned by the cue vectors $\mathbf{c}$ for $C1$ and $V1$ phonemes (239 UFr, 219 VFr, 458 V), is shown in Fig. 3. A similar pattern is observed for the feature space defined by the cue vectors $\mathbf{c}$ for $C2$ and $V2$ phonemes. This mapping indicates that $\overline{\kappa}$ acts as a voicing identifier distinguishing voiced phonemes (V and VFr; lower $\overline{\kappa} \lesssim 0.2$) from UFr (higher $\overline{\kappa} \gtrsim 0.2$), whereas A_{rms} separates vowels (V; higher $A_{\text{rms}} \gtrsim 0.3$) from fricatives (VFr and UFr; lower $A_{\text{rms}} \lesssim 0.3$). To further illustrate this, Sec. III C evaluates voicing-contrast classification for the three phoneme classes (V, VFr, and UFr) based on the cue vectors $\mathbf{c}$.

C. Voicing-contrast classification: *V*, VFr, and UFr

1. Classifiers and metrics: Mahalanobis and SVM

a. MD. For each phoneme class $q \in \{\text{UFr, VFr, } V\}$, let $\mathcal{C}_q = \left\{\mathbf{c}_1^{(q)}, \ldots, \mathbf{c}_{N_q}^{(q)}\right\}$ denote the set of all cue vectors. For

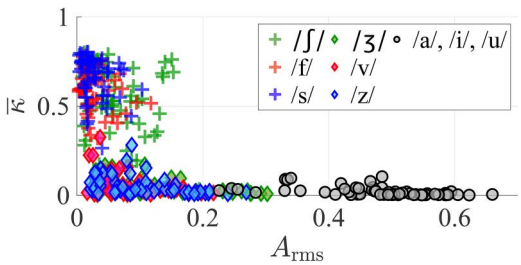

FIG. 3. Phoneme feature space of $C1$ and $V1$ cue vectors $\mathbf{c}$, comprising mean aperiodicity $\overline{\kappa}$ and RMS amplitude A_{rms}.

https://doi.org/10.1121/10.0044096

each bootstrap iteration $b = 1, \ldots, B$, a reference set $\mathcal{R}_q^{(b)} \subset \mathcal{C}_q$ containing $\lceil 0.9 N_q \rceil$ vectors was randomly selected, with the test set defined as $\mathcal{T}_q^{(b)} = \mathcal{C}_q \backslash \mathcal{R}_q^{(b)}$, where $\lceil \cdot \rceil$ denotes the ceiling function. The reference mean vector, $\boldsymbol{\mu}_q^{(b)}$, and covariance matrix, $\boldsymbol{\Sigma}_q^{(b)}$, for each class were then computed from $\mathcal{R}_q^{(b)}$ as

$$\boldsymbol{\mu}_q^{(b)} = \frac{1}{|\mathcal{R}_q^{(b)}|} \sum_{\mathbf{c} \in \mathcal{R}_q^{(b)}} \mathbf{c}, \tag{6}$$

$$\boldsymbol{\Sigma}_q^{(b)} = \frac{1}{|\mathcal{R}_q^{(b)}| - 1} \sum_{\mathbf{c} \in \mathcal{R}_q^{(b)}} (\mathbf{c} - \boldsymbol{\mu}_q^{(b)})(\mathbf{c} - \boldsymbol{\mu}_q^{(b)})^\top. \tag{7}$$

Each test vector $\mathbf{c} \in \mathcal{T}_q^{(b)}$ was classified by assigning it to the class minimizing the Mahalanobis distance d_M:

$$\hat{q}^{(b)} = \arg\min_q \, d_M(\mathbf{c}, \boldsymbol{\mu}_q^{(b)}) \tag{8}$$

with[30]

$$d_M(\mathbf{c}, \boldsymbol{\mu}_q^{(b)}) = \sqrt{(\mathbf{c} - \boldsymbol{\mu}_q^{(b)})^\top (\boldsymbol{\Sigma}_q^{(b)})^{-1} (\mathbf{c} - \boldsymbol{\mu}_q^{(b)})}. \tag{9}$$

