Evaluating prosodic variation in English and Icelandic synthesized speech

Caitlin Richter, Naizeth Núñez Macías, Stefán Ólafsson, Jón Guðnason, Hannes Högni Vilhjálmsson
Reykjavik University
Reykjavik, Iceland
{caitlinr, naizeth23, stefanola, jg, hannes}@ru.is

Abstract

Text to speech (TTS) systems still struggle with the prosodic one-to-many problem, wherein the same utterance text can require different prosodic realisations according to its context. We propose and demonstrate a method of automatic prosodic variant assessment, which may be used to evaluate how well speech synthesisers achieve a range of prosodic targets.

1 Introduction

The same string of words can require different prosodic contours depending on the discourse context in which it occurs, but speech synthesis systems struggle to generate appropriately conditioned variation in outputs. To facilitate TTS development, we automate interpretable evaluation of prosodic variants, proposing a method for benchmarking against a database of reference utterances. Contributions include a pilot study of prosodic variant assessment with English human speech, and a demo of its application to assess TTS voices in Icelandic, which is our target language for development.[1]

2 Background and motivation

Prosody can provide the only cue to distinguish different semantic, syntactic, or pragmatic acts, and infelicitous prosody increases listener effort and impairs comprehension (Klopfenstein, 2009; Wester et al., 2016). For instance, the Icelandic words in Example 1 are spoken differently for either of two syntactic, and consequently semantic, interpretations (Thráinsson, 2007).

Example 1

Hundar	*sem*	*borða*	*ís*	*gelta*	*hátt*
dogs	that/which	eat	icecream	bark	loudly

(A) “Dogs that eat ice cream bark loudly.”
(B) “Dogs, which eat ice cream, bark loudly.”

For an AI in an operating room (Example 2), the emphasis produced in utterance 03 is appropriate following context A, but could lead to real uncertainty by the human if spoken in context B.

Example 2
01-AI: (A) Pick up the **sinus probe**.
OR (B) Pick up the **iris scissors**.
02-Human: Ok.
03-AI: I will have the sinus **scissors** ready.
04-Human: (A) Ok.
(B) [doubts if iris scissors was recalled correctly, or if they should have the sinus probe]

Contrastive focus like this remains an unresolved challenge for TTS systems, even for English (Latif et al., 2021; Suni et al., 2020). The use of speech synthesis markup language (SSML) tags is established for controlling prosody, but often fails to generate acceptable modifications (Kim et al., 2021), and languages including Icelandic lack complete implementations (Schnell and Nikulásdóttir, 2024).

Costly evaluation is a barrier to improvement. Human TTS evaluations like Mean Opinion Scores (MOS) are more expensive for utterances in contexts than for isolated speech, but these contexts consistently influence judgements (Clark et al., 2019), including observable benefits when prosody is manipulated to align with them (Wang et al., 2024). Automated prosody evaluation and prosodic prominence prediction often rely on expensively annotated data and/or massive resources unavailable in many languages (Ekstedt et al., 2023; Dai et al., 2022; Stephenson et al., 2022; Talman et al., 2019; Suni et al., 2020; Chan and Kuang, 2024).

3 Methods

We propose an evaluation of prosody per context, extended from a method originally assessing goodness of non-native (human) speakers’ prosody through comparison to a native speaker reference database (Cheng, 2011):

[1] Code: https://github.com/catiR/ParParC-pcluster
Demo: https://huggingface.co/spaces/clr/pce/

1. Extract acoustic features like F0, energy, or spectral centroids, from reference and test speech.
2. Compute pairwise acoustic distances across all utterance-pairs' features. Cheng sampled 25 time points per word, while we use dynamic time warping (DTW) alignment cost without resampling.
3. Cluster reference data with 3 clusters per label. Human language can have multiple natural acoustic realisations of the same prosodic function (Cheng, 2011); this also helps our evaluation accommodate different segmental content in syntactically comparable texts. We use k-medoids clustering because it is compatible with DTW pairwise dissimilarity.
4. Measure similarity from test utterances to each cluster, defined as the average similarity to the cluster's members.
5. Quality Assessment. The test utterance is taken as an instance of the prosodic cluster it most closely matched, and its distance to this cluster reflects expected goodness (as a non-native speaker or TTS).

4 Experiments

4.1 Data

Initial validation uses the Paralinguistic Paraphrase Corpus (PPC; Suzuki and Nakamura, 2022), as there is no similarly balanced Icelandic corpus. Speakers repeat short sentences, varying prosodic focus across each valid position in turn. For our experiments, broadly comparable sets were extracted from PPC source texts with the following syntactic structures: **DNV:** Determiner-Noun-Verb; **DNVDMN:** Determiner-Noun-Verb-Determiner-{Modifier}-Noun; **DNVPDN:** Determiner-Noun-Verb-Preposition-Determiner-Noun.

4.2 Evaluation

An inventory of 3 prosodic clusters was learned for each possible focus position, per text type, with a random 20% of utterances held-out for testing. Then combining clusters from all emphasis conditions, distance from test utterances to each prosodic cluster was measured, producing a ranking from most to least similar cluster for each test utterance. Root mean square energy (RMSE) is the single acoustic feature for results in Table 1; although this is a deficient representation of speech prosody, it establishes the minimal baseline.

4.3 Results

	DNV	**DNVDMN**	**DNVPDN**
N	48	206	1131
Correct	80%	52%	54%
Mean rank	1.3	2.1	2.0

Table 1: N: total utterances; Correct (%): test utterances whose closest cluster is the correct label; Rank: average rank of closest correct cluster (top-1 match = rank 1).

