RESEARCH REPORT

EXPLORING INTONATIONAL CONTOURS IN HUMAN AND SYNTHETIC SPEECH: AN F0-BASED STUDY OF VENEZUELAN SPANISH

João Paulo Moraes LIMA DOS SANTOS
Department of Education – Federal Institute of Sertão Pernambucano (IFSertãoPE)
Floresta, PE, Brazil

Eugenia SAN SEGUNDO
Instituto de Lengua, Literatura y Antropología – Spanish National Research Council (CSIC)
Madrid, Spain

EDITORS
- Leônidas Silva Jr. (UEPB)
- Plinio Barbosa (UNICAMP)
- Sandra Madureira (PUC-SP)
- Renata Passetti (UFSCar)

REVIEWERS
- Carolina Silva (UFPB)
- Amaury Silva (FATEC)

ABOUT THE AUTHORS
- João Paulo Moraes Lima dos Santos
  Conceptualization; Methodology; Formal Analysis; Data Curation; Visualization; Writing - Original Draft.
- Eugenia San Segundo
  Conceptualization; Methodology; Supervision; Funding Acquisition; Writing - Review & Editing.

Received: 05/02/2026
Accepted: 03/06/2026
Published: 17/07/2026

HOW TO CITE
LIMA DOS SANTOS, J. P. M.; FERNÁNDEZ, E. S. S. (2026). Exploring Intonational Contours in Human and Synthetic Speech: An F0-Based Study of Venezuelan Spanish. *Cadernos de Linguística*, v. 7, n. 7, e976.

ABSTRACT

Comparisons between human and synthetic speech reveal that prosodic differences manifest primarily in how pitch develops over time rather than in static acoustic properties. This study investigates how human and synthetic intonation differ in Venezuelan Spanish by analyzing the temporal organization of fundamental frequency (F0) contours in declarative and interrogative sentences. The analysis is based on data from the HABLA corpus, which contains utterances produced by human speakers paired with multiple synthetic realizations derived from identical sentence materials, allowing controlled comparisons across voice types. Time-normalized F0 contours were extracted and analyzed using generalized additive mixed models in order to capture differences in contour shape and temporal organization. In addition, linear mixed-effects models were employed to examine complementary token-level acoustic measures, including pitch range, global variability, and slope-based metrics. Across both sentence modalities, the results show that synthetic speech reproduces general intonational configurations compatible with sentence modality. However, systematic differences emerge in the temporal implementation of these patterns. Human speech exhibits greater internal

differentiation and localized modulation across the utterance, particularly in linguistically salient sentence-final regions. Synthetic speech, by contrast, displays smoother and more temporally regularized contours, with reduced local variability and attenuated slope changes. Global measures of pitch range and whole-utterance variability show considerable overlap between the two voice types, indicating that divergence is not driven by global pitch amplitude or aggregate dispersion. Instead, the results demonstrate that differences are localized in the temporal organization of F0 across the utterance. These findings highlight the importance of contour-based, time-sensitive approaches for assessing prosodic realism in synthetic speech.

RESUMO

As comparações entre a fala humana e a fala sintética revelam que as diferenças prosódicas se manifestam principalmente na forma como o *pitch* se desenvolve ao longo do tempo, e não em propriedades acústicas estáticas. Este estudo investiga como a entoação humana e a sintética diferem no espanhol venezuelano por meio da análise da organização temporal dos contornos da frequência fundamental (F0) em sentenças declarativas e interrogativas. A análise baseia-se em dados do corpus HABLA, que contém enunciados produzidos por falantes humanos emparelhados com múltiplas realizações sintéticas derivadas dos mesmos materiais frasais, o que permite comparações controladas entre tipos de voz. Contornos de F0 normalizados no tempo foram extraídos e analisados por meio do modelo aditivo generalizado misto, com o objetivo de capturar as diferenças na forma do contorno e em sua organização temporal. Além disso, os modelos lineares de efeitos mistos foram empregados para examinar medidas acústicas complementares no nível de *token*, incluindo o *pitch range*, a variabilidade global e métricas baseadas nos *slope*s. Em ambas as modalidades oracionais, os resultados mostram que a fala sintética reproduz configurações entoacionais gerais compatíveis com a modalidade. No entanto, diferenças sistemáticas emergem na implementação temporal desses padrões. A fala humana apresenta maior diferenciação interna e modulação localizada ao longo do enunciado, particularmente em regiões finais linguisticamente salientes. A fala sintética, por sua vez, exibe contornos mais suaves e temporalmente mais regularizados, com menor variabilidade local e mudanças de *slope*s atenuadas. As medidas globais de *pitch range* e de variabilidade ao longo do enunciado apresentam sobreposição considerável entre os dois tipos

de voz, indicando que a divergência não é impulsionada pela amplitude global do *pitch* nem pela dispersão agregada. Em seu lugar, os resultados demonstram que as diferenças se localizam na organização temporal da F0 ao longo do enunciado. Esses resultados ressaltam a importância de abordagens baseadas em contornos e sensíveis ao tempo para a avaliação do realismo prosódico na fala sintética.

KEYWORDS
Synthetic Speech; Prosody; F0 Contours; Temporal Organization; Venezuelan Spanish.

PALAVRAS-CHAVE
Fala Sintética; Prosódia; Contornos de F0; Organização Temporal; Espanhol Venezuelano.

INTRODUCTION

Intonation is fundamental to the perception of naturalness and communicative adequacy in spoken language and is therefore a critical component of both human communication and speech technologies (Hirschberg, 2002; Nussbaum *et al.*, 2025). Beyond encoding linguistic structure, intonation is essential for signaling sentence modality, speaker attitude, and interactional meaning (Pierrehumbert; Hirschberg, 1990; Wagner; Watson, 2010).

In recent years, these classical phonetic concerns have gained renewed relevance with the rapid development of neural text-to-speech (TTS) systems. Although contemporary synthetic voices achieve high levels of segmental intelligibility, converging evidence indicates that prosody remains a major locus of divergence between synthetic and human speech, with fundamental frequency (F0) patterns emerging as a key factor in assessments of naturalness and communicative adequacy (Galdino *et al.*, 2025a; Galdino *et al.*, 2025b; Nussbaum *et al.*, 2025). In particular, the reproduction of intonation patterns that depend on fine-grained pitch modulation and precise temporal coordination continues to pose challenges for current synthesis systems (Xu, 2011; Niebuhr; Skarnitzl, 2019).

Beyond implications for perceived naturalness, limitations in the temporal realization of intonation also bear directly on how listeners distinguish human from synthetic speech. Recent perceptual evidence indicates that judgments of voice authenticity rely primarily on suprasegmental and higher-level features rather than on segmental detail. For instance, San Segundo *et al.* (2025) show that listeners rely primarily on suprasegmental and higher-level or extralinguistic cues when judging whether a voice is human or synthetic, highlighting the relevance of temporally structured features such as F0 contours. From this perspective, understanding how intonational patterns are temporally organized in human and synthetic speech is not only relevant for modeling prosody, but also for explaining perceptual judgments of voice authenticity.

Despite this growing body of research, direct acoustic comparisons between human and synthetic intonation remain limited for specific languages, such as Spanish. As a result, it remains unclear to what extent current TTS systems are able to reproduce the temporal and dynamic properties of F0 contours that characterize intonation in this variety.

In this context, the present study addresses this gap by comparing time-normalized F0 contours in human and synthetic speech in Venezuelan Spanish, with particular attention to sentence modality (declarative vs. interrogative). Using a contour-based acoustic approach, the analysis examines whether synthetic speech preserves the temporal variability characteristic of human intonation and whether modality-related contrasts are maintained across voice types. Specifically, the study asks (i) how F0 contours differ between human and synthetic speech in Venezuelan Spanish and (ii) how such differences are manifested in declarative and interrogative sentences. Based on recent research on intonation and speech synthesis across different languages, which consistently reports systematic differences between human and synthetic prosody (Galdino *et al.*, 2025a; Nussbaum *et al.*, 2025),

four hypotheses are advanced: *H1*: Human speech is expected to exhibit greater dynamic variation in F0 contours than synthetic speech; *H2:* Differences between human and synthetic speech are expected to be reflected not only in overall contour shape, but also in more pronounced and more locally differentiated changes in F0 slope, particularly in linguistically salient sentence-final regions; *H3:* Global F0 measures, such as pitch range and whole-utterance variability, are expected to show substantial overlap between human and synthetic speech; *H4:* Local measures and contour-based analyses are expected to provide greater discriminatory power than global metrics, revealing systematic, region-specific differences in the temporal organization of F0. Differences between voice types are further expected to be especially apparent in interrogative sentences, given the central role of utterance-final F0 modulation and its temporal organization in signaling interrogative meaning.

