# A Non-intrusive Binaural Speech Intelligibility Prediction for Clarity-2023

The 4th Clarity Workshop on Machine Learning Challenges for Hearing Aids (Clarity-2023)

Katsuhiko Yamamoto
AI Lab, CyberAgent, Inc.

CyberAgent AI Lab

# 1st Clarity Prediction Challenge (CPC1)

Barker+ (2022)

Clarity

## Tasks

| data | speech enhancement method for CEC1 | listeners' characteristics (e.g., audiogram) |
|---|---|---|
| Track1 | known | known |
| Track2 | unknown | unknown |

## Results

machine learning-based models

- [E30] baseline + classification + non-linear regression
- [E32&E29] Transformer-based ASR
- [E36] baseline + Conformer + classification + SSL
- [E33&E16] CNN + BLSTM + self-attention + SSL (MBI-Net)
- [Baseline] a "better-ear" model of STOI with MSBG hearing loss (HL) model

Table 1: *Evaluation of 15 submitted systems plus baseline for RMS prediction error (RMSE) and ground-truth vs prediction* c… *are shown for closed set (Track 1)* a… *tr' Yes indicates an intrusive sys-* te… *ndly always guessing the mean of* *the tra… data intelligibility.*

Yes: intrusive
No: non-intrusive

| Entrant | Intr. | Track 1 (closed) RMSE ↓ | Track 1 (closed) Corr ↑ | Track 2 (open) RMSE ↓ | Track 2 (open) Corr ↑ |
|---|---|---|---|---|---|
| E30 [22] | Yes | **22.5 ± 0.5** | 0.79 | – | – |
| E32 [23] | Yes | 23.1 ± 0.5 | 0.77 | **23.5 ± 0.9** | 0.76 |
| E29 [24] | No | 23.3 ± 0.5 | 0.77 | 24.6 ± 1.0 | 0.73 |
| E36 [25] | Yes | 24.0 ± 0.5 | 0.76 | 29.2 ± 1.2 | 0.60 |
| E33 [26] | No | 24.1 ± 0.5 | 0.75 | 28.9 ± 1.1 | 0.65 |
| E16 [26] | No | 24.7 ± 0.5 | 0.74 | 30.7 ± 1.2 | 0.59 |
| E22 [27] | No | 25.9 ± 0.5 | 0.70 | 32.1 ± 1.2 | 0.54 |
| E19 [28] | Yes | 27.5 ± 0.6 | 0.66 | 28.1 ± 1.1 | 0.63 |
| Base. [1] | Yes | 28.5 ± 0.6 | 0.62 | 36.5 ± 1.4 | 0.53 |

CyberAgent AI Lab

# MBI-NET

Edo-Zerario+ (2022)

Clarity

## A non-intrusive SI prediction model for each ear

- Pre-processing
  - MSBG hearing loss (HL) model
- Input features for the DNN
  - spectrogram (STFT)
  - learnable filterbank (LFB)
  - self-supervised learned model (SSL)
- Outputs
  - frame-level SI (Left)
  - frame-level SI (Right)
  - frame-level SI (avg. of Left & Right)
  - SI (avg. of overall frames)

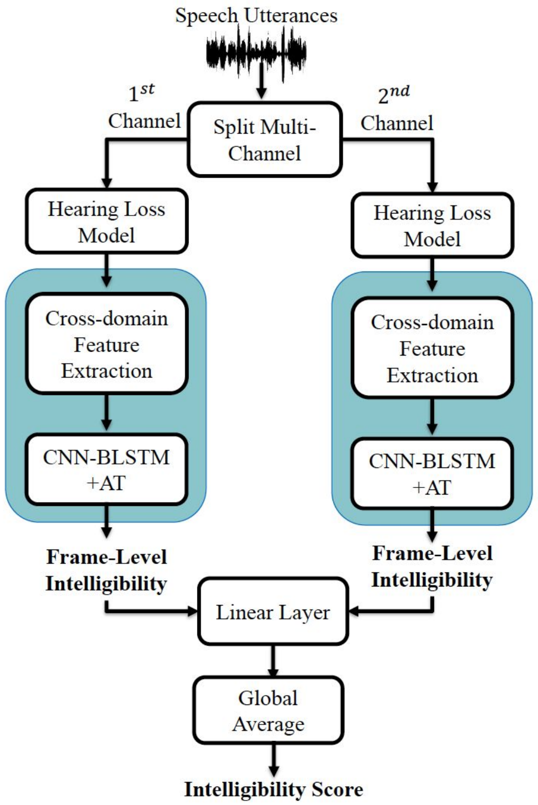



Figure 1: *Architecture of the MBI-Net model.*

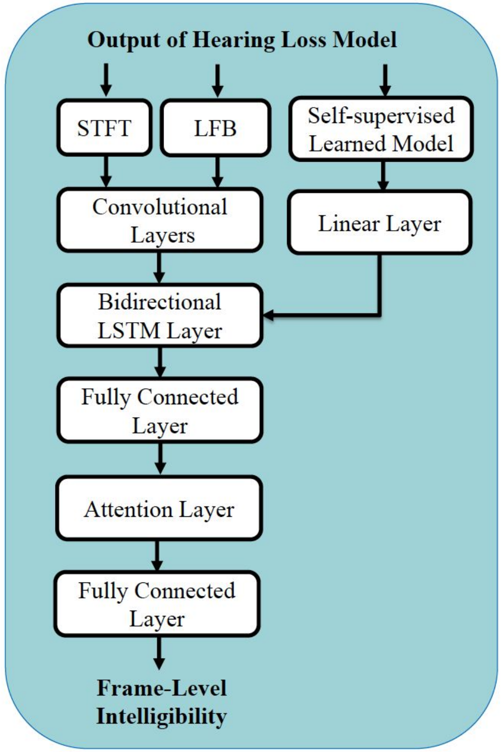



Figure 2: *Illustration of extraction cross-domain feature and obtaining frame-level intelligibility score on CNN-BLSTM+AT architecture.*

