Evaluating the Prosodic Diversity of TTS Models for L2 Prosody Assessment

Mariana Julião[1]
HLT, INESC-ID, Lisbon, Portugal

Alberto Abad
HLT, INESC-ID, Lisbon, Portugal
Instituto Superior Técnico, Lisbon, Portugal

Helena Moniz
HLT, INESC-ID, Lisbon, Portugal
Faculdade de Letras da Universidade de Lisboa, Portugal
CLUL, Lisbon, Portugal

Abstract

Recent advances in text-to-speech (TTS) have led to synthetic speech that is often indistinguishable from natural speech at the level of individual utterances. However, it remains unclear whether such systems reproduce the prosodic variability observed in natural speech in a large native-speaker population. This question is particularly relevant for applications in computer-assisted language learning (CALL), as variability is a central property of prosody and a prerequisite for robust assessment.

In this work, we investigate whether TTS can approximate the distribution of prosodic patterns found in native speech, and whether it can serve as a reference for evaluating second language (L2) productions. To this end, we first compare different TTS models to native speakers, and then compare L2 speakers to synthetic speech by the model previously seen as the closest to native speakers.

Our results show that regardless of TTS achieving high perceptual quality, its prosodic variability substantially differs from that of native speakers. As a consequence, comparisons between L2 speech and TTS-based references reveal both the potential and the limitations of using synthetic speech for prosody assessment.

2012 ACM Subject Classification Applied computing → Computer-assisted instruction; Applied computing → Education

Keywords and phrases Speech synthesis, TTS, prosody, L1, L2

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.2

Funding Work supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: `https://doi.org/10.54499/UID/50021/2025`) and UID/PRR/50021/2025 (DOI: `https://doi.org/10.54499/UID/PRR/50021/2025`), by CLUL UID/214/2025 (DOI: `https://doi.org/10.54499/UID/00214/2025`) and by the Portuguese Recovery and Resilience Plan and NextGenerationEU European Union funds under project C644865762-00000008 (ACCELERAT.AI).

1 Introduction

Synthetic speech has reached a level of quality at which individual utterances are often indistinguishable from natural speech. Modern text-to-speech systems can generate highly natural signals across a range of voices and speaking styles. However, this success is typically

[1] Corresponding author

evaluated at the level of isolated utterances or individual speakers. It remains unclear whether these systems exhibit the same properties at a population level. In other words, when considered collectively, do multiple synthetic utterances with diverse timbres behave like a population of natural speakers, or do they reveal systematic biases and reduced variability?

The rapid development of TTS has already led to successful applications in second language learning, as reported in previous work [8, 14, 2, 20]. In these contexts, synthetic speech is primarily used as a production model: learners are exposed to target utterances and are encouraged to imitate them. While effective for pronunciation training, this paradigm does not address a complementary and equally important task, namely the assessment of learners' spontaneous productions.

In the field of CALL, prosody assessment for L2 speech remains a challenging problem. We identify three key limitations in current approaches.

First, many systems rely on imitation-based paradigms, where learners reproduce a reference utterance. While such approaches can be useful in specific cases, particularly for languages with markedly different prosodic systems, to which the learner needs to get acquainted (e.g., tonal languages), they do not capture a learner's ability to produce prosody in a natural communicative context. Mastery of prosody involves not only the correct realization of acoustic patterns, but also the ability to deploy them appropriately depending on linguistic and pragmatic context. Assessment frameworks should therefore move beyond imitation and support the evaluation of unconstrained productions.

Second, prosody exhibits inherent variability. A single sentence with a given communicative intent can be realized through multiple valid prosodic patterns. Effective assessment systems must therefore account for this one-to-many mapping between form and function, rather than enforcing a single canonical realization.

Finally, transparency is a critical requirement for language learning technologies. Learners benefit from feedback that is specific, interpretable, and actionable. Systems that rely solely on global scores or opaque quality metrics provide limited pedagogical value. Instead, prosody assessment should identify which aspects of a learner's production require improvement and how they relate to the target patterns.

We therefore argue that, if TTS systems can faithfully reproduce the variability of native human speech, they may provide a useful basis for overcoming the aforementioned limitations in L2 prosody assessment. Rather than relying on a single human reference or on fixed imitation-based tasks, synthetic speech could make it possible to generate multiple acceptable realizations of the same utterance, better reflecting the one-to-many nature of prosody. However, this potential depends on whether current TTS systems are sufficiently close to native-speaker populations, whether combining different systems can better approximate native variability, and whether synthetic speech can effectively replace human references when assessing learner productions.

