Predicting Speech Intelligibility using the Spike Activity Mutual Information Index

Franklin Alvarez¹, Waldo Nogueira¹

¹Medizinische Hochschule Hannover and Cluster of Excellence “Hearing4all”

Abstract

The spike activity mutual information index (SAMII) is presented as a new intrusive objective metric to predict speech intelligibility. A target speech signal and speech-in-noise signal are processed by a state-of-the-art computational model of the peripheral auditory system. It simulates the neural activity in a population of auditory nerve fibers (ANFs), which are grouped into critical bands covering the speech frequency range. The mutual information between the neural activity of both signals is calculated using analysis windows of 20 ms. Then, the mutual information is averaged across these analysis windows to obtain SAMII. SAMII is also extended to binaural scenarios by calculating the index for the left ear, right ear, and both ears, choosing the best case for predicting intelligibility.

SAMII was developed based on the first clarity prediction challenge training dataset and compared to the modified binaural short-time objective intelligibility (MBSTOI) as baseline. Scores are reported in root mean squared error (RMSE) between measured and predicted data using the clarity challenge test dataset. SAMII scored 35.16%, slightly better than the MBSTOI which obtained 36.52%. This work leads to the conclusion that SAMII is a reliable objective metric when “low-level” representations of the speech, such as spike activity, are used.

Index Terms: speech intelligibility, mutual information, peripheral auditory models

1. Introduction

Speech intelligibility (SI) prediction algorithms are useful to evaluate the performance of speech enhancement technologies used in hearing devices. These algorithms provide an objective index that can be mapped into behavioural responses from humans. Complexity of SI predictor algorithms comes in different levels, but in general, these consist of a front-end where the sound waves are processed to obtain representative features, and a back-end where those features are translated into a SI index [1, 2, 3, 4].

Nowadays, there are well-established SI prediction algorithms. The modified binaural short-time objective intelligibility (MBSTOI) calculates the correlation between a clean and a degraded speech signal [5]. It counts with an equalization cancellation (EC) stage where interaural time differences (ITDs) and interaural level differences (ILDs) between both ears are calculated to obtain a binaural representation. Also, it is computed in short-time windows to account for non-stationary noises. Additionally, because it is correlation based, the MBSTOI performs well with non-linearly distorted speech signals. Another well-established algorithm is the multi-resolution speech-based envelope power spectrum model (mr-sEPSM) [6], which computes the signal-to-noise ratio (SNR) between the degraded signal and the noise alone. In this case, a gammatone filterbank is implemented for every frequency band. It also counts with a binaural version where an EC stage similar to the MBSTOI is added to the algorithm [7]. Finally, it has been shown that mutual information performs successfully as a speech intelligibility metric, compared to its previously mentioned counterparts SNR and correlation [8, 9]. The proposed model by Jensen and Haal (2014) uses a front-end similar to the MBSTOI, but for monaural speech signals only. The mutual information between clean and degraded speech signals is computed for every short-time frame and frequency band. Then, it is averaged only in those frames where the speech is detected by a voice activation detector (VAD).

All the previous mentioned algorithms use “high-level” representations of the sound as features since they correspond to an abstraction of the sound somewhere in the central auditory system. “Lower-level” representations can be obtained with more complex models. An example is the model presented by Bruce et al. (2018) [10], referred to as BEZ2018. It simulates the spike activity, which is a representation of the action potentials, also called spikes, that are produced in a population of auditory nerve fibers (ANFs). Therefore, the representation of the speech corresponds only to the peripheral auditory system. 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 trains” which are unique for each ANF.

In the context of the first clarity prediction challenge [11], it is presented the spike activity mutual information index (SAMII) as a new intrusive objective algorithm to predict SI. The motivation for developing SAMII goes beyond hearing aid applications and it is to offer a reliable SI metric for more physiologically inspired auditory models like the BEZ2018. Such models can be useful to infer aspects in the human peripheral auditory system that contribute to speech understanding like inner hair cells (IHCs) and outer hair cells (OHCs) loss, or the synapses between IHCs and ANFs, or the status of the ANFs. Applications could extend to cochlear implant (CI) models where the neural excitation is obtained from electrical stimulation of a population of ANFs [12, 13, 14], instead of using objective measures with vocoded versions of the speech [15].