b. Support vector machine classifier (SVM). In order to limit the influence of outliers on phoneme classification, a standard soft-margin linear SVM was trained on the subsets $\mathcal{R}_q^{(b)}$, where superscript b denotes again the bootstrap iteration (Sec. III C 1 a).[31,32] The full training set was $\mathcal{R}^{(b)} = \cup_q \mathcal{R}_q^{(b)} = \{(\mathbf{c}_i, q_i)\}_{i=1}^{N_t}$, comprising N_t samples, where $\mathbf{c}_i \in \mathbb{R}^2$ denotes the two-dimensional feature vector and q_i its corresponding class label. A one-vs-all scheme was used, in which each binary classifier distinguished a single target class from all remaining classes. For each binary subproblem, the labels were defined as $z_i \in \{+1, -1\}$, with $z_i = +1$ for samples belonging to the target class q and $z_i = -1$ for all other samples. The corresponding optimization problem was

$$\begin{aligned} \min_{\mathbf{w}_{(b)}, a^{(b)}, \{\xi_i\}_i} \quad & \frac{1}{2} ||\mathbf{w}^{(b)}||^2 + C \sum_{i=1}^{N_t} \xi_i \\ \text{s.t.} \quad & z_i \left(\mathbf{w}^{(b)\top} \mathbf{c}_i + a^{(b)} \right) \geq 1 - \xi_i, \quad i = 1, \ldots, N_t, \\ & \xi_i \geq 0, \end{aligned}$$

where $\mathbf{w}^{(b)} \in \mathbb{R}^2$ and $a^{(b)} \in \mathbb{R}$ define the separating hyperplane, ξ_i are slack variables, and the regularization parameter is $C > 0$. A sample $\mathbf{c}$ from the test set $\mathcal{T}^{(b)} = \cup_q \mathcal{T}_q^{(b)}$ was then assigned a binary prediction by the sign $\hat{z}_q^{(b)}(\mathbf{c}) = \text{sign}(f_q^{(b)}(\mathbf{c})) \in \{+1, -1\}$, with $f_q^{(b)}(\mathbf{c}) = \mathbf{w}_q^{(b)\top} \mathbf{c} + a_q$ being the decision function for the qth binary classifier. This results in a binary vector

$$\hat{\mathbf{z}}^{(b)} = \left(\hat{z}_{\text{UFr}}^{(b)}, \hat{z}_{\text{VFr}}^{(b)}, \hat{z}_{\text{V}}^{(b)} \right),$$

from which $\hat{q}^{(b)}$ was obtained based on the separating hyperplanes.

c. Performance metrics. Classification performance for the MD (Sec. III C 1 a) and SVM (Sec. III C 1 b) classifiers, applied to the cue vectors $\mathbf{c}$ of the three phoneme classes ($q \in \{\text{UFr}, \text{VFr}, V\}$), associated with different datasets (C1V1, C2V2, CVCV), was assessed using up to $B = 150$ bootstrap resamples. For each bootstrap iteration b, a 3×3 confusion matrix $\left\{ \eta_{i,j}^{(b)} \right\}_{i,j=1,\ldots,3}$ was computed, from which class-specific precision $\text{Prec}_q^{(b)}$ and recall $\text{Rec}_q^{(b)}$ for class q were obtained as

$$\text{Prec}_q^{(b)} = \frac{\eta_{q,q}^{(b)}}{\sum_i \eta_{i,q}^{(b)}}, \qquad \text{Rec}_q^{(b)} = \frac{\eta_{q,q}^{(b)}}{\sum_j \eta_{q,j}^{(b)}}.$$

Global performance markers: accuracy (Acc), precision (Prec), and recall (Rec) [$\text{Acc}^{(b)}$, $\text{Prec}^{(b)}$, and $\text{Rec}^{(b)}$] for bootstrap iteration b were then computed as

$$\text{Acc}^{(b)} = \frac{\sum_q \eta_{qq}^{(b)}}{\sum_{i,j} \eta_{ij}^{(b)}},$$

$$\text{Prec}^{(b)} = \frac{1}{3} \sum_q \text{Prec}_q^{(b)},$$

$$\text{Rec}^{(b)} = \frac{1}{3} \sum_q \text{Rec}_q^{(b)}.$$

Convergence of the bootstrap estimates was assessed using the cumulative mean of each performance metric, as shown in Fig. 4 for the global accuracy Acc (SVM, MD) and the class-specific precisions Prec_q (MD). For each metric, the final performance estimate across classes was defined as the cumulative mean after $B = 150$ bootstrap replicates.