The Research Questions (RQs) we address in this article are: (1) How closely does the prosodic diversity generated by different TTS models match that of a native speaker population? (2) Can the combination of multiple TTS models achieve prosodic variability that more closely matches that of a native speaker population? (3) Can synthetic speech be used in the place of human references for L2 prosody assessment?

2 Related Work

The use of TTS in Computer-Assisted Language Learning has been discussed since at least the late 2000s. Early work, such as that of Handley [8], examined its potential for learning French and concluded that, largely due to limited expressiveness, the available TTS systems were not yet suitable for such applications. A decade later, however, the outlook had become markedly more positive.

Bione et al. [1] evaluated the use of TTS for pronunciation instruction with native speakers of Brazilian Portuguese and found that, across most assessment measures, the synthetic voice performed comparably to a human voice, achieving high intelligibility. Similarly, Liakin et al. [14] investigated the use of TTS for learning French liaison and concluded that it constitutes a valuable complement to classroom instruction. Tejedor [18] employed TTS to provide corrective feedback on minimal pairs in a pronunciation training tool, showing that learners using the system significantly outperformed a control group. Overall, these studies primarily focus on segmental aspects of speech.

Other works using TTS included prosodic aspects. The study of Bonces et al. [2] focused on understanding the effect of TTS on the reading of L2. Results indicated a favourable effect, namely in aspects such as linking sounds, pronounciation accuracy, and reading timing. The work of Xiao et al. [20] intended to assess L2 prosody using a native speaker or a TTS reference, by comparing its F0 contours and breaks, to classify speakers as native or non-native. Results showed that the accuracy of the system trained using native speakers is 91.7%, while the accuracy of the system trained using TTS speech is 88.1%. This points towards TTS being sufficiently close to native speech to replace it for this purpose, given its practical advantages. A different application of TTS in prosody-related assessment is presented in the work of Korzekwa [13], who propose a system for pronunciation and lexical stress error detection. The authors start from the observation that annotated data containing lexical stress errors is often scarce, which limits effective model training. To address this issue, they use TTS to generate synthetic examples of such errors.

Hu et al. [11] have investigated the pragmatic capabilities of OpenAI's TTS across three pragmatic conditions: non-contrastive, contrastive, and corrective. The authors found that while TTS is able to produce a variety of F0 contours, these are diverse from the ones realized by humans in all three pragmatic conditions investigated.

3 Methodology

The present study comprises two main components. The first examines the prosodic variability of different TTS models. The second evaluates whether the model exhibiting the highest variability correlates more strongly with high-performing L2 speakers than with low-performing ones, in terms of prosody.

Our methodology involves generating synthetic speech and comparing it with native speech. This comparison is carried out using two approaches. The first computes the similarity between the prosodic features of each TTS utterance and those of all L1 utterances corresponding to the same prompt. The average similarity per speaker is then used as an indicator of prosodic distance from native English. The second approach defines a reference interval based on the prosodic features of L1 speakers and measures the proportion of TTS feature values that fall outside this interval.

For the second component, the methodology consists of comparing L2 utterances with TTS-based references. The validity of this approach is assessed by examining the extent to which the resulting scores align with the proficiency labels assigned to the L2 utterances.

3.1 Speech Synthesis

In this work, we compare three different TTS models: Coqui (XTTS v2 model)[2] [3], CosyVoice (Fun-CosyVoice 3.0 model)[3] [4, 5] and VoxCPM (model 1.5)[4] [23].

For each text, we generated 50 utterances, each leveraging a different reference audio from a unique VCTK (Section 4.1.2) utterance, thereby emulating the voice characteristics of distinct speakers and utterances. Reference audios were randomly sampled from the subset of recordings with a minimum duration of 6 seconds, as required by some of the models. The choice of VCTK is motivated by the fact that it includes a diversity of native speakers of English from different varieties. For CosyVoice, we also used the instruction modality, and randomly assigned a prosody-eliciting prompt from the list in Appendix A.

The synthesized utterances retained this diversity and naturalness. For the same text, the generated utterances exhibited a range of timbres, ages, styles, and dialectal variations. Even at the phonetic level, some synthesized outputs demonstrated pronounced rhoticity, while others did not. This variability is particularly valuable, as it enables the assessment of speakers without bias toward a single specific variety of English.

