

Multi-Modal Speech Intelligibility Prediction Using Gradient Boosting for the Cadenza CLIP1 Challenge

Roman Vygon
SmartNews, Inc.
roman.vygon@gmail.com

Abstract—We present a multi-modal system for predicting speech intelligibility in music-speech mixtures for hearing-impaired listeners as part of the Cadenza CLIP1 Challenge. Our approach combines *whisper-large-v3-turbo* audio embeddings extracted from four audio variants (processed/unprocessed, full-mix/vocals-only), RoBERTa-large text embeddings of ground truth prompts, and hand-crafted acoustic features. Vocal separation is performed using HTDemucs. An ensemble of XGBoost regressors predicts the word recognition correctness ratio. Our system achieves an RMSE of 26.59 on the challenge validation set. The entire pipeline runs on a consumer GPU (NVIDIA RTX 3070).

Index Terms—speech intelligibility prediction, hearing impairment, Whisper, gradient boosting, source separation

I. INTRODUCTION

The Cadenza CLIP1 Challenge [2] focuses on predicting *correctness*, the proportion of words correctly recognized by hearing-impaired listeners from processed music-speech signals, using the CLIP dataset [1]. Building on recent work showing that pre-trained speech representations effectively model intelligibility [3], our system combines Whisper embeddings from multiple audio conditions, text embeddings of ground truth prompts, and hand-crafted acoustic features including pairwise quality metrics.

II. SYSTEM ARCHITECTURE

A. Overview

Figure 1 illustrates our pipeline. Both submitted systems (T072a and T072b) use the **processed signal** (hearing-loss-simulated mixture), **unprocessed signal** (original clean mixture), and **hearing loss severity** (No Loss, Mild, or Moderate). T072a (intrusive) additionally uses the **reference text** (ground truth prompt), while T072b (non-intrusive) does not. Features are extracted from three modalities and fed to an XGBoost ensemble.

B. Whisper Embeddings

We use *whisper-large-v3-turbo* to extract encoder embeddings from four audio variants: processed mix, processed vocals, unprocessed mix, and unprocessed vocals. Vocal separation uses HTDemucs [4]. Audio is resampled to 16 kHz mono. We apply mean pooling across the temporal dimension, yielding 1280 dimensions per variant (5120 total).

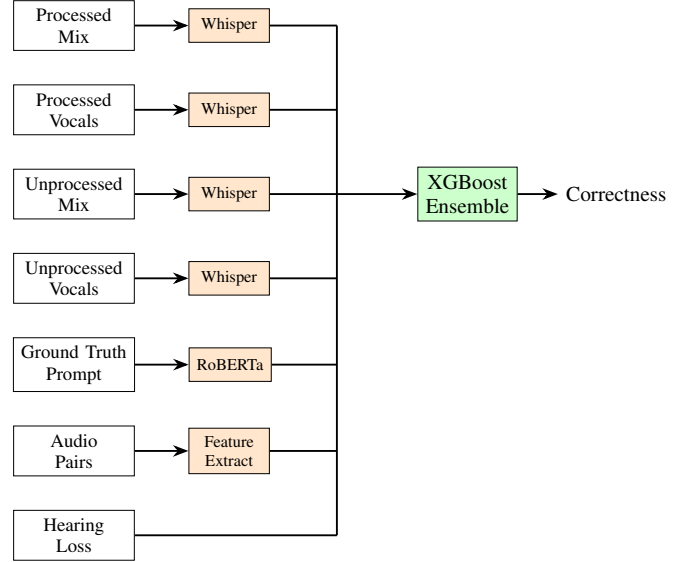


Fig. 1. System architecture showing multi-modal feature extraction and fusion. Audio variants are processed through *whisper-large-v3-turbo* to obtain embeddings. Text embeddings from RoBERTa and acoustic features are concatenated before XGBoost prediction.

C. Text Embeddings

We encode the ground truth prompt using RoBERTa-large, extracting the [CLS] token (768 dimensions). This captures contextual predictability—some words may be easier to guess from context even under degraded conditions.

D. Acoustic Features

We extract ~ 300 hand-crafted features from four audio files per sample (clean/processed \times mix/vocals).

