SPECTRAL AUGMENTED SELF-ATTENTIVE INTELLIGIBILITY PREDICTION

TEAM T036: ICASSP 2026 CADENZA CLIP1 CHALLENGE TECHNICAL REPORT

Fredrik Cumlin [1]*, Xinyu Liang* [2]*, Victor Ungureanu* [3]*,*
Chandan K. A. Reddy [3]*, Christian Schüldt* [3]*, Saikat Chatterjee* [1]

[1] KTH Royal Institute of Technology, Stockholm, Sweden [2] HCLTech AB [3] Google LLC

ABSTRACT

This report presents our submission to the ICASSP 2026 Cadenza challenge for lyric intelligibility prediction (CLIP1). We propose a non-intrusive intelligibility prediction system that combines spectral representations with self-supervised learning (SSL) representations from wav2vec 2.0. We call our system spectral augmented self-attentive intellibility predictor (SASI-P). The architecture combines the encoders from DNSMOS Pro and MultiGauss. SASI-P takes 44.1 kHz sampling rate and stereo signals as input, and predicts an intelligibility score. SASI-P achieves a Pearson correlation coefficient (PCC) of 0.53 and root mean square error (RMSE) of 30.98 on the validation data.

***Index Terms*—** Speech intelligibility assessment, speech quality assessment, self-supervised learning, generalization ability

1. INTRODUCTION

Understanding the lyrics in music is fundamental to music enjoyment. The Cadenza Lyric Intelligibility Prediction Challenge (CLIP1) addresses the need for automatic metrics to evaluate lyric intelligibility in popular Western music by proposing a challenge in automatic lyric intelligibility prediction. The challenge can be compared to the VoiceMOS/AudioMOS challenges, which address the need for automatic speech quality predictors [1].

The CLIP1 challenge proposes unique challenges: (1) the lyrics are embedded in music, (2) the voiced speech differs significantly from natural speech, and (3) the data is of high sampling rate and stereo. Furthermore, the data is processed through hearing loss simulators representing normal, mild, and moderate hearing loss conditions. SSL-based approaches typically only support monochannel and low bandwidth (8 kHz) signals, causing challenges for intelligibility prediction of the CLIP1 dataset.

Our contribution is a non-intrusive intelligibility predictor that combines spectral representations with deep SSL representations. Both the spectral and the SSL representations can be multichannel. We use convolutional layers to fuse the spectral representations, and self-attention to fuse the SSL representations. We demonstrate higher performance measures compared to the non-intrusive baseline, and slightly worse than the intrusive baseline.

1.1. Problem formulation

Given an intelligibility-rated dataset $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ where $\mathbf{x}_n$ are audio features (e.g., spectrograms, SSL representations) and y_n is the intelligibility score, the task is to learn a neural network regressor $f_{\boldsymbol{\theta}} : \mathbf{x}_n \rightarrow y_n$.

2. SYSTEM DESCRIPTION

2.1. Architecture Overview

Our system, termed Spectral Augmented Self-attentive SSL Intelligibility Prediction (SASI-P), consists of three main components: (1) a spectrogram encoder processing magnitude spectrograms, (2) an SSL-based encoder with cross-channel attention processing wav2vec 2.0 features, and (3) a dense prediction head mapping fused representations to intelligibility scores.

2.2. Spectrogram Encoder

The spectrogram encoder is based on the DNSMOS Pro architecture [2]. DNSMOS Pro is a non-intrusive end-to-end speech quality assessment predictor, designed for quality prediction of 16 kHz speech signals. We adapt DNSMOS Pro for 44.1 kHz signals by increasing the window and hop length of the short-time Fourier transform (STFT) of the input signals. The window length is 1024 and the hop length is 512. We compute the log-magnitude of the STFT-transformed signal input, and clip to -7 and 7. If there are several channels of the audio, each channel is STFT-transformed separately and stacked.

The Spectrogram encoder consists of five convolutional layers, 2D max pooling layers, batch normalization, and dropout. After the convolutional layers, a global max pooling layer is used, mapping the ‘time x frequency x convolutional channel‘ representation to a ‘convolution channel‘ representation. This means that we take the maximum value for each convolutional channel representation. The output of the Spectrogram encoder is a ‘convolution channel‘ representation of the audio.

2.3. SSL Encoder

The SSL encoder takes SSL-based representations as input, and outputs 128 dimensional representations. The SSL-model used is the wav2vec 2.0 [3]. Wav2vec 2.0 takes 16 kHz monochannel signals as inputs; hence, the input audio is resampled to this. If multiple channels exist in the audio, the channels are processed separately. After a small search, we use the output from layer 40 as representations to be used for subsequent processing.

The architecture processing the SSL-based representations is inspired by the MultiGauss architecture [4]. Assume we have audio that consists of two channels, and hence, two SSL-based representations. Then, two MultiGauss encoders are initialized. A MultiGauss encoder consists of 1D convolutional layers, layer normalization, pooling, and dropout. After processing an SSL-based representation with a MultiGauss encoder, we compute the $1D$ global max pooling, which gives a representation of dimension equal to the number

of kernels of the last convolutional layer. In MultiGauss, we use 128 kernels in the last convolutional layer; hence, the output representations are 128-dimensional vectors.