3.2 Features

We extracted prosodic features and applied perceptually motivated normalization techniques for energy and pitch. Word and phone boundaries were identified using the Montreal Forced Aligner (MFA) [15], which demonstrated robust results for L1, L2 and TTS speech. The type of speech representation considered is envisioned for the comparison of utterances of the same text (or of texts of the same prosodic structure). Therefore, it is expected that the n^{th} point of the feature vectors of two utterances compared corresponds to the same portion of the same word.

3.2.1 Pitch

To extract pitch, we used pyYAAPT[5] [22], which extracts F0 at 10 ms steps. We use pyYAAPT for pitch extraction because it provides a robust implementation of the YAAPT algorithm, which combines time-domain and frequency-domain information to estimate fundamental frequency. This hybrid approach makes it less sensitive to common pitch-tracking errors, such as octave jumps, voicing misclassification, and irregularities in the acoustic signal.

We aimed to follow the methodology of Elvira [6], which involved prosody comparison by extracting three points per vowel – at the beginning, middle, and end. However, this approach proved unfeasible in our case due to frequent vowel reductions, particularly in the speech of advanced speakers, which often precluded the collection of three distinct points. To address this, we opted to extract $N \times 3$ points per word as a proxy, where N represents the number of vowels. A potential advantage of this approach is that it includes voiced consonants, which carry relevant prosodic information, particularly through pitch-bearing sonorants such as nasals and liquids, as well as through duration and intensity cues.

Since our focus is on perceived pitch rather than absolute F0 values, we converted F0 measurements into semitones. This transformation provides a more accurate representation of perceived intonation and enables fair comparisons between utterances from speakers with

[2] https://docs.coqui.ai/en/latest/models/xtts.html
[3] https://huggingface.co/FunAudioLLM/Fun-CosyVoice3-0.5B-2512
[4] https://github.com/OpenBMB/VoxCPM
[5] http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html

different vocal pitch ranges. To compare the overall intonation of entire utterances, we applied utterance-wise pitch normalization by subtracting the lowest pitch value across the utterance.

3.2.2 Loudness

For loudness, we applied the same principles as for pitch. Specifically, we used the Loudness LLD from GeMAPS [7], which provides an "estimate of perceived signal intensity based on an auditory spectrum." We chose loudness over other energy metrics to ensure relevance from a perceptual perspective. As with pitch, we extracted $N \times 3$ points per word, where N represents the number of vowels. Utterance-wise normalization was performed by dividing the loudness values by the maximum value within each utterance.

3.2.3 Duration

Using the alignment boundaries, we measured the duration of individual words as well as the intervals between them. We also examined phone- and vowel-level durations; however, word-level durations proved to be more informative.

3.3 Metrics

In the first part of this work, we assessed TTS models having L1 speakers as reference; in the second part, we assessed L2 speech having TTS speech as a reference. Therefore, in the next subsections, we refer to different types of utterances as test and reference.

3.3.1 Similarity Value

Each pair of utterances was compared using Pearson's correlation coefficient, ρ, in an approach similar to that of Elvira-García et al. [6]. For each probe utterance utt_i, a value of Similarity S_i is computed as

$$S_i = \frac{1}{M} \sum_{j=1}^{M} \rho(utt_i, utt_j)\,, \tag{1}$$

where M is the number of reference utterances utt_j with the same text as utt_i.

3.3.2 Reference Interval

We also computed a reference interval (RI) for each utterance and each feature, with the aim of characterizing the range of acceptable feature values for a given utterance. In a production setting, this metric could be used to indicate where and how a learner's production diverges from the reference distribution. We adopted a non-parametric approach, defining the lower and upper limits of each RI as the 2.5th and 97.5th percentiles, respectively [12]. We define the Outlier Ratio R as the ratio of feature points falling within the RI. This ratio provides an additional measure of how closely the prosodic profile of a test utterance aligns with that of the reference utterances.

4 Experiments

4.1 Data

4.1.1 Speech Accent Archive

The Speech Accent Archive (SAA) [19] is a publicly accessible database that contains audio samples of speakers from various linguistic backgrounds reading the same scripted passage in English, the phonetically-rich paragraph *Please call Stella*.