2. Classification performance

a. Influence of dataset: C1V1, C2V2, CVCV. Global and class-specific performance metrics for phonemes from datasets C1V1, C2V2, and CVCV are presented in Tables I and II, respectively. Classification performance was robust, with metrics essentially invariant across classifiers (SVM, MD). Moreover, the classifiers performed consistently

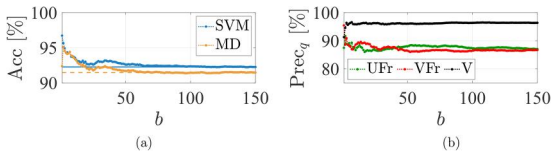

FIG. 4. Cumulative mean across bootstrap iterations $b \in \{1, \ldots, B\}$ (with $B = 150$) for C1V1 data. (a) Global accuracy Acc (SVM, MD), (b) class-specific precisions Prec_q (MD). (a) $\text{Acc}(b)$ (SVM and MD), (b) $\text{Prec}_q(b)$ (MD).

TABLE I. Global performance metrics (accuracy Acc, precision Prec, and recall Rec in %) for the SVM and MD classifiers, computed after $B = 150$ bootstrap iterations for datasets C1V1, C2V2, and CVCV.

Data	C1V1		C2V2		CVCV	
Classifier	SVM	MD	SVM	MD	SVM	MD
Acc	92%	91%	77%	75%	85%	83%
Prec	90%	90%	73%	75%	82%	82%
Rec	91%	90%	75%	76%	83%	83%

across all classes, as global metrics for each dataset (C1V1, C2V2, CVCV) exhibited similar magnitudes. For C1V1, cue vectors $\mathbf{c}$ yielded accurate classification, with all global (Acc, Prec, Rec) and class-specific (Prec_q, Rec_q) metrics near 90%. For C2V2, global performance decreased to $\approx 75\%$, primarily due to VFr misclassifications, with $\text{Prec}_{q=\text{VFr}}$ and $\text{Rec}_{q=\text{VFr}}$ dropping to 30%–70%.

Probability distributions (Pr) of the cue vectors $\mathbf{c}$ $[\Pr(\overline{\kappa}), \Pr(A_{\text{rms}})]$ for phonemes (UFr, VFr, V) in C1V1 and C2V2 (Fig. 5) show that the proposed acoustic cue $\overline{\kappa}$ exhibits nearly overlapping distributions across datasets. In contrast, $\Pr(A_{\text{rms}})$ for V shifts from higher values (peak $\approx$0.5) in C1V1 to lower values (peak ≈ 0.2) in C2V2, increasing inter-class overlap, particularly between VFr and V. These results indicate that $\overline{\kappa}$ is robust to CVCV-signal normalization, whereas normalization affects A_{rms} for vowels, consistent with the qualitative behavior observed in Fig. 2 for a representative CVCV utterance. Consequently, CVCV classification metrics fall within the range observed for C1V1 and C2V2, with global metrics near 83%.

b. Influence of added standard white noise. A noisy signal $s_n(t)$ was generated according to

$$s_n(t) = s_a(t) + \alpha_c n(t), \tag{10}$$

where $s_a(t)$ denotes the normalized acoustic CVCV signal, and $n(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is a white Gaussian noise process. The scalar coefficient α_c controls the noise amplitude and thereby the resulting SNR. Concretely, α_c was varied from 0 to 0.5 in increments of 0.05, yielding progressively lower SNRs, approximately 12–0.2 dB for the $C1$ phonemes and 21–2 dB for $V1$. The effect of additive noise on the proposed aperiodicity cue $\overline{\kappa}$ and the RMS amplitude A_{rms} is shown in

TABLE II. Class-specific performance metrics (precision Prec_q and recall Rec_q in %) for the SVM and MD classifiers, computed after $B = 150$ bootstrap iterations for datasets C1V1, C2V2, and CVCV.

