(REPORT DRAFT) Predicting Speech Intelligibility using SAMII: Spike Activity Mutual Information Index

Franklin Alvarez[1], *Waldo Nogueira*[1]

[1]Medizinische Hochschule Hannover and Cluster of Excellence Hearing4All

Alvarez.Franklin@mh-hannover.de, NogueiraVazquez.Waldo@mh-hannover.de

1. Introduction

In the context of the first clarity prediction challenge [1], it is presented the spike activity mutual information index (SAMII) as a new intrusive objective metric to predict speech intelligibility. I has been shown that mutual information performs successfully as a speech intelligibility metric, compared to its more commonly used counterpart metrics, such as the signal-to-noise ratio (SNR) and correlation [2].

The motivation for developing SAMII goes beyond hearing aid applications, and it is to offer a reliable speech intelligibility metric for more physiologically inspired auditory models. Such models are capable of simulating the spike activity produced by an acoustic stimulus in a population of auditory nerve fibers (ANFs). This models can be useful to infer aspects in the human periphery that contributes to speech understanding.

The spike activity is a representation of the action potentials, also called spikes, that are produced in a population of ANFs. Spikes can be represented as binary variables where the value “one” means that a spike has occurred. Also, spikes can be concatenated in time to form spike “spike trains” which are unique for each ANF.

2. Methodology

2.1. Peripheral auditory model

In this work, the auditory peripheral model presented by Bruce et al. (2018) [3] is used. It uses a population of ANFs, grouped by critical bands centered at different center frequencies, to simulate the spike activity from any sound stimulus. Additionally, it is capable of simulating the hearing loss from a subject audiogram.

Because of the required computational resources, a “light” version of the BEZ2018 model was configured to work with 25 critical bands with center frequencies distributed logarithmically between 250 Hz and 8 kHz to cover the whole speech frequency range. The number of ANFs was limited to five per critical band, giving a total population of 125 ANFs. This is the minimum possible number of fibers that preserve the original ratio of 30-10-10 for high, medium and low spontaneous firing rate ANFs, respectively [3]. This version of the model is referred to as BEZ2018_L.

Another “hight fidelty” version of the BEZ2018 model, that runs in a GPU, has been developed as well. It was configured to work with 40 critical bands with center frequencies distributed between 125 Hz and 16 kHz. Each critical band counts with 100 ANFs giving a total population of 4000 ANFs. In this case the ratio is 61-23-16 for high, medium and low spontaneous firing rate ANFs, respectively, which is another well known ratio [4]. This version of the model is referred to as BEZ2018_H.

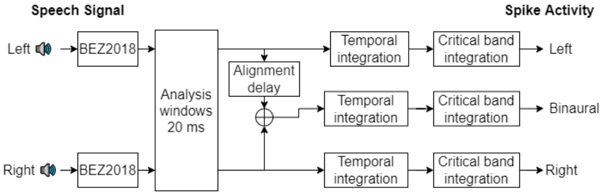

Figure 1: *Signal path from audio to spike activity.*

2.2. Spike Activity Mutual Information Index

SAMII is defined as the averaged mutual information $I\left(S|R\right)$ between the spike activity of the clean speech target S, and the spike activity of the corresponding noisy speech R.

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

where j and k are the analysis window and critical band indices, respectively. $|Z|$ is the number of (j,k) frames where the clean signal is detected while Z_I is a subset of Z where the mutual information is greater than a threshold.

2.2.1. *Temporal an spatial integration*

Figure 1 shows a block diagram of how the spike activity (S or R) is obtained from the speech signal. Left and right ear audio signals are processed independently with BEZ2018 to obtain the spike trains for each ANF. Then, analysis windows of 20 ms with an overlap of 10 ms are used. For every analysis window, an additional binaural representation is obtained by grouping together the delayed version of the left ear spike trains and the right ear spike trains. The alignment delay is selected as the value between -1 ms and 1 ms that results in the lowest root mean square error (RMSE) between the left and right spike trains.

To obtain the spike activity, the spike trains are integrated in windows of 200 µs, grouped by center frequency, and added together. The result is a matrix of size $N_{CB} \times N_I$, where N_{CB} is the number of critical bands and N_I is the number of integration windows ($N_I = \frac{20ms}{200\mu s} = 100$).

2.2.2. *Mutual information*

As seen in equation (1), the mutual information between spike activities S and R is computed for every (j,k) frame. For practical reasons, the indices of the analysis window j and critical band k are removed in the following equations.

