NON-INTRUSIVE PREDICTION OF LYRIC INTELLIGIBILITY USING A CRNN AND GRADIENT BOOSTING ENSEMBLE

Matthias Wellershoff & Vincent Wellershoff

ABSTRACT

We present a robust, non-intrusive system for lyric intelligibility prediction developed for the ICASSP 2026 Cadenza challenge. Addressing intelligibility as a multi-dimensional phenomenon, our approach integrates heterogeneous feature streams ranging from low-level acoustic degradation (Fisher information, extended STOI) to high-level semantic coherence (Whisper embeddings, LLM assessment). The architecture operates in two stages: a lightweight temporal CRNN ($\sim$8 million parameters) models moment-by-moment intelligibility, followed by a gradient boosting ensemble that captures global cognitive and prosodic characteristics. On the challenge evaluation set, this method achieves an RMSE of 27.31 and a correlation of 0.64. These results demonstrate that systematic coverage of acoustic, linguistic, and cognitive failure modes allows for accurate intelligibility estimation without reliance on ground truth lyrics.

1. INTRODUCTION

Predicting lyric intelligibility from degraded audio is an inherently multi-dimensional challenge. A signal may be rendered unintelligible by acoustic interference (poor source separation), linguistic complexity (rare vocabulary), phonological hurdles (difficult consonant clusters), or cognitive load (semantic ambiguity). Because no single metric captures this full spectrum, effective prediction requires a synthesis of complementary features.

In response to the ICASSP 2026 Cadenza challenge [1], we propose a two-stage architecture: a temporal neural network models the moment-by-moment evolution of intelligibility, while a scalar ensemble captures global, cross-sectional characteristics. Together, these stages answer different aspects of the same underlying question: *how much of the content can a listener actually recover?*

Our neural network architecture draws inspiration from [2, 3]. Further refinements of our full system are inspired by systematic diagnosis of failure modes. As an example, we introduced confidence gating — modulating Whisper embeddings by ASR certainty — which effectively teaches the ensemble to weigh linguistic analysis more heavily when acoustic cues became ambiguous.

Crucially, our approach is non-intrusive. The system predicts intelligibility relying solely on the degraded and clean audio signals provided in the challenge dataset [4], without access to ground truth lyric transcriptions. To facilitate reproducibility and future research, we make our complete feature extraction and training pipeline available via our public repository.[1]

2. METHODOLOGY

2.1. Temporal neural architecture

A core challenge in temporal modeling is that intelligibility depends on heterogeneous information streams. The convolutional recurrent neural network (CRNN) stage of our system processes five complementary streams, each addressing a distinct aspect of the problem:

1. *Information-theoretic separability.* Before attempting separation, we quantify whether it is even possible. Using learned source distributions for vocals and accompaniment, we compute the Fisher information matrix (FIM) across mel-scaled frequency bands inspired by the French–Steinberg articulation index [5]. This provides frame-by-frame bounds on achievable separation precision — a fundamental limit that constrains any downstream model.
2. *Achieved separation quality.* Using HDemucs-separated stems [6], we measure frequency-domain competition: spectral overlap, mutual masking, vocal-to-accompaniment signal-to-noise ratio, transient characteristics, zero-crossing rates, and per-band reconstruction error. These features capture whether vocals remain entangled with residual instrumentation.
3. *Frame-wise extended short-term objective intelligibility (ESTOI).* Rather than collapsing ESTOI [7] to a scalar, we compute it frame-by-frame across all pairings of processed/unprocessed and separated vocals/signals. This produces temporal features showing when and how degradation impacts intelligibility.
4. *Deep semantic representation.* Inspired by [3], we extract raw embeddings from Whisper's [8] encoder without forcing transcription. This provides a learned semantic representation that can interact with acoustic features.
5. *Raw time-frequency content.* Short-time Fourier transform power spectra serve as a baseline.

Each stream is processed through a dedicated four-layer convolutional neural network (CNN) stack. After resampling to a common temporal resolution, features are concatenated into a unified representation.

This representation passes through a bidirectional long short-term memory (BLSTM) as well as a self-attention mechanism. The result is pooled to produce an estimate of scalar word correct rate (WCR). The architecture treats intelligibility as an emergent property of temporal dynamics, not a static signal characteristic.

The CRNN network has $\sim$8 million trainable parameters. This relatively compact architecture allows for exceptional computational efficiency. The model was trained in about one hour on a single NVIDIA L4 GPU. Furthermore, the lightweight design ensures rapid inference, achieving approximately $20\times$ real-time (RT) processing speed on a single Apple M3 Pro (without using the Metal Performance Shaders (MPS) framework).

M. Wellershoff acknowledges the use of the Gemini LLM (Google) for technical writing and structuring assistance. V. Wellershoff acknowledges the use of the Claude LLM (Anthropic) for the same.

