Predicting Speech Intelligibility using the Spike Activity Mutual Information Index

Franklin Alvarez and Waldo Nogueira

Auditory Prosthetic Group, Hannover Medical School

Cluster of Excellence "Hearing4All"

1st Clarity Prediction Challenge (2021-2022)

29/06/2022

Information Theory

- **Shannon Entropy:**

It measures the amount of ***uncertainty*** or ***information*** that we are receiving by ***looking*** at a random event.

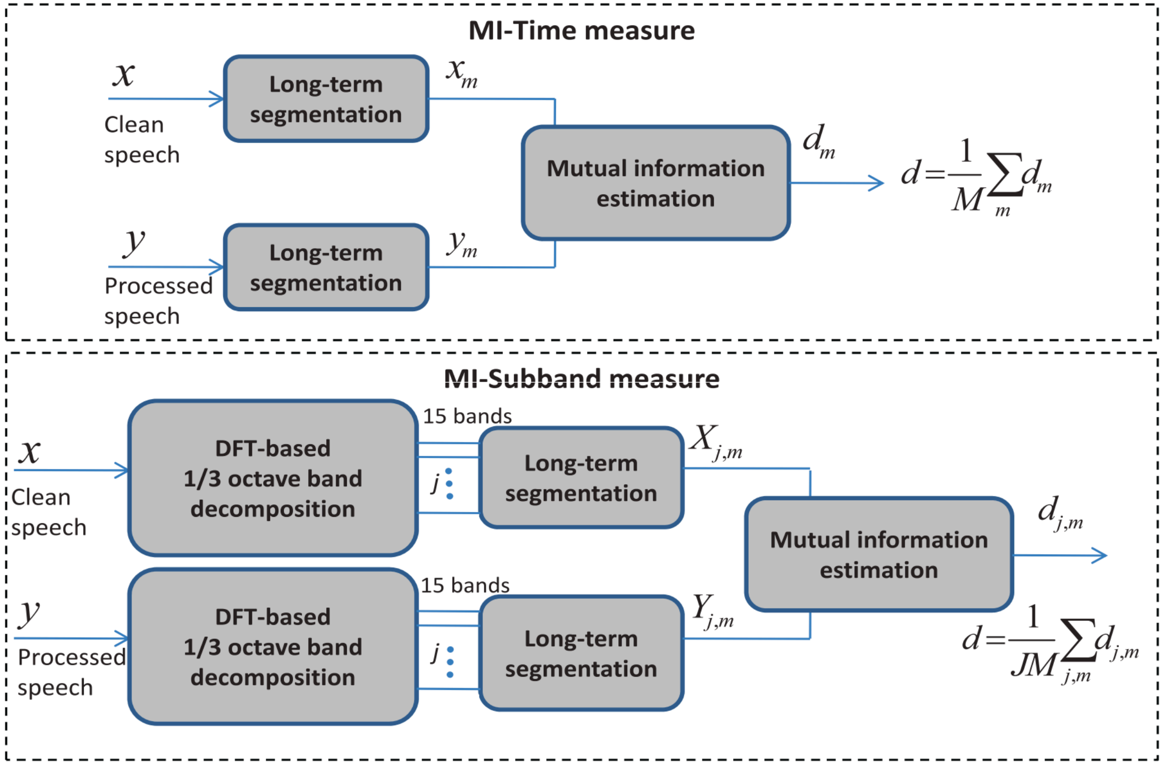
- **Mutual Information:**

Given two random events “S” and “R”, it measures how much ***information*** you will get off “S” by ***looking*** at “R”, and vice versa.

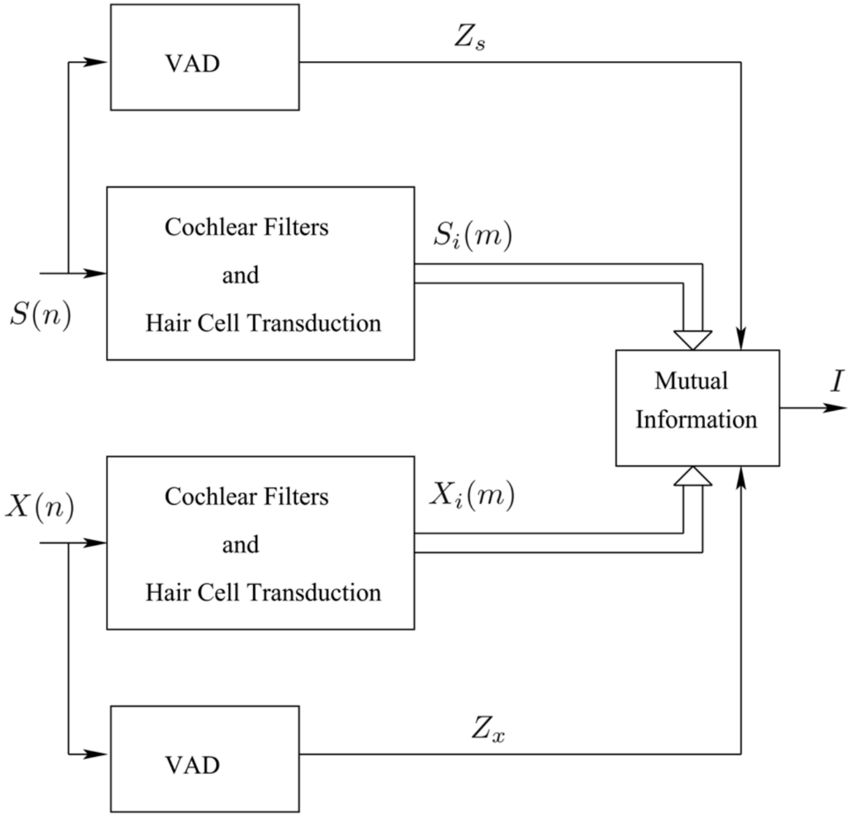
- Information is measured in bits!

Mutual Information as Speech Intelligibility Metric

Taghia et al. (2012)



Jensen and Taal (2014)

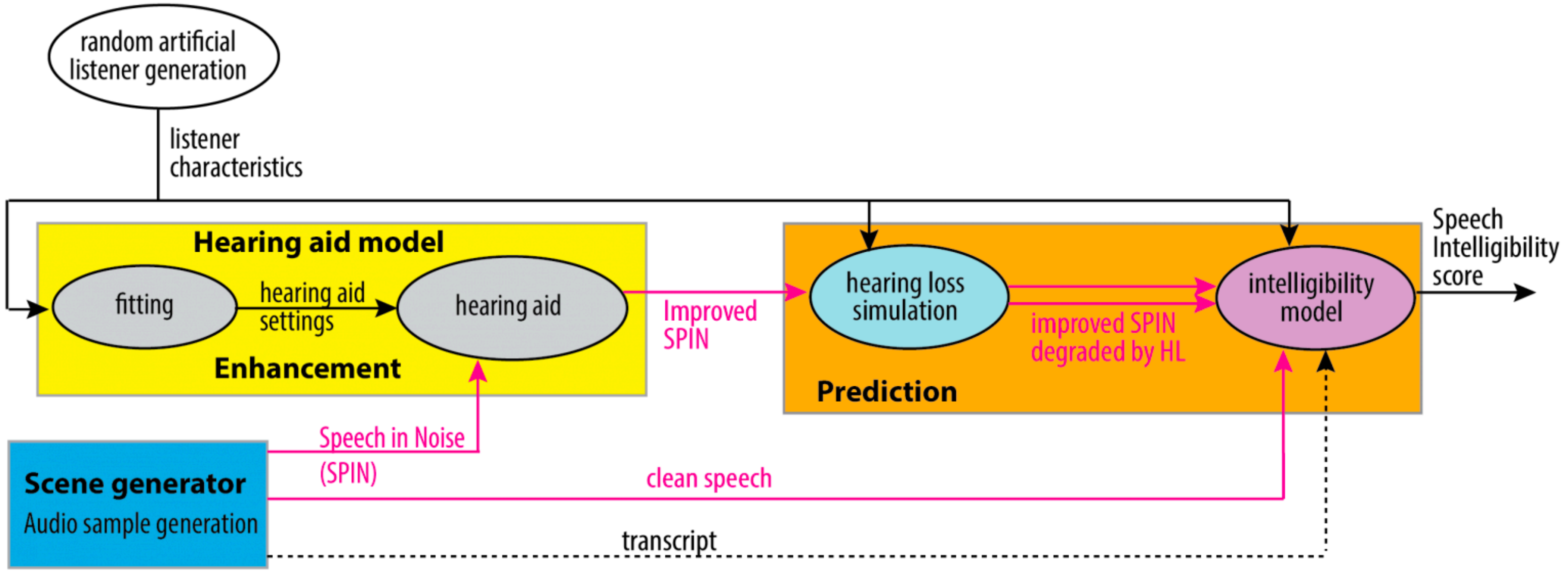


Motivation

- What if we add a peripheral auditory model to get a “lower-level” representation of the perceived sound (neural activity)?
- This would allow us to study the effects of more physiological aspects in speech intelligibility (neural health conditions, damage in the middle ear, different pathologies).
- For cochlear implant (CI) users, the objective speech intelligibility is computed with vocoders, not taking into account any physiological aspect of the implantation.

