CADENZA LYRIC INTELLIGIBILITY PREDICTION CHALLENGE

Jack Webb

Dyson School of Design Engineering, Imperial College London, UK

ABSTRACT

Hearing aid users often struggle to distinguish sung lyrics, with the link between audio signal features and lyric saliency remaining poorly understood. This report presents two dual-branch neural networks combining fine-tuned speech model embeddings with spectro-temporal features to predict lyric intelligibility.

1. INTRODUCTION

Difficulties deciphering lyrics are one of the most common problems reported by hearing aid users when listening to music [1]. Root causes are shared between the perceptual effects of hearing loss, as well as the reduced bandwidth, dynamic range and smeared spectro-temporal cues introduced by the compressive amplification in a modern hearing aid [2]. In the communications domain, automatic assessments of speech intelligibility have led to improved processing without the expenses associated with running a listening test with human participants [3]. However, there are no existing models for predicting lyric intelligibility in hearing aids. This work introduces two neural network models to estimate lyric intelligibility. Embeddings from a self-supervised learning (SSL) speech model are combined with acoustic features extracted by a convolutional neural network (CNN) to capture lyrics against competing acoustic information. A reference-free model that uses only the processed signal is presented alongside a reference-based counterpart that also uses vocals estimated from the unprocessed signal via Demucs [4].

2. METHODS

Since both models share frontend feature extraction modules, these are outlined first before the model details are discussed in turn. In order to reduce task complexity, all input audio is collapsed to mono and downsampled to 16 kHz.

2.1. Speech Model Features

To obtain cues relevant to word recognition, the WavLM pretrained SSL speech model forms the basis of the frontend [5]. A lightweight fine-tuning is used to specialise the embeddings for sung lyrics and the specific degradations that occur in hearing aids. Learnable weights are obtained for each hidden state of the WavLM model, and used to compute a final weighted sum. This final fine-tuned embedding is then projected from a dimension of 768 to 128 using a multilayer perceptron (MLP) with a single hidden layer to reduce the capacity required for downstream pooling and regression.

Initially, this strategy was tested with Whisper [6], using token probability divergences to estimate intelligibility in a manner similar to [7]. However, Whisper's training prioritizes transcription robustness over capturing perceptual degradations, yielding ambiguous relationships between the token probability distributions and lyric saliency. WavLM was therefore chosen for speech feature extraction.

2.2. Acoustic Features

Equally, while WavLM will capture some acoustic context from the input audio, it is also not explicitly trained for this purpose, meaning the model may discard cues that are perceptually relevant. Accordingly, a strided CNN is used to capture frequency-dependent acoustic cues from a log Mel spectrogram of the input audio. These features complement the WavLM embeddings and capture masking acoustic information from other sound sources. The CNN output dimension is restricted to 48, as high-level semantic embeddings are likely to be better modelled by WavLM.

2.3. Reference-based Model (A)

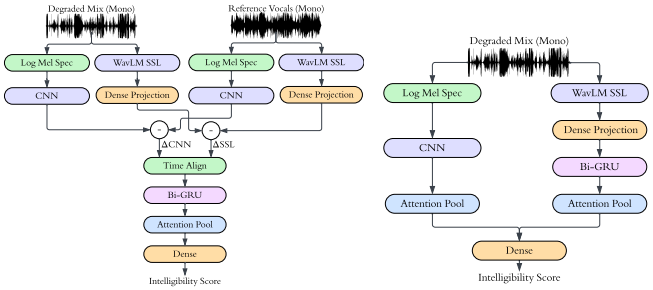
The reference-based model (Fig. 1a) uses a vocal reference extracted from the unprocessed signal using Demucs. Both the vocal reference and processed signal are passed through the SSL speech model and CNN, producing reference and degraded embeddings for each input frame. Framewise signed differences are then calculated between the reference and degraded embeddings for each branch, giving Δ_{CNN} and Δ_{SSL}.

Following time-alignment, Δ embeddings are concatenated and undergo temporal pooling. This is achieved through a bi-directional gated recurrent unit (GRU) with a hidden dimension of 64, and an attention-based pooling mechanism. This results in a single pooled tensor per extract. The final intelligibility prediction is computed with a dense MLP consisting of two hidden layers.

Work supported by an EPSRC CASE conversion PhD scholarship, co-funded by Sonova.

2.4. Reference-free Model (B)

Substantially simpler, the reference-free model (Fig. 1b) passes the processed signal through the SSL speech model and CNN independently. The resulting time series are pooled separately: the SSL series using a small GRU and attention pooling combination, and the lighter CNN features using attention pooling only. Pooled tensors are concatenated and fed into a similar MLP to the reference-based model for the intelligibility prediction, albeit with a smaller input dimension.



(a) Reference-based Model (b) Reference-free Model

Fig. 1. Model Architectures.

2.5. Training and Validation

Both prediction models were trained using a weighted combination of mean squared error, and a differentiable loss metric based on Spearman's rank correlation coefficient to discourage the model from collapsing to mean regression. Generalisation was monitored using an 85%/15% train-test split of the training data. On an i5 CPU, inference runs at approximately $0.91\,\mathrm{s\,it^{-1}}$ for the reference-free model, and $1.88\,\mathrm{s\,it^{-1}}$ for the reference-based model if source separation is processed offline.

3. EVALUATION

On unseen validation and evaluation datasets, moderate positive correlations between the model predictions and expected scores are achieved. Both models exceed the STOI baseline, but do not outperform the intrusive Whisper baseline. Pearson correlation and RMSE are reported for both datasets in Table 1.

	Val. r	Eval. r	Val. RMSE	Eval. RMSE
Ref. Based	0.44	0.48	34.22	32.77
Ref. Free	0.51	0.45	31.45	32.20
Whisper	0.59	0.58	29.32	29.08
STOI	0.14	0.21	36.11	34.89

Table 1. Pearson correlation and RMSE of different models, validation and evaluation datasets.

4. DISCUSSION & CONCLUSION

While the proposed models outperform the non-intrusive baseline, this advantage is limited. Potentially, this is a result of lightweight architectures that do not fully exploit linguistic temporal structures in the audio-lyric pairs. More comprehensive transformer-based backends may better capture prosodic patterns and semantic predictability, particularly in the case of the reference-based model, whose difference embeddings may fail to preserve important auditory cues. There is also likely an empirical ceiling to model performance, due to high variance in target intelligibility scores. Nonetheless, the moderate positive correlations suggest that along with the Whisper baseline, both models capture some cues relevant to lyric intelligibility in hearing aids.

5. REFERENCES

[1] Alinka Greasley, Harriet Crook, and Robert Fulford, "Music listening and hearing aids: perspectives from audiologists and their patients," *International Journal of Audiology*, vol. 59, no. 9, pp. 694–706, Sept. 2020.

[2] B. C. J. Moore, "Effects of sound-induced hearing loss and hearing aids on the perception of music," *Journal of the Audio Engineering Society*, vol. 64, no. 3, pp. 112–123, Mar. 2016.

[3] James M Kates and Kathryn H Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[4] Simon Rouard, Francisco Massa, and Alexandre Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[7] Angel Mario Castro Martinez, Constantin Spille, Jana Roßbach, Birger Kollmeier, and Bernd T Meyer, "Prediction of speech intelligibility with DNN-based performance measures," *Computer Speech & Language*, vol. 74, pp. 101329, 2022.