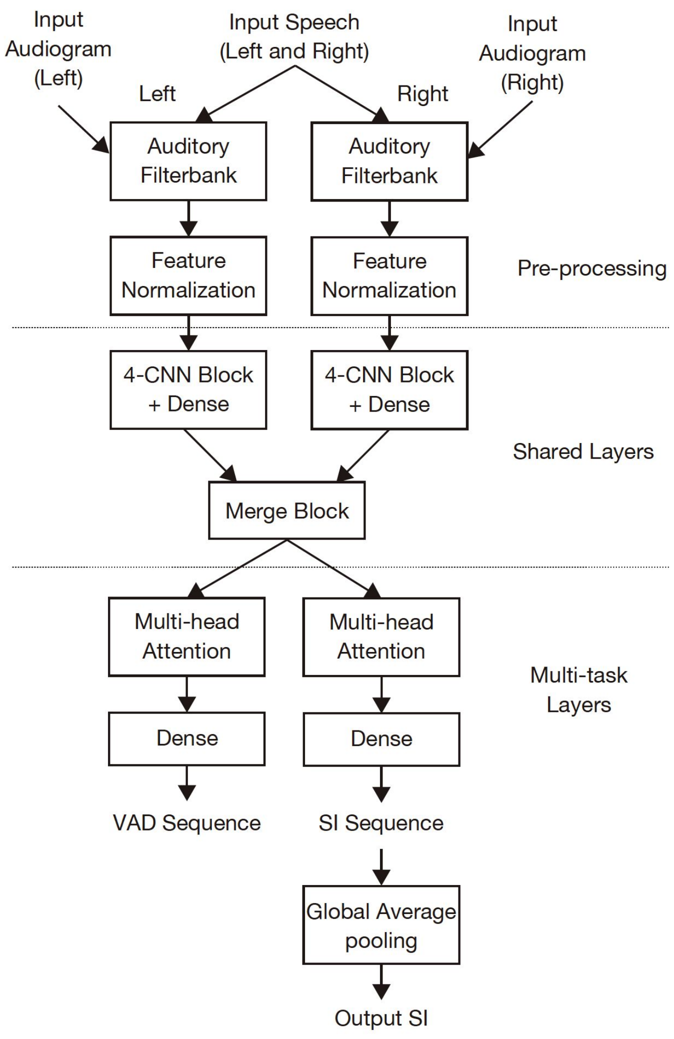
CyberAgent AI Lab

Clarity

# Overview

## A non-intrusive SI prediction using binaural information

- Pre-processing
  - auditory filterbank with listeners' characteristics
  - feature normalizations
- Input features for the DNN
  - normalized auditory spectrogram
- Outputs
  - frame-level SI (Left)
  - frame-level SI (Right)
  - frame-level SI (avg. of Left & Right)
  - SI (avg. of overall frames)



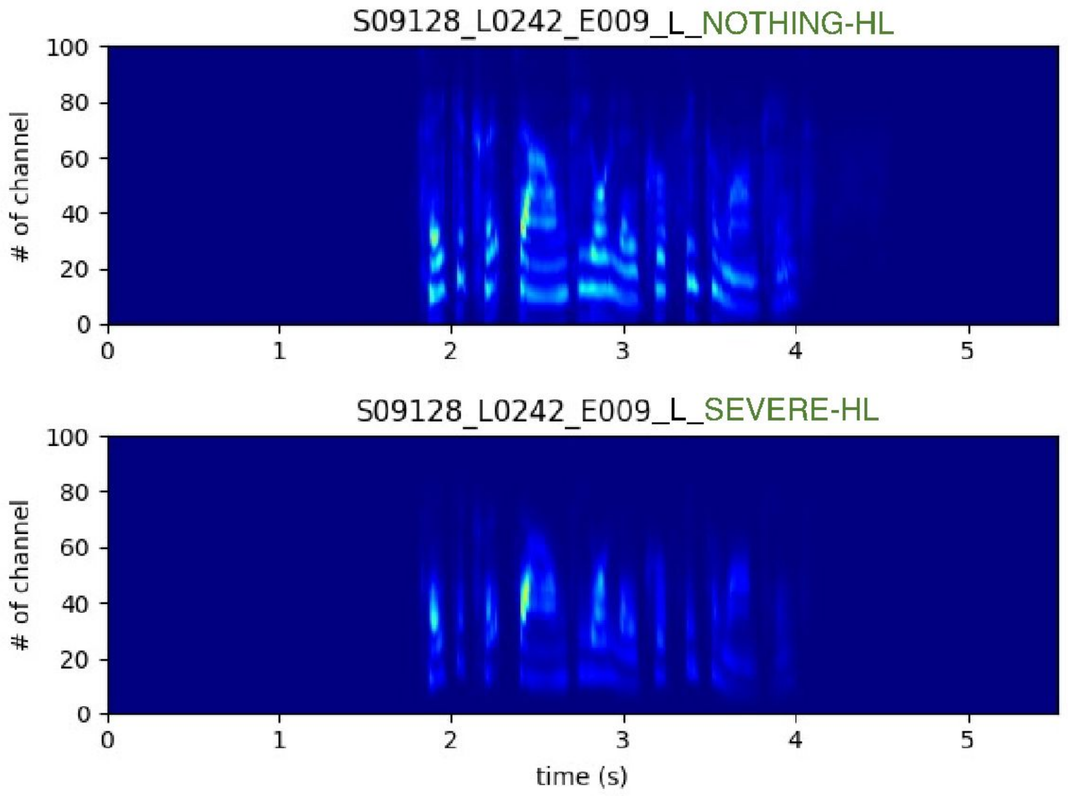
CyberAgent AI Lab

# Pre-processing

Clarity

## A new version of Gammachirp filterbank (Irino, 2023)

- Level-dependent and non-linear processing
  - asymmetric filter shape
  - compression
- Two parameters for listeners' characteristics
  - audiogram
  - healthiness of the compression




## Feature normalization (Andersen+, 2018)

- Down-sampling to 10,000 Hz
- Normalizing with long-time frames (≈384 ms)
  - mean to 0
  - variance to 1

CyberAgent AI Lab

Clarity

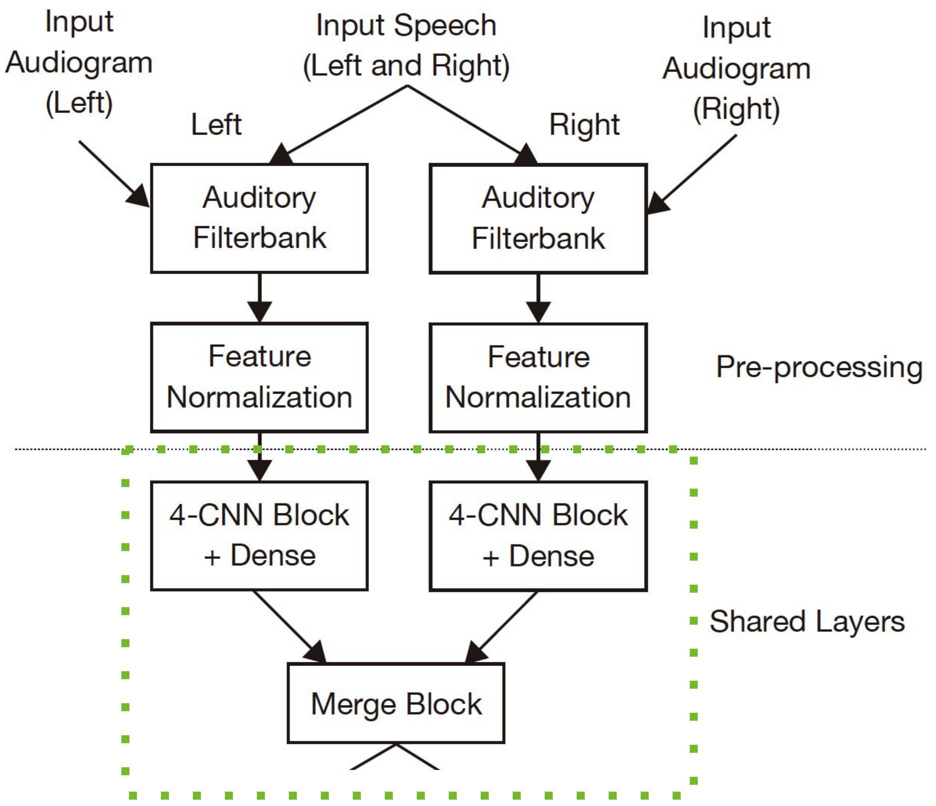
# Shared Layers

## 4-CNN block + Dense (Zezario+, 2023)

- 2-D convolutional layers
  1. kernel: 3×3, strides: 1×1
  2. : 3×3, : 1×1
  3. : 3×3, : 1×3