First Clarity Prediction Challenge

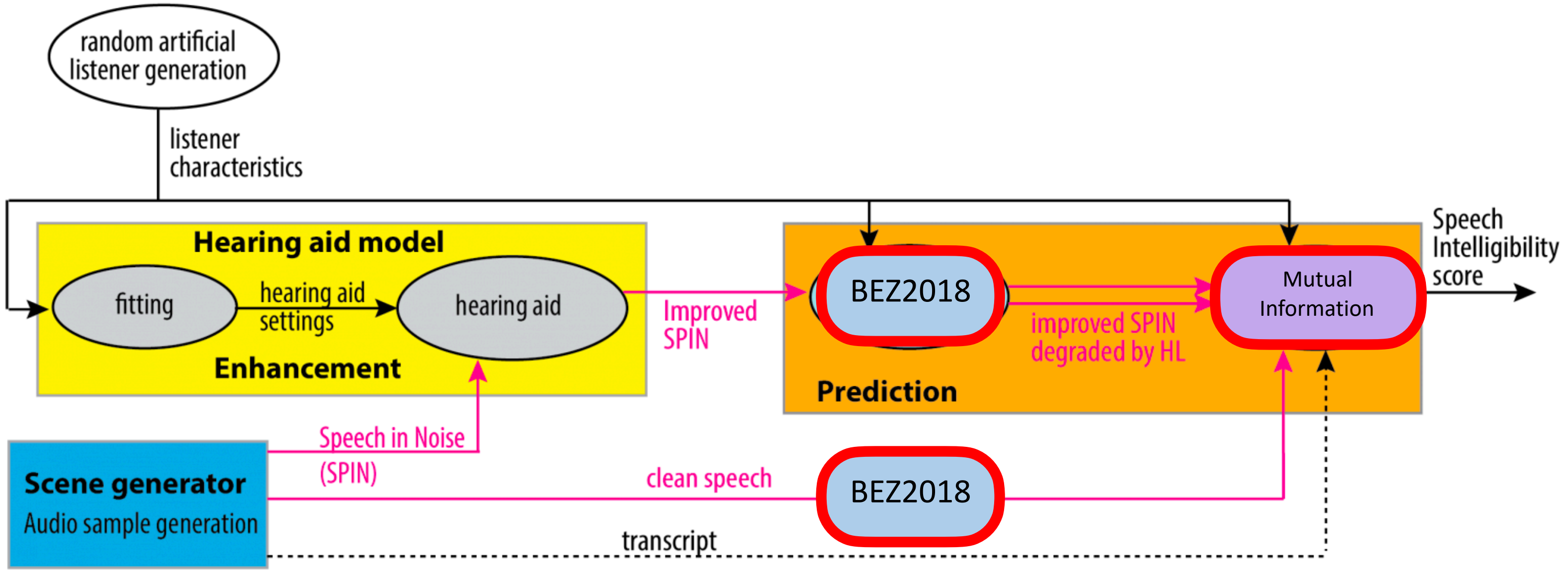
- Our proposal uses a peripheral auditory model which output is the spike train produced in the auditory nerve fibers (Bruce et al., 2018). Referred to as BEZ2018 model.
- The BEZ2018 is able to simulate the physiological damage causing the hearing loss from the listener audiogram.
- The intelligibility model in our proposal is based on the mutual information between the spike trains of the clean speech and the improved speech-in-noise (SPIN) degraded by the hearing loss (HL).



https://claritychallenge.github.io/clarity_CC_doc/docs/cpc1/cpc1_baseline

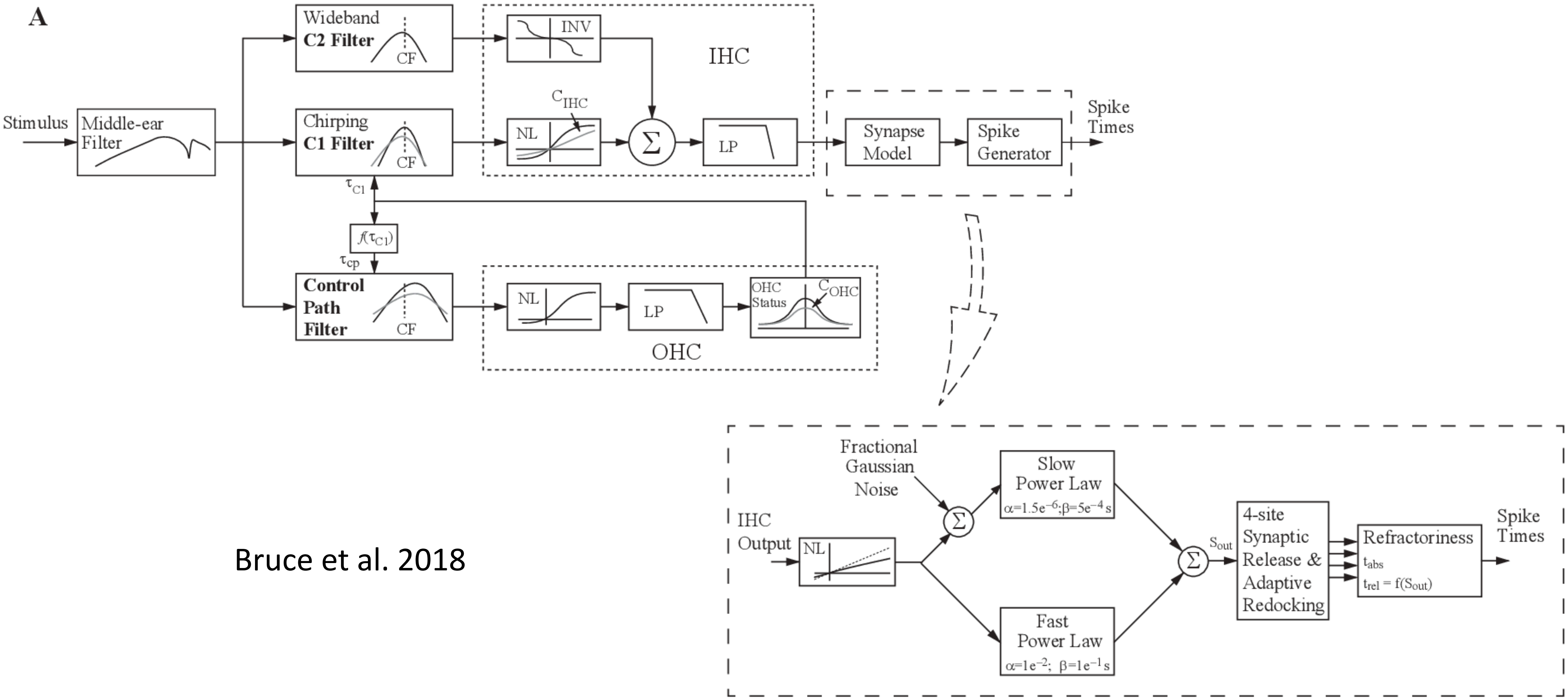
First Clarity Prediction Challenge

- Our proposal uses a peripheral auditory model which output is the spike train produced in the auditory nerve fibers (Bruce et al., 2018). Referred to as BEZ2018 model.
- The BEZ2018 is able to simulate the physiological damage causing the hearing loss from the listener audiogram.
- The intelligibility model in our proposal is based on the mutual information between the spike trains of the clean speech and the improved speech-in-noise (SPIN) degraded by the hearing loss (HL).



