Can self-supervised speech models predict the perceived acceptability of prosodic variation?

Sarenne Wallbridge*, Adaeze Adigwe*, Peter Bell
Centre for Speech Technology Research, University of Edinburgh, UK

***Abstract*—Though producing an appropriate prosodic realisation of text is a one-to-many problem, modern speech generation often focuses on identifying the "best" or "most likely" output, overlooking acceptable variation across realisations. How listeners perceive such variation–and whether models capture it–is unaccounted for in current evaluation paradigms. In this study, we present exploratory analyses of whether self-supervised models encode acceptable prosodic variation. Using a new dataset of relative acceptability ratings across carefully controlled, high-quality synthetic utterances, we show that SSL representations contain information predictive of such judgments. By introducing a novel method for deriving probability-based uncertainty from autoregressive speech models, we examine whether this information is available in an unsupervised setting, highlighting the complexity of prosodic perception and the value of more human-centric evaluation paradigms.**

***Index Terms*—self-supervised speech representation, speech synthesis evaluation**

I. Introduction

PSYCHOLINGUISTIC research shows that language users make gradient judgements about how well an utterance matches their expectations [1], [2]. This sensitivity allows for variation in how utterances are realised. For example, prosodic variation—shifts in pitch, rhythm, and intensity—can yield distinct realisations of the same utterance that are equally acceptable in a given discourse context or for conveying the same intent [3], [4], [5]. Yet most paradigms for evaluating language generation ignore this permissible variability. As synthetic speech nears human-like quality, understanding the prosodic variation is increasingly important to improve generation methods and models of human language comprehension.

Recent studies of text-to-speech (TTS) evaluation recognise that widely used subjective paradigms like Mean Opinion Score (MOS) or Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)—which aim to capture fidelity or fluency—cannot assess whether a particular prosodic variant is more, less, or equally acceptable to another [6], [7], [8]. A new subjective task of *relative acceptability judgements*, where listeners compare multiple prosodic candidates of the same text, instead places different prosodic realisations along a continuum of acceptability. Their relative positions allow us to charaterize the wider distribution of permissible prosodic variation and its boundaries [9].

In this paper, we present an exploratory analysis of the dataset of relative acceptability judgements collected by [9]. The stimuli have been carefully synthesised to control for confounding factors of lexical content, speaker, and generation quality, thus isolating the effect of prosodic variation on percieved acceptability. Models trained with self-supervised learning (SSL) such as wav2vec and wavLM encode complex patterns in their training data spanning phonetic details, speaker traits, and aspects of prosodic information [10], [11], [12]. They have also been shown to be strongly predictive of MOS preference scores; however, it is unclear whether prosodic acceptability contributes to this performance [13], [14]. Do these models reflect the permissible variation across prosodic realisation that human listeners exhibit when the quality of underlying samples is high and the main component of variation is prosody?

To answer this question, we conduct two exploratory analyses: in the first, we demonstrate that SSL-derived features are useful for predicting relative acceptability scores, suggesting that SSL models encode aspects of permissible prosodic variation; in the second, we ask whether this information can be accessed in a zero-shot manner through SSL probability assignments. For this, we propose a new method for extracting probabilities from autoregressive SSL models to score speech samples. Though such scores correlate well with human MOS preference judgements, their relationship to relative-acceptability judgements is weaker and more complex. Our findings have implications for TTS evaluation, and contribute to the growing body of research investigating whether SSL models reflect human speech perception by offering a new method to explore the statistical patterns encoded through SSL training.

II. Background

A. Evaluating human perception of prosody

A critical element in determining whether listeners perceive speech as human-like is the appropriateness of

* for co-authorship.

its prosodic realisation—in other words, does it match their expectations? Perceptual studies have demonstrated that listeners are tolerant of prosodic variation over fixed lexical content. For example, multiple realisations can be judged as achieving the same communicative function or as acceptable in a given discourse context [15], [2], [4] and listeners can imagine contexts in which diverse prosodic renditions are valid [16]. This tolerance is mirrored in speech production and re-enactment paradigms where numerous prosodic interpretations can be licensed by the same underlying sentence [17], [5].