Table 1 reports the percentage of test utterances whose closest measured cluster is the correct label, and the average rank of the best-ranking correct cluster, evaluated with stratified 5-fold cross validation. All performance is highly above chance.

5 Application: Icelandic TTS

Figure 1 illustrates an interpretable evaluation of Icelandic TTS for the text *En hvað veldur þá þessari miklu fjölgun snjógæsa?* "But what is causing this great increase in snow geese?". Voices `Álfur` and `Rósa` are both closest to the same cluster, but Álfur's distance is low (0.55) with fair match to human data; conversely, Rósa is subjectively less natural, and has higher distance (0.66) with oddities like high energy in the second syllable of *snjógæsa*.

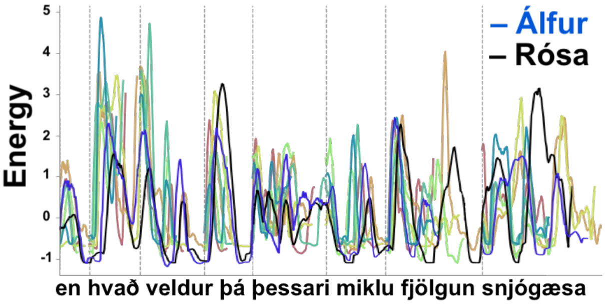

Figure 1

6 Discussion

Pilot validation (§4) establishes that, despite acoustic variability within each prosodic context as well as across segmental content, the proposed method yields a signal about the fit of an utterance to particular prosodic intents. This can be applied to assess synthesised utterances (§5). We anticipate reaching greater accuracy with enriched acoustic representations, such as eGeMAPS and local speech rate cues (Eyben et al., 2015; Tilsen and Tiede, 2023).

Feature learning with a cluster purity objective may likewise improve prosodic comparison. Ultimately, we aim to learn and validate representations that will support unsupervised prosodic clustering, while textually comparable sets like §4.1 may be automatically extracted from parsed or POS-tagged speech corpora. Therefore, while an automated benchmark from hand-annotated data is already more efficient than human MOS judgement, we envision a path away from this dependence as well.

7 Limitations

One limitation of this study is the use of artificial prosodic contexts in the validation. Speakers from the PPC were told to emphasise a certain word without sufficient context information to distinguish between broad, narrow or contrastive focus (Latif et al., 2021). Further validation with audio excerpted from real conversations is necessary in order to evaluate whether a specific prosodic realisation is adequate in the context it appears in.

Acknowledgments

The authors acknowledge Magnús Freyr Morthens with gratitude for collaboration on design and implementation of the Icelandic application. This work was supported by The Icelandic Centre for Research (RANNÍS), under the Icelandic Student Innovation Fund project *Better Icelandic speech synthesis: Prosody roadmap* (2311052-1101), and the Strategic Research and Development Programme for Language Technology project *Spoken Dialogue Framework for Icelandic* (200103-5301).

References

Cedric Chan and Jianjing Kuang. 2024. Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models. In *Speech Prosody 2024*, pages 27–31.

Jian Cheng. 2011. Automatic assessment of prosody in high-stakes English tests. In *Proc. Interspeech*, pages 1589–1592.

Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. 2019. Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*. ISCA.

Ziqian Dai, Jianwei Yu, Yan Wang, Nuo Chen, Yanyao Bian, GuangZhi Li, Deng Cai, and Dong Yu. 2022. Automatic prosody annotation with pre-trained text-speech model. In *Proc. Interspeech 2022*, pages 5513–5517.

Erik Ekstedt, Siyang Wang, Éva Székely, Joakim Gustafsson, and Gabriel Skantze. 2023. Automatic evaluation of turn-taking cues in conversational speech synthesis. In *Proc. Interspeech 2023*, pages 5481–5485. ISCA.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and 1 others. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers characterize naturalness in voice user interfaces: their goals, practices, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Marie Klopfenstein. 2009. Interaction between prosody and intelligibility. *International Journal of Speech-Language Pathology*, 11(4):326–331.

Siddique Latif, Inyoung Kim, Ioan Calapodescu, and Laurent Besacier. 2021. Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 544–551, Online. Association for Computational Linguistics.

Daniel Schnell and Anna Björk Nikulásdóttir. 2024. Icelandic TTS for android (24.04). CLARIN-IS.

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. 2022. BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model. In *Proc. Interspeech 2022*, pages 3383–3387. ISCA.

Antti Suni, Sofoklis Kakouros, Martti Vainio, and Juraj Šimko. 2020. Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis. In *Speech Prosody 2020*, pages 940–944.

Naoaki Suzuki and Satoshi Nakamura. 2022. Representing 'how you say' with 'what you say': English corpus of focused speech and text reflecting corresponding implications. In *Proc. Interspeech 2022*, pages 4980–4984.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations. *NoDaLiDa 2019*, page 281.

Höskuldur Thráinsson. 2007. The syntax of Icelandic. Cambridge University Press.

Sam Tilsen and Mark Tiede. 2023. Looking within events: Examining internal temporal structure with local relative rate. *Journal of Phonetics*, 100:101264.

Siyang Wang, Éva Székely, and Joakim Gustafson. 2024. Contextual Interactive Evaluation of TTS Models in Dialogue Systems. In *Interspeech 2024*, pages 2965–2969. ISCA.

Mirjam Wester, Oliver Watts, and Gustav Eje Henter. 2016. Evaluating comprehension of natural and synthetic conversational speech. In *Speech Prosody 2016*, pages 766–770.