OSQA-SI: A Lightweight Non-Intrusive Analysis Model for Speech Intelligibility Prediction

Hsing-Ting, Chen[1], Po-Hsun Sung[1]

[1] Merry Electronics Co., Ltd.

{samuel.chen, peter.sung}@merry.com.tw

Abstract

This report proposes a non-intrusive speech intelligibility prediction model named Objective Sound Quality Analysis for Speech Intelligibility (OSQA-SI). The model adopts a simple sequential architecture, trained with a minimal number of parameters, and its performance is compared across two types of input acoustic features. Due to its extremely low parameter count, the model is suitable for real-time speech intelligibility assessment in real-world environments on mobile devices.

Index Terms: speech clarity, speech intelligibility, non-intrusive, hearing aid, hearing loss

1. Introduction

For individuals with hearing loss, the clarity of sound or the perception of semantic understanding is of paramount importance when using hearing assistance devices. Evaluating the intelligibility of various algorithms and subsequently adjusting them based on these evaluations has become a common optimization approach in modern development.

Analysis methods for measuring intelligibility, such as the Short-Time Objective Intelligibility (STOI) [1] and the Modified Binaural STOI (MBSTOI) [2], have effectively aided product technology development by assessing speech clarity. From a user perspective, individuals with hearing loss are particularly sensitive to these speech intelligibility issues. To accurately assess speech intelligibility while accounting for individual hearing levels, the Hearing-Aids Speech Perception Index (HASPI) [3] comprehensively considers the impact of hearing loss curves and inner ear hair cell damage on speech clarity.

However, the aforementioned analysis metrics are all based on Intrusive Analysis. They obtain corresponding intelligibility scores by comparing the processed signal with an ideal reference signal. While accurate, this method is challenging to apply in real-world user scenarios. Non-Intrusive Analysis methods don't require comparison with an ideal reference signal during analysis. Instead, they map the characteristics of the analyzed signal itself to a corresponding evaluation score, making them highly suitable for real-world testing. Most non-intrusive analysis systems are implemented through deep learning models, which learn to capture features that influence changes in speech clarity.

This report proposes a lightweight model designed for real-time speech intelligibility assessment on mobile devices. We also compare the model's correlation with subjective scores and its error when using two different acoustic features as input.

2. Model

The OSQA-SI model is primarily divided into three main components: Hearing Loss Simulation, Acoustic Feature Extraction, and Scoring Model.

2.1. Hearing Loss Simulator

We utilize the officially provided MSBG (Moore, Stone, Baer and Glasberg) hearing loss simulator [4-7] to preprocess the binaural hearing aid recordings. This processing references the listener's audiogram (e.g., mild, moderate, moderate-severe hearing loss) to simulate the perceived signal for that listener. The processed signal is then subjected to acoustic feature extraction.

2.2. Acoustic Features

We compared the model's performance using two types of input features: Mel-spectrograms and vocoder features derived from the WORLD [8] vocoder. Vocoder features have historically been used in text-to-speech or speech-to-text tasks as a converter between speech and acoustic characteristics. Input signals were uniformly resampled to a 48 kHz sampling rate, and acoustic features were extracted with a 5 ms hop size. Mel-spectrograms were extracted with 48 dimensions, while vocoder features were extracted with 63 dimensions, comprising 60 Mel-generalized cepstral (MGC) coefficients, log fundamental frequency (lf0), band aperiodicity (bap), and voiced/unvoiced (vuv). Both sets of binaural acoustic features were then fed into the scoring model for speech intelligibility score prediction.

2.3. Scoring Model Structure

The framework of the scoring model is illustrated in Figure 1. The model adopts a Sequential Model framework, which includes a Temporal Processing Model and a Score Mapping Model. The feature-transformed binaural signal features are first analyzed by the Temporal Processing Model, then passed through the Score Mapping Model to generate individual scores for each ear. Finally, the higher score between the two ears is selected as the predicted binaural score output.