Though its role has been well-established, acceptable variation in prosodic realisation is not well-accounted for in standard subjective evaluations of synthetic speech. Perhaps the most widely used metric, MOS, reflects aggregate listener impressions of quality or naturalness. However, it is sensitive to the sample set [18], susceptible to range-equalizing bias [19], and prone to score-saturation for high-quality samples [8]. As such, it is too coarse to investigate nuanced prosodic differences [8], [7]. Moreover, it remains unclear whether listeners prioritize naturalness, intelligibility, or audio quality in their ratings [8], leading some papers to claim that synthetic samples are "more natural" than human speech [20]. MUSHRA testing uses side-by-side comparisons and fine-grained scoring, making it more sensitive than MOS; however, this paradigm relies on judging proximity to reference samples, meaning it cannot capture permissible variability when multiple prosodic realisations are equally felicitous. Other reference-based paradigms like ABX-testing suffer from the same problem [21], [7]. Relative acceptability judgments, where participants rank or rate multiple realisations of the utterance without a reference, can provide novel insights into how acceptability is distributed across prosodic variation.

B. Prosodic information in SSL representations

Benchmarks like SUPERB-prosody [11] demonstrate that SSL representations can be used for tasks that involve prosody such as prominence, emotion, and sarcasm detection, while [22] shows that SSL models encode abstract suprasegmental categories (e.g., tone, stress) that don't directly map to surface acoustics like F0.

Recent work has begun to explore whether these representations align with more explicit human judgments of prosody. For example, [23], [24] report moderate correlations between SSL embeddings and listener ratings for prosodic and pragmatic similarity across utterances with shared lexical content, while the ProsAudit benchmark provides direct comparisons between human and model judgements of structural prosody [25]. These studies primarily focus on whether SSL models reflect human judgments of perceived similarity. However, the question of whether SSLs reflects human perception of acceptability across *prosodic variation* (i.e., reference-free evaluation) has yet to be addressed.

C. Modeling listener judgments with SSL models

Recent work has shown that SSL-based models are effective at predicting MOS scores for synthetic speech samples, with lightweight prediction architectures using SSL representations as features performing well in benchmarks like the VoiceMOS Challenge [14], [26]. This effectiveness extends to zero-shot settings where the prediction task has been framed as out-of-distribution detection, showing correlation between uncertainty scores of pre-trained SSL models and MOS scores [27], [28].

However, predicted MOS scores inherit the noise and biases of the human MOS labels discussed in Section II-A. As such, it remains unclear what aspects of human judgements are being leveraged by MOS prediction models or encoded by SSL models. Targeted evaluations of how subtle characteristics of speech affect human judgements are required to assess which aspects of speech are represented in modern speech encodings.

D. Estimating uncertainty from SSL models

The probabilities assigned to upcoming units of text by self-supervised language models are used extensively in NLP. They form the basis of perplexity—a standard evaluation metric for text-model quality [29]—and numerous studies show a strong relationship between model-estimates of word-level surprisal and human language comprehension behaviours like reading times [30], [31], [32]. However, extracting probability assignments from models of speech is more complicated. For example, autoregressive SSL models of text are trained to predict upcoming units from a token vocabulary that maps directly to their textual input, such as wordpieces of BPE units [33], [34]; speech cannot be discretised as easily.

Only a handful of works have explicitly attempted to extract probabilities from SSL speech models, employing different methods and models. [28] extract probabilities from the logits of frame-level prediction in wav2vec models as an unsupervised proxy for MOS prediction, showing that poor quality samples can be detected as anomalies from high model uncertainty. [27] take an approach closer to text-base probability estimation: using SSL models like HuBERT to encode speech into a sequence of tokens, they train a language model on such token sequences and can then compute generation probabilities under the model. They find correlations between MOS scores and such probabilities across model architectures and encoders. This approach has also been applied to probe whether SSL models encode phonetic, lexical, syntactic, and prosodic information [35], [25]. Given that the underlying token sequences stem from bi-directional contextual representations, we note that these

are pseudo-probabilities [36]. It is unclear what aspects of the speech signal are encoded through the different extraction methods. In this work, we propose a novel method for obtaining true probability estimates to better understand what each method may encode.