Mutual information is obtained with equation (2):

$$I(S|R) = H(S) + H(R) - H(S,R), \tag{2}$$

where $H(S)$ and $H(R)$ are the individual entropy of both spike activities, and $H(S, R)$ is their joint entropy. The individual entropy of a generic spike activity T is obtained with equation (3):

$$H(T) = -(\rho \cdot \log_2(\rho) + (1-\rho) \cdot \log_2(1-\rho)), \quad (3)$$

where T could be substituted by S or R, and ρ is the probability of a spike occurring. It is obtained with equation (4):

$$\rho = \frac{N_{spikes,T}}{N_F \cdot N_I}, \quad (4)$$

where N_F is the number of ANFs per critical band.

For the joint entropy between the spike activities S and R, it is necessary to obtain their joint probability distribution. The joint probability distribution is obtained with the probabilities $\sigma(s, r)$ of all possible events (s, r), which are the absence (0), or presence (1), of a spike within an integration window l. In example, $\sigma(0, 1)$ is the probability of a spike occurring in R, but not in S, during the same integration window of 200 µs. Equations (5), (6), (7), and (8) show how those probabilities are computed.

$$\sigma(1,1) = \frac{\sum_l^{N_1} \min(S_l, R_l)}{N_F \cdot N_I}. \quad (5)$$

$$\sigma(1,0) = \frac{\sum_l^{N_1} \max(0, S_l - R_l)}{N_F \cdot N_I}. \quad (6)$$

$$\sigma(0,1) = \frac{\sum_l^{N_1} \max(0, R_l - S_l)}{N_F \cdot N_I}. \quad (7)$$

$$\sigma(0,0) = 1 - \sigma(1,1) - \sigma(1,0) - \sigma(0,1). \quad (8)$$

Then, the joint entropy is obtained with the following equation (9):

$$H(S,R) = -\sum_{(s,r)} \sigma(s,r) \cdot \log_2[\sigma(s,r)]. \quad (9)$$

2.3. Dataset

The open-set of the training data provided by the first clarity prediction challenge was chosen to assess the performance of SAMII and MBSTOI. To perform predictions, a sigmoid function was fitted to map SAMII and MBSTOI with the correctness score (percentage of correctly guessed words in a sentence) provided with the training data.

3. Preliminary Results

The training dataset was divided into a fitting set, and a validation set. The fitting set was a random selection of 90% of the training data, living the remaining 10% for the validation set. The score used to evaluate the proposed speech intelligibility prediction algorithm was the RMSE between the predictions and the correctness of the validation set. Figure 2 shows the obtained fitted curve.

Scores obtained using the two versions of the BEZ2018 model (L and H) are shown in table 1.

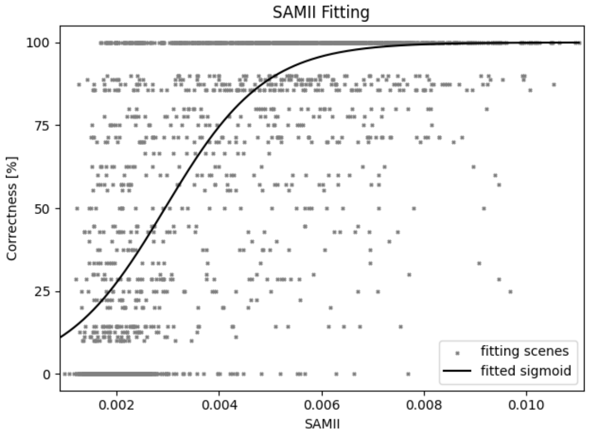

Figure 2: *Spike activity mutual information index (SAMII) fitting curve.*

Table 1: *Preliminary root mean square error (RMSE) for each algorithm*

Algorithm	**preliminary score**
MBSTOI (Baseline)	27.35
SAMII + BEZ2018_L	30.36
SAMII + BEZ2018_H	28.76

4. Preliminary Discussion

Results show that the scores obtained with SAMII and the baseline MBSTOI are similar. This is a favorable finding for the proposed new metric since it is reaching a similar performance than a well established speech intelligibility metric. Having a closer look, it is shown the SAMII is very reliable when predicting high scores, but at low SAMIIs there is more uncertainty. This is evident in Figure 2, where SAMIIs around 0.002 correspond to a wide variety of correctness scores. Moreover, the results show no evidence for a relation between density of ANFs and performance since both BEZ2018_L and BEZ2018_H obtained similar scores as well.

5. References

[1] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, “Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, pp. 1181–1185, 2021.

[2] J. Jensen and C. H. Taal, “Speech intelligibility prediction based on mutual information,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 430–440, 2014.

[3] I. C. Bruce, Y. Erfani, and M. S. Zilany, “A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites,” *Hearing Research*, vol. 360, pp. 40–54, 2018. [Online]. Available: https://doi.org/10.1016/j.heares.2017.12.016

[4] B. C. Moore, “Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants,” *Otology and Neurotology*, vol. 24, no. 2, pp. 243–254, 2003.