Per-file features (~ 35 per file): RMS energy (mean, std), zero-crossing rate, spectral descriptors (centroid, rolloff, bandwidth, flatness, contrast), 13 MFCCs (mean, std), F0 statistics via pYIN, voiced frame ratio, voice activity ratio, and modulation band power.

Pairwise features: SI-SDR, SNR, envelope SNR, spectral overlap, STOI/Extended STOI [5], PESQ [6], and envelope correlation. For STOI and PESQ, we use **unprocessed vocals** (extracted via HTDemucs) as reference and **processed vocals** as degraded signal.

TABLE I
OFFICIAL CHALLENGE RESULTS: VALIDATION AND EVALUATION SETS

System	Validation		Evaluation	
	RMSE	Corr.	RMSE	Corr.
Baseline STOI	36.11	0.14	34.89	0.21
Baseline Whisper	29.32	0.59	29.08	0.58
T072b (Non-Intrusive)	26.94	0.67	26.84	0.66
T072a (Intrusive)	26.59	0.68	26.68	0.66

Derived features: Vocal-to-music power ratio and voiced ratio difference. Hearing loss is encoded as ordinal (No Loss, Mild, Moderate). The final feature set consists of 6465 dimensions.

III. MODEL TRAINING

We train XGBoost regressors with hyperparameters optimized via Optuna [7] using 3-fold cross-validation. Our ensemble uses 7 models: six trained with one feature group masked, plus one on all features, combined by averaging. All experiments used an NVIDIA RTX 3070 (8GB VRAM); the complete pipeline takes approximately 1 hour.

IV. RESULTS

Table I presents official results. Using only Whisper embeddings, larger encoders improve performance (Base: 29.38, Medium: 28.04, Large: 26.9 RMSE). Our full systems substantially outperform baselines, with T072a (intrusive, using reference text) slightly improving over T072b (non-intrusive). Performance generalizes well from validation to evaluation.

V. DISCUSSION AND CONCLUSION

The strong performance of Whisper embeddings suggests pre-trained speech models effectively capture degradation patterns relevant to intelligibility. Text embeddings improved performance, supporting our hypothesis that contextual predictability aids word recognition.

Limitations: Our system lacks sequential pattern modeling—Whisper embeddings are averaged temporally, and text embeddings use only the sentence-level token. Future work could incorporate listener-specific modeling to handle noisy ground truth scores.

We presented a multi-modal system combining Whisper embeddings, RoBERTa text embeddings, and acoustic features within an XGBoost ensemble. Both systems use hearing loss severity, processed and unprocessed signals; T072a additionally uses reference text. Our approach achieved competitive performance (T072a: eval RMSE 26.68, correlation 0.66) while remaining efficient on consumer hardware.

REFERENCES

- [1] G. Roa-Dabike, T. J. Cox, J. P. Barker, B. M. Fazenda, S. Graetzer, R. R. Vos, M. A. Akeroyd, J. Firth, W. M. Whitmer, S. Bannister, and A. Greasley, “The Cadenza Lyric Intelligibility Prediction (CLIP) Dataset,” *Data in Brief*, 2025.
- [2] G. Roa-Dabike, J. P. Barker, T. J. Cox, M. A. Akeroyd, S. Bannister, B. Fazenda, J. Firth, S. Graetzer, A. Greasley, R. R. Vos, and W. M. Whitmer, “Overview of the ICASSP 2026 Cadenza Challenge: Predicting Lyric Intelligibility,” in *Proc. IEEE ICASSP*, 2026, to appear.
- [3] R. Buragohain, J. Ajaybhairam, A.K. Singh, K. Nathwani, and S.K. Kopparapu, “Non-Intrusive Speech Intelligibility Prediction Using Whisper ASR and Wavelet Scattering Embeddings for Hearing-Impaired Individuals,” in *Proc. The 6th Clarity Workshop on Improving Speech-in-Noise for Hearing Devices (Clarity-2025)*, pp. 18–21, 2025, doi: 10.21437/Clarity.2025-6.
- [4] A. Défossez *et al.*, “Hybrid Transformers for Music Source Separation,” in *Proc. ICASSP*, 2023.
- [5] C. H. Taal *et al.*, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, 2011.
- [6] ITU-T Rec. P.862, “Perceptual evaluation of speech quality (PESQ),” 2001.
- [7] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-Generation Hyperparameter Optimization Framework,” in *Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.