III. Datasets

We use two datasets of preference judgements. Both reflect listener perceptual judgments–*VoiceMOS'22* involves absolute quality ratings across diverse systems while *PrefRank* provides relative preference rankings of prosodic variants with comparable quality from a single TTS model. As such, they provide complementary insights for modelling human speech preferences.

PrefRank. In a study of TTS evaluation for stochastic generative models, [9] collected listener judgments of synthetic utterances produced by by a SOTA generative model, ParlerTTS [37]. Listeners ranked four different generations of the same underlying lexical content, including utterances from both read and conversational material ($N_{text} = 120$). The data collection pipeline for *PrefRank* provided a high degree of control over speaker identity, speech quality, and intelligibility, ensuring that listener judgements were primarily driven by variation in prosodic realisation. Listener rankings were aggregated using a probabilistic Bradley–Terry–Luce (BTL) model, which estimates a relative score θ_i for each rendition based on its ranking outcomes [38]. The resulting *BT scores* reflect the relative acceptability of prosodic realizations across renditions.

VoiceMOS'22. Initiatives like the VoiceMOS Challenge [26] aim to advance the automatic prediction of listener preferences by benchmarking models against human ratings. We use the VoiceMOS 2022 challenge as a point of comparison for the *PrefRank* dataset. *VoiceMOS'22* comprises crowdsourced MOS evaluations of synthetic speech samples from systems submitted to the Blizzard Challenge over a decade [39], [40].

IV. Experiment A: Do SSLs encode acceptable variation across prosodic realisations?

We begin by examining whether SSL representations are predictive of the perceptual acceptability of prosodic variations in synthetic speech: if a simple supervised model can successfully predict humans' relative rankings, it suggests that SSL models indeed encode cues relevant to perceived prosodic acceptability. We then test whether popular MOS predictors are useful for predicting acceptability judgments.

A. Experimental Set-up

Predicting Relative Acceptability. We implement a listwise ranking model (ListNet [41]) to predicts scores $\{\theta_1, \theta_2, \theta_3, \theta_4\}$ over sets of renditions $\{r_1, r_2, r_3, r_4\}$ rather than scoring renditions individually. This approach models the probability that each rendition is ranked highest. It converts both the true normalized BT scores (θ_i) and the model's predicted (s_i) into probability distributions over the set using a softmax function and learns to minimize the cross-entropy between them.

Following MOS-SSL-Net [14], we use a lightweight predictor architecture to probe the value of SSL features more directly. Given the relatively small size of *PrefRank*, we perform 5-fold cross-validation with an 80/20 split for training and testing in each fold. All models are trained for 10 epochs, and performance metrics are reported as the average across all five folds.

We compute correlation metrics per stimulus set between the predicted and ground truth BT scores; results are averaged across all stimulus sets in the test set and across all K-folds. We report widely used rank correlation metrics: Spearman's Rank Correlation Coefficient (SRCC), Kendall's Tau (τ), and Top-1 Accuracy. Additionally, to account for variance/spread in BT scores, we also report Weighted Kendall's Tau (τ_w) which penalizes misrankings more heavily for samples with higher human acceptability ratings.

SSL Representations. We test several speech foundation models, which are commonly used to benchmark various downstream speech processing tasks: HuBERT [42], WavLM [43], vq-wav2vec [44], wav2vec2.0-960H [45]and Wav2vec2.0-SWBD . Before feature extraction, all audio samples were downsampled to 16 kHz. For each utterance, we extract feature vectors from the last hidden layer of the pre-trained SSL model with dimensions $D \times L$, where D is the hidden layer dimension (e.g., 1024 for all models apart from vq-wav2vec) and L is the sequence length. Following previous work of deriving SSL features for previous downstream speech tasks [11], [24], we applied average-pooling across the time dimension to obtain a fixed-size utterance-level representation. Though mean pooling across the time axis is a relatively crude operation, such representations have been shown to maintain prosodic information. As an acoustic baseline, we used the ComParE-EGEMAPS acoustic features, a comprehensive set of 88 low-level descriptors related to pitch, temporal information, energy, and spectral characteristics extracted via the openSMILE toolkit [46]. We do not expect these features to reflect the complex relationship between prosodic variation and acceptability.