1. THEORETICAL BACKGROUND

1.1. INTONATION AND F0 CONTOURS

Intonation constitutes a central dimension of spoken language, articulating linguistic structure with the physiological mechanisms of speech and with communicative intent. Pitch movements aligned with stress and prosodic boundaries are the main resources for encoding sentence modality and pragmatic stance (Ladd, 2008; Prieto; Roseano, 2010). Phonetic research has consistently identified fundamental frequency (F0) as one of the primary acoustic correlates of intonation, emphasizing that intonational contrasts emerge from the dynamic organization and temporal coordination of pitch movements across the utterance rather than from static tonal targets or isolated pitch points. From this perspective, intonation is best understood as a time-continuous acoustic phenomenon, whose linguistic relevance is distributed across the contour and constrained by both linguistic structure and temporal organization (Ladd, 2008; Xu, 2011). Experimental approaches to intonation have focused on the analysis of F0 contours across the utterance, treating intonation as a long-term organizational property of speech and distinguishing it from more localized pitch phenomena such as tone or microprosody (Barbosa, 2019). This temporal perspective motivates analytical approaches that examine the development of F0 across the utterance rather than relying solely on static pitch descriptors. By focusing on contour form and temporal organization, analyses can capture dimensions of prosodic realization that are perceptually relevant and that are often obscured by simplified representations (Xu, 2011).

Advances in speech synthesis have reframed classical questions about intonation and prosodic naturalness. Although contemporary text-to-speech systems achieve high levels of segmental intelligibility, early work on statistical parametric synthesis showed that synthetic speech parameters are often generated as average trajectories derived from acoustic models trained on large speech

corpora (Tokuda *et al.*, 2000). Because these models represent multiple realizations through shared mean values, the resulting speech may lack detailed acoustic and prosodic variation. Zen *et al.* (2009) discussed this tendency as a central limitation of statistical parametric speech synthesis, highlighting reduced variance and trajectories that are too smooth, with consequences for perceived naturalness across languages. Ren *et al.* (2022) referred to this phenomenon as *over-smoothing*, formalizing it as a persistent problem in neural text-to-speech systems.

From the perspective of speech technology, recent work has shown that model architectures and generation strategies can systematically affect the temporal dynamics of prosodic parameters such as F0, reinforcing overly averaged contours and reducing temporal detail in synthetic speech, particularly in non-autoregressive text-to-speech systems (Ren *et al.*, 2022). A contour-based, time-resolved analysis therefore provides a particularly appropriate foundation for investigating how well synthetic speech reproduces the temporal richness of human intonation and for identifying systematic divergences that arise from statistical or neural generation processes.

1.2. HUMAN VS. SYNTHETIC SPEECH

The growing presence of neural text-to-speech systems has highlighted the need for systematic comparisons between human and synthetic speech. Despite substantial advances in intelligibility and segmental realization, synthetic speech often diverges from human speech in the temporal organization and variability of prosodic patterns, with intonation remaining a particularly sensitive domain.

From a phonetic perspective, synthetic speech is often described as exhibiting reduced variability relative to human speech. Human speakers naturally modulate F0 as a function of linguistic structure, discourse context, and communicative intent, resulting in contours that display local fluctuations, microprosodic effects, and adaptive timing. Synthetic speech, by contrast, tends to rely on statistically learned averages that smooth over this variability, reflecting temporal constraints that limit local pitch variation (Taylor, 2009; Zen *et al.*, 2009).

From a perceptual and technological perspective, Henter *et al.* (2014) demonstrate that common modelling assumptions in statistical parametric speech synthesis — such as mean-based parameter generation and independence constraints — lead to systematic reductions in perceived naturalness, even when stimuli are derived from repeated recordings of natural speech. Complementing these findings, Veilleux *et al.* (2012) show that listeners are sensitive to the global temporal organization and overall form of F0 contours, and that overly regular or simplified pitch trajectories negatively affect perceived naturalness. Further recent work demonstrates that intonational meaning is encoded in temporally distributed properties of F0 contours across the utterance, motivating time-resolved, contour-based analyses (Veilleux *et al.*, 2024).

Intonational realization varies systematically across sentence modalities. Wagner and Watson (2010) argue that prosodic meaning is grounded in the temporal organization and scaling of pitch

movements relative to prosodic structure, with sentence-final regions being central to pragmatic interpretation. In interrogative sentences, this organization typically involves complex F0 movements in sentence-final regions, where rises, falls, or fall–rise configurations encode meanings such as confirmation-seeking, stance, or epistemic uncertainty. Because these patterns rely on fine-grained temporal coordination, they remain particularly challenging for synthetic speech systems, even when global pitch characteristics and segmental quality are well matched to human speech (Henter *et al.*, 2014).

Evidence from neural text-to-speech makes this limitation explicit. Zou *et al.* (2021) show that conventional neural TTS architectures systematically favor averaged prosodic realizations, which undermines the reproduction of temporally structured intonation unless explicit, linguistically informed representations are introduced. Similarly, O'Mahony *et al.* (2024) demonstrate that contour-based, hierarchical modeling of F0 captures phrase-level slopes and word-level patterns, leading to substantial improvements in the time-resolved realization of intonation. Human–synthetic speech comparisons therefore require contour-based analyses that directly target the temporal organization of F0 contours, where biological control mechanisms and algorithmic constraints most clearly diverge.

1.3. VENEZUELAN SPANISH INTONATION

Spanish intonation has traditionally been described within the Autosegmental-Metrical framework as exhibiting phonological contrasts associated with sentence modality, including the recurrent association of declaratives with falling contours and yes/no interrogatives with rising or high boundary tones (Hualde, 2014). Studies on Spanish prosody demonstrate that fine-grained differences in peak height and the timing of tonal events contribute to the signaling of pragmatic and information-structural distinctions (Face, 2006; Face, 2007), while cross-dialectal work shows that such differences are systematically encoded in the alignment and scaling of F0 contours across Spanish varieties (Colantoni; Gurlekian, 2004). From a cross-dialectal perspective, Prieto and Roseano (2010) further show that, while Spanish shares a core set of phonological intonational contrasts across dialects, intonational meaning is systematically determined by the phonetic implementation of F0 contours, including their scaling and temporal organization, which vary in regular ways across varieties.

Within this broader context, Venezuelan Spanish provides a particularly clear illustration of how shared phonological contrasts in Spanish are realized phonetically, especially in the sentence-final region. Early descriptive work by Mora *et al.* (1997) documents that, in Venezuelan Spanish, intonational patterns associated with sentence modality are not realized through rigid categorical configurations, but instead exhibit gradient phonetic implementation conditioned by discourse and pragmatic factors. In contrast, Sosa (1999) situates Venezuelan Spanish within the broader typology

of Spanish intonation, describing it as conforming to the general pattern of falling contours in declaratives and rising or high-ending contours in yes/no interrogatives observed across many Spanish varieties. More recent analyses further refine these accounts by demonstrating that the observed phonetic variability is not random, but systematically distributed across speakers, sentence types, and discourse contexts (Díaz *et al.*, 2019). These findings underscore that, in Venezuelan Spanish, modality contrasts are maintained through differences in the phonetic realization of F0 rather than through rigid tonal configurations.

1.3.1. DECLARATIVE INTONATION

Declarative intonation in Venezuelan Spanish exhibits a gradual decline in fundamental frequency across the sentence, culminating in a low final boundary tone. This declination pattern is consistent with observations across Spanish varieties more generally, where declaratives are characterized by a progressive lowering of pitch accents followed by a final fall (Sosa, 1999; Prieto; Roseano, 2010). Comparative work on declarative intonation across multiple Spanish dialects, including Venezuelan Spanish, further demonstrates that this global decline may be accompanied by differences in pre-nuclear pitch levels and degrees of pitch compression, resulting in contours that vary systematically across varieties (Sosa, 1999; Díaz Campos; Tevis McGory, 2002). As a result, declarative contours may display localized modulations, such as delayed falls or partial pitch resets, without compromising their overall declinational profile.

Detailed phonetic analyses of Venezuelan Andean Spanish further refine this description by showing that declarative nuclei are frequently realized with high or downstepped high pitch accents followed by low boundary tones, rather than with uniformly low nuclear targets (Astruc *et al.*, 2010). Importantly, these studies indicate that the declarative profile is defined not by an absolute low endpoint, but by the temporal relation between the nuclear pitch accent and the subsequent decline toward the utterance boundary.