https://claritychallenge.github.io/clarity_CC_doc/docs/cpc1/cpc1_baseline

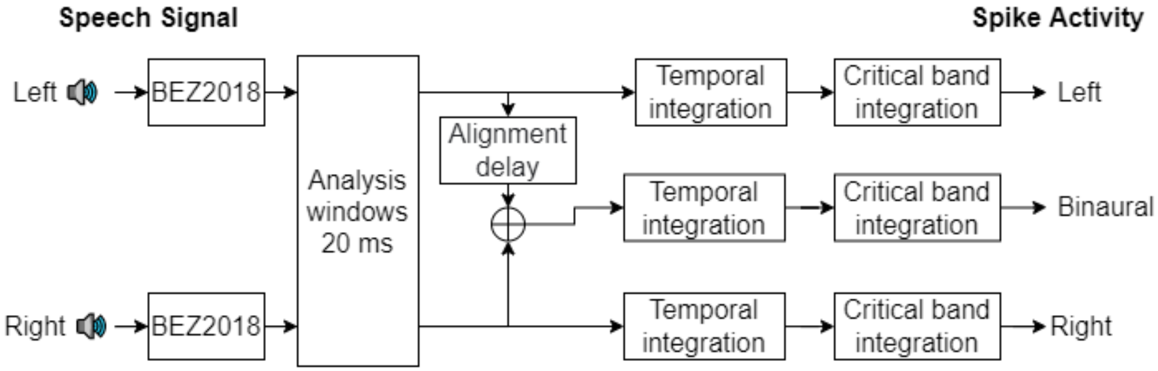
BEZ2018 Model



Bruce et al. 2018

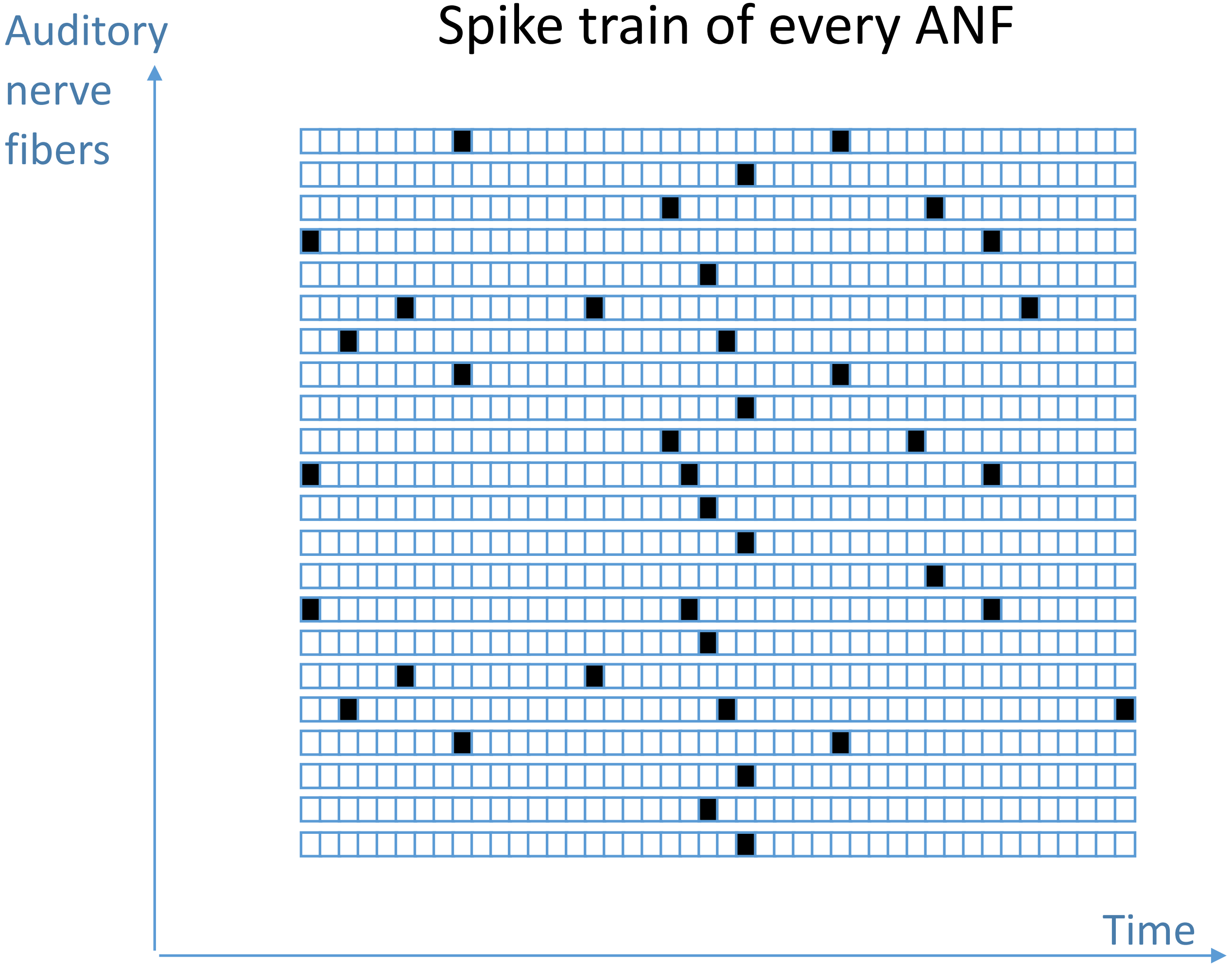
SAMII: Front-end

- **Front-end:**
  - The BEZ2018 model was configured to simulate the spike trains of 125 auditory nerve fibers (ANFs) distributed equally in 25 critical bands.
  - The whole speech signal is divided into overlapping analysis windows of 20 ms.
  - For each analysis window, a binaural representation was found by applying an alignment delay to the left spike trains. Then left and right spike trains are concatenated together.
  - The spike trains are added together by critical band and integrated in temporal windows of 200 μs.