MOS Prediction Models. Additionally, we investigate whether off-the-shelf MOS prediction models can generalize to *PrefRank* acceptability judgments. For each stimulus set, we predict MOS scores (s_i) for each rendition using various automatic MOS models implemented in [47]. We convert scores to rankings and compute their

Model	Top-1	τ	τ_w	SRCC
WAV2VEC 2.0	0.325	0.190	0.189	0.227
WAV2VEC2.0-SWBD	**0.500**	0.329	0.351	0.376
WAVLM	0.458	0.335	0.344	0.394
HUBERT	0.492	0.282	0.316	0.313
VQ-WAV2VEC	0.492	**0.353**	**0.374**	**0.400**
EGEMAPS	0.292	-0.005	-0.011	-0.006

TABLE I: Listwise model performance of SSL models and EGEMAPS on predicting relative acceptability in *PrefRank*.

correlation with ground-truth Bradley-Terry (BT) scores. Since we do not conduct any fine-tuning, we perform our evaluation on all stimulus sets in *PrefRank*.

B. Predictive Power of SSL Representations

The listwise learning presented in Table I show that SSL representations are indeed predictive of relative acceptability, indicating that these models encode information relevant for relative acceptability. We find moderate and comparable correlations across all SSL models except wav2vec 2.0 which achieves slightly weaker correlation. This is an encouraging finding, especially considering our relatively small training dataset and simple predictor architecture. We observed that the dispersion in BT scores was subtly reflected in the reported Weighted Tau τ_w values, highlighting the importance of selecting informative evaluation metrics.

Despite being a less parameterized model than wav2vec2.0, vq-wav2vec exhibits comparable performance to the other SSL models. Future analysis should consider whether correlation is related to factors such as model capacity, training data characteristics, or the specific underlying pre-training task (e.g., wav2vec2.0's contrastive loss versus HuBERT's masked prediction objective). We also find a difference in performance between wav2vec2.0 trained on LibriSpeech and wav2vec2.0-SWBD trained on conversational Switchboard data, suggesting that domain differences in the training data for speech foundation models may influence the prediction task. As anticipated, the EGEMAPS acoustic baseline showed only marginal performance $\tau_w \approx 0$. The prosodic variation exhibited in *PrefRank* depends on complex prosodic structures that are not present in low-level acoustic measures alone.

Though SSL representations are strongly predictive of absolute MOS scores, our findings corroborate the expectations expressed in [8] that the high performance of SSL models on *VoiceMOS'22* may be attributed to the diversity of synthesis systems and quality levels in that dataset [28], [14]. Using the more controlled *PrefRank* stimuli, we confirm their hypothesis that SSL features are less predictive predictors and may struggle to capture fine-grained listener preferences.

Model	Top-1	τ_w	SRCC
UTMOS	**0.283**	0.059	**0.087**
DNS_OVERALL	0.217	0.005	0.026
PLCMOS	0.158	-0.093	-0.076
SHEETMQA	0.267	0.057	0.061

TABLE II: Correlation between MOS prediction models' outputs and relative acceptability rankings in *PrefRank*

C. Performance of MOS Models on Predicting Relative Acceptability

As shown in Table II, MOS-prediction models perform poorly on the *PrefRank* dataset. While they effectively capture absolute quality, these models do not effectively transfer to the more nuanced relative acceptability judgments in *PrefRank*. Prior work has demonstrated that some MOS-prediction models generalise across out-of-domain datasets [14], suggesting that poor performance is not solely a reflection of model limitations. Instead, it may reflect a deeper misalignment between information that is useful for predicting the broad quality differences reflected by MOS and for the fine-grained preferences reflected in *PrefRank*. This raises important questions regarding the limitations of MOS predictors for evaluation, especially of high-quality, prosody-focused speech.