For the two channels processed by their respective MultiGauss encoders, we fuse them in the following way. First, we stack the representations and then apply positional embeddings. Then we apply a multihead attention layer on the stacked representations, along the 128-dimensional axis. These are then passed through a linear layer predicting importance weights for each audio channel. These are then summed according to the importance weights. This is then passed through a residual connection, adding the attention and weight-processed representations with the average of the 128-dimensional representations. Finally, we apply a layer normalization to the output. This means the final output is 128-dimensional.

2.4. Prediction Head

The Spectrogram encoder and the SSL encoder produce 128-dimensional representations of the input audio, respectively. Note that the Spectrogram encoder got information about higher frequencies, while the SSL encoder did not. In the prediction head, these are then merged.

We merge the representations by concatenating them, giving a 256 dimensional representation. We then process this representation with a dense layer, predicting a mean and variance of a Gaussian distribution. The variance is guaranteed positive by the sofplus activation function. Softplus is given by $x \mapsto \log(1 + e^x)$. The reason for predicting a Gaussian is to have a probabilistic system maximising the log-likelihood of the data, as per [5, 2, 4].

2.5. Training

Training is done with batches, minimising the Gaussian negative log-likelihood (GNLL). Assume we have a regressor $f_{\boldsymbol{\theta}} : \mathbf{x}_n \to (\mu_n, \sigma_n^2)$; this means the regressor predicts the parameters of a Gaussian distribution over the intelligibility score. Thus, $p_{\boldsymbol{\theta}}(y_n|\mathbf{x}_n) = \mathcal{N}(y_n; \mu_n, \sigma_n^2)$. The training objective is to minimize the GNLL:

$$\arg\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} \frac{1}{2} \left[\log \sigma_n^2 + \frac{(\mu_n - y_n)^2}{\sigma^2} \right]. \tag{1}$$

3. EXPERIMENTAL RESULTS

3.1. Dataset

The dataset used is the dataset provided by Cadenza. It is comprised of excerpts from Western songs in English. The clips are kept as is, or synthetically degraded with mild or moderate distortions. The duration varies from 1 s to 25 s, and the sampling rate is 44.1 kHz. The labels are given as correct-word-scores, where humans have transcribed the song excerpt, normalized to the interval $[0, 1]$. The dataset is split into train, validation, and evaluation/test.

3.2. Training Details

SASI-P were trained for 30 epochs using Adam optimizer with learning rate 1×10^{-4}, batch size 64, and gradient clipping (max norm 5). Audio signals were cropped or repeated to 10 s duration and resampled to 44.1 kHz. Training used 99% of the training set, with 1% held out for validation. The model at epoch 30 was selected for testing.

3.3. Results

Following the challenge protocols, we evaluate using **Pearson Correlation Coefficient (PCC):** Measures linear correlation between predicted and true intelligibility scores; and **Root Mean Square Error (RMSE):** Quantifies prediction accuracy. The results are shown in Tab. 1.

Table 1: Performance comparison on CLIP1 validation set

Method	PCC ↑	RMSE ↓
Validation		
Baseline Whisper	0.59	29.32
Baseline STOI	0.14	36.11
SASI-P (Ours)	0.53	30.98
Evaluation		
Baseline Whisper	0.58	29.08
Baseline STOI	0.21	34.89
SASI-P (Ours)	0.50	30.97

We can see that SASI-P beat the non-intrusive baseline, but not the intrusive baseline. This might be expected, since the intrusive baseline receives more information than SASI-P. We can also see that the correlation values are significantly lower than what we typically see for non-intrusive speech quality models (see [2] or [4]). The reason for this could be two things. First, intelligibility and overall quality do not measure the same thing, and potentially, intelligibility is more difficult to measure for a non-intrusive system. Second, there is only one transcription on each clip, suggesting that the quality might be low. The speech quality datasets typically have 4 or more ratings per clip, to reduce noise. A low number of raters per clip increases noise, which limits what the performance measures can achieve.

4. CONCLUSION

We presented SASI-P, a non-intrusive lyric intelligibility prediction system for the Cadenza CLIP1 Challenge. SASI-P combines spectral representations with SSL representations for high-bandwidth stereo signal input. We demonstrated a correlation of 0.53 and an RMSE value of 30.98 on the validation data.

5. REFERENCES

[1] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-min Wang, Tomoki Toda, and Junichi Yamagishi, "The voicemos challenge 2022," 09 2022, pp. 4536–4540.

[2] Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan KA Reddy, Christian Schüldt, and Saikat Chatterjee, "Dnsmos pro: A reduced-size dnn for probabilistic mos of speech," in *Proc. Interspeech 2024*, 2024, pp. 4818–4822.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[4] Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan KA Reddy, Christian Schüldt, and Saikat Chatterjee, "Multivariate probabilistic assessment of speech quality," in *Proc. Interspeech 2025*, 2025.

[5] Xinyu Liang, Fredrik Cumlin, Christian Schüldt, and Saikat Chatterjee, "Deepmos: Deep posterior mean-opinion-score of speech," in *Interspeech 2023*. Aug 2023, ISCA.