The Temporal Processing Model comprises two sub-models: Bi-directional Long Short-Term Memory (Bi-LSTM) and Self-Attention. Bi-LSTM is used for hidden feature transformation and dimensionality reduction of temporal forward and backward information. Self-Attention then focuses on salient features across the entire signal to extract critical information from the input binaural signals. The Score Mapping Model includes Attention-Pooling, Max-Pooling, and a Regression Model. Attention-Pooling assigns different weights to various temporal features and outputs the corresponding

predicted scores. Max-Pooling then outputs the score from the better ear. Finally, a Regression Model maps this score to the subjective intelligibility scale.
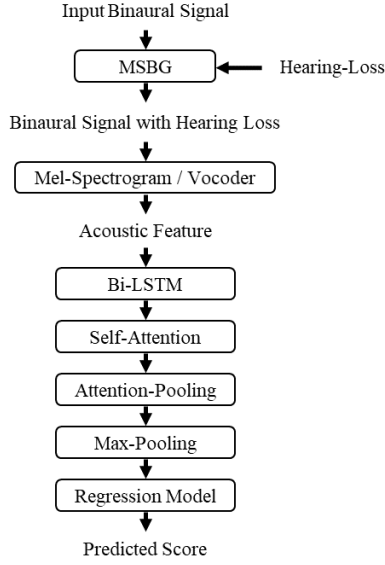
Input Binaural Signal
↓
MSBG ← Hearing-Loss
↓
Binaural Signal with Hearing Loss
↓
Mel-Spectrogram / Vocoder
↓
Acoustic Feature
↓
Bi-LSTM
↓
Self-Attention
↓
Attention-Pooling
↓
Max-Pooling
↓
Regression Model
↓
Predicted Score

Figure 1: *Schematic diagram of speech production.*

2.4. Training Score

Since subjective scores are derived from the overall binaural perception, to ensure the reliability of model training, we calculated HASPI scores for each ear's audio separately. The score from the better ear was used as the subjective ground truth, while the other ear was assigned a score proportionally. During training, the model was iteratively trained on single-ear signals. In the prediction phase, Max-Pooling was applied to select the score from the better ear as the final output.

2.5. Training Setup

Mean Squared Error (MSE) was used as the loss function during training. The Adam optimizer was employed with a learning rate of 10^{-6} for iterative training.

3. Result

We compared two OSQA-SI models, each trained with a different acoustic feature set. Their root mean squared error (RMSE) and correlation coefficients (PCC) during training and development are shown in the table below. The model trained with vocoder features consistently outperformed the one trained with Mel-spectrograms across all metrics.

Table 1: *OSQA-SI training and develop performance result.*

Input Type	Model Param	Train RMSE	Train PCC	Dev RMSE	Dev PCC
Vocoder	55.6k	29.1	0.70	31.1	0.66
Mel-Spec	53.1k	34.3	0.65	33.1	0.65

4. Limitation and Future Work

Our proposed model is limited by its framework design, as it does not explicitly account for binaural interaction, making it challenging to accurately predict scores based on complex inter-aural characteristics. Furthermore, while Attention-Pooling provides time-weighted pooling for the overall signal, it cannot estimate the inherent vocabulary size or the duration of pronunciations within the input signal. We aim to address these issues in our future model designs. Given the model's remarkably low parameter count, we will also explore its deployment on edge computing devices such as mobile phones, to facilitate more convenient real-world measurements.

5. References

[1] C. H. Taal et al., "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[2] A. H. Andersen et al., "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[3] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[4] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *JASA*, vol. 94, no. 3, pp. 1229–1241, 1993.

[5] ——, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *JASA*, vol. 95, no. 4, pp. 2277–2280, 1994.

[6] B. C. J. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *JASA*, vol. 94, no. 4, pp. 2050–2062, 1993.

[7] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing"

[8] Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. Inf. Syst., 99-D*, 1877-1884.