Class q		UFr		VFr		V	
Classifier		SVM	MD	SVM	MD	SVM	MD
C1V1	Prec_q	92%	87%	80%	86%	98%	96%
	Rec_q	88%	91%	87%	80%	97%	98%
C2V2	Prec_q	88%	86%	31%	68%	94%	74%
	Rec_q	84%	86%	57%	50%	79%	89%
CVCV	Prec_q	90%	87%	69%	79%	90%	83%
	Rec_q	87%	89%	69%	62%	91%	95%

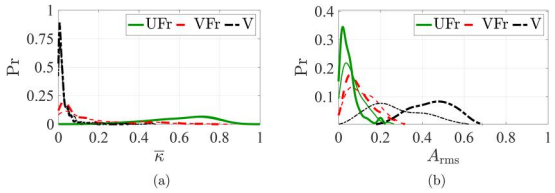

FIG. 5. Pr of cue vectors $\mathbf{c} = [\overline{\kappa}, A_{\text{rms}}]^\top$ for phonemes (UFr, VFr, V) from datasets C1V1 (thick) and C2V2 (thin). (a) $\Pr(\overline{\kappa})$, (b) $\Pr(A_{\text{rms}})$.

Fig. 6, which plots the Pr for the phoneme classes (UFr, VFr, V) in the C1V1 dataset for $\alpha_c \in \{0, 0.05, 0.25, 0.5\}$. The Pr of RMS amplitude A_{rms} for vowels remained largely invariant across tested α_c-values. In contrast, the distributions associated with UFr and VFr substantially overlapped and shifted toward the vowel distribution as the noise amplitude increased. Despite this shift, the plotted $\Pr(A_{\text{rms}})$ curves indicate that A_{rms} continued to separate V from fricatives (UFr, VFr) for all examined noise levels, highlighting its robustness to additive Gaussian noise. This robustness was not observed for the proposed aperiodicity cue $\overline{\kappa}$. As the noise level increased, the probability distribution for VFr shifted from overlapping with V in the noise-free condition to overlapping with UFr, indicating a reduced or lost ability of $\overline{\kappa}$ to differentiate between voiced and unvoiced phoneme classes under noisy conditions.

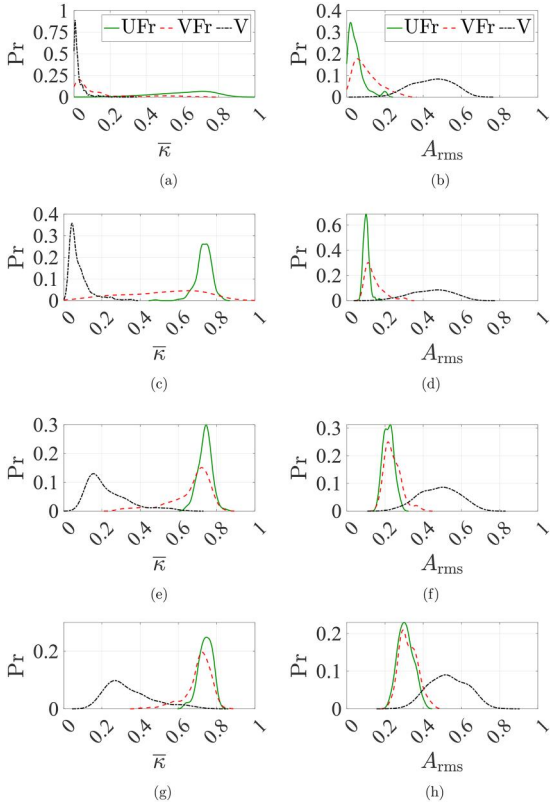

FIG. 6. Influence of added noise amplitude $\alpha_c \in \{0, 0.05, 0.25, 0.5\}$ on Pr of cue vectors $\mathbf{c} = [\overline{\kappa}, A_{\text{rms}}]^\top$ for phonemes (UFr, VFr, V) from C1V1. (a) $\Pr(\overline{\kappa})$, $\alpha_c = 0$, (b) $\Pr(A_{\text{rms}})$, $\alpha_c = 0$, (c) $\Pr(\overline{\kappa})$, $\alpha_c = 0.05$, (d) $\Pr(A_{\text{rms}})$, $\alpha_c = 0.05$, (e) $\Pr(\overline{\kappa})$, $\alpha_c = 0.25$, (f) $\Pr(A_{\text{rms}})$, $\alpha_c = 0.25$, (g) $\Pr(\overline{\kappa})$, $\alpha_c = 0.5$, (h) $\Pr(A_{\text{rms}})$, $\alpha_c = 0.5$.