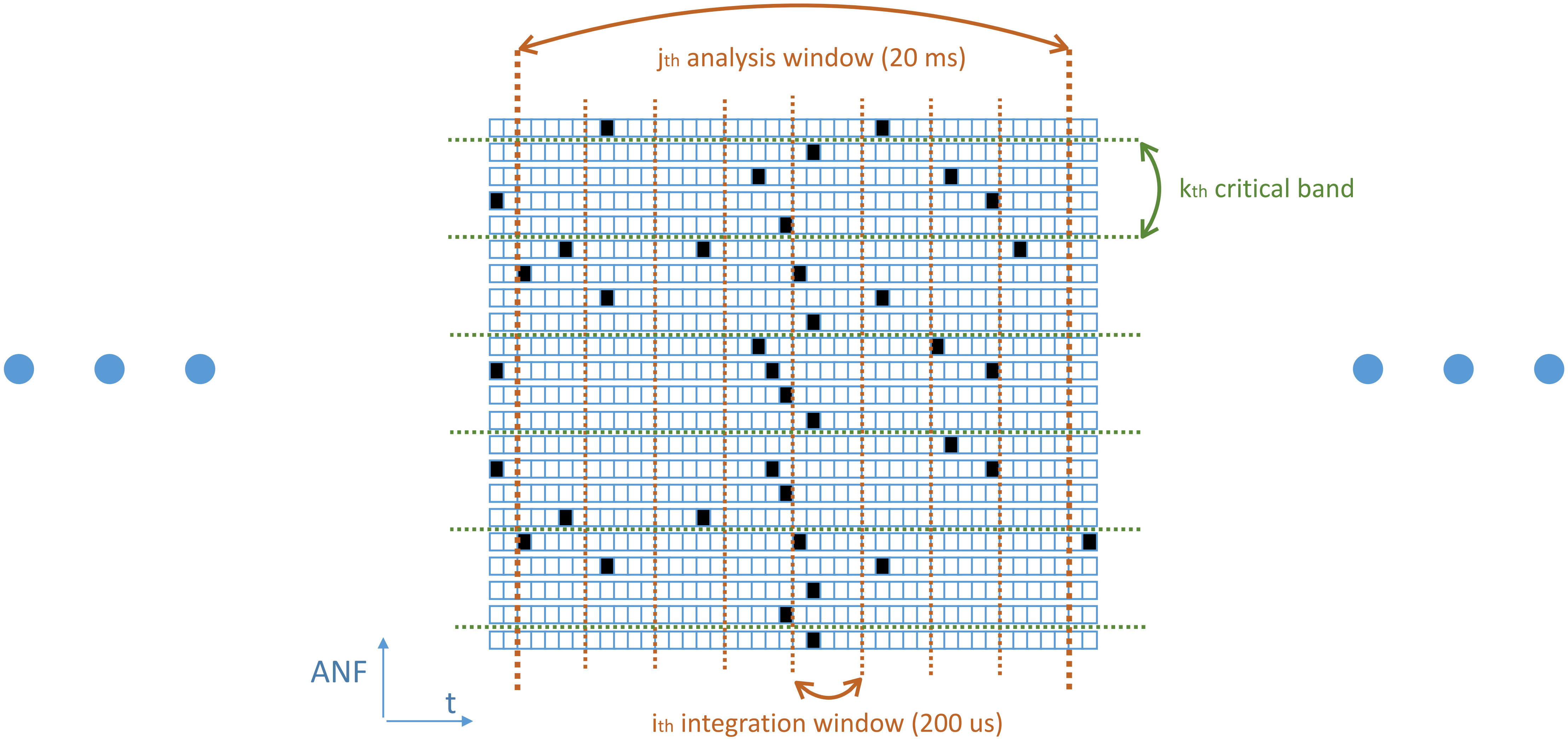


SAMII: Temporal and critical band integration

Spike train of every ANF

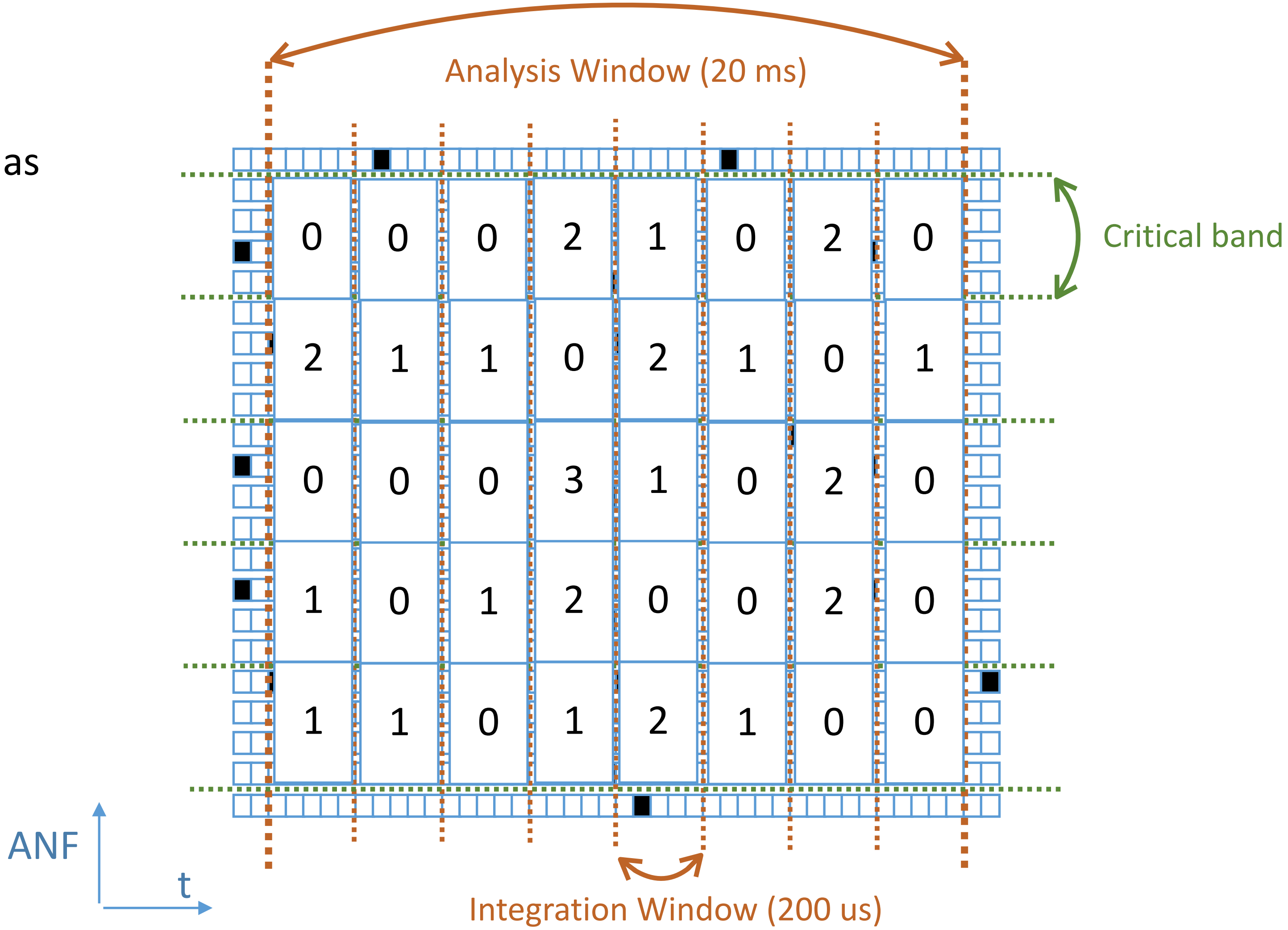


SAMII: Temporal and critical band integration



SAMII: Temporal and critical band integration

We are referring to this as the spike activity

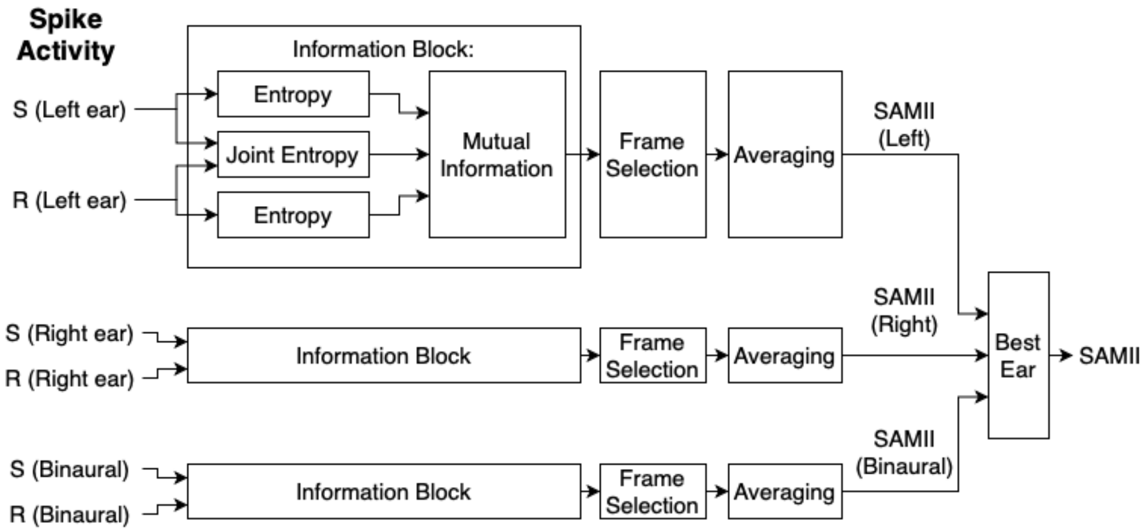


SAMII: Back-end

- **Back-end:**
  - The mutual information between the spike activity of the clean speech “S” and the degraded speech “R” is computed in the Information block.
  - Then a frame selection is performed to average the mutual information only in those analysis window and critical band frames “(j,k)” where the speech is present.
  - SAMII is then the average mutual information “I(S|R)” in those frames:

$$SAMII = \frac{1}{|Z|} \sum_{(j,k)\in Z_I} I_{j,k}\left(S|R\right)$$

  - SAMII is obtained for the left ear only, right ear only and binaural. Best value is used for prediction.



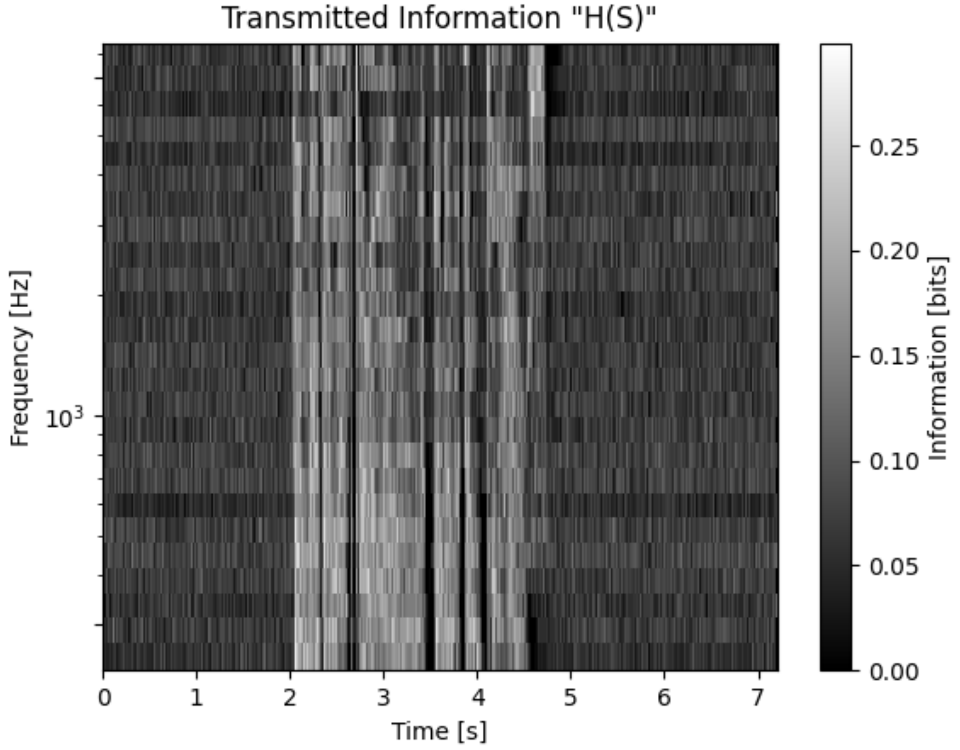
SAMII: Information Block

- **Entropy:**
  - The entropy H of a spike activity “T” is computed as:

$$H(T) = -\left(\rho \cdot \log_2(\rho) + (1-\rho) \cdot \log_2(1-\rho)\right)$$

  - With ρ being the probability of a spike occurring:

$$\rho = \frac{N_{\text{spikes},T}}{N_F \cdot N_I}$$



SAMII: Information Block

- **Joint Entropy:**
  - The joint entropy is calculated between the spike activity of "S" and "R".
  - The pair (s,r) are realisations of "S" and "R". e.g. (1,0) means that a spike occurred in "S" but not in "R". The joint entropy is then computed as:

$$H(S,R) = -\sum_{(s,r)} \sigma(s,r) \cdot \log_2[\sigma(s,r)]$$

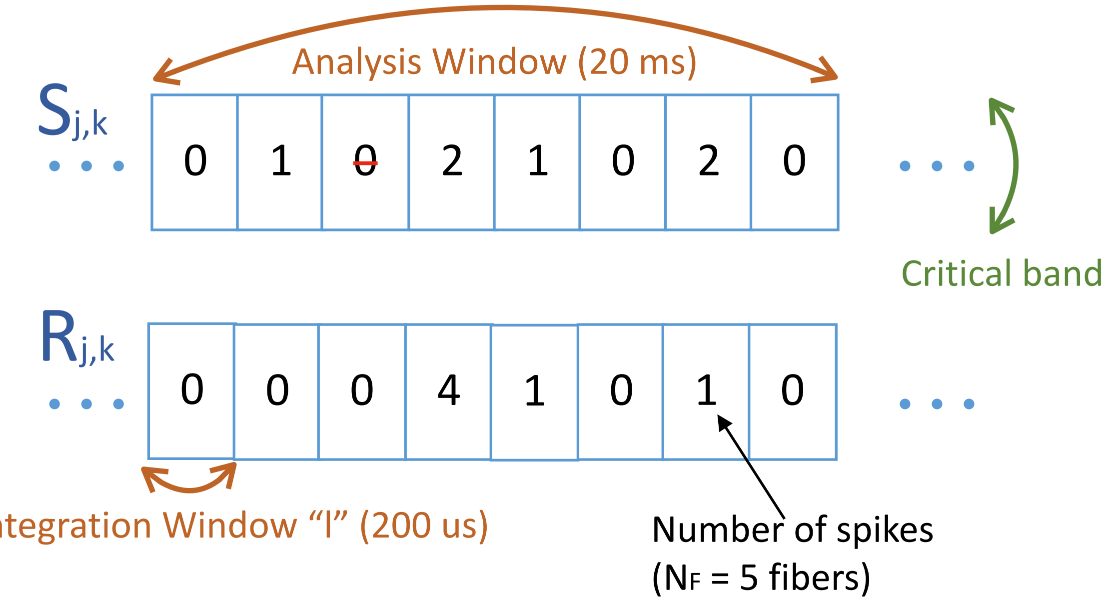
  - There are four possible combinations of (s,r), and their probability distribution is computed as:

$$\sigma(1,1) = \frac{\sum_l \min(S_l, R_l)}{N_F \cdot N_I}$$

$$\sigma(1,0) = \frac{\sum_l \max(0, S_l - R_l)}{N_F \cdot N_I}$$

$$\sigma(0,1) = \frac{\sum_l \max(0, R_l - S_l)}{N_F \cdot N_I}$$

$$\sigma(0,0) = 1 - \sigma(1,1) - \sigma(1,0) - \sigma(0,1)$$



$$\sigma(1,1) = \frac{0+0+0+2+1+0+1+0}{5 \cdot 8} = \frac{4}{40} \qquad \sigma(1,0) = \frac{0+1+0+0+0+0+1+0}{5 \cdot 8} = \frac{2}{40}$$

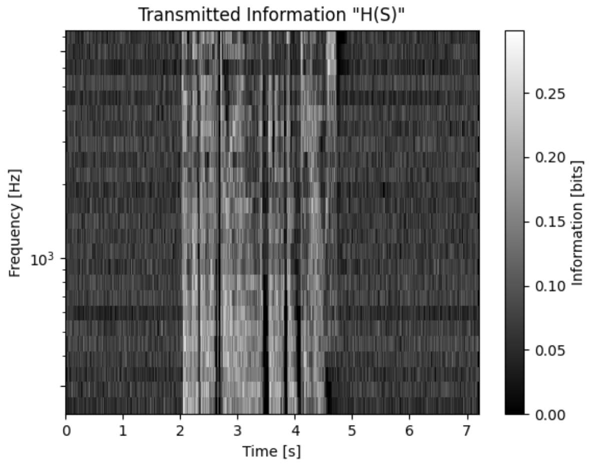
$$\sigma(0,1) = \frac{0+0+0+2+0+0+0+0}{5 \cdot 8} = \frac{2}{40} \qquad \sigma(0,0) = 1 - \frac{4}{40} - \frac{2}{40} - \frac{2}{40} = \frac{32}{40}$$

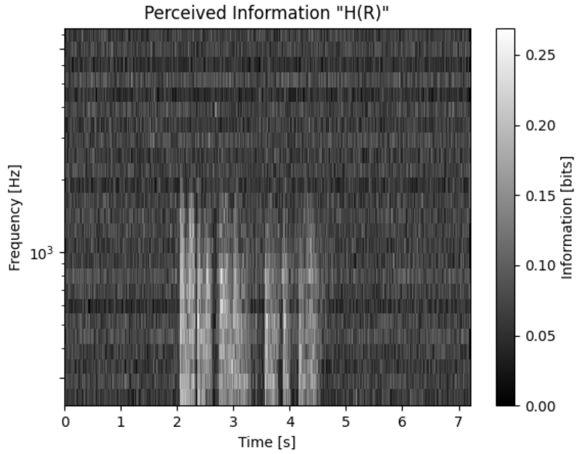
SAMII: Information Block

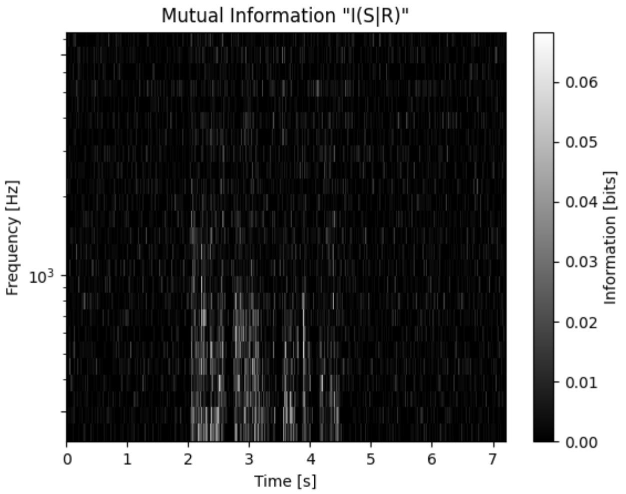
- **Mutual Information:**
  - This is an example of how the information evolves in time.
  - Notice that the mutual information only rises in those frames where the speech is transmitted and perceived.
  - The mutual information is:

$$I(S\,|\,R) = H(S) + H(R) - H(S,R)$$

  - The mutual information ranges from 0 to min(H(S), H(R))





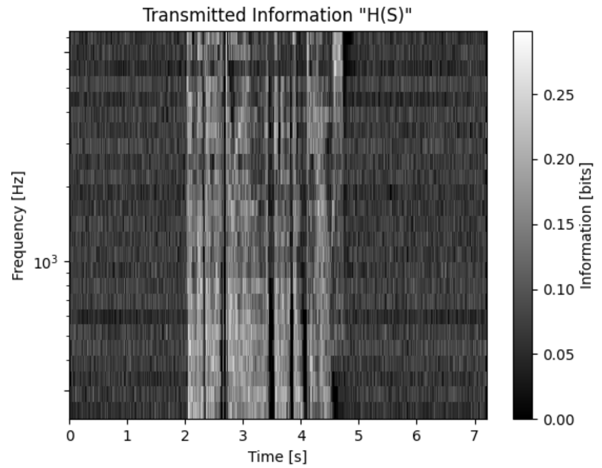


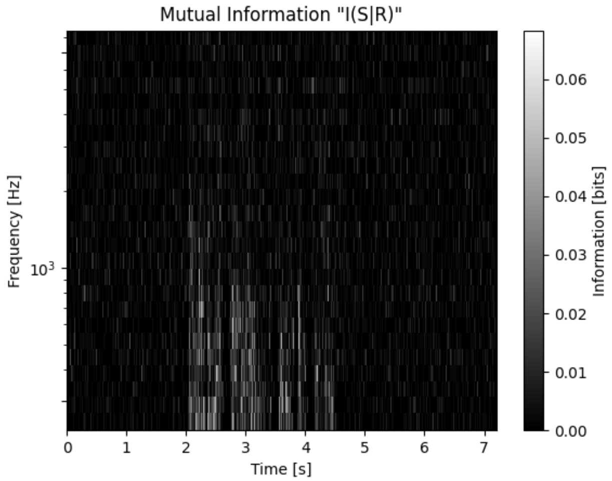
SAMII: Frame Selection

- **Z Frames:**
  - The pre-speech entropy is low and corresponds to the spontaneous activity of the auditory nerve fibers.
  - A rise in the entropy means that the voice is present. The entropy itself is used as a VAD.