[1] Code available at: `https://github.com/wellersm/CLIP1`

2.2. Scalar feature ensemble

The temporal network is not designed to effectively capture summary statistics or holistic signal properties. We therefore construct a complementary gradient boosting ensemble operating on cross-sectional features spanning nine conceptual domains:

1. *Learned prior.* The CRNN prediction is used directly as a learned prior.
2. *Acoustic summaries.* Includes STOI (unprocessed vocals separated using HDemucs compared to the processed signal), spectral flatness, RMS energy, and peak-to-RMS ratio.
3. *Temporal and prosodic structure.* Captures time-pressure effects via BPM, speech rate, and duration.
4. *Phonological difficulty.* Consonant-class features from hearing-science literature [9] reflect phoneme-level difficulty independent of acoustics.
5. *Lexical accessibility.* Word-frequency features estimate vocabulary difficulty.
6. *Semantic coherence (LLM-based).* An LLM (Haiku 4.5) assesses whether ASR transcriptions form coherent text.
7. *Deep linguistic robustness.* Whisper-embedding similarity between clean and processed speech measures how much of the meaning of the speech is preserved.
8. *Cross-model consensus.* Agreement between Whisper and Meta's oASR [10] indicates transcription ambiguity.
9. *Recognition confidence.* ASR confidence and the consistency of word sequences determine how much the system trusts each segment of speech.

The final prediction is produced by an ensemble comprising ten diverse gradient-boosted models (XGBoost, LightGBM, CatBoost) trained with perturbed hyperparameters. The average prediction is then discretized to the Farey sequence F_{20}. This sequence contains all 351 possible rational WCR values for sentences up to length 20, respecting the discrete nature of word correctness while reducing overfitting.

3. EVALUATION AND RESULTS

We measure the performance of our system using two metrics: the root mean squared error (RMSE) and the correlation (ρ) of the sequence of true and predicted WCRs. Evaluation was conducted on three configurations using the designated validation and evaluation sets: the non-intrusive baseline, the temporal network alone, and the full system.

	Validation		Evaluation	
System	RMSE	ρ	RMSE	ρ
Full system	**27.43**	**0.66**	**27.31**	**0.64**
Temporal network	30.95	0.53	30.12	0.53
Baseline	36.11	0.14	34.89	0.21

We note that the final system architecture achieves a significant improvement over the official non-intrusive baseline. Additionally, the integration of the scalar feature ensemble greatly improves performance compared to the temporal network alone.

4. CONCLUSION

Our two-stage design reflects a core insight: intelligibility has both temporal structure (how competing sounds evolve) and atemporal structure (whether vocabulary is accessible, whether content is coherent). The neural network excels at the former; gradient boosting excels at the latter. By processing them separately and combining their outputs, we avoid forcing either model to learn patterns outside its natural regime.

Critically, we measure separability *before* attempting separation (Fisher information), separation quality *after* attempting it (stem metrics), and semantic preservation at multiple levels (frame-wise ESTOI, holistic embedding correlation, transcription agreement, LLM coherence assessment). Each metric captures different failure modes: spectral overlap does not imply semantic confusion; clean separation does not guarantee lexical accessibility; high confidence does not ensure correctness.

The result is a mosaic representation of the audio signal, where each tile addresses a specific aspect of the intelligibility problem. The models learn how these aspects interact — when phonological difficulty compounds with poor separation, when high tempo compensates for clear acoustics, when semantic coherence survives severe distortion. This interaction learning, rather than any individual feature, is designed to promote effective generalization

5. REFERENCES

[1] G. Roa-Dabike et al., "Overview of the ICASSP 2026 Cadenza challenge: Predicting lyric intelligibility," ICASSP (submitted), 2025.

[2] B. Sharma and Y. Wang, "Automatic evaluation of song intelligibility using singing adapted STOI and vocal-specific features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, 2019, doi: 10.1109/TASLP.2019.2955253.

[3] R. E. Zezario et al., "Non-intrusive multi-branch speech intelligibility prediction using multi-stage training," in *Proc. Clarity Workshop*, 2025, doi: 10.21437/Clarity.2025-4.

[4] G. Roa-Dabike et al., "The Cadenza lyric intelligibility prediction (CLIP) dataset," Data in Brief (submitted), 2025.

[5] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 90, 1947, doi: 10.1121/1.1916407.

[6] A. Défossez, "Hybrid spectrogram and waveform source separation," doi: 10.48550/arXiv.2111.03600, 2022.

[7] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, 2016, doi: 10.1109/TASLP.2016.2585878.

[8] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, vol. 202 of *PMLR*.

[9] S. A. Phatak et al., "Consonant recognition loss in hearing impaired listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 5, 2009, doi: 10.1121/1.3238257.

[10] G. Keren et al., "Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages," doi: 10.48550/arXiv.2511.09690, 2025.