Evidence from semispontaneous speech supports and extends these findings. Using a Map Task corpus covering multiple Venezuelan regions, Dorta and Díaz (2021) show that declarative intonation preserves a clear global declinational trend while exhibiting systematic variation in nuclear configuration across speakers and regions. In particular, declaratives may display either fully descending contours or circunflex nuclear patterns, especially in male speakers, with the perceptual relevance of the nuclear peak increasing in more spontaneous speech contexts. These results indicate that declarative intonation in Venezuelan Spanish combines a stable global decline with locally salient temporal modulation, reinforcing the view that modality-related contrasts are encoded through the organization and scaling of F0 over time rather than through categorical tonal endpoints alone.

1.3.2. INTERROGATIVE INTONATION

With respect to interrogative intonation, Venezuelan Spanish exhibits a rich set of sentence-final pitch configurations that are fundamental to the expression of interrogative meaning. Within the broader typology of Spanish intonation, yes/no interrogatives are often associated with rising or high-ending contours (Sosa, 1999). However, Sosa's cross-dialectal analysis shows that interrogative contours are not uniformly realized as final rises but instead display systematic variation across Spanish varieties. While many mainland dialects tend to favor rising nuclear configurations, Caribbean varieties – including Puerto Rican, Cuban, and Venezuelan Spanish – frequently exhibit falling or circumflex final contours, reflecting stable dialect-specific patterns in the phonetic implementation of interrogative meaning. These differences involve not only contour shape but also pitch scaling, temporal alignment, and the interaction between pitch accents and boundary tones.

Subsequent work has shown that interrogatives often involve wider pitch excursions and higher overall pitch levels than their declarative counterparts, even when final boundary tones remain low (Díaz, 2023). Detailed phonetic analyses further indicate that interrogative meaning in Venezuelan Spanish emerges from the global organization of F0 in the nuclear region, rather than from a single categorical tonal event. Mora (1997) demonstrates that yes/no questions may begin at higher pitch levels, show expanded pitch range, and display more pronounced pitch movements in the nuclear accent, without necessarily exhibiting a final rise. In this sense, interrogatives are better characterized by enhanced pitch excursions and temporal reorganization of the nuclear contour than by the presence of a specific boundary tone.

Research conducted within the AMPER framework reinforces this view. Analyses of Venezuelan Spanish across different regions reveal that interrogatives frequently display circumflex or complex nuclear contours, whose contrast with declaratives lies primarily in pitch range, alignment, and scaling, rather than in categorical tonal opposition (Astruc *et al.*, 2010; Dorta; Díaz, 2021). Focusing on the Venezuelan variety of Zulia, Díaz (2023) demonstrates that interrogatives may share the same globally descending nuclear configuration as declaratives, while differing systematically in overall pitch height, tonal distance, and melodic prominence. Based on AMPER-based acoustic analyses, the study shows that interrogatives maintain a consistently higher tonal register across the utterance and exhibit greater perceptual salience of nuclear pitch movements, despite converging on a low final boundary tone.

Importantly, these converging findings align with the broader characterization of Spanish intonation outlined in Section 1.3, according to which sentence modality is associated with shared phonological contrasts that are realized through systematic phonetic implementation. In Venezuelan Spanish, interrogative meaning is thus not encoded by a fixed or invariant tonal configuration, but by regular differences in pitch scaling, contour shape, and temporal organization of F0, which may result in rising, high-ending, or even globally descending interrogative contours depending on dialectal and

discourse conditions. As argued by Sosa (1999) and further demonstrated in AMPER-based analyses (Dorta; Díaz, 2021; Díaz, 2023), the declarative-interrogative distinction is maintained through gradient phonetic properties of the F0 contour, rather than through rigid tonal endpoints. This perspective directly motivates analyses that focus on time-normalized F0 trajectories, allowing interrogative modality to be examined as an emergent property of contour dynamics over time, fully consistent with cross-dialectal models of Spanish intonation.

2. METHOD

2.1. MATERIALS AND DATA

The data analyzed in this study were drawn from the HABLA corpus (Tamayo Flórez *et al.*, 2023), a large-scale, multi-varietal resource that integrates recordings of natural speech with multiple types of automatically generated speech. The corpus provides parallel materials in which human utterances are paired with synthetic realizations derived from the same linguistic content, enabling controlled comparisons between natural and machine-generated speech. HABLA includes data from several varieties of Spanish spoken in South America. For the purposes of the present study, only Venezuelan Spanish was selected. Focusing on a single regional variety allows for a detailed examination of intonational patterns while minimizing cross-dialectal variability that could obscure differences between human and synthetic productions.

In the synthetic domain, the analysis was restricted to utterances generated using CycleGAN and diffusion-based generative models. CycleGAN implements an adversarial training framework in which a generator and a discriminator are jointly optimized to transform voice characteristics while preserving linguistic content, whereas diffusion models generate speech through iterative probabilistic reconstruction rather than adversarial training.

The analysis is based on a set of human utterances produced by native speakers of Venezuelan Spanish, paired at the sentence level with multiple synthetic realizations generated from the same orthographic materials. This design ensures strict lexical and syntactic comparability between human and synthetic speech while allowing intonational patterns in the synthetic data to emerge from each system's internal modeling strategies, without manual prosodic manipulation. The resulting dataset comprises repeated observations nested within speakers, sentence items, and voice types, reflecting the fact that each speaker contributes multiple utterances and that each utterance gives rise to multiple synthetic realizations. Detailed information on dataset composition is provided via the OSF repository (see the Data Availability Statement).

2.2. ACOUSTIC ANALYSIS

All acoustic measures were extracted using Praat (Boersma; Weenink, 2025) via the Parselmouth (Jadoul *et al.*, 2018) in Python. Fundamental frequency (F0) was extracted from each utterance using Praat's autocorrelation-based algorithm. Extraction parameters were selected to accommodate inter-speaker variability while remaining consistent across human and synthetic voices. All F0 tracks were visually inspected to identify potential tracking errors, such as octave jumps or spurious values, and unvoiced frames were excluded from further analysis. For each utterance, a reference F0 value was computed as the median F0 across voiced frames, providing a stable normalization anchor.

To facilitate comparison across utterances with different absolute durations, all F0 trajectories were time-normalized. Each utterance was resampled to a fixed number of equidistant points spanning the sentence from onset to offset, producing a normalized time axis. This procedure preserves the relative temporal organization of pitch movements while removing absolute duration as a confound. F0 values were subsequently converted to a semitone scale relative to reference F0, allowing comparisons across speakers and voice types while preserving relative pitch variation. The resulting time-normalized F0 contours were treated as functional data, representing the trajectory of pitch over the course of the sentence. Global properties, such as overall F0 declination, were derived from the full contour, whereas localized measures, including sentence-final F0 slope, were computed over the final portion of the normalized time axis. Pitch scaling and range were assessed using normalized F0 values, reducing the influence of physiological differences between speakers and voices.

2.3. STATISTICAL ANALYSIS

All statistical analyses were conducted in R (R Core Team, 2025) using the RStudio integrated development environment (Posit Team, 2025). For scalar outcome measures derived from the acoustic analysis, including global F0 declination, sentence-final F0 slope, and pitch range, linear mixed-effects models were fitted separately for declarative and interrogative sentences using the *lme4* package (Bates *et al.*, 2015). Within each modality, voice type (human vs. synthetic) was included as a fixed effect. To account for clustering in the data, token-level models included a random intercept for speaker, reflecting repeated observations per informant.

Time-continuous F0 contours were analyzed using generalized additive mixed models (GAMMs) fitted with the *mgcv* package (Wood, 2017). In these models, F0 (expressed in semitones relative to a reference value) was modeled as a smooth function of normalized time, with separate smooths specified for each voice type. Declarative and interrogative sentences were analyzed in separate models. Random intercepts for *token_id* (and for *file* in the interrogative models) were

included to account for clustering across utterances. Model specification, visualization, and diagnostic procedures were supported by *itsadug* (van Rij *et al.*, 2015), facilitating interpretation of smooth terms and difference curves.

Model estimation and comparison followed an information-theoretic approach. Competing models were evaluated using Akaike's Information Criterion (AIC), with differences in AIC (ΔAIC) used to assess the incremental explanatory value of predictors such as voice type and sentence modality. Model diagnostics included inspection of residual distributions and assessment of smoothness parameters to ensure appropriate model fit.