Repeat 4 times with different hidden layers

- Flatten layer
- Dense layer




## Merge block

- Concatenates Left and Right channels
- Fuses by a dense layer with 128 ReLU nodes
- Regularizes by a dropout layer

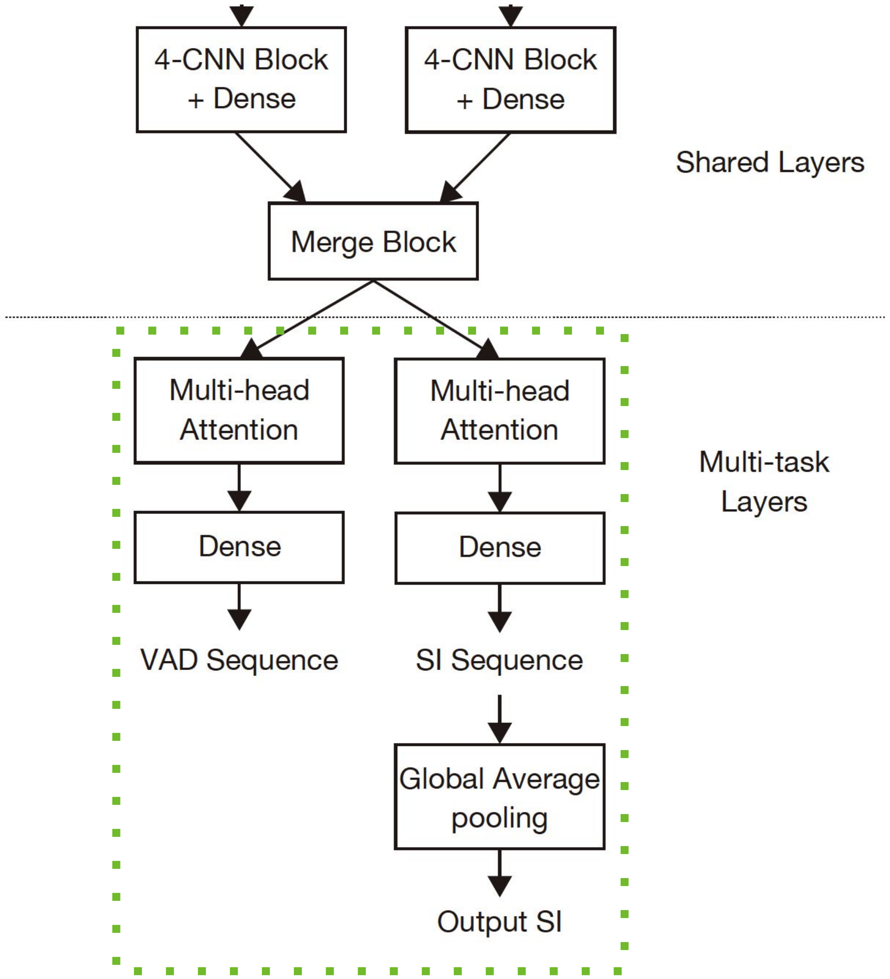
# Multi-task Layers

Clarity

## Multi-task learning (MTL)

- MTL improves the prediction accuracy of each task
- Previous study uses for SI prediction models with:
  - speech quality
  - other objective metrics

Chiang+ (2021)




## MTL for the proposed model

- Tasks:
  - speech intelligibility (SI)
  - voice activity detection (VAD)
- Architectures:
  - multi-head attention
  - dense layer for output sequences
  - global average pooling for the single output SI

CyberAgent AI Lab

Clarity

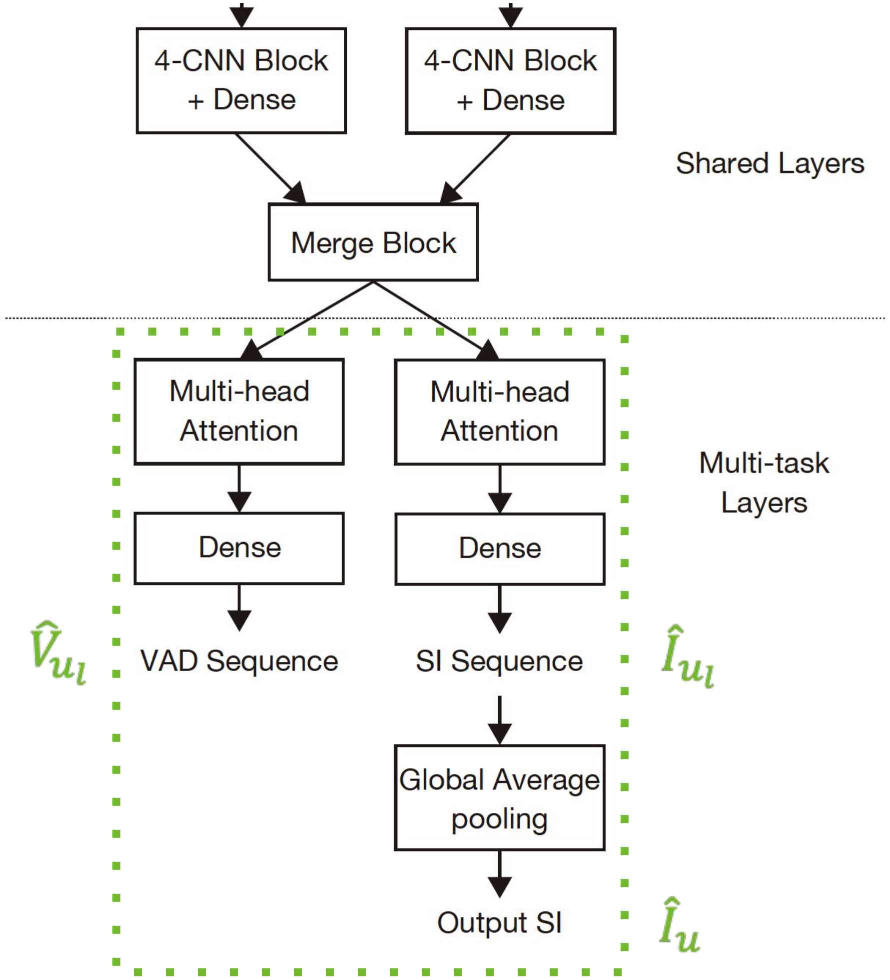
# Objective Function

## A Combination of SI and VAD

- Correct Labels
  - SI: speech intelligibility（utterance: $I_u$）
  - VAD: ideal duration of the target utterance（each frame: $V_l$
- Output of the model
  - SI: speech intelligibility（utterance: $\hat{I}_u$ & each frame: $\hat{I}_{u_l}$ ）
  - VAD: probability of the duration（each frame: $\hat{V}_{u_l}$ ）