- **Z_I Frames:**
  - The pre-speech mutual information is also low and corresponds to the noise and the spontaneous activity.
  - A rise in the mutual information means that the voice is being perceived. Therefore, these frames are averaged to compute SAMII.

$$SAMII = \frac{1}{|Z|} \sum_{(j,k)\in Z_I} I_{j,k}\left(S \mid R\right)$$





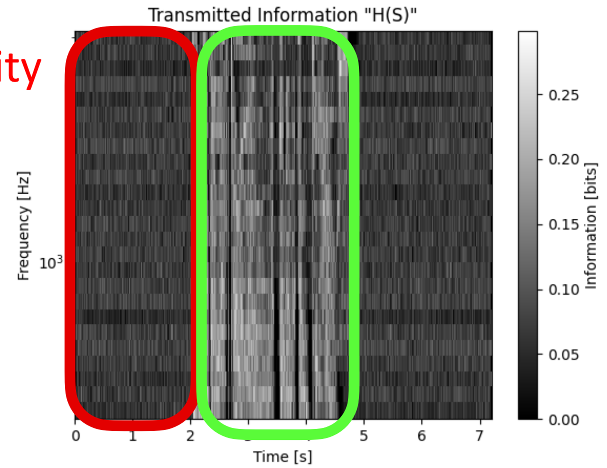
SAMII: Frame Selection

- **Z Frames:**
  - The pre-speech entropy is low and corresponds to the spontaneous activity of the auditory nerve fibers.
  - A rise in the entropy means that the voice is present. The entropy itself is used as a VAD.
- **ZI Frames:**
  - The pre-speech mutual information is also low and corresponds to the noise and the spontaneous activity.
  - A rise in the mutual information means that the voice is being perceived. Therefore, these frames are averaged to compute SAMII.

$$SAMII = \frac{1}{|Z|} \sum_{(j,k)\in Z_I} I_{j,k}(S|R)$$

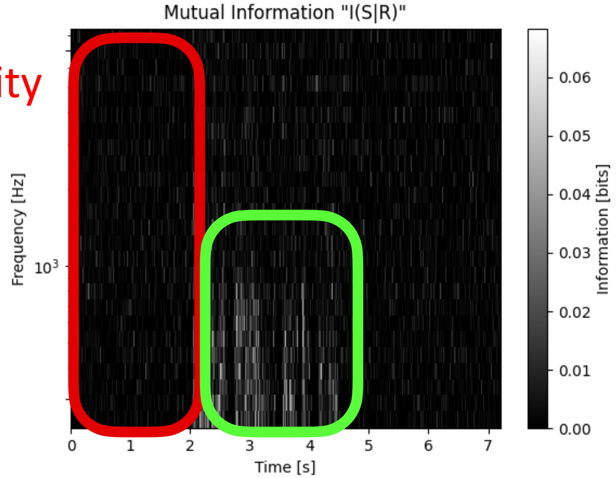
Spontaneous activity

Z frames



Noise & spontaneous activity

ZI frames



Methods

- **Baseline algorithm:**
  - The MBSTOI.

- **Dataset:**
  - It consists of various scenes where a spoken sentence is presented in a noisy and reverberant environment using a simulated binaural room impulse response (BRIR).
  - The listeners had mild to severe hearing loss and are bilateral hearing aid users.
  - The **open-set** data provided was used, with 3580 scenes for training and 632 for testing.

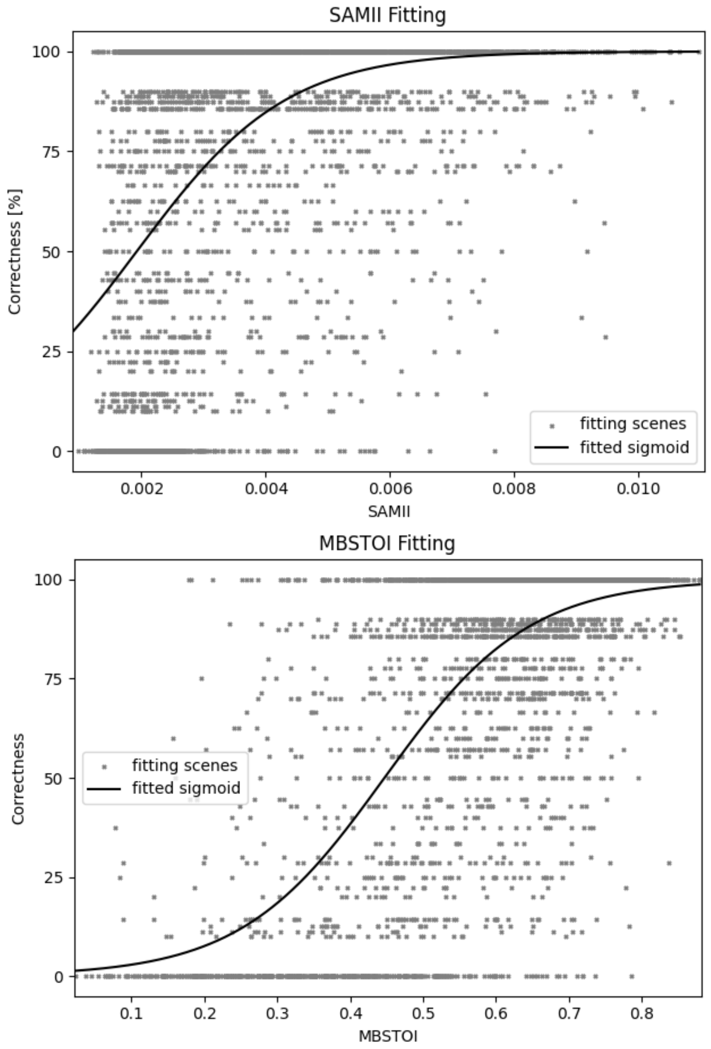
Methods

- **Fitting:**
  - The training scenes were divided in two groups. 90% of the scenes were used to fit a sigmoid function as a transfer function between SAMII and correctly guessed words. The remaining 10% were used to validate the transfer function.
  - The same fitting was performed with the MBSTOI as baseline.
  - Root mean square error (RMSE) was used as validation score.

- **Testing:**
  - Once the testing data was published by the challenge organizers, SAMII was computed and the transfer function used to predict the correct guessed words for each scene.
  - Predictions were submitted and the challenge organizers used the RMSE to evaluate the proposed algorithm SAMII and provided the score obtained by the baseline.

Results

- The transfer function (sigmoid) seems to be imprecise in both metrics.
- In SAMII, the imprecision is higher at low values, while high values are a good indication of better intelligibility.
- MBSTOI is imprecise a high values.



Results

- Performances of the MBSTOI and SAMII were similar.
- With the validation data, MBSTOI obtained better scores than SAMII.
- With the **open-set** testing data, SAMII performed slightly better than MBSTOI.