Inference focused on identifying systematic differences between human and synthetic speech in terms of (i) overall F0 declination patterns, (ii) sentence-final pitch behavior associated with modality, and (iii) contour complexity and temporal organization as reflected in time-continuous F0 trajectories. Visualizations of fitted smooths and difference curves were used to support interpretation of time-localized effects, particularly in the sentence-final region. Together, these analyses provide a principled statistical framework for evaluating the phonetic adequacy of synthetic intonation relative to human speech.

3. RESULTS

3.1. DECLARATIVE SENTENCES

3.1.1. F0 CONTOUR DIFFERENCES

Figure 1 presents the overall (fixed-effects) fitted F0 contours for declarative sentences produced by human and synthetic voices, estimated using a generalized additive mixed model that captures time-varying patterns across the normalized utterance. In both conditions, declaratives exhibit a general decline in F0, consistent with expected patterns of declination. Beyond this shared tendency, however, the fitted contours reveal systematic differences in the temporal organization of F0 between human and synthetic speech.

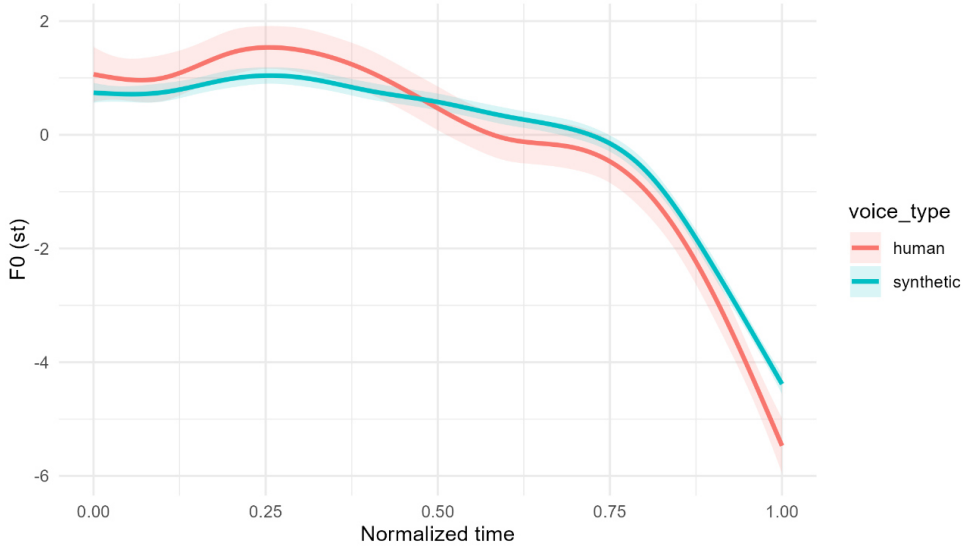

Figure 1. Fitted F0 contours for declarative sentences.

Human declarative contours show greater temporal differentiation across the utterance, with more clearly defined changes in curvature between early, medial, and final regions. In contrast, synthetic contours display smoother trajectories with reduced curvature, resulting in a more uniform temporal profile. These differences are reflected in the smooth interaction between normalized time and voice type, indicating that divergence between human and synthetic speech is expressed at the level of contour shape rather than through a uniform shift in pitch height. Importantly, this pattern is robust across tokens, as evidenced by the inclusion of random effects that account for token-specific variation.

To further characterize where these differences emerge, Figure 2 illustrates the time-resolved difference curve between the fitted human and synthetic contours (human minus synthetic), together with associated uncertainty bands. The magnitude of the difference varies across the utterance and is not evenly distributed over time. Instead, divergence increases toward the latter portion of the contour, suggesting that the sentence-final region contributes disproportionately to the overall contrast between voice types. This pattern is made more explicit in the zoomed view of the final interval shown in Figure 3, which highlights differences in terminal contour implementation.

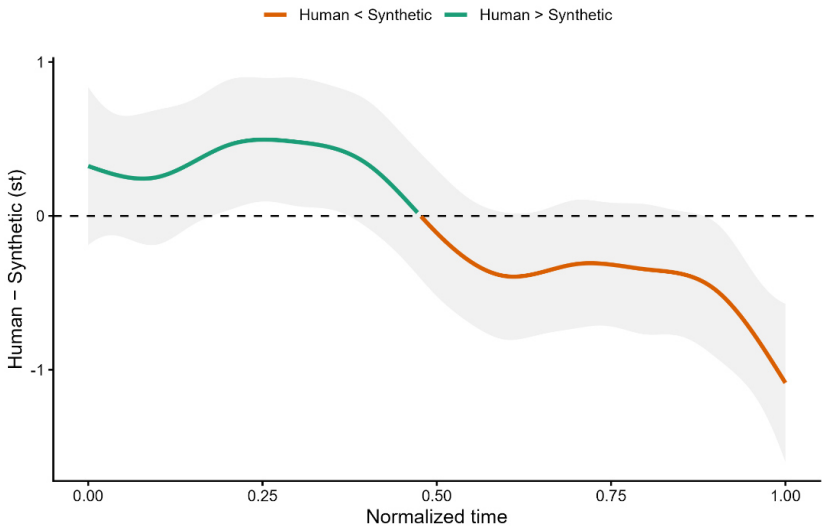

Figure 2. Time-resolved difference curve for declarative F0 contours.

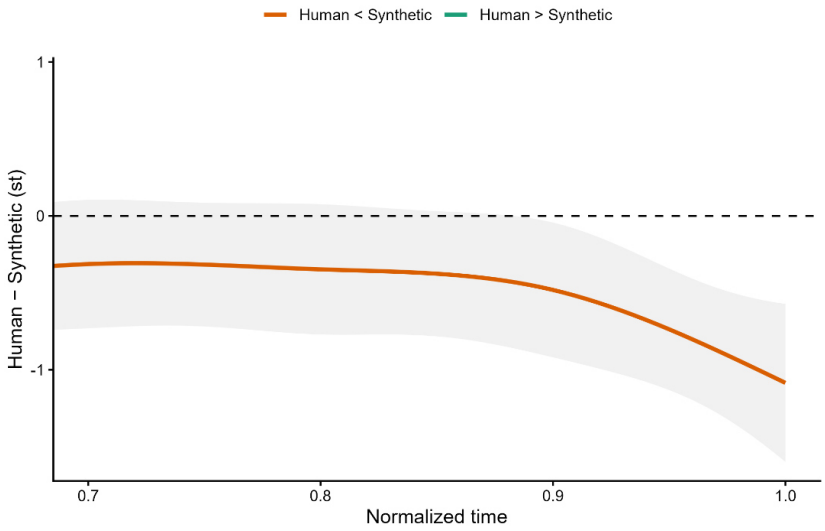

Figure 3. Zoomed difference curve (human versus synthetic) for declarative sentences.

In general, these results indicate that while synthetic speech reproduces the overall declining trend characteristic of declarative intonation, it differs from human speech in the detailed temporal structuring of F0 across the utterance. The contrast between voice types is therefore expressed primarily in the organization and evolution of the contour over time, rather than in the presence or absence of declination itself. This contour-level perspective provides a basis for the analysis of global F0 metrics reported in Section 3.1.2, which quantify how these temporal differences are reflected in aggregate acoustic measures.

3.1.2. GLOBAL F0 METRICS

To complement the contour-based analysis presented in Section 3.1.1, global F0 metrics were derived for each declarative token by summarizing the time-normalized contours into scalar measures. These metrics capture overall variability and slope properties of F0, providing a compact description of pitch behavior that is directly comparable across voice types. Figure 4 shows the distributions of these metrics for human and synthetic speech, while Table 1 reports estimates from linear mixed-effects models controlling for utterance duration and clustering by file.