The remainder of this work is organized as follows. Section 2 describes the proposed SI prediction algorithm by analysing each processing block composing it. Section 3 presents the dataset used to validate SAMII. Section 4 shows the results obtained in the first clarity challenge with a discussion. Section 5 presents the conclusions.

2. The spike activity mutual information index (SAMII)

The SI prediction algorithm presented in this work consists of a front-end that computes the spike activity and a back-end that uses mutual information as the SI metric.
2.1. Front-end

Figure 1 shows a block diagram of how the spike activity is obtained from the speech signal arriving to the two ears (left and right) of a human.

2.1.1. Peripheral auditory model

In this work, the BEZ2018 model was selected as the peripheral auditory model to simulate the neural activity produced by the speech and speech-in-noise signal. Nevertheless, any peripheral auditory model that reproduces the spike trains from a population of ANFs can be used instead. 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 preserves the original ratio of 30-10-10 for high, medium, and low spontaneous firing rate ANFs, respectively. Additionally, the BEZ2018 model is capable of simulating the hearing loss of a subject from its audiogram.

2.1.2. Analysis window, temporal integration and critical band integration

Left and right ear audio signals are processed independently with the BEZ2018 model. The spike trains generated for each ANF are divided into analysis windows of 20 ms with an overlap of 10 ms. 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. This process is equivalent to the EC stage found in algorithms like MBSTOI or mr-eEPSM, but grouping the spike trains instead of subtracting them.

To obtain the spike activity, the spike trains for each critical band are temporally integrated in windows of 200 µs. The temporal integration window is below the minimum possible refractory period for any ANF in the BEZ2018 model, meaning that two spikes of the same ANF will never be integrated together. This is important for the upcoming calculations of probability distribution of the spike occurrence.

For every analysis window, a matrix of size $N_{CB} \times N_I$ is obtained, where $N_{CB}$ is the number of critical bands and $N_I$ is the number of integration windows ($N_I = \frac{20 \text{ms}}{200 \mu \text{s}} = 100$).

2.2. Back-end

The back end of SAMII is based on the average mutual information $I(S|R)$ between the spike activity of the clean speech target $S$, and the spike activity of the corresponding noisy speech $R$. As shown in Figure 2, the mutual information is calculated in the information block for the left, right and binaural signals.

In the end, the best index obtained is the one chosen for SI predictions.

Equation (1) shows how the SAMII is obtained for each ear:

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

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, and $Z_1$ is a subset of $Z$ where the mutual information is greater than a threshold.

2.2.1. Information Block

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):

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

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):

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

where $T$ could be substituted by $S$ or $R$ and $\rho$ is the probability of a spike occurring. It is obtained with Equation (4):

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

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 $l$. For 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 these probabilities are computed.

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

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

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)

Figures 3, 4 and 5 are representations of the entropy of the target signal \( H(S) \), the entropy of the perceived signal \( H(R) \), and the mutual information between them \( I(R|S) \).

2.2.2. Selection of frames

SAMII uses only a subset \( Z \) of the whole \( (j,k) \) frames, as shown in Equation (1). The reason behind this, is that we are only interested on those frames where the speech is present in the clean signal \( S \). The entropy \( H_{j,k}(S) \) during a silent period corresponds to the spontaneous activity of the ANFs, therefore, when speech is present, the spike activity rises along with the entropy. The selection of \( Z \) was done by setting a threshold \( H_{j,k}(S) \) during this period, and for every critical band. This threshold is equal to the average entropy, caused by the spontaneous activity, plus three times its standard deviation.

On the other hand, the mutual information during the pre-silence period corresponds to the information between the perceived noise and the spontaneous activity. These mutual information values depend completely on noise, therefore, it is expected that the more prominent is the target speech in the perceived signal \( R \), the greater will be the mutual information. To take into account only those frames where the mutual information rises, another subset \( Z_I \) of frames is selected from \( Z \), where \( I_{j,k}(S|R) \) is greater than the threshold \( I_{j,k}(S|R) \). Similarly, the threshold is the mutual information average between noise and spontaneous activity plus three times its standard deviation.

3. Dataset

The dataset used in this work was provided for the first clarity prediction challenge [11]. 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 correctness score (number of words correctly understood from the sentence) and audiometry is also provided. Every scene contains the anechoic version of the target speech \( S \) with two seconds of preliminary silence and one second of post silence. Also, it contains the improved speech-in-noise signal coming from the hearing aid device \( R \).