$$O = \frac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \frac{1}{L_u}\sum_{l=1}^{L_u}(I_u - \hat{I}_{u_l})^2$$
$$- \frac{1}{L_u}\sum_{l=1}^{L_u}\{(V_{u_l}\log\hat{V}_{u_l}) + (1 - V_{u_l})\log(1 - \hat{V}_{u_l})\}]$$

binary cross-entropy



# Experimental Set-Up

Clarity

## Dataset

- CEC2 (target of CPC2)
- CEC1

## Separation of dataset

- Training: 90%
- Validation (Development): 10%

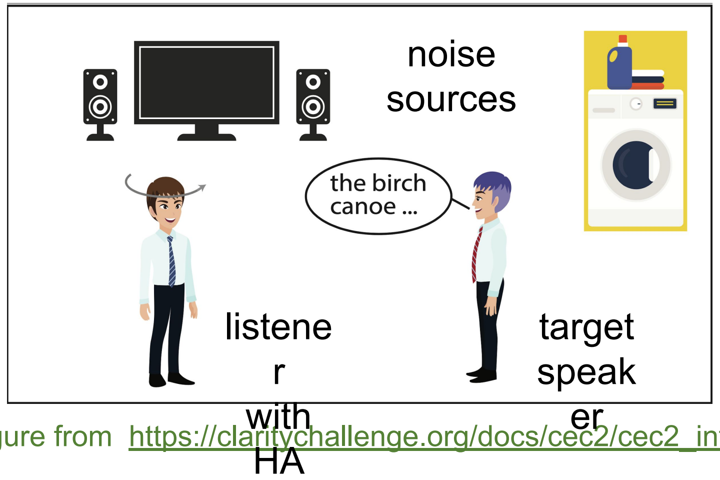
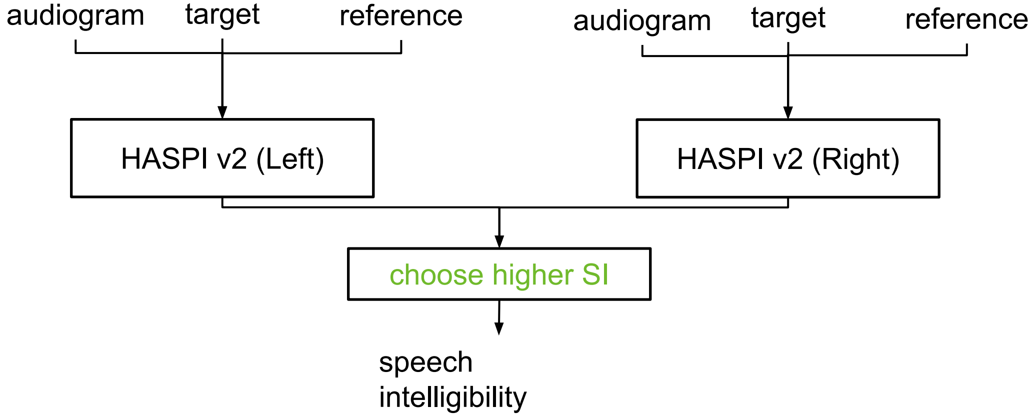



Figure from https://claritychallenge.org/docs/cec2/cec2_intro

## Baseline: a ”better ear” model of HASPI version 2 (Kates & Arehart, 2021)

- inputs (left & right)
  - target speech signals
  - reference signals (clean speech)
  - audiogram
- output
  - higher SI chosen in left/right channels



# Validation Sets

## Prediction Models

- Baseline (HASPI version 2)
- Proposed (ours)

## Evaluation Metrics

- RMSE (root-mean-squared error)
- NCC (normalized correlation coefficient)
- KT (Kendall's tau)

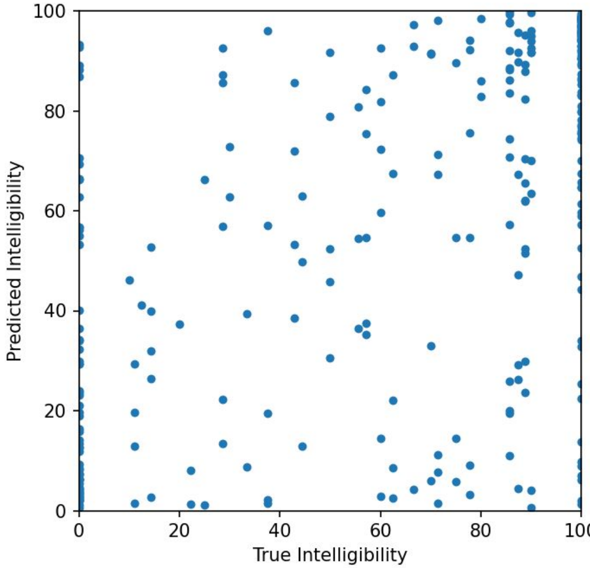
Our proposed model predicted SI with less RMSE than the baseline system.

| Dataset | Model | RMSE ↓ | NCC ↑ | KT ↑ |
|---|---|---|---|---|
| CEC2.train.1 | Baseline | 29.82 | 0.66 | 0.50 |
| | Proposed | **28.23** | **0.73** | **0.56** |
| CEC2.train.2 | Baseline | 30.06 | 0.68 | 0.51 |
| | Proposed | **27.47** | **0.76** | **0.58** |
| CEC2.train.3 | Baseline | 30.35 | 0.67 | 0.50 |
| | Proposed | **27.09** | **0.75** | **0.52** |
| CEC1.train.1 | Baseline | 26.56 | 0.68 | **0.43** |
| | Proposed | 21.62 | 0.68 | 0.36 |
| CEC1.train.2 | Baseline | 26.62 | **0.69** | **0.43** |
| | Proposed | 22.63 | 0.56 | 0.34 |
| CEC2.train.3 | Baseline | 26.48 | **0.67** | **0.43** |
| | Proposed | 22.19 | 0.58 | 0.29 |

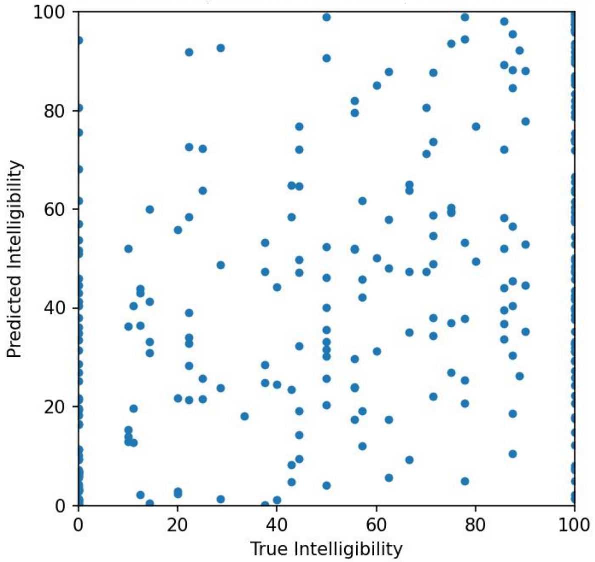
# Test Sets for CPC2

| Dataset | Model | RMSE ↓ | NCC ↑ | KT ↑ |
|---|---|---|---|---|
| CEC2.test.1 | Proposed | 34.88 | 0.59 | 0.45 |
| CEC2.test.2 | Proposed | 38.70 | 0.45 | 0.44 |
| CEC2.test.3 | Proposed | 31.09 | 0.74 | 0.58 |