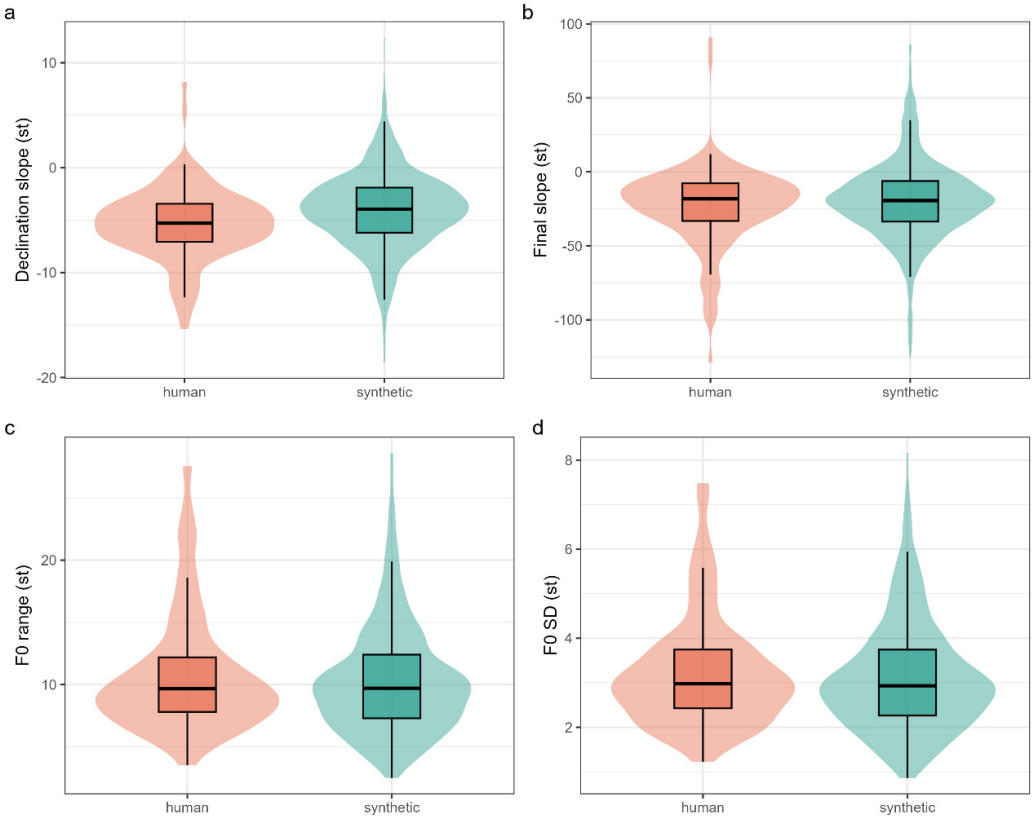

Figure 4. Distribution of global F0 measures for declarative sentences in human and synthetic voices, shown as violin plots including boxplots: (a) global declination slope, (b) sentence-final slope, (c) F0 range, and (d) F0 standard deviation.

Metric	Estimate	Std. Error	Statistic	*p* value
decl_slope	1.34	0.326	4.10	0.0000437
final_slope	4.07	2.54	1.60	0.11
f0_range	-0.483	0.413	-1.17	0.243
f0_sd	-0.113	0.115	-0.983	0.326

Table 1. Linear mixed-effects model estimates for global F0 metrics (declaratives).

As shown by the violin plots in Figure 4, human and synthetic speech exhibit broadly overlapping distributions in terms of global F0 range and overall F0 variability (standard deviation), as illustrated in Figures 4c and 4d. Mixed-effects models confirm that differences in these dispersion measures are limited in magnitude once duration is considered. This indicates that both voice types cover comparable pitch intervals and display similar overall levels of variability when considered at the level of entire utterances. In contrast, clearer differences emerge for slope-based measures that index the temporal organization of F0. The global declination slope, computed across the full normalized contour, is systematically less negative for synthetic speech than for human speech (Figure 4a), indicating a reduced rate of F0 decline over the course of the utterance. This effect is consistent in the mixed-effects analysis reported in Table 1. Sentence-final slope, computed within the 0.8-1.0 interval, shows a similar directional pattern (Figure 4b), with synthetic speech exhibiting weaker terminal movement on average. However, this effect is more sensitive to duration control and displays greater variability across tokens.

Overall, the global metrics reveal a consistent asymmetry between dispersion-based and slope-based measures. While overall pitch range and variability are broadly comparable across voice types, measures that capture directional change over time differentiate human and synthetic declaratives more clearly. These findings converge with the contour-based results in Section 3.1.1, indicating that the primary contrast between human and synthetic speech in declaratives consists in the temporal structuring of F0 rather than in aggregate variability alone.

3.2. INTERROGATIVE SENTENCES

3.2.1. F0 CONTOUR DIFFERENCES

Figure 5 shows the overall (fixed-effects) fitted F0 contours for interrogative sentences produced by human and synthetic voices, estimated to use the same generalized additive mixed modeling framework applied to declaratives. Across voice types, interrogatives exhibit broadly comparable global trajectories, with both contours displaying a marked change in F0 over the second half of the time-normalized interval. This overall similarity indicates that synthetic voices capture the general intonational configuration associated with interrogative modality.

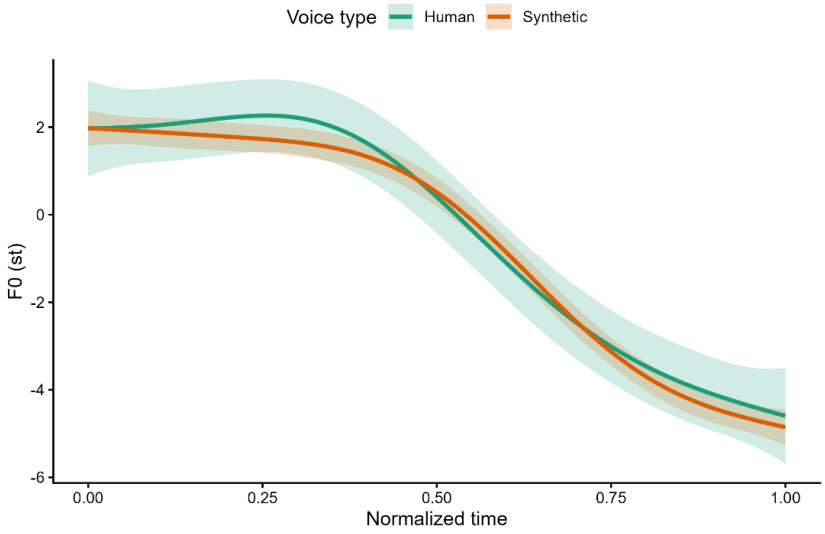

Figure 5. Fitted F0 contours for interrogative sentences.

At the same time, the fitted trajectories reveal systematic differences in temporal shaping. Human interrogative contours display greater internal modulation across the utterance, with more pronounced curvature and clearer differentiation between early, medial, and late portions of the contour. In contrast, synthetic interrogative contours are smoother and more regularized, exhibiting more gradual and evenly distributed changes in F0 over time. Rather than reflecting a strongly differentiated temporal organization, the synthetic trajectory appears temporally homogenized, with reduced localized modulation.

Importantly, these contour-level differences are captured by the smooth interaction between normalized time and voice type, indicating that divergence between human and synthetic interrogatives reflects consistent, time-dependent properties of the modeled trajectories rather than isolated pitch differences at individual time points.

The time-resolved difference curve shown in Figure 6 (human minus synthetic) further clarifies how these contrasts unfold across the utterance. Differences between voice types are not uniform over time but instead exhibit an alternating pattern. In an early interval, human interrogatives show higher F0 values than synthetic interrogatives. This relationship reverses in the medial-to-late portion of the utterance, where synthetic contours temporarily exceed human ones. Toward the utterance-final interval, the difference becomes positive again, with human interrogatives exhibiting higher F0 than synthetic interrogatives.

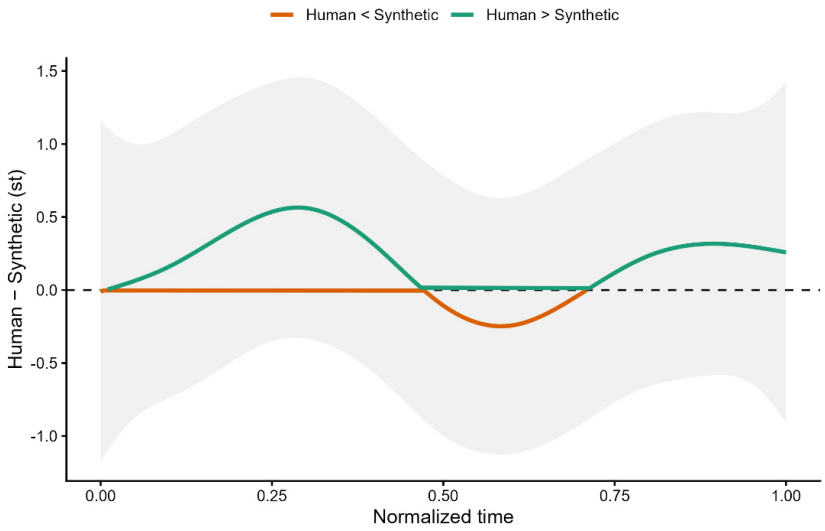

Figure 6. Time-resolved difference curve for interrogative F0 contours.