Table 1: *Score obtained in root mean square error (RMSE).*

Algorithm	Validation data	Testing data
MBSTOI (Baseline)	27.35%	36.52%
SAMII + BEZ2018	30.36%	35.16%

Discussion and Conclusion

- SAMII may be better at generalizing than MBSTOI.
- A high SAMII is a good indication that the speech is clearly understood while a low SAMII is not conclusive.
- Contrary to SAMII, MBSTOI is generally good at predicting low intelligible speech, but scenes with an MBSTOI greater than 0.3 are spread all over the correctness axis.
- Misalignments between the spike activity of “S” and “R” are the possible cause of the imprecision at low SAMIIs.
- Although this imprecision, SAMII performed similar to the baseline MBSTOI, which is a state-of-the-art algorithm. With future improvements, SAMII could be a reliable SI objective metric that works at “low-level” representations of the perceived sound

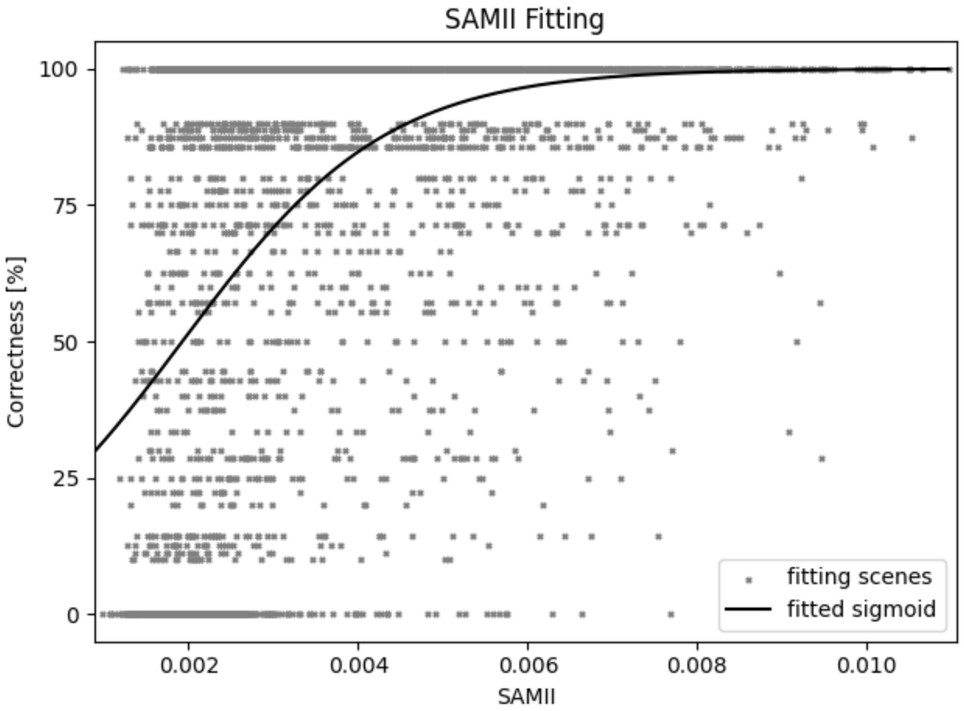
Thank you!

I’m happy to answer your questions!

Or you can find me at:

Alvarez.Franklin@mh-hannover.de

Results



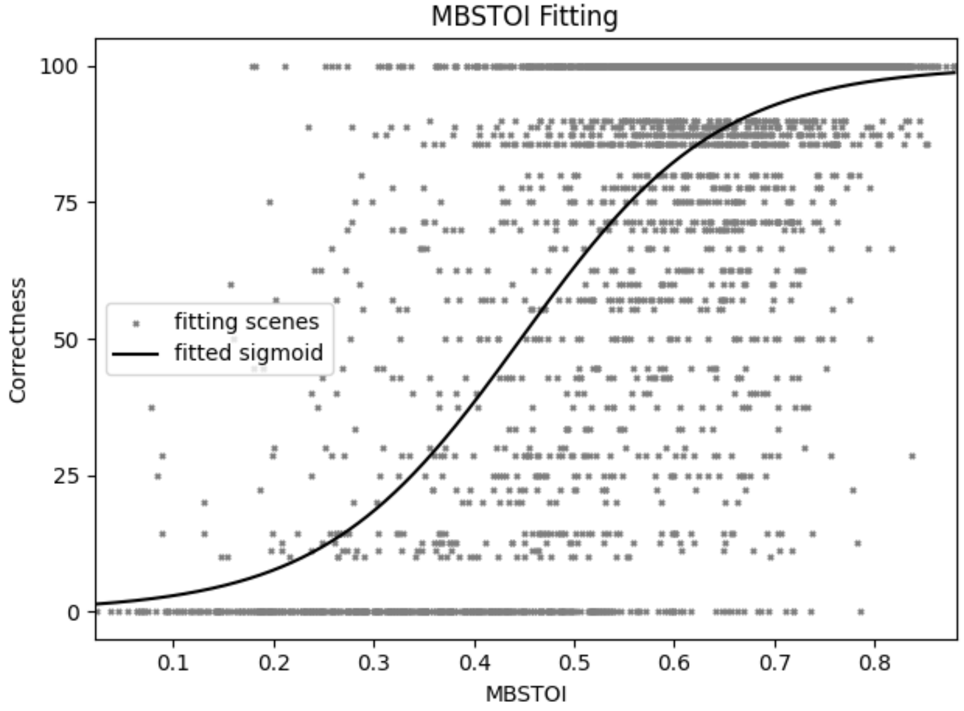
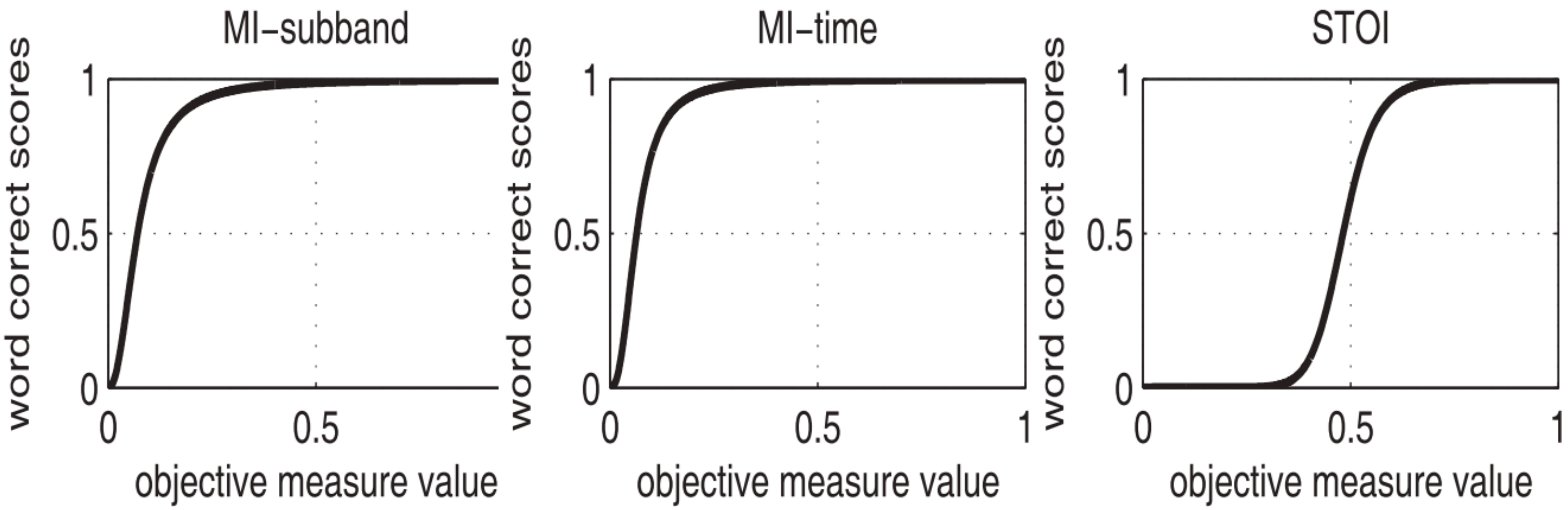


Table 1: *Score obtained in root mean square error (RMSE).*

Algorithm	**Validation data**	**Testing data**
MBSTOI (Baseline)	27.35%	36.52%
SAMII + BEZ2018	30.36%	35.16%

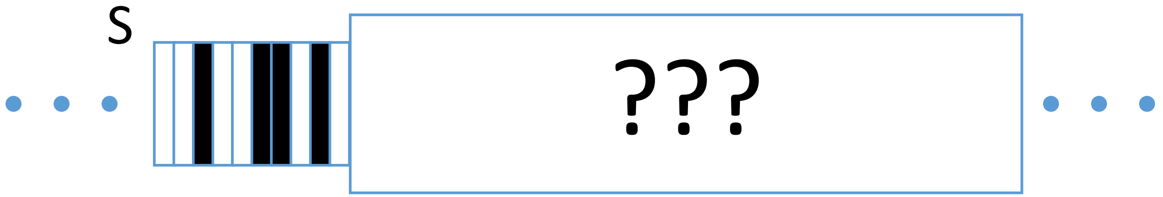
Mutual Information as Speech Intelligibility Metric



Taghia et al. (2012)

Information Theory (Example)

- Shannon Entropy



- Mutual Information

Information Theory (Example)

Shannon Entropy

S

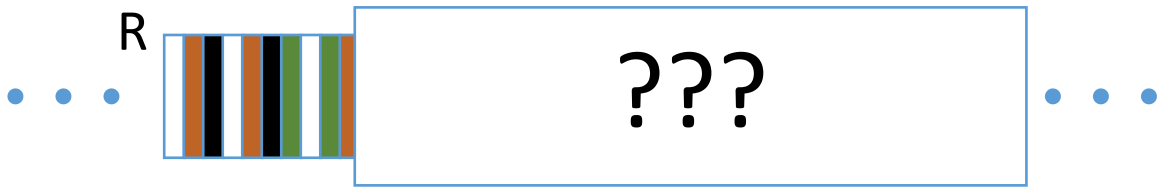
A signal with a 50% probability of getting a white or a black square carries 1 bit of information

$$H(S) = 1 \text{ bit}$$

Mutual Information

Information Theory (Example)