Due to the large number of native English speakers reading this passage, the corpus enables the characterization of the range of acceptable realizations of a single text. We sampled 200 speakers to represent a native population.

4.1.2 VCTK

The *CSTR VCTK* corpus [21] is a corpus for voice cloning that comprises 108 English speakers, with average age of 22.7 years, of different varieties of English, with approximately 400 utterances per speaker. These correspond to a phonetically-rich paragraph, common to all speakers, and a newspaper text which is different for all speakers.

4.1.3 BNA

The Brave New Approach corpus (BNA) is a combination of the L2 corpora AUWL [9], ISLE [16] (train set), and C-AuDiT [10] (test set) specifically curated for the Paralinguistics Sub-Challenge of Interspeech 2015 [17]. AUWL and ISLE have been annotated with the same method: each speech file was annotated by five phoneticians with respect to its prosody (defined as sentence melody and rhythm) on a five-level scale, where 1 is *normal* for a native speaker and 5 is *very unusual* for a native speaker, which has then been averaged. The annotation of C-AuDiT, on the other hand, ranges from 0 to 2, where 0 is *good* and 2 is *bad*. The train set includes 3,980 utterances, while the test set includes 999 utterances. Only the train set is considered in the current work.

4.2 Model Comparison

We compared the TTS models by computing their Similarity and Outlier Ratio, having the native speech obtained from the SSA database as a reference. We also combined the utterances from different models and computed the Outlier Ratio, R. TTS models use the speakers of the VCTK corpus as target speakers to generate the samples.

4.3 TTS for L2 assessment

After selecting the TTS model or combination of models yielding the best results, we evaluate its potential for L2 prosody assessment. This is achieved by computing a similarity score for each L2 utterance with respect to the selected model, and subsequently comparing these scores to the corresponding prosodic labels using Spearman's rank correlation coefficient. This metric has been widely used in prior work on the BNA corpus [17].

The experiments are conducted at both the utterance level and the aggregated label level. In the latter case, similarity scores are averaged across utterances sharing the same label.

Table 1 Similarity values (↑) between each model and the set of L1 spekers, for each feature. CosyVoice_P – CosyVoice with prosodic instructions.

Models	Coqui	CosyVoice	CosyVoice_P	VoxCPM
Pitch	0.540	**0.541**	0.527	0.476
Loudness	**0.644**	0.634	0.643	0.634
Duration	0.968	**0.974**	0.973	0.973
Average	***0.718***	*0.717*	*0.714*	*0.694*

Table 2 Outlier ratio (R) (↓) for each model or model combination, based on the set of L1 speakers, for each feature. Cq – Coqui; Cv – CosyVoice; Cvp – CosyVoice with prosodic instructions; VoxCPM – V.

Models	Cq	Cv	Cvp	V
Pitch	0.062	0.049	0.040	0.068
Loudness	0.046	0.128	0.125	0.094
Duration	0.029	0.008	0.015	0.025
Average	***0.046***	*0.062*	*0.060*	*0.062*

Models	Cq+Cv	Cq+Cvp	Cq+V	Cv+Cvp	Cv+V	Cvp+V
Pitch	0.049	0.040	0.068	0.040	0.068	0.068
Loudness	0.128	0.125	0.094	0.125	0.094	0.094
Duration	0.008	0.015	0.025	0.015	0.025	0.025
Average	*0.062*	*0.060*	*0.062*	*0.060*	*0.062*	*0.062*

Models	Cq+Cv+Cvp	Cq+Cv+V	Cq+Cvp+V	Cv+Cvp+V	All
Pitch	0.040	0.068	0.068	0.068	0.068
Loudness	0.125	0.094	0.094	0.094	0.094
Duration	0.015	0.025	0.025	0.025	0.025
Average	*0.060*	*0.062*	*0.062*	*0.062*	*0.062*

5 Results and Discussion

5.1 Model Comparison

The results of TTS model evaluation regarding similarity S values with L1 are in Table 1. The results of TTS model and model combination evaluation regarding Outlier Ratio R values with L1 are in Table 2. For clarity, in both cases, we also provide the average of the values of the three features considered.

The results indicate that, although certain models outperform others for specific features, overall performance differences remain relatively limited. For instance, Coqui achieves the best results for loudness, while CosyVoice performs best for duration.