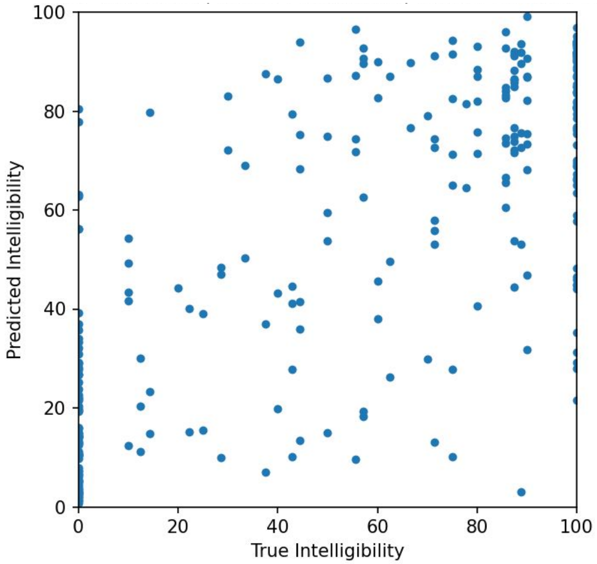
CEC2.test.1




CEC2.test.2




CEC2.test.3



# Discussions

Clarity

## Using output of auditory filterbank with audiograms

- It makes individual excitation patterns in humans’ cochlea.
- The normalization process is also crucial for inputs of DNN-based models.

## Combining binaural information in latent representations

- It is effective for the CEC2 dataset, including temporal changes due to head motions.
- It may be helpful to predict SI in more realistic environments.

## Setting Multi-task Learning

- Other information (e.g., speech direction) may enable for more speech-focused learning.

## To improve the prediction accuracy for test data:

- The training dataset should be manually separated for unknown conditions.
- The individual listeners’ characteristics should be Embedded into the DNN.

# Conclusions

CyberAgent AI Lab

Clarity

## A non-intrusive binaural speech intelligibility prediction model

- Auditory filterbank with hearing-impaired listeners' audiogram
- Combination of latent representations from left and right channel
- Multi-task learning with:
  - speech intelligibility prediction task
  - voice activity detection task

## Experimental Results of the CPC2 Datasets

- Validation: The proposed model predicts SI with less RMSE than the baseline.
- Test: RMSEs of the predicted SI are 4-11% points higher than RMSEs for validation sets.

# References


- Barker+ (2022)
  Barker, J., Akeroyd, M., Cox, T.J., Culling, J.F., Firth, J., Graetzer, S., Griffiths, H., Harris, L., Naylor, G., Podwinska, Z., Porter, E., Munoz, R.V., The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction. *Proc. Interspeech 2022*, 3508-3512, 2022, doi: 10.21437/Interspeech.2022-10821

- Edo-Zezario+ (2022)
  Edo Zezario, R., Chen, F., Fuh, C.-S., Wang, H.-M., Tsao, Y., MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids. *Proc. Interspeech 2022*, 3944-3948, 2022, doi: 10.21437/Interspeech.2022-10838

- Irino (2023)
  T. Irino, "Hearing impairment simulator based on auditory excitation pattern playback: WHIS," IEEE Access, doi: 10.1109/ACCESS. 2023.3298673, 2023.

- Andersen+ (2018)
  A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1925–1939, 2018.

- Chiang+ (2021)
  H.-T. Chiang, Y.-C. Wu, C. Yu, T. Toda, H.-M. Wang, Y.-C. Hu, and Y. Tsao, "HASA-Net: A Non-Intrusive Hearing-Aid Speech Assessment Network," in Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 907–913, 2021.

- Kates & Arehart (2021)
  J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," Speech Communication, vol. 131, pp. 35–46, 2021.

# Introduction

Clarity

## Clarity Project

- **Clarity Enhancement Challenge (CEC)** to improve speech intelligibility (SI) for hearing aids (HAs)
- **Clarity Prediction Challenge (CPC)** to improve prediction accuracies of SI processed by HAs

## 2nd Clarity Prediction Challenge (CPC2)

- Objective: predicting correct SI in CEC2 dataset
  - More varied noise sources
  - The listener turns their head during the talking
- Two types of system:
  - Intrusive system with a clean speech reference
  - Non-intrusive system without any reference

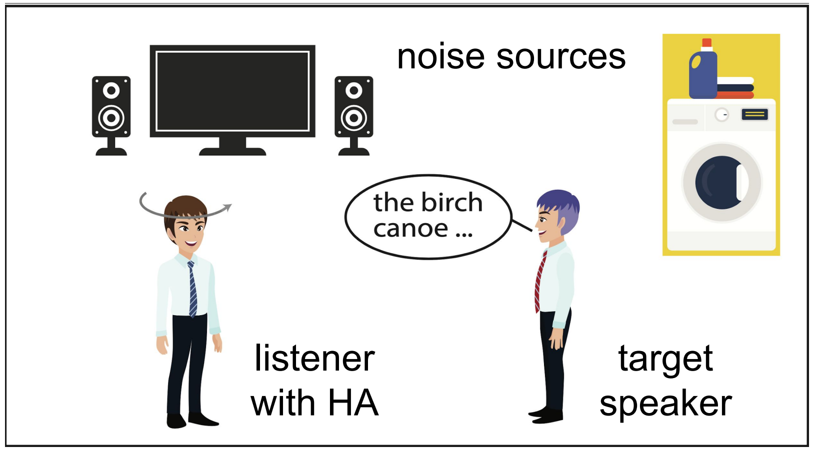



Figure from https://claritychallenge.org/docs/cec2/cec2_intro

CyberAgent AI Lab

Clarity

# Architecture of the MBI-NET

Edo-Zerario+ (2022)

- Left/Right ear
  1. MSBG Hearing Loss Model
  2. Extraction cross-domain features
     - Short-time Fourier's Transform (STFT)
     - Learnable filter banks (LFB)
     - Self-supervised learning model (SSL)
  3. Frame-level SI prediction
     - 4 CNN-block
     - BLSTM
     - Self-attention (AT)
- Fuse Left/Right ear
  1. Linear Layer
  2. Global Average Pooling
  3. Utterance-level SI prediction