This pattern indicates that human and synthetic interrogatives are not separated by a simple global pitch offset. Instead, they differ in how F0 is dynamically distributed across the utterance, with relative advantages alternating across temporal regions. Colored segments in the difference curve indicate intervals where the estimated difference is credibly different from zero, while uncolored portions correspond to intervals where differences are not reliably distinguishable.

The zoomed view of the utterance-final interval in Figure 7 highlights a small but consistent positive difference between voice types. In this region, human interrogatives maintain higher F0 values than synthetic interrogatives by several fractions of a semitone. Although the magnitude of this difference is modest, its consistency suggests systematic underestimation of late-utterance F0 in synthetic speech.

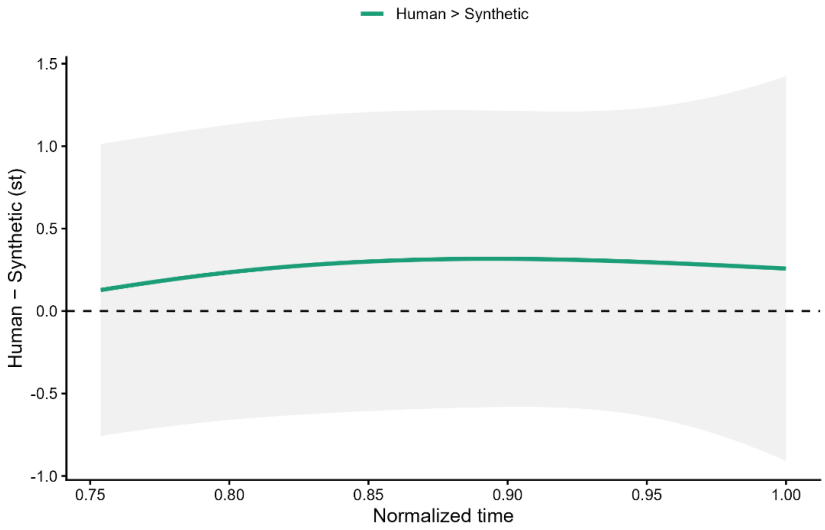

Figure 7. Zoomed view of the time-resolved difference curve in the utterance-final interval.

Together, Figures 5 and 7 indicate that while synthetic interrogatives approximate the overall interrogative contour shape, they differ from human speech in temporal organization. Human interrogatives exhibit greater internal differentiation and localized modulation, whereas synthetic interrogatives are characterized by smoother, more regularized trajectories with reduced temporal contrast.

3.2.2. FINAL RISE AND VARIABILITY

To complement the time-resolved contour analysis, token-level metrics targeting variability in interrogative F0 patterns were examined, with particular attention to differences between global dispersion and variability localized to the utterance-final region. Figure 8 summarizes these measures for human and synthetic interrogatives.

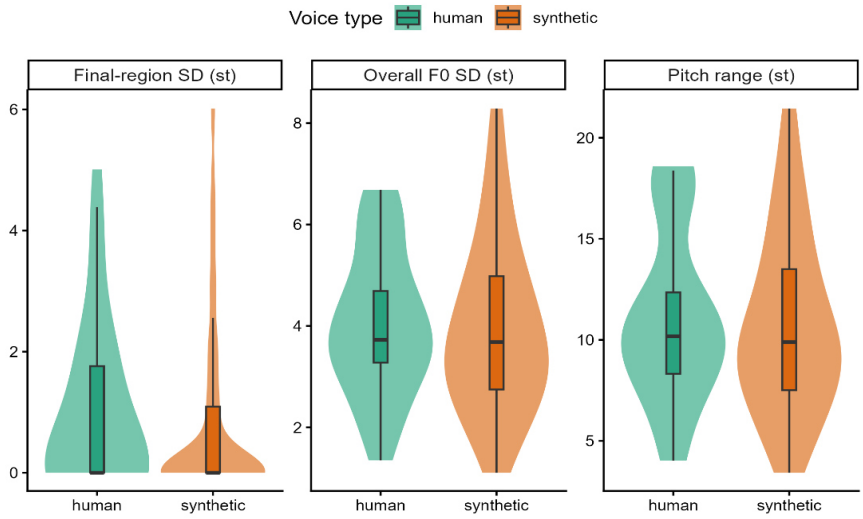

Figure 8. Token-level variability metrics for interrogative sentences for human and synthetic voices. Slope-based measures are omitted because temporal differences are more directly captured by the contour-level analyses shown in Figures 5–7.

Broad measures of variability, including overall F0 standard deviation and pitch range, show substantial overlap across voice types. In both metrics, synthetic speech exhibits a wider upper tail, indicating that synthetic interrogatives are not characterized by a uniform reduction in global variability. Instead, global dispersion across the entire contour appears comparable between human and synthetic productions, with synthetic speech occasionally exhibiting greater extremes.

Clearer differentiation emerges when variability is restricted to the utterance-final region. Final-region F0 standard deviation is more concentrated near zero for synthetic interrogatives, whereas human interrogatives display a broader distribution. This pattern indicates reduced local modulation

in the final portion of synthetic interrogatives, consistent with the smoother and more temporally regularized contours observed in the population-level fits.

Overall, these results suggest that differences between human and synthetic interrogatives are not primarily driven by overall pitch range or whole-utterance variability. Rather, they are localized to how variability is deployed over time. Human interrogatives exhibit richer token-to-token modulation in the utterance-final region, while synthetic interrogatives tend to regularize late-utterance dynamics, concentrating variability away from the region most closely associated with interrogative signaling.

4. DISCUSSION

4.1. INTONATIONAL DYNAMICS AND TEMPORAL ORGANIZATION

Across both declarative and interrogative sentences, the results demonstrate that intonational distinctions are closely related to the temporal organization of F0 trajectories across the utterance. Instead of being adequately captured by static pitch targets or global summary measures alone, intonation emerges from the dynamic shaping and temporal coordination of F0 over time, consistent with methodological approaches that emphasize contour-based and time-resolved analyses of speech prosody (Xu, 2011).

This interpretation is also consistent with previous descriptions of Venezuelan Spanish intonation, which show that sentence modality is distinguished through differences in contour configuration, nuclear development, and melodic prominence, even in cases where yes/no interrogatives do not exhibit categorical final rises (Mora, 1997; Sosa, 1999; Astruc *et al.*, 2010; Dorta; Díaz, 2021).

In human speech, F0 contours exhibit clear internal differentiation across early, medial, and final regions of the utterance. Declarative sentences are characterized by structured declination patterns, while interrogatives display temporally concentrated sentence-final movements associated with modal marking. In particular, this internal organization is not uniform: linguistically salient regions, especially toward the utterance end, tend to show stronger local modulation and clearer temporal differentiation. This pattern is consistent with descriptions of Venezuelan Spanish intonation, emphasizing the role of alignment, scaling, and timing of F0 movements across the contour (Sosa, 1999; Astruc *et al.*, 2010; Dorta; Díaz, 2021).

Importantly, the absence of a categorical utterance-final rise in some interrogative contours is compatible with established descriptions of Venezuelan Spanish intonation. As noted by Sosa (1999), yes/no interrogatives across Spanish varieties are not uniformly realized with final rises, and interrogative meaning may instead be conveyed through differences in overall pitch level, nuclear

configuration, and contour shape. Previous work within AMPER-based research also shows interrogative differences in scaling, alignment, and the prominence of nuclear movements. In Map task data, for instance, interrogatives show a strong preference for a circumflex nuclear configuration, while still allowing structured variation across speakers and sentence conditions (Dorta; Díaz, 2021). Similarly, descriptions of Venezuelan Andean Spanish report limited evidence for complex boundary tones and document interrogative patterns that often end in low boundary tones, consistent with interrogative meaning being expressed without a categorical final rise (Astruc *et al.*, 2010). From this perspective, interrogative marking need not surface as a sharply delimited final rise. It can emerge through region-specific differences in F0 timing, slope, and distribution over the utterance, consistent with *hypothesis 2* of this study.

Overall, the results demonstrate that the most informative distinctions between sentence modality are concentrated in the nuclear and utterance-final regions of the contour. Differences between human and synthetic speech emerge primarily in the timing, scaling, and slope of F0 movements rather than in global pitch range or variability. Intonational contrasts in Venezuelan Spanish are therefore best understood as properties of time-resolved F0 organization, confirming that temporal dynamics constitute a central dimension of intonational structure.