We also observe that combining models, which increases the number of samples used to define the reference interval, does not generally lead to improved performance. Although combining outputs from different TTS systems led to small changes in the Outlier Ratios, these differences did not indicate a consistent or substantial improvement over the best individual systems. This suggests that aggregating several generative models may increase the number of synthetic samples without substantially expanding the prosodic space in the dimensions that matter for native-like variability. In other words, model diversity at the system level does not automatically translate into prosodic diversity at the acoustic-feature level.

Table 3 Spearman's correlation coefficient between similarity scores S and prosody labels, for utterance level and aggregated-label level. p-values are represented by asterisks according to significance: * : $p > 5 \times 10^{-2}$; ** : $1 \times 10^{-5} < p \leq 5 \times 10^{-2}$; *** : $p \leq 1 \times 10^{-5}$.

	Utterance Level	**Aggregated Label Level**
Pitch	0.062^{**}	0.745^{**}
Loudness	0.214^{***}	0.973^{***}
Duration	0.205^{***}	0.982^{***}

Furthermore, CosyVoice with prosodic instruction stands out only in achieving the highest correlation for pitch, with no comparable gains for other features. Notably, the best overall average results are obtained with Coqui, the oldest model considered in this study. This finding cautions against assuming that more recent or larger TTS models are necessarily better suited for modelling native prosodic variability. One possible explanation is that newer systems may produce more stable or regularized outputs, reflecting optimization for naturalness, intelligibility, or speaker consistency rather than for variability across acceptable prosodic realizations. Coqui's comparatively stronger performance may therefore result from its preserving a wider range of acoustic variation, especially in energy-related dimensions. Nevertheless, this explanation should be interpreted cautiously, since the models differ in multiple respects, including architecture, training data, synthesis pipeline, and default generation settings.

These findings suggest that, despite substantial advances in TTS quality, recent models do not appear to improve in terms of prosodic variability. Instead, progress seems to be primarily driven by improvements in the quality of individual utterances rather than in the diversity of their realizations.

5.2 TTS for L2 assessment

To address whether synthetic speech can be used as a reference for L2 prosody assessment, we selected Coqui, the best-performing model in the previous experiments. We evaluated the relationship between similarity scores (S) – computed via the average Pearson's correlation between TTS and L2 utterances – and the human-assigned prosodic labels. This was computed using Spearman's correlation coefficient, and assessed both at the utterance level and at the aggregated label level. Results are in Table 3.

On an utterance-by-utterance basis, the correlation between TTS similarity and proficiency was relatively low. For loudness and duration, correlation values hovered around 0.2, while pitch showed almost no significant relationship (near 0.0). This indicates that a high similarity to a TTS reference does not always guarantee a high human rating for a specific utterance.

When similarity scores were aggregated by label (averaging scores for all utterances of a given label), the correlations became strikingly strong. For loudness and duration, the absolute Spearman's rank correlation reached 0.973 and 0.982, respectively. For pitch, it reached 0.745. This suggests that, while TTS cannot reliably assess the quality of individual utterances, it accurately captures the general prosodic trends associated with increasing L2 proficiency.

Table 4 Average Similarity Value (S) between L2 utterances and TTS references, and number of outlier points (# Outliers) to the reference interval, for Utt1.0, Utt2.4 and Utt3.2.

		Pitch	Loudness	Duration
S	Utt1.0	0.478	0.800	0.969
	Utt2.4	0.617	0.593	0.968
	Utt3.2	0.346	0.817	0.950
# Outliers	Utt1.0	8	5	3
	Utt2.4	0	12	6
	Utt3.2	4	2	8

5.3 Reference Intervals for Local Analysis of L2 Speech

To better understand the results previously presented, we focus on utterances corresponding to the prompt *I'd like a one-way ticket to London, please*: one with highest rating (Utt1.0), one with average-good rating (Utt2.4) and one with average-low rating (Utt3.2). We plotted the feature sequences alongside the TTS references in Figure 1. Table 4 presents the values of Average Similarity and # Outliers for these utterances.

For pitch, Utt1.0 has 8 outliers: twice as much as Utt3.2. Utt2.4, however, does not have any outliers. The different sequences and the centroid appear to have mostly the same monotony – Utt1.0's outliers being mostly due to major shifts upward and downward, which can be seen as just an emphasized intonation by a proficient speaker. Utt3.2, nevertheless, presents jumps that are in the opposite direction as the rest, which is mostly visible right at the beginning.