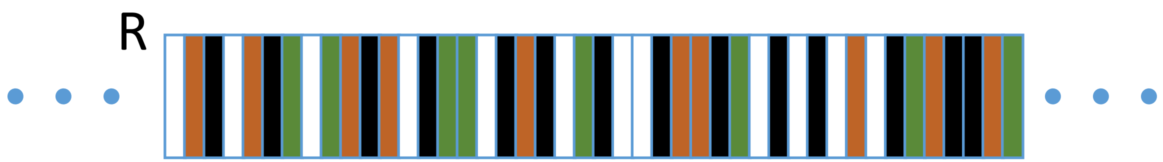
- Shannon Entropy



- Mutual Information

Information Theory (Example)

- Shannon Entropy



A signal with a 25% probability of getting a white, black, brown or green square carries 2 bits of information

$H(R) = 2$ bits

- Mutual Information

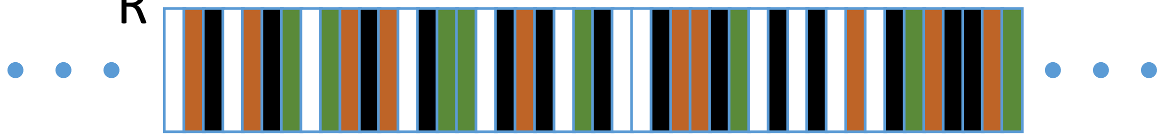
Information Theory (Example)

Shannon Entropy

S

A signal with a 50% probability of getting a white or a black square carries 1 bit of information

R



A signal with a 25% probability of getting a white, black, brown or green square carries 2 bits of information

Number of expected outcomes → $E = 2^{n}$ ← Bits to represent those outcomes

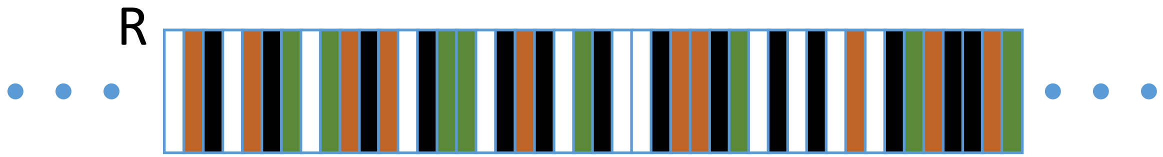
Only when all outcomes are equally probable!

Mutual Information

Information Theory (Example)

Shannon Entropy

A signal with a 50% probability of getting a white or a black square carries 1 bit of information



A signal with a 25% probability of getting a white, black, brown or green square carries 2 bits of information

Number of expected outcomes → $E = 2^n$ ← Bits to represent those outcomes

Only when all outcomes are equally probable!

Mutual Information

When S is white, R is white or brown
When S is black, R is black or green

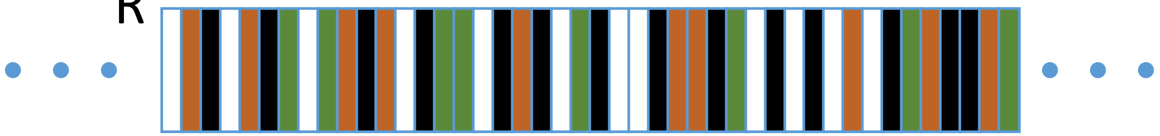
Information Theory (Example)

Shannon Entropy

S

A signal with a 50% probability of getting a white or a black square carries 1 bit of information.

R



A signal with a 25% probability of getting a white, black, brown or green square carries 2 bits of information.

Number of expected outcomes

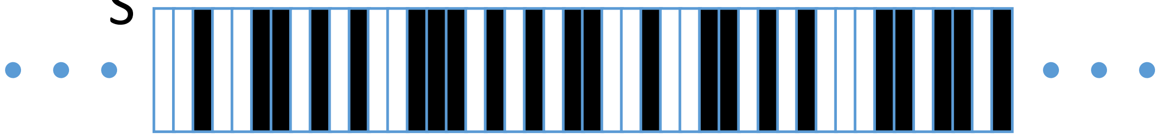
$$E = 2^n$$

Bits to represent those outcomes

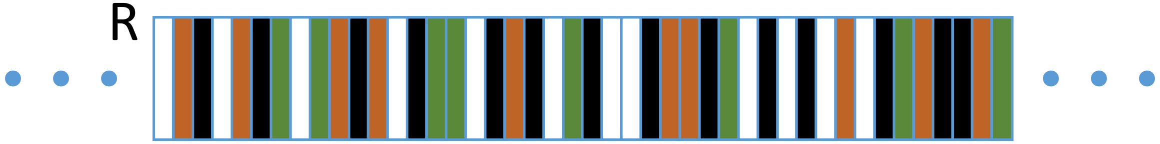
Only when all outcomes are equally probable!

Mutual Information

S



R



When S is white, R is white or brown
When S is black, R is black or green

In this case the mutual information is 1 bit because looking at S, there is 50% probability of correctly guessing the outcome in R.

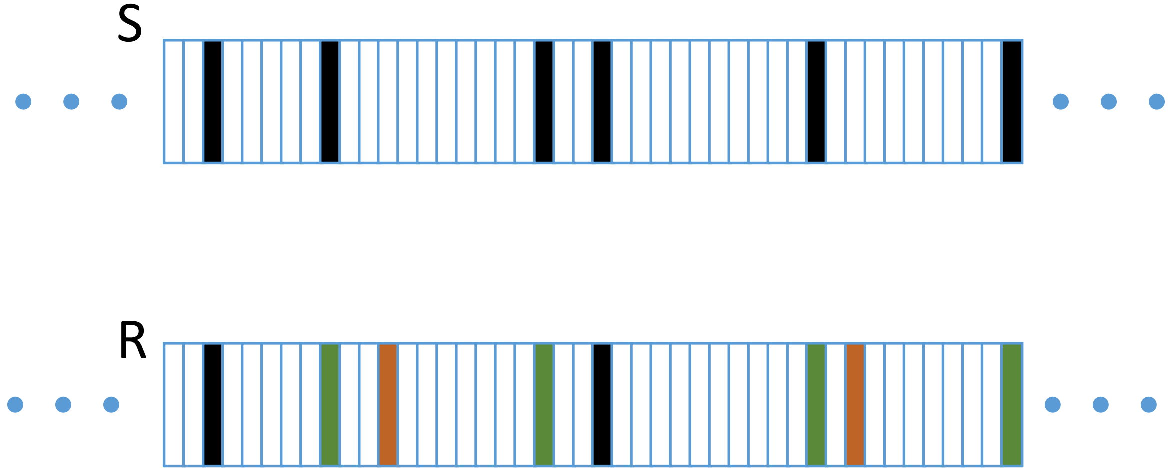
$$I(S|R) = H(S) = 1 \text{ bit}$$

Mutual information ranges from 0 to min(H(S), H(R))

Information Theory (Example)

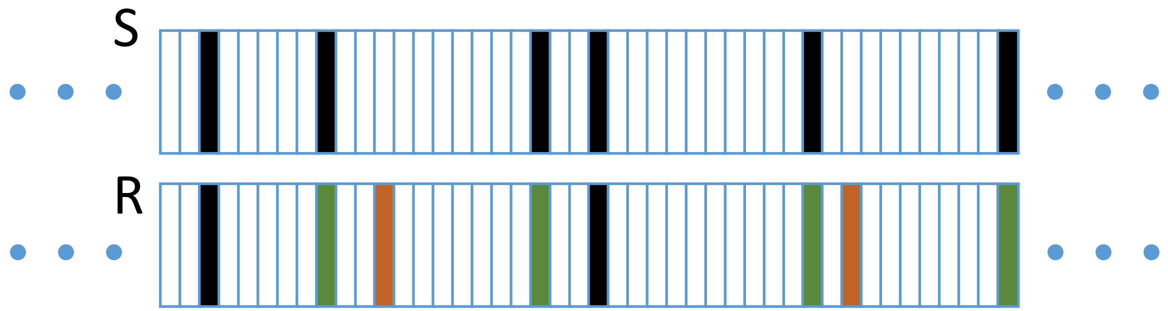
Hannover Medical School

Shannon Entropy



In these cases the Shannon entropy is reduced considerably because the probability distribution has changed

Mutual Information



Nevertheless, the mutual information will be relatively high because looking at S you can get plenty of information about R

$$I(S|R) \sim H(S)$$