https://doi.org/10.1121/10.0044096

These probability distribution trends explain the decline in global performance metrics with increasing α_c for both the SVM and MD classifiers, as observed in the accuracy $\mathrm{Acc}(\alpha_c)$ (70%–92%) and precision $\mathrm{Prec}(\alpha_c)$ (65%–90%) shown in Fig. 7. The class-specific precision, $\mathrm{Prec}_q(\alpha_c)$, and recall, $\mathrm{Rec}_q(\alpha_c)$, plotted for the SVM classifier in Fig. 7, show that metrics associated with the V class remain essentially invariant with respect to α_c. The precision of the UFr class was largely preserved, whereas that of the VFr class decreased sharply to below 10%, indicating that most VFr phonemes were misclassified for $\alpha_c > 0$. The recall for both fricative classes (UFr, VFr) decreased (40%–100%) with increasing α_c, demonstrating reduced classifier sensitivity and increased misclassification in noisy conditions.

IV. COMPARISON WITH OTHER ACOUSTIC VOICING CUES

Figure 8 presents the Pr for voicing cues derived from the CVCV dataset other than $\overline{\kappa}$ and A_{rms} (cf. Fig. 5), as introduced in Sec. I. These include segment duration T_{ms}, CPP, low-frequency COG computed up to 500 Hz and 2 kHz, and the HNR. The distributions of T_{ms}, CPP, and COG up to 500 Hz exhibit significant overlap across phoneme classes (UFr, VFr, V). In contrast, COG computed up to 2 kHz and HNR show reduced overlap for the UFr class relative to the voiced classes (VFr, V), a behavior previously observed for $\overline{\kappa}$ [Fig. 5(a)]. These observations suggest that effective voicing-contrast classification may be achieved using cue vectors that combine A_{rms} with either HNR or COG up to 2 kHz, rather than with T_{ms}, CPP, and COG, up to 500 Hz. This expectation is confirmed by the global performance metrics of the SVM and MD classifiers, computed after $B = 150$ bootstrap iterations on the CVCV dataset. The results for cue vectors combining A_{rms} with the proposed aperiodicity cue $\overline{\kappa}$, HNR, and COG up to 2 kHz are reported in Table III. These configurations yield comparable performance levels (approximately 80%–86%). By contrast, combinations involving CPP, COG up to 500 Hz, or T_{ms} result in lower performance (60%–79%) and are therefore omitted from Table III.

Table III indicates that, in addition to the proposed aperiodicity cue $\overline{\kappa}$, both HNR and COG up to 2 kHz are effective features for detecting voicing contrasts under clean

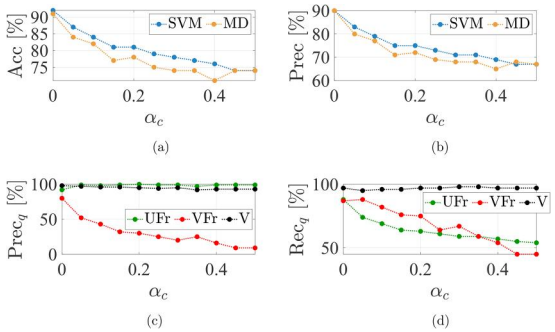

FIG. 7. (a) and (b) Global metrics accuracy $\mathrm{Acc}(\alpha_c)$ and precision $\mathrm{Prec}(\alpha_c)$ for the SVM and MD classifiers. (c) and (d) Class-specific metrics, precision $\mathrm{Prec}_q(\alpha_c)$ and recall $\mathrm{Rec}_q(\alpha_c)$, for the SVM classifier.