4.2. WHERE DO HUMAN AND SYNTHETIC INTONATION DIVERGE?

The contour-based analyses reveal that synthetic intonation is markedly smoother and more temporally regularized than human intonation. Internal differentiation across the utterance is attenuated, and localized modulations, particularly in sentence-final regions, are reduced. This pattern is consistent with previous work showing that statistical and neural speech synthesis systems tend to produce overly smooth acoustic trajectories and reduced prosodic variability as a consequence of statistical parameter generation and modeling assumptions, with consequences for perceived naturalness (Zen *et al.*, 2009; Henter *et al.*, 2014). The present results extend these observations by showing that such regularization is not uniform but is concentrated precisely in regions that are most relevant for linguistic interpretation, providing direct support for *hypothesis 1*, which predicted systematic differences in the temporal realization of F0 contours between human and synthetic speech.

The contrast between contour-based and metric-based analyses further clarifies the nature of this divergence. Global measures such as pitch range and whole-utterance variability show substantial overlap between human and synthetic speech, indicating that both voice types occupy comparable pitch spaces at a general level. This finding reinforces the view that global summary measures are insufficient for capturing prosodic structure when temporal organization is the primary carrier of linguistic information (Xu, 2011; Veilleux *et al.*, 2012). In some cases, synthetic speech even exhibits equal or greater dispersion in global metrics, underscoring that reduced variability per se is

not a defining property of synthetic intonation, in line with *hypothesis 3*, which predicted that global summary measures would be insufficient to distinguish between voice types. More specifically, differences emerge most clearly when intonation is analyzed in its temporal structure. Difference curves and region-specific measures reveal systematic divergence in how directional change and variability are expressed over time. Human speech shows greater heterogeneity and sharper modulation in linguistically salient regions, whereas synthetic speech distributes changes more evenly across the utterance, generating smoother and less punctuated trajectories. These findings show that the divergence between human and synthetic intonation is not a matter of missing modality-appropriate patterns, but of how those patterns are temporally instantiated. While synthetic speech preserves broad pitch characteristics, it systematically reduces localized modulation and temporal differentiation in linguistically salient regions, resulting in smoother and more regularized contours. This temporal redistribution of F0 movement explains why global summary measures fail to capture the core differences between voice types and underscores the explanatory advantage of contour-based, time-resolved analyses, in direct support of *hypothesis 4*.

5. CONCLUSION

This study examined human and synthetic intonation in Venezuelan Spanish using a contour-based acoustic approach that integrates time-normalized F0 modeling with complementary token-level measures. Across both declarative and interrogative sentences, the results show that synthetic speech reproduces modality-appropriate intonational configurations while differing systematically from human speech in the temporal organization and local variability of F0.

Rather than reflecting reduced pitch span or the absence of modality-specific patterns, differences between human and synthetic speech are concentrated in how intonation is structured over time. Human speech exhibits greater internal differentiation, sharper localized modulation, and richer token-to-token variability in linguistically salient regions, particularly toward the end of the utterance. Synthetic speech, by contrast, displays smoother, more regularized contours, with reduced local modulation and a tendency to distribute F0 change more evenly across time.

These findings have direct implications for the evaluation and development of speech technologies. For text-to-speech and voice conversion systems, they suggest that improving naturalness requires moving beyond accurate reproduction of global contour shapes toward models that better capture localized temporal dynamics and controlled variability. Incorporating mechanisms that allow for region-specific modulation may be essential for approximating the phonetic richness of human intonation.

Beyond synthesis quality, the results also have relevance for voice forensics and deepfake detection. The temporal regularization observed in synthetic speech represents a systematic

acoustic signature that distinguishes it from human speech, even when global pitch properties overlap. Time-resolved analyses of F0 contours, particularly in linguistically salient regions, may therefore provide valuable features for detecting synthetic or manipulated voices in forensic and security contexts.

Ultimately, this study demonstrates that the critical distinction between human and synthetic intonation lies not in whether modality-specific patterns are present, but in how those patterns are temporally implemented. By highlighting the dynamic organization of F0, the results contribute to a more nuanced understanding of prosodic realism and highlight the importance of temporal structure as a key dimension in both phonetic theory and speech technology.

ADDITIONAL INFORMATION

CONFLICT OF INTEREST

The author declares no conflicts of interest.

STATEMENT OF DATA AVAILABILITY

The full reproducible research package for this study is available at the Open Science Framework (OSF) under DOI: https://doi.org/10.17605/OSF.IO/F6WUE. The repository includes:

(i) derived acoustic datasets in CSV format;
(ii) R analysis scripts;
(iii) statistical model outputs;
(iv) figures reproduced from the analyses; and
(v) a step-by-step README describing folder structure, required R version and packages, and execution order.

All materials are shared under the CC-BY 4.0 license. The package enables reproduction of all analyses and figures from a clean R session using the included instructions.

AI USE STATEMENT

The AI tool ChatGPT (OpenAI, GPT-5.2, accessed January 2026) was used only for language revision and very minor code adjustments. No AI tools were used in the design of the study, data analysis, interpretation of the results, or in drawing the conclusions.

ETHICS AND CONSENT

This study uses speech data from the publicly available HABLA corpus (Tamayo Flórez *et al.*, 2023), available at Zenodo (https://doi.org/10.5281/zenodo.7370805). The corpus consists of previously collected recordings that were anonymized and made available for research purposes under specified licensing conditions. The present study involves secondary analysis of these data and does not include direct interaction with human participants. No personally identifiable information was accessed. Data sharing is limited to the terms established by the original corpus repository. Derived datasets and analysis materials generated in this study are shared separately as indicated in the Data Availability Statement.

FUNDING SOURCES

This research was supported by Grant PID2024-161495OB-I00, funded by MICIU/AEI/10.13039/501100011033 and by the European Regional Development Fund (ERDF), EU.

REVIEW AND AUTHORS' REPLY

Review: https://doi.org/10.25189/2675-4916.2026.V7.N7.ID976.R
Author's Reply: https://doi.org/10.25189/2675-4916.2026.V7.N7.ID976.A

REFERENCES

ASTRUC, L.; MORA, E.; REW, S. Venezuelan Andean Spanish intonation. In: PRIETO, P.; ROSEANO, P. (ed.). *Transcription of intonation of the Spanish language*. München: LINCOM Europa, 2010. p. 285-316.

BARBOSA, P. *Prosódia*. São Paulo: Parábola, 2019.

BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, v. 67, n. 1, 2015. DOI: https://doi.org/10.18637/jss.v067.i01.

BOERSMA, P.; WEENINK, D. *Praat*: doing phonetics by computer. Version 6.4.49. Amsterdam: University of Amsterdam, 2025. Software. Available at: https://www.fon.hum.uva.nl/praat/. Accessed on: 23 Dec. 2025.

COLANTONI, L.; GURLEKIAN, J. Convergence and intonation: historical evidence from Buenos Aires Spanish. *Bilingualism: Language and Cognition*, v. 7, n. 2, p. 107-119, 2004. https://doi.org/10.1017/s1366728904001488.

DÍAZ, C. Prominencia melódica en la entonación descendente de interrogativas y declarativas de Zulia (Venezuela). *Loquens*, v. 10, n. 1-2, 2023. https://doi.org/10.3989/LOQUENS.2023.E101.

DÍAZ, C.; DORTA, J.; MORA, E.; MUÑETÓN, M. Intonation across two border areas in the North Andean region: Mérida (Venezuela) and Medellin (Colombia). *Spanish in Context*, v. 16, n. 3, p. 329-352, 2019. https://doi.org/10.1075/sic.00042.dia.

DÍAZ CAMPOS, M.; TEVIS MCGORY, J. La entonación en el español de América: un estudio acerca de ocho dialectos hispanoamericanos. *Boletín de Lingüística*, n. 18, p. 3-26, 2002.

DORTA, J.; DÍAZ, C. Caracterización de la entonación venezolana a partir de un corpus obtenido con Map task. *Boletín de Filología*, v. 1, p. 329-354, 2021. Available at: http://w3.u-grenoble3.fr/dialecto/. Accessed on: 2 Dec. 2025.

FACE, T. L. Rethinking Spanish L*+H and L+H*. In: DÍAZ-CAMPOS, M. (ed.). *Selected proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology*. Somerville: Cascadilla Proceedings Project, 2006. p. 75-84.

FACE, T. L. The role of intonational cues in the perception of declaratives and absolute interrogatives in Castilian Spanish. *Journal of Experimental Phonetics*, v. 16, p. 185-225, 2007.