For loudness, however, Utt2.4 is the one with the most outliers, which often correspond to an exaggeration of the tendency peaks of the distribution. Utt3.2 yields the best performance having the lowest number of outliers.

For duration, the number of outliers correlates to the labels. The fact that the intervals between words are considered allows for differentiation between good and worse sequences – normally, due to hesitations, worse speakers can spend more time between words. This is the case of Utt3.2 between the words *London* and *please*. However, this example seems to also confirm the idea that less proficient speakers have a lower speech rate and spend more time uttering words, as pointed out by the outliers at the top of the plot.

6 Conclusion

We first examined the similarity between the prosodic feature distributions of different TTS models and those of native speakers. We then evaluated the potential of the best-performing model to serve as a reference for L2 prosody assessment.

Our results show that trends observed at the aggregate level do not consistently hold at the utterance level. Individual utterances with higher proficiency scores do not systematically exhibit stronger alignment with the TTS distribution, nor do they consistently present fewer deviations from the reference interval. A more detailed analysis further reveals that local behavior is highly variable, limiting the reliability of TTS-based comparisons at the utterance level.

These findings are consistent with our qualitative observations. While synthetic speech exhibits some variability, it also displays noticeable regularities across utterances generated from the same prompt. In contrast, low-scoring L2 productions tend to diverge substan-