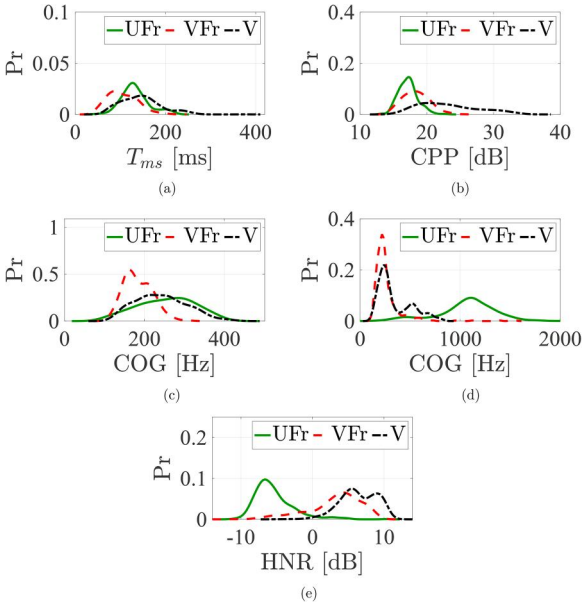

FIG. 8. Pr of voicing cues for phonemes (UFr, VFr, V) from datasets CVCV. (a) Segment duration T_{ms}, (b) CPP, (c) COG up to 500 Hz, (d) COG up to 2 kHz, (e) HNR.

speech conditions. The effect of additive white Gaussian noise on SVM and MD classifiers using cue vectors based on HNR or COG up to 2 kHz is evaluated following the same procedure described in Sec. III C 2 for cue vectors including $\overline{\kappa}$. The global accuracy metric $\mathrm{Acc}(\alpha_c)$ is plotted in Fig. 9 for HNR and COG and may be compared with the corresponding results for $\overline{\kappa}$ in Fig. 7(a). Although classification accuracy decreases with increasing noise amplitude α_c for all cue vectors (reaching approximately 70%), the degradation is gradual with increasing α_c for $\overline{\kappa}$ but markedly more abrupt (from $\alpha_c \geq 0.05$) for HNR and COG. A similar trend is observed for the other performance metrics. These results suggest that the proposed aperiodicity cue $\overline{\kappa}$ exhibits greater robustness to moderate additive white noise than other voicing cues, such as HNR or COG.

V. CONCLUSION

A new acoustic aperiodicity cue for fricative voicing-contrast classification is introduced. The cue is derived from the minimum of the cumulative mean normalized difference function of a normalized acoustic signal, an inherent

TABLE III. Global performance metrics (accuracy Acc, precision Prec, and recall Rec; in %) for the SVM and MD classifiers, computed after $B = 150$ bootstrap iterations for the CVCV dataset. Results are reported for different cue vectors combining A_{rms} with either the proposed aperiodicity cue $\overline{\kappa}$, the HNR, or the COG computed up to 2 kHz.

Cues	$[\overline{\kappa}, A_{\mathrm{rms}}]^\top$		$[\mathrm{HNR}, A_{\mathrm{rms}}]^\top$		$[\mathrm{COG}, A_{\mathrm{rms}}]^\top$	
Classifier	SVM	MD	SVM	MD	SVM	MD
Acc	85%	83%	84%	83%	80%	86%
Prec	82%	82%	83%	83%	78%	86%
Rec	83%	83%	81%	82%	79%	85%

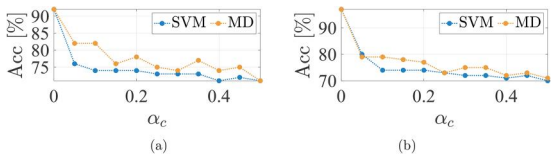

FIG. 9. Global metric accuracy $\mathrm{Acc}(\alpha_c)$ for the SVM and MD classifiers with cue vector. (a) $[\mathrm{HNR}, A_{\mathrm{rms}}]^\top$, (b) $[\mathrm{COG}, A_{\mathrm{rms}}]^\top$ up to 2 kHz.

component of YIN-based fundamental frequency estimators. Combined with RMS amplitude, this yields two-dimensional acoustic cue vectors whose Pr indicate that UFr, VFr, and *V* can be reliably distinguished. This is confirmed by global classification performance metrics of approximately 91% using standard classifiers (Mahalanobis distance, linear SVM) applied to the first CV sequence of French CVCV utterances recorded under noise-free laboratory conditions (SNR $>$ 10 dB for all three phoneme classes).

Analysis of the second CV sequence further shows that the proposed aperiodicity cue is invariant to normalization, whereas the RMS amplitude is not, resulting in a reduced global classification performance of approximately 83% for the full CVCV dataset. In comparison to other voicing cues, the HNR and the COG computed up to 2 kHz exhibit voicing-contrast behavior similar to that of the proposed aperiodicity cue, unlike commonly used features as segment duration, CPP, or COG, up to 500 Hz. Accordingly, for the full CVCV dataset, classification based on HNR or COG up to 2 kHz yields performance comparable to that obtained with the proposed cue, whereas the use of other features leads to reduced performance (60%–80%). These findings confirm the relevance of the proposed cue for voicing-contrast detection under clean speech conditions.

