

Simon Rosen

Baltimore, MD — simon.rosen17@gmail.com — [linkedin.com/in/simon-rosen-linguist](https://www.linkedin.com/in/simon-rosen-linguist)

Professional Summary

Linguist specializing in phonetics, phonology, and prosody, working on speech and voice AI evaluation at ServiceNow since 2023. I design TTS and voice agent evaluations, including multilingual vendor comparisons, write the rubrics and annotation guidelines behind them, and build the internal labeling tools linguists use to run them. Co-first author of *Beyond Naturalness* (EMNLP 2026 Main), a dimension-level TTS evaluation benchmark.

Core Competencies

Speech & Linguistics: Phonetics & Phonology, Prosody & Tone, Segmental/Suprasegmental Analysis, IPA Transcription, Acoustic Analysis (Praat), Forced Alignment (Montreal Forced Aligner), Pragmatics & Discourse

Voice AI Evaluation: TTS & ASR Evaluation, TTS Vendor Comparison, Text Normalization Test Design, MOS Predictor & Audio-LLM Judge Benchmarking, Voice Agent Conversation Rating, Failure Taxonomies, LLM-as-Judge Design

Experiment Design & Statistics: Benchmark Construction, Rubric & Annotation Guideline Design, Inter-Annotator Agreement (Krippendorff's α , Bootstrap CIs), Significance Testing (Kendall's τ , Mann-Whitney U , McNemar)

Engineering: Python (pandas, scikit-learn, statsmodels, matplotlib/seaborn), Internal Web Tool Development, LLM Applications & Prompt Engineering, Structured Output Schemas, AI-Assisted Development (Claude Code), JSON/YAML, Git

Languages: English (fluent); university coursework in Mandarin Chinese and Indonesian

Professional Experience

Senior ML Data Linguist

Feb 2026 – Present

ServiceNow, LUNAR (Linguistic Unit for NLU Analysis & Research) Team, Santa Clara, CA

- Co-designed a 10-dimension, linguistically grounded annotation schema that breaks TTS “naturalness” into word-level, prosodic, and paralinguistic dimensions, and built an 860-utterance benchmark labeled by trained linguist raters (EMNLP 2026 Main, co-first author).
- Compared eight automated speech evaluators (four MOS predictors, four audio-LLM judges) against human labels under four prompting conditions, showing per dimension where each evaluator type can and cannot replace human listeners.
- Designed and ran a multilingual TTS vendor comparison: wrote test sentences drawn from ITSM and CSM language and normalization edge cases (symbols, math, units, URLs), defined a rubric pairing holistic 1–5 ratings with per-category head-to-head judgments, and adapted it with a native-speaker linguist for Japanese.
- Built a rating framework for AI voice agent conversations (speech fluency, tool calling, task completion, helpfulness) and wrote its annotation guidelines.
- Ran the labeling efforts that produced the human benchmarks for customer-facing voice fluency judges: wrote the guidelines, onboarded labelers, and presented results across agents, separating ASR-driven failures (such as repeated confirmations) from conversation-flow and voice-style problems.
- Leading success labeling for document question answering and building the matching LLM judge.
- Built and maintain the internal labeling tools used on these projects, plus the team’s project-tracking app.

ML Data Linguist

Jan 2023 – Feb 2026

ServiceNow, LUNAR Team, Santa Clara, CA

- Wrote LLM-as-judge prompts and structured output schemas for internal evaluation tooling; authored a prompt design guide adopted as the team reference and led training sessions so product teams could build their own evaluations.
- Cataloged recurring failure patterns (hallucination, under-specification, intent misclassification) and worked with ML engineers on prompt changes, retraining, and production model selection.
- Built a ground-truth benchmark through paired human–model evaluation and checked it against LLM-as-judge and automatic metrics to find scoring discrepancies.

Publications

O. Bamgbose*, S. Rosen*, et al. *Beyond Naturalness: Probing Automated Text-to-Speech Evaluators on Linguistically Grounded Dimensions*. EMNLP 2026 (Main Conference). [arXiv:2608.09930](https://arxiv.org/abs/2608.09930)

*Equal contribution. Designed the annotation schema and led the annotation effort; dataset, schema, and evaluation code are publicly released.

Education

B.A. in Linguistics

2022

Princeton University, Princeton, NJ

Thesis: *Tone Sandhi and Constituency in Classifier Phrases Across Sinitic Languages*

Relevant Coursework: Phonetics & Phonology, Syntax, Semantics, Pragmatics, Java Programming, Data Structures & Algorithms