V. Experiment B: Do probability assignments from SSL models of speech correlate with prosodic acceptability judgements?

The supervised task presented in the preceding section demonstrates that SSL representations encode information relevant to human assessments of prosodic acceptability. However, supervised prediction methods don't generalise well across domains. As such, we take an unsupervised approach to investigate how this information may be encoded. We examine the statistical regularities that SSL models encode through their probability assignments to speech inputs, and how strongly they correlate with both MOS judgements and acceptability scores. We propose a novel method for obtaining probability estimates and compare the scores derived from this approach to a previously established method.

A. Extracting probability scores from autoregressive SSL models of speech.

We extract probability-based utterance scores from vq-wav2vec, one of the few autoregressive SSL models of speech that operates over a discrete vocabulary of units, and compare to pseudo-probability estimates from more powerful non-autoregressive SSL of interest from Section IV. We describe the models and their probability estimates below.

The **vq-wav2vec** model consists of a convolutional acoustic encoder, a quantization module, and a convolutional context network [44]. It is trained to predict

quantized latent representations, learning both a set of codebook embeddings and a way to score how well each code fits a given audio frame. Though its training objective is contrastive (i.e., distinguishing between latent representations of input segments), the task of predicting *future* latent representations from past context encourages it to learn conditional probability distributions.

We extract frame-level probability assignments from this model with two methods. The first is our proposed method—**representational uncertainty** U_{rep}. We extract probabilities from the logits of the learned quantization module for every frame of input speech. Taking the softmax of these logits reflects a distribution of the model's uncertainty over its discrete codebook vocabulary at every time step. The second method, which we refer to as **predictive uncertainty** U_{pred}, was proposed by [28]. Probabilities are obtained by taking a softmax over the contrastive predictive logits of the vq-wav2vec context network at every timestep. These logits stem from the contrastive training loss (i.e., distinguishing the true future code from sampled distractors). As such, predictive uncertainty reflects the certainty of the models prediction within the contrastive learning objective while representational uncertainty is more analogous to classical notions of surprisal from text models as it reflects the model's uncertainty in representing its input.

Wav2vec2.0 and its extension **wavLM** are popular SSL models which involve encoding raw audio into latent features [45], [43]. Operating over these latent features, a Transformer-based context network is trained to predict representations of masked time steps in its training input [48]. As such, these models don't learn to encode conditional probability directly. We instead extract frame-level pseudo-probabilities from an additional head fine-tuned with a CTC loss to predict a distribution over a token vocabulary at each time step; we apply a softmax to the logits over this prediction at every frame.

To compute utterance-level scores for each method, we apply two summarisation operations to the frame-level probability distributions before mean-pooling across all frames in each utterance: entropy, and maximum. Entropy is a direct quantification of uncertainty over a distribution, while maximum offers more coarser description by only providing information about a single unit prediction. However, it mirrors the method for extracting surprisal from text models more closely.

B. Experimental Set-up

We follow the audio preprocessing described in Section IV-A. For wav2vec2.0 and wavLM, we use model checkpoints with fine-tuned CTC heads and for vq-wav2vec, we use the model detailed in Section IV-A.

Score	vq-wav2vec U_{rep}	vq-wav2vec U_{pred}	wav2vec2.0 CTC	wavLM CTC
Entropy	-0.422	-0.690	-0.383	-0.434
Max prob	0.374	0.674	0.348	0.421

TABLE III: Spearman ρ correlation coefficients between SSL uncertainty scores and MOS ratings from *VoiceMOS'22*. All correlations are significant at p< 0.001.

Domain	Score	vq-wav2vec U_{rep}	vq-wav2vec U_{pred}	wavLM CTC	wav2vec2.0 CTC
Overall	Entropy	*-0.140*	-0.061	0.083	0.063
Overall	Max	0.068	0.050	-0.062	-0.085
Read	Entropy	*-0.223*	*-0.144*	-0.091	0.007
Read	Max	0.135	0.144	0.062	-0.045
Conv.	Entropy	-0.086	-0.005	*0.200*	0.099
Conv.	Max	0.023	-0.013	-0.144	-0.112

TABLE IV: Spearman ρ correlations between *PrefRank* acceptability scores and SSL model probability score. Correlations are averaged over all, read, and conversational stimuli. Only *italicised* elements are significant with $p < 0.05$; p-values are obtained through permutation testing.