In the presence of additive white Gaussian noise, the aperiodicity cue for VFr increases, leading to degraded—or under severe noise conditions, lost—fricative voicing-contrast classification, as reflected in reduced performance metrics. Nevertheless, the degradation remains gradual for the proposed cue at moderate noise levels, in contrast to the more abrupt decline observed for HNR and COG up to 2 kHz. These results motivate further investigation of the proposed aperiodicity cue across additional languages, more diverse speech materials, and varied noise environments.

ACKNOWLEDGMENTS

The authors acknowledge CNRS through the 80|Prime program (InteracSon).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX: THEORETICAL APERIODICITY CUE

The signal in Fig. 1(a) is defined as

$$s(t) = A\sin(2\pi f_0 t) + \alpha(t)n(t), \tag{A1}$$

where $n(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ denotes standard white Gaussian noise, i.e., the samples of $n(t)$ are independent and identically distributed (i.i.d.) standard normal random variables, and

$$\alpha(t) = \alpha_0 e^{\gamma t} \tag{A2}$$

represents an exponentially increasing noise-amplitude envelope.

Since the sinusoid and zero-mean noise are uncorrelated, the ACF is

$$r_k(\tau) = \frac{A^2}{2}\cos(2\pi f_0 \tau) + \alpha_k \alpha_{k+\tau}\delta(\tau), \tag{A3}$$

where the Dirac delta function, $\delta(\tau)$, accounts for the noise contribution at zero lag. The time-varying noise scaling is defined as

$$\alpha_k = \alpha_0 e^{\gamma t_k}, \qquad \alpha_{k+\tau} = \alpha_k e^{\gamma \tau}, \tag{A4}$$

with k denoting the discrete-time frame index.

The corresponding difference function [Eq. (2)] is

$$d_k(\tau) = A^2(1-\cos(2\pi f_0\tau)) + \alpha_k^2(1+e^{2\gamma\tau}) - 2\alpha_k\alpha_{k+\tau}\delta(\tau). \tag{A5}$$

The cumulative mean for $\tau > 0$ is expressed as

$$M(\tau) = \frac{1}{\tau}\sum_{j=1}^{\tau} d_k(j) \tag{A6}$$

$$= A^2\left[1 - \frac{\sin(2\pi f_0\tau)}{2\pi f_0\tau}\right] + \alpha_k^2\left[1 + \frac{e^{2\gamma\tau}-1}{2\gamma\tau}\right], \tag{A7}$$

from which the theoretical expression of the cumulative mean normalized difference function [Eq. (3)] follows:

$$d'_k(\tau) = \begin{cases} 1, & \tau = 0, \\ \dfrac{d_k(\tau)}{M(\tau)}, & \tau > 0. \end{cases} \tag{A8}$$

Furthermore, the instantaneous signal-to-noise ratio [in decibels (dB)] for the signal in Eq. (A1) decreases linearly with time and is given by

$$\mathrm{SNR}_k = 10\log_{10}\left(\frac{A^2/2}{\alpha_k^2}\right). \tag{A9}$$

The parameters are $A = 1$, $f_0 = 200$ Hz, $\alpha_0 = 0.01$, and $\gamma = 5\,\mathrm{s}^{-1}$.