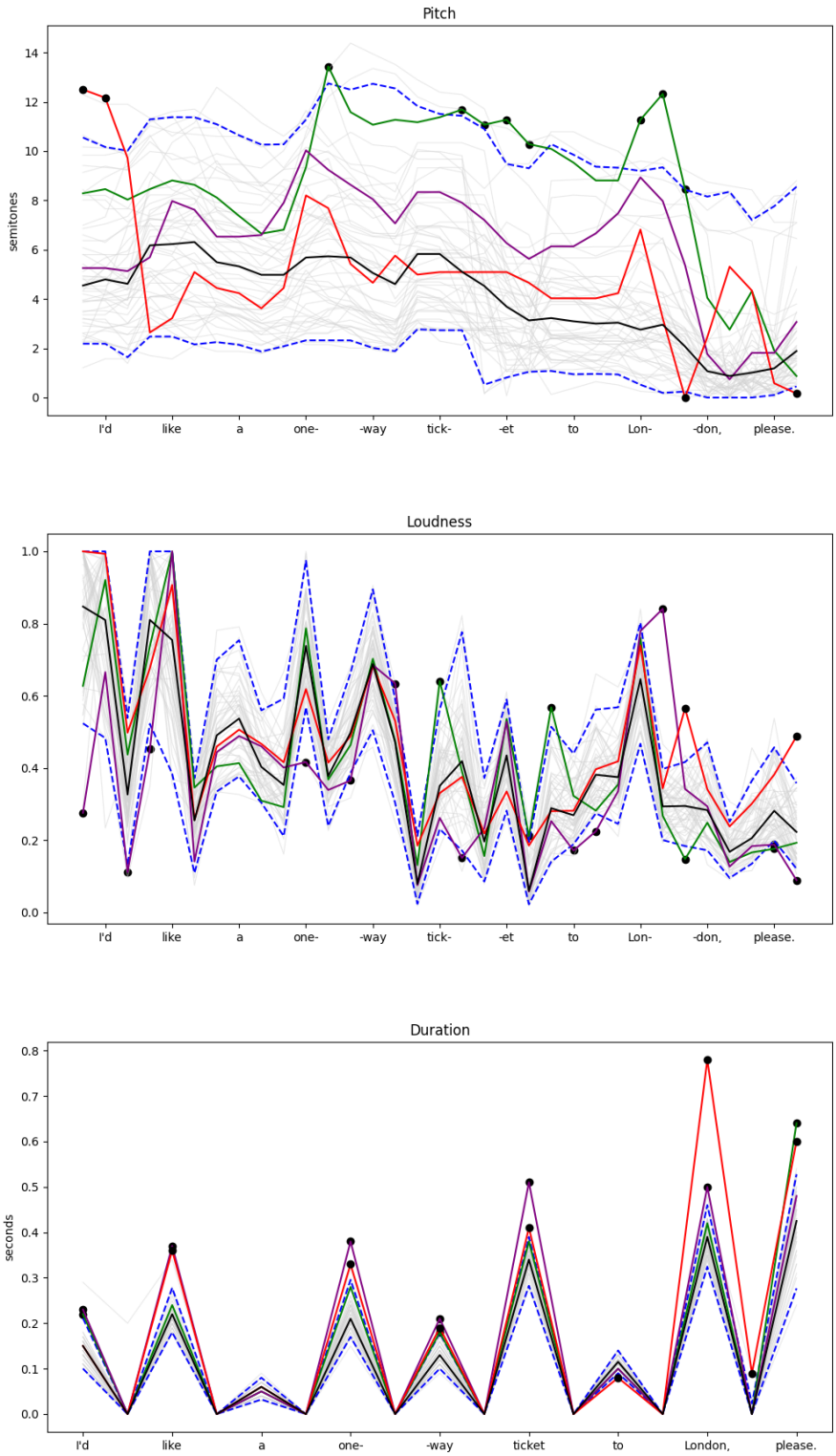

Figure 1 Pitch, Loudness and Duration sequence representation for prompt *I'd like a one-way ticket to London, please.* – TTS samples; — **centroid**; - - 2.5% and 97.5% reference intervals; – Utt1.0; – Utt2.4; – Utt3.2; • point falling outside reference interval.

tially from synthetic speech, whereas high-scoring productions, although generally closer to TTS, often achieve correctness through patterns that differ from the synthesized trend. In some cases, high-scoring utterances even exaggerate these patterns, resulting in outliers, as illustrated by the pitch distribution in Figure 1.

With respect to our research questions, we find that (RQ1) different TTS models do not differ substantially in their overall proximity to native speech, although individual models may perform better for specific prosodic features. Regarding RQ2, combining outputs from multiple models does not increase prosodic diversity in a way that more closely approximates native variability. Finally, for RQ3, our results suggest that synthetic speech cannot yet reliably replace human references for L2 prosody assessment, as its limitations at the utterance level undermine its effectiveness for this purpose.

We argue that these limitations stem from the design objectives of current TTS systems. Such systems are primarily optimized for perceptual naturalness at the level of individual utterances, rather than for capturing the full range of variability present in natural speech. Consequently, the variability observed in synthetic speech appears to be an incidental by-product rather than a controlled or representative feature.

References

1 Tiago Bione, Jennica Grimshaw, and Walcir Cardoso. An evaluation of TTS as a pedagogical tool for pronunciation instruction: the 'foreign'language context. *CALL in a climate of change: adapting to turbulent global conditions–short papers from EUROCALL*, pages 56–61, 2017.

2 Jeisson Alonso Rodríguez Bonces. Implementing text-to-speech technology as a means of enhancing L2 reading fluency. *EDU REVIEW. International Education and Learning Review/Revista Internacional de Educación y Aprendizaje*, 1(2):61–73, 2019.

3 Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. *arXiv preprint arXiv:2406.04904*, 2024. arXiv:2406.04904.

4 Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024. arXiv:2407.05407.

5 Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training. *arXiv preprint arXiv:2505.17589*, 2025. arXiv:2505.17589.

6 Wendy Elvira-García, Simone Balocco, Paolo Roseano, and Ana Ma Fernandez-Planas. ProDis: A dialectometric tool for acoustic prosodic data. *Speech Communication*, 97:9–18, 2018. doi:10.1016/J.SPECOM.2017.12.013.

7 Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015. doi:10.1109/TAFFC.2015.2457417.

8 Zöe Handley. Is Text-to-Speech synthesis ready for use in Computer-Assisted Language Learning? *Speech Communication*, 51(10):906–919, 2009. doi:10.1016/J.SPECOM.2008.12.004.

9 Florian Hönig, Anton Batliner, and Elmar Nöth. Automatic assessment of nonnative prosody–annotation, modelling and evaluation. In *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, pages 21–30, 2012.

10 Florian Hönig, Anton Batliner, Karl Weilhammer, and Elmar Nöth. Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners. In *International Workshop on Speech and Language Technology in Education*, 2009.

11 Na Hu, Jiseung Kim, Riccardo Orrico, Stella Gryllia, and Amalia Arvaniti. Can OpenAI's TTS model convey information status using intonation like humans? In *Speech Prosody 2024*, pages 32–36, 2024. doi:10.21437/SpeechProsody.2024-7.