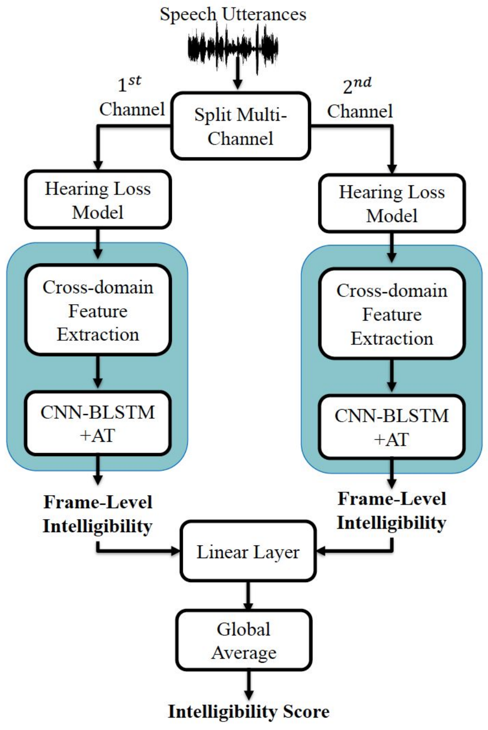



Figure 1: *Architecture of the MBI-Net model.*

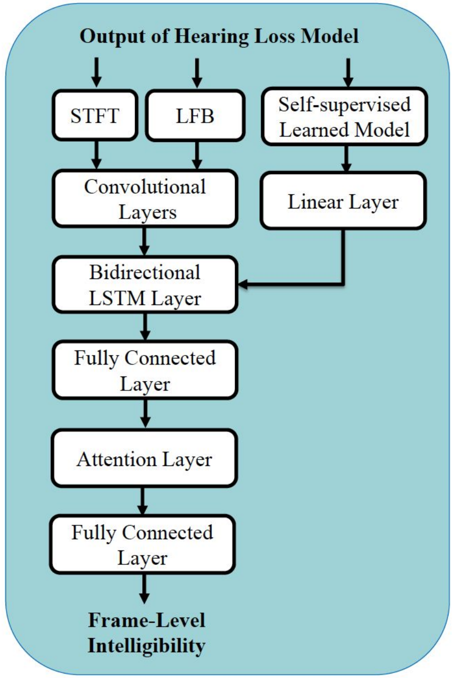



Figure 2: *Illustration of extraction cross-domain feature and obtaining frame-level intelligibility score on CNN-BLSTM+AT architecture.*

# Experiments

Clarity

## Dataset

- CEC2 (target of CPC2)
- CEC1

## Separation of dataset

- Training: 90%
- Development: 10%

| Dataset | Training | Development |
|---|---|---|
| CEC2.train.1 | 2449 | 272 |
| CEC2.train.2 | 2501 | 277 |
| CEC2.train.3 | 2494 | 277 |
| CEC1.train.1 | 5191 | 576 |
| CEC1.train.2 | 4774 | 530 |
| CEC2.train3 | 4598 | 510 |

## Baseline: a ”better ear” model of HASPI version 2 (Kates & Arehart, 2021)

- inputs
  - speech signals (left & right)
  - reference signals (clean speech)
  - audiogram (left & right)
- output
  - higher SI chosen in left/right channels

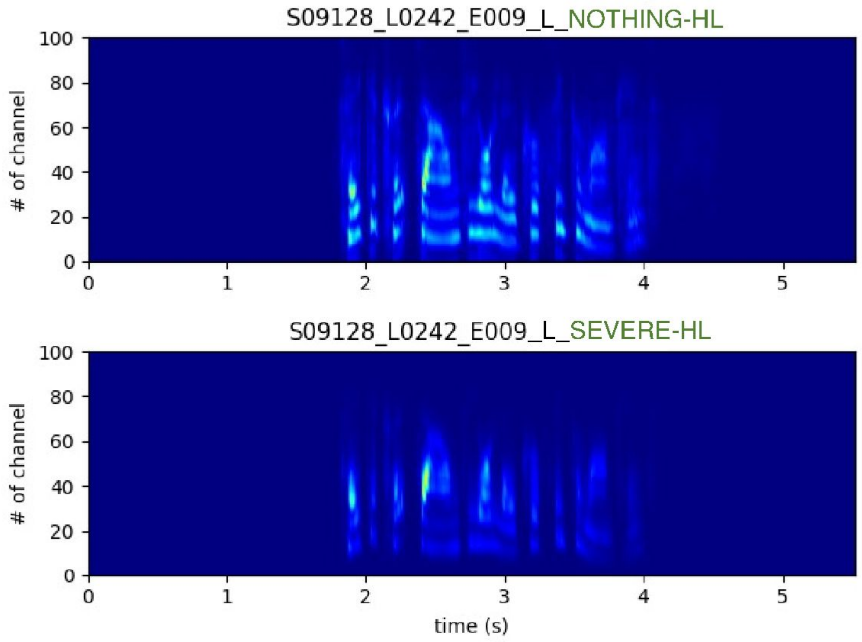
# Pre-processing

Clarity

## A new version of Gammachirp filterbank (Irino, 2023)

- Level-dependent and non-linear processing
  - asymmetric filter shape
  - compression
- Two parameters for listeners' characteristics
  - audiogram
  - health factor of the compression (0.00~1.00)

| HI Listener's Class | Avg. of Listener's Audiogram (dB) | Health Factor of Compression |
|---|---|---|
| NOTHING | 0.0~14.9 | 1.00 |
| MILD | 15.0~34.9 | 0.75 |
| MODERATE | 35.0~55.9 | 0.50 |
| SEVERE | 56.0~ | 0.25 |




## Feature normalization (Andersen+, 2019)

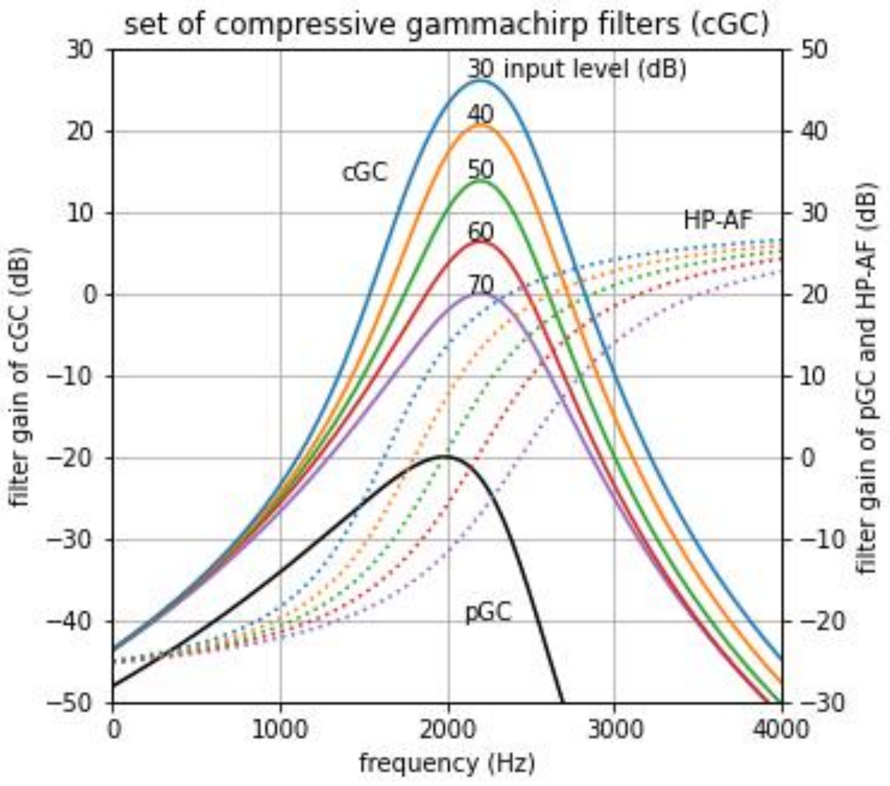
- Down-sampling to 10,000 Hz
- Normalizing with long-time frames (≈384 ms)
  - mean to 0
  - variance to 1