C. Correlation with MOS judgements.

As expected, Table III shows a positive relationship between maximum probability estimates and MOS scores while entropy is negatively correlated; low frame-level entropy and high probabilities both indicate high model certainty. Supervised models trained to predict MOS from SSL features, such as the baseline from [49], obtain $\rho \approx 0.92$, demonstrating how much performance can be improved through supervision.

As we hypothesized, entropy consistently outperforms maximum probability, confirming that a more detailed description of distribution uncertainty is useful.

Both probability scores extracted from vq-wav2vec outperform those from wav2vec2.0; although it is a much smaller model, vq-wav2vec is better-suited to this task. The highest correlations is obtained from the predictive uncertainty U_{pred} of vq-wav2vec, suggesting that uncertainty computed across contrastive logits aligns well with perceptual features that drive average MOS ratings. As these logits are related to predicting future frames, they likely encode temporal coherence and may therefore be more sensitive to artifacts of generation quality.

D. Correlation with acceptability judgements.

We now examine the relationships with *PrefRank* relative acceptability judgements to examine whether SSL mechanisms encode elements of prosodic realisation in ways that mirrors human perception. As was done in Section IV-A, correlation is computed at the stimulus level and reported as the mean over all stimuli Table IV. The same trends of significance were found for weighted-τ; only ρ correlation is reported for brevity.

Table IV shows that the correlations with BT scores are much weaker than MOS for both SSL models—only a few scores show a significant correlation with a p-values < 0.05. Although both wav2vec2.0 and wavLM scores produced moderate correlations with MOS judgements, none of their probability scores show any correlation with relative acceptability judgements. The CTC fine-tuning for these models encourages encoding of features linked to intelligibility—which may explain the correlation to MOS scores—but likely discards prosodic information used for judging relative acceptability.

Though the statistically significant correlations displayed by the vq-wav2vec estimation methods are very weak, we see slightly stronger relationships for representational uncertainty U_{rep} than predictive uncertainty U_{pred}. Although the strong correlation between U_{pred} and MOS highlights its sensitivity to features of generation quality, this score appears less attuned to prosodic variation within the bounds of acceptability. U_{rep} is computed over the codebook logits which may capture more phonetic encoding, and thus a greater sensitivity to fine-grained prosodic differences.

Interestingly, we find that correlation strength depends on the domain from which the lexical content of the *PrefRank* stimuli were drawn. Both uncertainty metrics from vq-wav2vec show weak but significant correlation for stimuli generated from read lexical content and no correlation for conversational stimuli. We hypothesise that the lack of relationship for conversational stimuli may reflect a domain mismatch with the LibriSpeech training data of vq-wav2vcec, or could be a function of the acceptability ratings themselves. Conversational speech is more prosodically varied than read speech; as such, listeners may tolerate or even expect greater variation in this domain [50].

VI. Discussion & Conclusions

As synthetic speech becomes even less distinguishable from natural speech, the shortcomings of current evaluation paradigms are also becoming more apparent [7], [8]. Traditional evaluation metrics like MOS offer only limited insight into how listeners perceive different prosodic realisations of high-quality renditions. We present a step toward addressing this limitation by investigating whether self-supervised speech representations encode a more nuanced perceptual signal: the relative acceptability of prosodic variation.

In Section IV, we show that SSL representations are predictive of relative acceptability scores among prosodic variants of the same text. Even with a small dataset, a simple supervised model using SSL features achieves promising performance. These results suggest that SSL models encode more nuanced perceptual cues of relevant to listener preference than those reflected by MOS. However, the correlation between SSL representations and acceptability rankings is markedly lower than their MOS prediction performance, indicating that prosodic acceptability is more complex and less readily encoded in current SSL representations. Furthermore, we find that existing MOS prediction models do not correlated with *PrefRank* prosodic acceptability scores, suggesting that the features important for predicting MOS may not overlap with those that drive relative acceptability judgements.