[1]K. Stevens, *Acoustic Phonetics* (MIT Press, Cambridge, MA, 2000), pp. 1–607.
[2]A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263 (2000).
[3]M. Gordon, P. Barthmaier, and K. Sands, "A cross-linguistic acoustic study of voiceless fricatives," J. Int. Phonetic Assoc. **32**, 141–174 (2002).
[4]B. McMurray and A. Jongman, "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," Psychol. Rev. **118**, 219–246 (2011).
[5]A. Ali, J. Van Der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," J. Acoust. Soc. Am. **109**, 2217–2235 (2001).
[6]J. Pincas and P. Jackson, "Amplitude modulation of turbulence noise by voicing in fricatives," J. Acoust. Soc. Am. **120**, 3966–3977 (2006).
[7]L. M. T. Jesus and P. J. B. Jackson, "Frication and voicing classification," in *Computational Processing of the Portuguese Language. PROPOR 2008*, Lecture Notes in Computer Science Vol. 5190, edited by A. Teixeira, V. L. S. de Lima, L. C. de Oliveira, and P. Quaresma (Springer, Berlin, Heidelberg, Germany, 2008), pp. 11–20.
[8]N. Silbert and K. De Jong, "Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production," J. Acoust. Soc. Am. **123**, 2769–2779 (2008).
[9]K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives," J. Acoust. Soc. Am. **125**, 3962–3973 (2009).
[10]V. Kharlamov, D. Brenner, and B. Tucker, "Temporal and spectral characteristics of conversational versus read fricatives in American English," J. Acoust. Soc. Am. **152**, 2073–2081 (2022).
[11]C. Bjorndahl, "Voicing and frication at the phonetics-phonology interface: An acoustic study of Greek, Serbian, Russian, and English," J. Phonetics **92**, 101136 (2022).
[12]C. Shadle, W. Chen, L. Koenig, and J. Preston, "Refining and extending measures for fricative spectra, with special attention to the high-frequency range," J. Acoust. Soc. Am. **154**, 1932–1944 (2023).
[13]P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages* (Blackwell, Oxford, UK, 1996), p. 426.
[14]J. Cisonni, K. Nozaki, A. Van Hirtum, X. Grandchamp, and S. Wada, "Numerical simulation of the influence of the orifice aperture on the flow around a teeth-shaped obstacle," Fluid Dyn. Res. **45**, 025505 (2013).
[15]R. Cole and W. Cooper, "Perception of voicing in English affricates and fricatives," J. Acoust. Soc. Am. **58**, 1280–1287 (1975).
[16]A. Jongman, "Duration of frication noise required for identification of English fricatives," J. Acoust. Soc. Am. **85**, 1718–1725 (1989).
[17]K. Stevens, S. Blumstein, L. Glicksman, M. Burton, and K. Kurowski, "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," J. Acoust. Soc. Am. **91**, 2979–3000 (1992).
[18]J. Hillenbrand, R. Cleveland, and R. Erickson, "Acoustic correlates of breathy vocal quality," J. Speech. Lang. Hear. Res. **37**, 769–778 (1994).
[19]P. Jackson and C. Shadle, "Frication noise modulated by voicing, as revealed by pitch-scaled decomposition," J. Acoust. Soc. Am. **108**(4), 1421–1434 (2000).
[20]H. Cho and M. Giavazzi, "Perception of voicing in fricatives," in *Proceedings of the 18th International Congress of Linguists*, Seoul, South Korea (2009), pp. 986–1006.
[21]A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002).
[22]B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference*, Austin, TX (2015), pp. 18–25.
[23]S. Strömbergsson, "Today's most frequently used f0 estimation methods, and their accuracy in estimating male and female pitch in clean speech," in *Proceedings Interspeech*, Dresden, Germany (2016), pp. 525–529.
[24]B. Quinn, J. Nielsen, and M. Christensen, "Fast algorithms for fundamental frequency estimation in autoregressive noise," Signal Process. **180**, 107860 (2021).
[25]T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech*, Florence, Italy (International Speech Communication Association, 2011), pp. 1973–1976.
[26]D. Jouvet and Y. Laprie, "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece (IEEE, New York, 2017), pp. 1614–1618.
[27]R. Vaysse, C. Astésano, and J. Farinas, "Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech," J. Acoust. Soc. Am. **152**, 3091–3101 (2022).
[28]W. Hess, "Pitch and voicing determination of speech with an extension toward music signals," in *Springer Handbook of Speech Processing*, edited by J. Benesty, M. M. Sondhi, and Y. A. Huang (Springer, Berlin, Heidelberg, Germany, 2008), pp. 181–212.
[29]D. Demolin, S. Hassid, C. Ponchard, S. Yu, and R. Trouville, "Speech aerodynamics database" (2019).
[30]P. Mahalanobis, "On the generalized distance in statistics," Sankhyā A **80**, S1–S7 (2018).
[31]N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods* (Cambridge University Press, Cambridge, UK, 2000), pp. 1–189.
[32]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer New York Inc., New York, 2009), pp. 1–737.