12 Kiyoshi Ichihara, James C Boyd, IFCC Committee on Reference Intervals, and Decision Limits (C-RIDL). An appraisal of statistical procedures used in derivation of reference intervals. *Clinical chemistry and laboratory medicine*, 48(11):1537–1551, 2010.

13 Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek. Computer-assisted pronunciation training — Speech synthesis is almost all you need. *Speech Communication*, 142:22–33, 2022. doi:10.1016/J.SPECOM.2022.06.003.

14 Denis Liakin, Walcir Cardoso, and Natallia Liakina. The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning*, 30(3-4):325–342, 2017.

15 Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502, 2017. doi:10.21437/Interspeech.2017-1386.

16 Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. The ISLE corpus of non-native spoken English. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, pages 957–964. European Language Resources Association, 2000.

17 Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, J. R. Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, Parkinson's & eating condition. In *Interspeech 2015*, pages 478–482, 2015. doi:10.21437/Interspeech.2015-179.

18 Cristian Tejedor-Garcıa, David Escudero-Mancebo, César González-Ferreras, Enrique Cámara-Arenas, and Valentín Cardeñoso-Payo. Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs. *Proc. SLaTE, Stockholm, Sweden*, pages 26–30, 2017.

19 Steven Weinberger. Speech Accent Archive. Retrieved from http://accent.gmu.edu., 2015.

20 Yujia Xiao and Frank K Soong. Proficiency Assessment of ESL Learner's Sentence Prosody with TTS Synthesized Voice as Reference. In *INTERSPEECH*, pages 1755–1759, 2017. doi:10.21437/INTERSPEECH.2017-64.

21 Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), 2019. [sound]. doi:10.7488/ds/2645.

22 Stephen A Zahorian and Hongbing Hu. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559–4571, 2008.

23 Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. VoxCPM: Tokenizer-Free TTS for Context-Aware Speech Generation and True-to-Life Voice Cloning. *arXiv preprint arXiv:2509.24650*, 2025. doi:10.48550/arXiv.2509.24650.

A List of Prosodic Prompts

1. Speak with quiet relief, as if a problem has just been resolved
2. Speak with mild frustration, while trying to remain polite
3. Speak with genuine enthusiasm and positive energy
4. Speak with a calm and reassuring tone
5. Speak with subtle anxiety beneath otherwise neutral wording
6. Speak with restrained excitement, trying not to overreact

7. Speak with quiet disappointment, keeping emotions controlled
8. Speak with gentle amusement, as if something is slightly funny
9. Speak with a serious and focused tone
10. Speak with a relaxed and informal attitude
11. Speak as if carefully choosing your words to avoid misunderstanding
12. Speak as if thinking aloud, including slight natural hesitations
13. Speak as if recalling something from memory with slight uncertainty
14. Speak with clear confidence and decisiveness
15. Speak as if double-checking your own reasoning while talking
16. Speak with slight hesitation, as if unsure about the statement
17. Speak as if explaining something you have just understood
18. Speak with a reflective and thoughtful tone
19. Speak with a sense of mild confusion
20. Speak with deliberate and precise articulation
21. Speak as if explaining something to a child
22. Speak as if giving instructions in a clear and direct way
23. Speak as if making a polite request
24. Speak as if defending a point in a discussion
25. Speak as if sharing a secret
26. Speak as if telling a short anecdote
27. Speak as if emphasizing an important point
28. Speak as if downplaying the importance of the message
29. Speak as if offering a suggestion rather than a command
30. Speak as if making a careful clarification
31. Speak as if talking to a close friend
32. Speak as if talking to someone you do not know well
33. Speak as if addressing a group of people
34. Speak as if speaking to someone you respect
35. Speak as if trying to be especially polite
36. Speak as if trying not to offend the listener
37. Speak with a slightly formal and professional tone
38. Speak with a warm and friendly tone
39. Speak with a neutral and detached tone
40. Speak with a slightly playful tone
41. Speak in a quiet environment, using a soft voice
42. Speak in a noisy environment, projecting your voice clearly
43. Speak as if you are slightly out of breath
44. Speak as if you are tired but still engaged
45. Speak with steady rhythm and even pacing
46. Speak with slightly faster-than-normal tempo
47. Speak with slightly slower-than-normal tempo
48. Speak with clear emphasis on key words
49. Speak with reduced emphasis, keeping the tone relatively flat
50. Speak with natural variation in pitch and rhythm