Section V explores whether the internal statistical structures learned by SSL models—reflected in their probability assignments—correlate with human preference judgments in an unsupervised setting. We introduce a novel method for computing model uncertainty from vq-wav2vec and compare it with a previously proposed approach. Scores derived from both methods correlate with MOS scores, suggesting that SSL models learn structural regularities aligned with perceived quality; however, correlations with acceptability rankings are weaker and less consistent. Code-based uncertainty scores show slightly stronger relationships than those of predictive uncertainty, but both still fall short of the supervised results in Section IV. Future work could explore whether alternative pooling strategies improve alignment with human judgments; however, these preliminary findings highlight the potential for using model uncertainty as a lens to understand how SSL mechanisms encode structure in speech. Interestingly, we also find that the domain of the speech material—conversational versus read—affects correlation strength. This may reflect domain-dependence in listener judgments, or sensitivity of SSL models to domain-biases in their training data. Disentangling these effects has implications for both model development and perceptual evaluation.

Our results contribute to the growing call to shift how speech synthesis is evaluated by moving toward a richer understanding of how listeners judge appropriateness. Achieving this will require new data, new metrics, and a closer integration with perceptual studies. The *PrefRank* dataset used in this study offers a novel opportunity to probe relative prosodic acceptability and is uniquely suited to testing how models and humans respond to subtle prosodic differences [9]. However, listener judgments of prosodic variation are shaped by many aspects of context—for example, the surrounding discourse or reference stimuli [21], [51], [52], [53]—which are not captured in this paradigm. Additionally, it remains unclear whether similarly scoring samples are equally acceptable or equally *un*acceptable. Integrating contextual features and grounding relatsive judgments in absolute acceptability will require new data collection paradigms. We believe these are fundamental for moving towards more human-centred evaluation.

REFERENCES

[1] F. Keller, "Gradience in grammar: Experimental and computational aspects of degrees of grammaticality," Ph.D. dissertation, University of Edinburgh, 2000.

[2] D. Goodhue, L. Harrison, Y. C. Su, and M. Wagner, "Toward a bestiary of english intonational contours," *The Proceedings of the North East Linguistics Society*, vol. 46, pp. 311–320, 2016.

[3] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," *Intentions in Communication*, vol. 271, p. 311, 1990.

[4] S. Wallbridge, P. Bell, and C. Lai, "It's not what you said, it's how you said it: Discriminative perception of speech as a multichannel communication system," in *Proceedings of Interspeech*. ISCA, 2021, pp. 2386–2390.

[5] C. Kurumada and T. B. Roettger, "Thinking probabilistically in the study of intonational speech prosody," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 13, no. 1, p. e1579, 2022.

[6] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. E. Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program," in *Proceedings of the Speech Synthesis Workshop*. ISCA, 2019, pp. 2019–19.

[7] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoustical Science and Technology*, vol. 45, no. 4, pp. 161–183, 2024.

[8] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[9] A. Adigwe, S. Wallbridge, Z. Tu, C. Lai, and S. King, "'can we cherry-pick?" investigating multiple renditions from a generative speech synthesis model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025, pp. 1–5.

[10] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proceedings of Interspeech*. ISCA, 2021, pp. 1194–1198.

[11] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop*. IEEE, 2023, pp. 1104–1111.

[12] B. R. Chernyak, A. R. Bradlow, J. Keshet, and M. Goldrick, "A perceptual similarity space for speech based on self-supervised speech representations," *The Journal of the Acoustical Society of America*, vol. 155, no. 6, pp. 3915–3929, 2024.

[13] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, "Utilizing self-supervised representations for MOS prediction," in *Proceedings of Interspeech*. ISCA, 2021.

[14] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 8442–8446.

[15] J. Hirschberg and G. Ward, "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english," *Journal of Phonetics*, vol. 20, no. 2, pp. 241–251, 1992.

[16] Z. Hodari, C. Lai, and S. King, "Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of f0," in *Speech Prosody*, 2020, pp. 965–969.

[17] H. Mixdorff, J. Cole, and S. Shattuck-Hufnagel, "Prosodic similarity–evidence from an imitation study," 2012.

[18] J. Chevelu, D. Lolive, S. Le Maguer, and D. Guennec, "How to compare TTS systems: a new subjective evaluation methodology focused on differences," in *Proceedings of Interspeech*. ISCA, 2015, pp. 3481–3485.