CyberAgent AI Lab

# Gammachirp Filterbank (GCFB$_{v23}$)

Irino (2023)

Clarity




https://github.com/kyama0321/gammachirpy

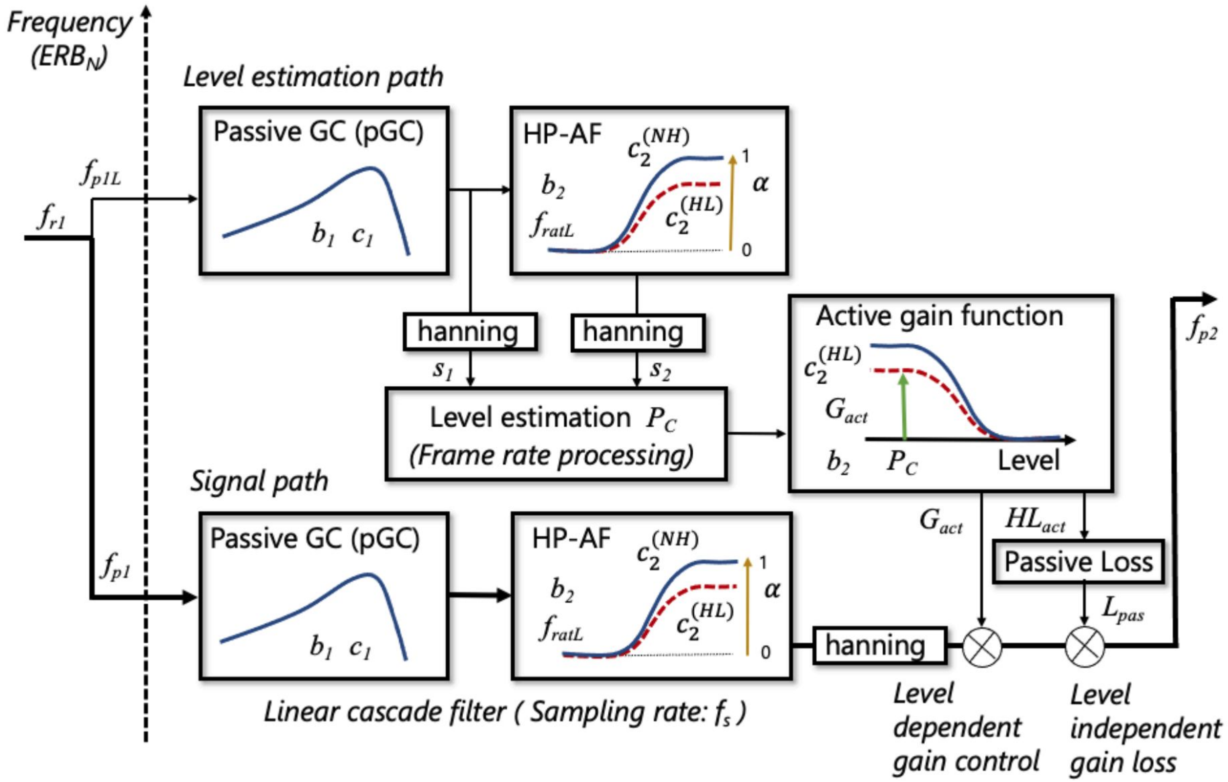


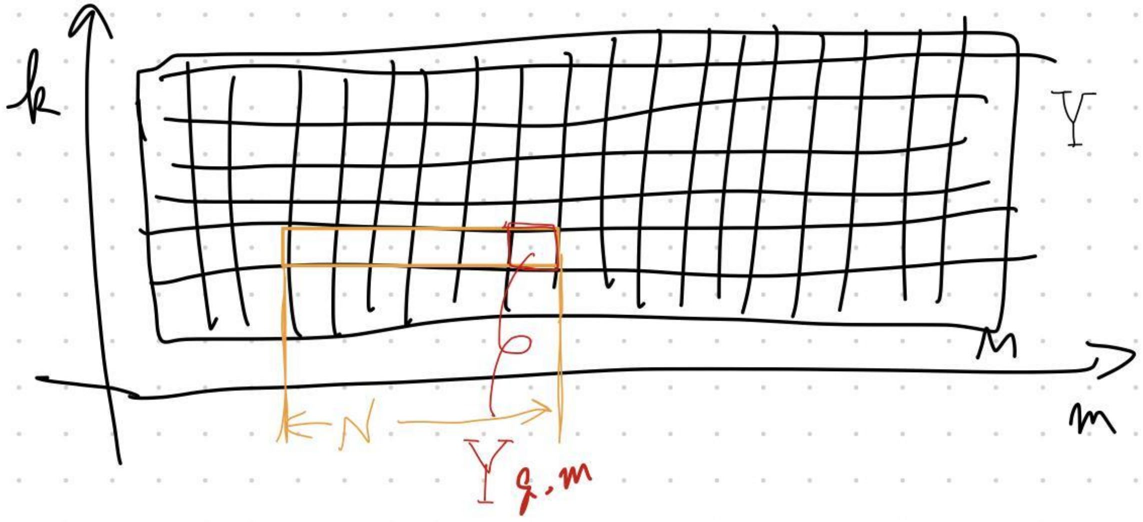

FIG. 1. Block diagram of one channel of the frame-based GCFB, GCFB$_{v23}$

CyberAgent AI Lab

Clarity

# Normalization Process

Andersen+ (2018)




an envelope spectrogram in time-frequency domain

The envelopes are mean- and variance normalized. We define the normalized envelope sample, $\bar{Y}_{q,m}$, by the two following steps:

$$\breve{Y}_{q,m} = Y_{q,m} - \frac{1}{N} \sum_{m'=m-N+1}^{m} Y_{q,m'}, \tag{2}$$