GALDINO, J. C.; LEAL, S. E.; DE SOUZA, L. G.; LIMA, R. de F.; MOREIRA, A. N. F. M.; JUNIOR, A. C.; OLIVEIRA, M.; CASANOVA, E.; ALUÍSIO, S. M. The impact of prosodic segmentation on speech synthesis of spontaneous speech. arXiv, 2025a. Preprint. Available at: http://arxiv.org/abs/2511.14779. Accessed on: 8 Dec. 2025.

GALDINO, J. C.; MATOS, A. N.; SVARTMAN, F. R. F.; ALUÍSIO, S. M. The evaluation of prosody in speech synthesis: a systematic review. *Journal of the Brazilian Computer Society*, v. 31, n. 1, p. 466-487, 2025b. https://doi.org/10.5753/jbcs.2025.5468.

HENTER, G. E.; MERRITT, T.; SHANNON, M.; MAYO, C.; KING, S. Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 15., 2014, Singapore. Proceedings of Interspeech 2014. Singapore: ISCA, 2014. p. 1504-1508. https://doi.org/10.21437/Interspeech.2014-361.

HIRSCHBERG, J. Communication and prosody: functional aspects of prosody. *Speech Communication*, v. 36, n. 1-2, p. 31-43, 2002. https://doi.org/10.1016/S0167-6393(01)00024-3.

HUALDE, J. I. *Los sonidos del español*. Cambridge: Cambridge University Press, 2014. DOI: https://doi.org/10.1017/CBO9780511719943.

JADOUL, Y.; THOMPSON, B.; DE BOER, B. Introducing Parselmouth: a Python interface to Praat. *Journal of Phonetics*, v. 71, p. 1-15, 2018. https://doi.org/10.1016/j.wocn.2018.07.001.

LADD, D. R. *Intonational phonology*. Cambridge: Cambridge University Press, 2008. https://doi.org/10.1017/CBO9780511808814.

MORA, E. División prosódica dialectal de Venezuela. *Omnia*, v. 2, n. 3, p. 93-99, 1997.

MORA, E.; HIRST, D.; DI CRISTO, A. Intonation features as a form of dialectal distinction in Venezuelan Spanish. In: INTONATION: THEORIES, MODELS AND APPLICATIONS, 1997, Athens, Greece. Proceedings of Intonation: Theories, Models and Applications. Athens, Greece: ISCA, 1997. p. 247-250. Available at: https://www.isca-archive.org/int_1997/mora97_int.html. Accessed on: 12 Dec. 2025.

NIEBUHR, O.; SKARNITZL, R. Measuring a speaker's acoustic correlates of pitch - but which? A contrastive analysis for perceived speaker charisma. In: INTERNATIONAL CONGRESS OF PHONETIC SCIENCES, 19., 2019, Melbourne. Proceedings of the 19th International Congress of Phonetic Sciences. Melbourne: Australasian Speech Science and Technology Association, 2019. p. 1774-1778. Available at: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1823.pdf. Accessed on: 5 Dec. 2025.

NUSSBAUM, C.; FRÜHHOLZ, S.; SCHWEINBERGER, S. R. Understanding voice naturalness. *Trends in Cognitive Sciences*, v. 29, n. 5, p. 467-480, 2025. https://doi.org/10.1016/j.tics.2025.01.010.

O'MAHONY, J.; CORKEY, N.; LAI, C.; KLABBERS, E.; KING, S. Hierarchical intonation modelling for speech synthesis using Legendre polynomial coefficients. In: INTERNATIONAL CONFERENCE ON SPEECH PROSODY, 12., 2024, Leiden, The Netherlands. Proceedings of Speech Prosody 2024. Leiden, The Netherlands: ISCA, 2024. p. 1030-1034. https://doi.org/10.21437/SpeechProsody.2024-208.

PIERREHUMBERT, J.; HIRSCHBERG, J. The meaning of intonational contours in the interpretation of discourse. In: COHEN, P. R.; MORGAN, J.; POLLACK, M. E. (ed.). *Intentions in communication*. Cambridge: MIT Press, 1990. p. 271-312. https://doi.org/10.7551/mitpress/3839.003.0016.

POSIT TEAM. RStudio: integrated development environment for R. Version 2026.1.0.392. Boston: Posit Software, PBC, 2026. Software. Available at: https://posit.co/products/open-source/rstudio/. Accessed on: 5 Jan. 2026.

PRIETO, P.; ROSEANO, P. (ed.). *Transcription of intonation of the Spanish language*. München: LINCOM Europa, 2010.

R CORE TEAM. *R*: a language and environment for statistical computing. Version 4.5.2. Vienna: R Foundation for Statistical Computing, 2025. Available at: https://www.R-project.org/. Accessed on: 5 Jan. 2026.

REN, Y.; TAN, X.; QIN, T.; ZHAO, Z.; LIU, T.-Y. *Revisiting over-smoothness in text to speech*. arXiv, 2022. Preprint. Available at: http://arxiv.org/abs/2202.13066. Accessed on: 11 Dec. 2025.

SAN SEGUNDO, E.; LÓPEZ-JAREÑO, A.; WANG, X.; YAMAGISHI, J. *Human perception of audio deepfakes: the role of language and speaking style*. arXiv, 2025. Preprint. https://doi.org/10.48550/arXiv.2512.09221.

SOSA, J. M. *La entonación del español*: su estructura fónica, variabilidad y dialectología. Madrid: Cátedra, 1999.

TAMAYO FLÓREZ, P. A.; MANRIQUE, R.; PEREIRA NUNES, B. HABLA: a dataset of Latin American Spanish accents for voice anti-spoofing. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 24., 2023, Dublin, Ireland. Proceedings of Interspeech 2023. Dublin, Ireland: ISCA, 2023. p. 1963-1967. https://doi.org/10.21437/Interspeech.2023-2272.

TAYLOR, P. *Text-to-speech synthesis*. Cambridge: Cambridge University Press, 2009. https://doi.org/10.1017/CBO9780511816338.

TOKUDA, K.; YOSHIMURA, T.; MASUKO, T.; KOBAYASHI, T.; KITAMURA, T. Speech parameter generation algorithms for HMM-based speech synthesis. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 25., 2000, Istanbul, Turkey. Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul, Turkey: IEEE, 2000. v. 3, p. 1315-1318. https://doi.org/10.1109/ICASSP.2000.861820.

VAN RIJ, J.; WIELING, M.; BAAYEN, R. H. *itsadug*: interpreting time series and autocorrelated data using GAMMs. [s. l.]: CRAN, 2015. R package. https://doi.org/10.32614/CRAN.package.itsadug.

VEILLEUX, N.; BARNES, J.; BRUGOS, A.; SHATTUCK-HUFNAGEL, S. Perceptual foundations for naturalistic variability in the prosody of synthetic speech. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 13., 2012, Portland, Oregon, USA. Proceedings of Interspeech 2012. Oregon, USA: ISCA, 2012. p. 2534-2537. https://doi.org/10.21437/Interspeech.2012-656.

VEILLEUX, N.; SHATTUCK-HUFNAGEL, S.; JEONG, S.; BRUGOS, A.; AHN, B. Machine learning facilitated investigations of intonational meaning: prosodic cues to epistemic shifts in American English utterances. In: INTERNATIONAL CONFERENCE ON SPEECH PROSODY, 12., 2024, Leiden, The Netherlands. Proceedings of Speech Prosody 2024. Leiden, The Netherlands: ISCA, 2024. p. 931-935. https://doi.org/10.21437/SpeechProsody.2024-188.

WAGNER, M.; WATSON, D. G. Experimental and theoretical advances in prosody: a review. *Language and Cognitive Processes*, v. 25, n. 7, p. 905-945, 2010. https://doi.org/10.1080/01690961003589492.

WOOD, S. N. Generalized additive models: an introduction with R. Boca Raton: Chapman and Hall/CRC, 2017. https://doi.org/10.1201/9781315370279.

XU, Y. Speech prosody: a methodological review. *Journal of Speech Sciences*, v. 1, n. 1, p. 85-115, 2011. https://doi.org/10.20396/joss.v1i1.15014.

ZEN, H.; TOKUDA, K.; BLACK, A. W. Statistical parametric speech synthesis. *Speech Communication*, v. 51, n. 11, p. 1039-1064, 2009. https://doi.org/10.1016/j.specom.2009.04.004.

ZOU, Y.; LIU, S.; YIN, X.; LIN, H.; WANG, C.; ZHANG, H.; MA, Z. Fine-grained prosody modeling in neural speech synthesis using ToBI representation. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 22., 2021, Brno, Czechia. Proceedings of Interspeech 2021. Brno, Czechia: ISCA, 2021. p. 3146-3150. https://doi.org/10.21437/Interspeech.2021-883.