[19] E. Cooper and J. Yamagishi, "Investigating range-equalizing bias in mean opinion score ratings of synthesized speech," in *Proceedings of Interspeech*. ISCA, 2023, pp. 1104–1108.

[20] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 19 594–19 621, 2023.

[21] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. Gales, "Speech intonation for TTS: study on evaluation methodology," in *Proceedings of Interspeech*. ISCA, 2014, pp. 2957–2961.

[22] A. de la Fuente and D. Jurafsky, "A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models," in *Proceedings of Interspeech*. ISCA, 2024, pp. 1290–1294.

[23] L. Qian, C. Figueroa, and G. Skantze, "Representation of perceived prosodic similarity of conversational feedback," in *Proceedings of Interspeech (to appear)*. ISCA, 2025.

[24] N. G. Ward, A. Segura, A. Ceballos, and D. Marco, "Towards a general-purpose model of perceived pragmatic similarity," in *Proceedings of Interspeech*. ISCA, 2024, pp. 4918–4922.

[25] M. de Seyssel, M. Lavechin, H. Titeux, A. Thomas, G. Virlet, A. S. Revilla, G. Wisniewski, B. Ludusan, and E. Dupoux, "ProsAudit, a prosodic benchmark for self-supervised speech models," in *Proceedings of Interspeech*. ISCA, 2023, pp. 2963–2967.

[26] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The VoiceMOS challenge 2024: Beyond speech quality prediction," in *2024 IEEE Spoken Language Technology Workshop*. IEEE, 2024, pp. 803–810.

[27] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "Speechlmscore: Evaluating speech generation using speech language model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[28] A. Ravuri, E. Cooper, and J. Yamagishi, "Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*. IEEE, 2024, pp. 580–584.

[29] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity—a measure of the difficulty of speech recognition tasks," *The Journal of the Acoustical Society of America*, vol. 62, no. S1, pp. S63–S63, 1977.

[30] C. Aurnhammer and S. L. Frank, "Evaluating information-theoretic measures of word prediction in naturalistic sentence reading," *Neuropsychologia*, vol. 134, p. 107198, 2019.

[31] A. Goodkind and K. Bicknell, "Predictive power of word surprisal for reading times is a linear function of language model quality," in *Proceedings of the workshop on Cognitive Modeling and Computational Linguistics*, 2018, pp. 10–18.

[32] E. G. Wilcox, T. Pimentel, C. Meister, R. Cotterell, and R. P. Levy, "Testing the predictions of surprisal theory in 11 languages," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1451–1470, 2023.

[33] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.

[34] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 66–75.

[35] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[36] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2699–2712.

[37] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv preprint arXiv:2402.01912*, 2024.

[38] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[39] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *Proceedings of the Speech Synthesis Workshop*. ISCA, 2021, pp. 183–188.

[40] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *The Blizzard Challenge 2011*, 2011, pp. 1–10.

[41] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the International Conference on Machine learning*, 2007, pp. 129–136.

[42] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[43] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[44] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

[45] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[46] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[47] J. Shi, H. jin Shim, J. Tian, S. Arora, H. Wu, D. Petermann, J. Q. Yip, Y. Zhang, Y. Tang, W. Zhang, D. S. Alharthi, Y. Huang, K. Saito, J. Han, Y. Zhao, C. Donahue, and S. Watanabe, "VERSA: A versatile evaluation toolkit for speech, audio, and music," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics – System Demonstration Track*, 2025.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[49] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proceedings of Interspeech*. ISCA, 2022, pp. 4536–4540.

[50] P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read and spontaneous speech material," *Speech Communication*, vol. 10, no. 2, pp. 163–169, 1991.

[51] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors affecting the evaluation of synthetic speech in context," *Proceedings of the Speech Synthesis Workshop*, pp. 148–153, 2021.

[52] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," in *Proceedings of the Speech Synthesis Workshop*. ISCA, 2019, pp. 99–104.

[53] J. Edlund, C. Tånnander, S. Le Maguer, and P. Wagner, "Assessing the impact of contextual framing on subjective tts quality," in *Proceedings of Interspeech*. ISCA, 2024, pp. 1205–1209.