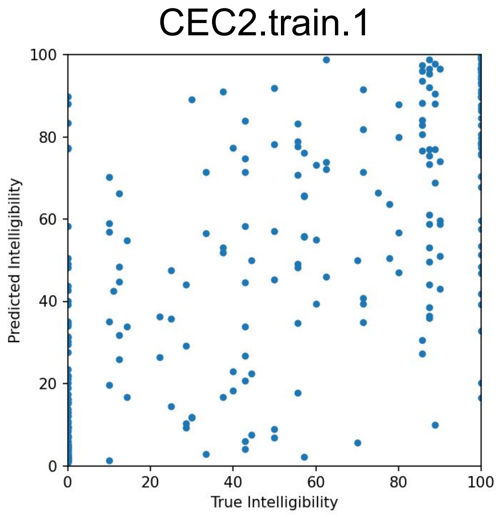
for $m = N, \ldots, M$, and:

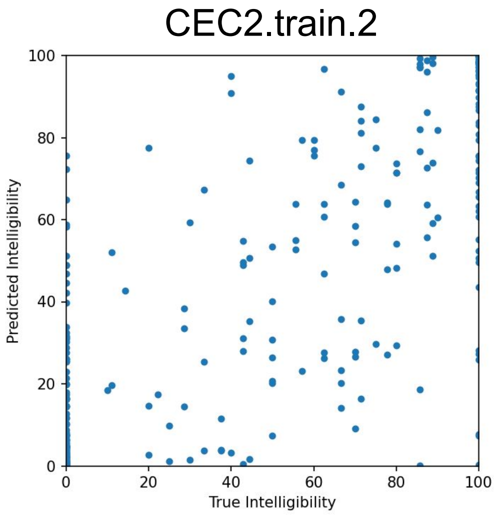
$$\bar{Y}_{q,m} = \frac{\breve{Y}_{q,m}}{\sqrt{\frac{1}{N} \sum_{m'=m-N+1}^{m} \breve{Y}_{q,m'}^2}}, \tag{3}$$

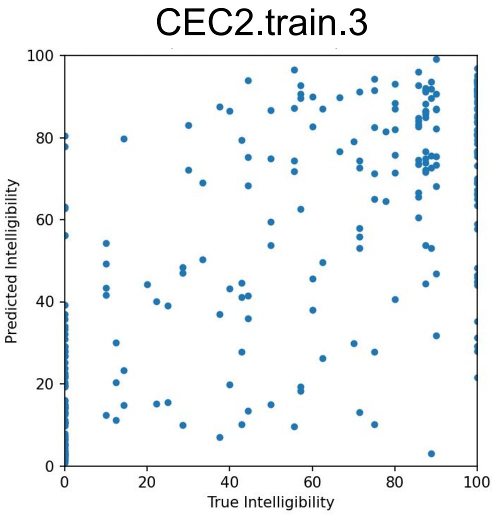
for $m = 2N - 1, \ldots, M$, where $\breve{Y}_{q,m}$ is a zero-mean intermediate variable, and $\bar{Y}_{q,m}$ is the normalized envelope. We use $N = 30$ envelope samples (corresponding to 384 ms) to estimate the mean and variance. The resulting normalized envelopes are defined for $Q = 15$ one-third octave bands, and for $L = M - 2N + 2$ time windows.

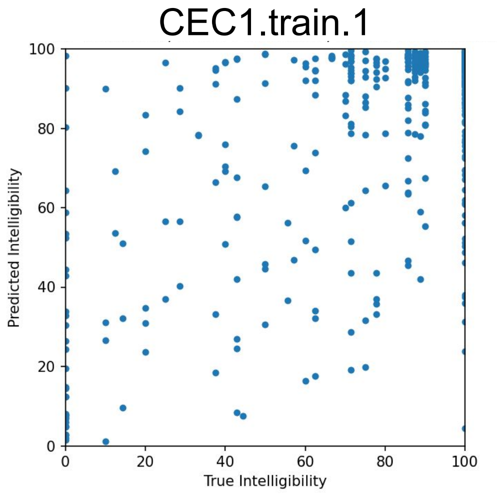
# Scatter Plots of the Proposed Model

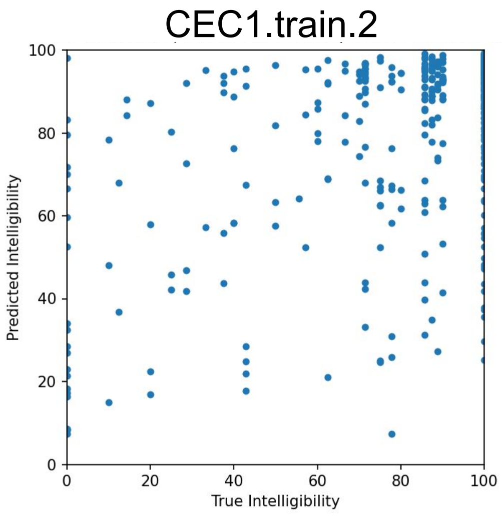
Clarity

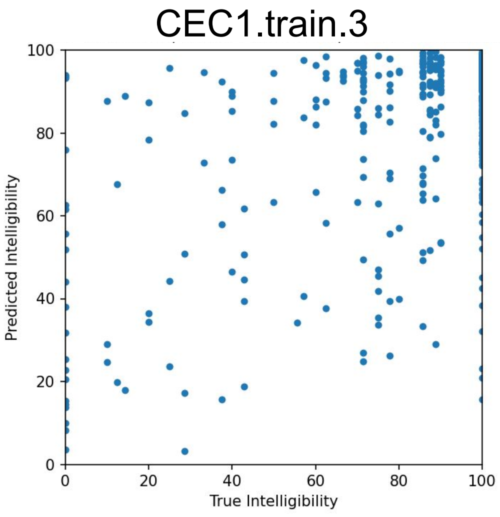


CyberAgent AI Lab

Clarity

# MBI-Net vs. Proposed Model

## Prediction Models

- MBI-Net with WavLM (Zezario+, 2022)
- Proposed: ours

## Evaluation Metrics

- RMSE (root-mean-squared error)
- NCC (normalized correlation coefficient)
- KT (Kendall's tau)

| Dataset | Model | RMSE ↓ | NCC ↑ | KT ↑ |
|---|---|---|---|---|
| CEC2.train.1 | MBI-Net | 29.47 | 0.70 | 0.55 |
| | Proposed | **28.23** | **0.73** | **0.56** |
| CEC2.train.2 | MBI-Net | 29.04 | 0.73 | 0.56 |
| | Proposed | **27.47** | **0.76** | **0.58** |
| CEC2.train.3 | MBI-Net | 28.25 | 0.72 | 0.52 |
| | Proposed | **27.09** | **0.75** | 0.52 |
| CEC1.train.1 | MBI-Net | **20.62** | **0.70** | **0.40** |
| | Proposed | 21.62 | 0.68 | 0.36 |
| CEC1.train.2 | MBI-Net | **20.48** | **0.62** | 0.32 |
| | Proposed | 22.63 | 0.56 | **0.34** |
| CEC2.train.3 | MBI-Net | **21.06** | **0.59** | **0.31** |
| | Proposed | 22.19 | 0.58 | 0.29 |

## Discussion

- SSL features help SI prediction accuracy for the CEC1 dataset (Zezario+, 2022).
- More implementation is needed for the CEC2 dataset (adaptation for temporal changes?)

CyberAgent AI Lab

Clarity

# Pre-processing part