The training dataset consisted of 3580 different scenes. During development of SAMII, it was divided into a fitting set and a validation set. The fitting set was a random selection of 90% of the training data, leaving the remaining 10% for the validation set. The testing dataset consisted of 632 scenes. It was chosen the open-set version of the data, meaning that the testing data contains scores for listeners, and enhancement algorithms, not seen in the training data.

4. Results and Discussion

To perform predictions, a sigmoid function is fitted to map indices obtained with SAMII and the baseline MBSTOI into the correctness score for the scenes in the fitting set. The score used to evaluate the proposed algorithm was the RMSE between the predictions and the correctness of the validation set.
Table 1: Score obtained in root mean square error (RMSE).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Validation data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBSTOI (Baseline)</td>
<td>27.35%</td>
<td>36.52%</td>
</tr>
<tr>
<td>SAMII + BEZ2018</td>
<td>30.36%</td>
<td>35.16%</td>
</tr>
</tbody>
</table>

Figure 6: SAMII sigmoid fitting with training dataset.

Once the testing dataset was published, the sigmoid fitting was performed combining the fitting and validation dataset. The predictions were submitted to the clarity prediction challenge and the score obtained is shown in Table 1.

Although having a worse score with the validation data, SAMII performed slightly better than the MBSTOI with the testing dataset. This may indicate that the MBSTOI (correlation based metric) performs better if the sigmoid is fitted to scenes with known listeners and enhancement algorithms while SAMII (mutual information based metric) is better at generalizing.

Looking at Figure 6, it is evident that SAMII is not a completely reliable SI metric yet. SAMII maps indices above 0.006 to a value close to 100% correctness, while being to optimistic in lower indices. Correctness under 30% are rarely predicted despite of being quite common in real measurements. This is because scenes in the training data that obtained a SAMII around 0.002 are widely dispersed along the correctness axis. In summary, a high SAMII is a good indication that the speech is clearly understood while a low SAMII is not conclusive.

Figure 7 shows the fitted sigmoid function for the MBSTOI, which shows high variability as well. Contrary to SAMII, MBSTOI is generally good at predicting low intelligible speech (MBSTOI less than 0.3), but scenes with an MBSTOI greater than 0.3 are spread all over the correctness axis.

Different adjustments, not shown in this work, were performed to address this issue without meaningful improvements in the results. A normalized version of the mutual information, shown in Equation (1), was implemented to overcome possible scaling problems. Mutual information $I(S|R)$ is lower in target signals with low entropy $H(S)$, therefore, low SAMIIIs with good score may be related to low target signal entropy. Nevertheless, this normalized version obtained worse RMSE with the validation data, probably because the normalization brought up mutual information values of noise and target speech as well. Additionally, the selection criteria of the frames $Z$ and $Z_t$ in Equation (1) was modified without success, meaning that the issue has to be addressed in lower computation levels.

A possible explanation of the high variability is that the mutual information $I(S|R)$ between highly correlated signals diminishes considerably when they are not aligned. This delay can be introduced by the enhancement algorithm in the hearing aid device or by the peripheral auditory model when simulating the hearing loss. Note that the dataset includes speech in noise processed through different speech enhancement algorithms, hence having different delays. This could also explain why this high variability in predicted scores only appears in low SAMIIIs, since they contain scenes where the mutual information is low because of the bad intelligibility and because of mis-aligned signals that are intelligible. On the other hand, high values of SAMII always occur because of a good intelligibility. Therefore, the solution could be to compute the mutual information using short alignment delays.

Another possible improvement is the addition of dynamic weighting values to the critical bands. These weights could be calculated by an artificial intelligence (AI) algorithm simulating more central auditory processes like memory or attention.

5. Conclusions

In this work is presented a new objective metric based on mutual information for intrusive speech intelligibility prediction. It differentiates from other metrics in that SAMII uses the spike activity generated in a population of ANFs, instead of “higher-level” representations of the stimulus commonly used by state-of-the-art algorithms.

Despite of having a high variance in the correctness mapped at low SAMIIIs, it is discussed the possible solutions to this issue that could make this metric suitable for SI objective measurements in future revisions. Nevertheless, having a slightly better score than the baseline metric in the prediction challenge is evidence of the potential of SAMII.

6. Acknowledgements

The authors would like to thank the German Academic Exchange Service (DAAD) and the Cluster of Excellence “Hearing4all” for funding this project. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy EXC 2177—Project ID 390895286.
7. References


