LIWhiz: A Non-Intrusive Lyric Intelligibility Prediction System for the Cadenza Challenge

RAM C. M. C. SHEKAR AND IVÁN LÓPEZ-ESPEJO*

*DEPT. OF SIGNAL THEORY, TELEMATICS AND COMMUNICATIONS, UNIVERSITY OF GRANADA, SPAIN

ramcharanmc@gmail.com, iloes@ugr.es

INTRODUCTION

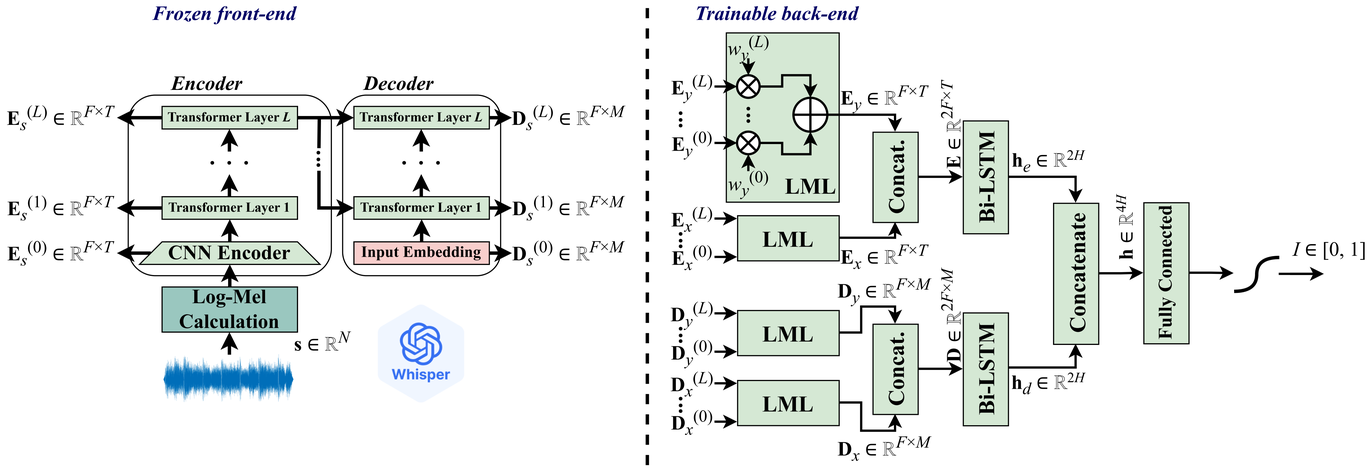
- Understanding lyrics is key to music enjoyment; however, listeners with hearing impairment often struggle to comprehend them
- The development of **lyric intelligibility prediction** (**LIP**) methods could enable new lyric enhancement technologies
- **LIWhiz**: A *non-intrusive* LIP system inspired by our prior work in no-reference speech intelligibility prediction (**wav2vec 2.0 adapted for ASR + lightweight back-end**)[1]

[1] H. Wang *et al.*, "No-Reference Speech Intelligibility Prediction Leveraging a Noisy-Speech ASR Pre-Trained Model," *in Proc. of Interspeech 2024*

EXPERIMENTAL SETUP

- To train the back-end, we use *only* the training partition of the **Cadenza Lyric Intelligibility Prediction** (**CLIP**) dataset
- Ground-truth lyric transcripts are **not** used
- The CLIP dataset is converted to mono and downsampled to 16 kHz
- k-fold cross-validation ($k = 10$) with early stopping (patience = 10 epochs)
- Optimizer: AdamW ($\texttt{lr} = 10^{-3}$)
- Loss function: Root mean square error (RMSE)

SYSTEM DESCRIPTION



$\mathbf{x} \in \mathbb{R}^N$: Original song excerpt

$\mathbf{y} \in \mathbb{R}^N$: Song excerpt with hearing loss simulation

$L = 32$, $F = 1{,}280$, $H = 512$

LML $\equiv$ Linear Mixing Layer

$I \in [0, 1]$: Lyric intelligibility score

- Motivated by **LyricWhiz**[2], a state-of-the-art automatic lyric transcription system (*an intrinsically related task*), we use a frozen Whisper Large v3 model for feature extraction
- Including $\mathbf{x}$ alongside $\mathbf{y}$ may help LIWhiz better adapt to the listener's degree of hearing loss

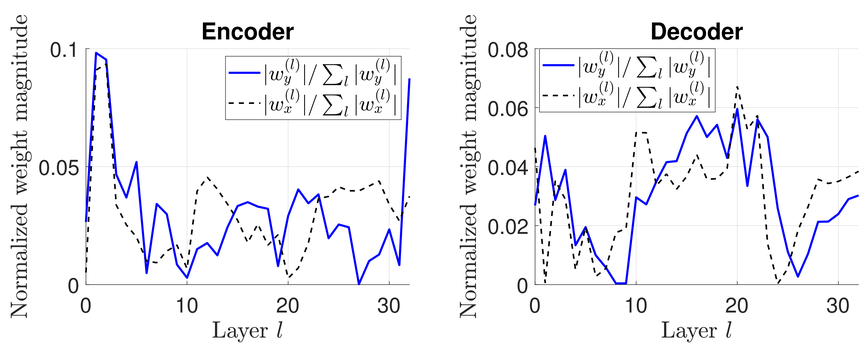
[2] L. Zhuo *et al.*, "LyricWhiz: Robust Multilingual Zero-Shot Lyrics Transcription by Whispering to ChatGPT," *in Proc. of ISMIR 2023*

RESULTS AND CONCLUSIONS

System	*CLIP validation set*		*CLIP evaluation set*	
	RMSE (%) ↓	**NCC** ↑	**RMSE (%)** ↓	**NCC** ↑
Baseline STOI	36.11	0.14	34.89	0.21
Baseline Whisper	29.32	0.59	29.08	0.58
LIWhiz w/o $\mathbf{x}$	27.36	0.66	27.34	**0.65**
LIWhiz	**27.13**	**0.67**	**27.07**	**0.65**

- LIWhiz substantially outperforms the baselines on both metrics and sets
- Including the original song excerpt $\mathbf{x}$ alongside $\mathbf{y}$ leads to a slight improvement in LIP performance

Normalized absolute values of the learned LML weights



1. Encoder: Shallowest layers → *low-level acoustic encoding*
2. Decoder: Middle layers → *encoder information integrated*