INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A Spectro-Temporal Glimpsing Index (STGI) for Speech Intelligibility Prediction

Amin Edraki[1], *Wai-Yip Chan*[1], *Jesper Jensen*[2,3], *and Daniel Fogerty*[4]

[1]Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada
[2]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[3] Oticon A/S, Smørum, Denmark
[4]Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

a.edraki@queensu.ca, chan@queensu.ca, jesj@demant.com, dfogerty@illinois.edu

Abstract

We propose a monaural intrusive speech intelligibility prediction (SIP) algorithm called STGI based on detecting *glimpses* in short-time segments in a spectro-temporal modulation decomposition of the input speech signals. Unlike existing glimpse-based SIP methods, the application of STGI is not limited to additive uncorrelated noise; STGI can be employed in a broad range of degradation conditions. Our results show that STGI performs consistently well across 15 datasets covering degradation conditions including modulated noise, noise reduction processing, reverberation, near-end listening enhancement, checkerboard noise, and gated noise.

Index Terms: speech intelligibility, speech quality model, spectro-temporal modulation, glimpsing

1. Introduction

In human-human communications mediated by machines and networks, speech signals are often distorted and not perfectly perceived by the listener. Machine algorithms can facilitate predicting the perceptual effect of different types of distortion/processing on speech signals. An ideal speech intelligibility prediction (SIP) algorithm would accurately predict a degraded/processed signal's average intelligibility as perceived by a group of listeners. This study aims to develop an intrusive or reference-based SIP algorithm for normal-hearing listeners that relies on inputting a clean reference signal.

Different approaches to computational modelling of speech perception in adverse listening conditions have been proposed [1–4]. In this regard, the glimpsing model of speech perception [4] was proposed based on two observations about the spectro-temporal energy distribution of the speech signal: the redundancy of the carried information and the sparse distribution of the regions with high energy. Based on these observations, the glimpsing model of speech perception hypothesizes that listeners process degraded speech by taking advantage of "glimpses" -time-frequency (TF) regions where the signal is least distorted. Using different glimpse detection criteria, Cooke [4] suggested that the proportion of the TF regions glimpsed -glimpse proportion (GP)- is a good predictor of intelligibility. GP has shown strong positive correlation with the intelligibility of speech degraded by additive uncorrelated noise (AUN) [4–7]. The glimpsing model proposed in [4] uses the local SNR as the glimpse detection criterion: the TF units whose local SNR exceeds 3 dB are detected as glimpses. While providing a simple and effective criterion for glimpse detection, using SNR limits the application of the method to speech degraded by AUN. In many common scenarios, the speech degradation cannot be treated as caused by AUN, e.g., speech processed by non-linear noise reduction algorithms and reverberated speech. To remove this limitation, here, we modify GP to extend its applicability to distortions beyond AUN. To this end, we use the normalized cross-correlation (NCC) between the spectro-temporal modulation (STM) envelopes of the clean and degraded/processed speech signals [6–10] as a measure of local signal quality and craft a glimpse detection criterion tailored to the measure. The NCC is calculated between the STM envelopes of the clean speech signal and their distorted versions and does not require a separate noise signal to be available. Therefore, the resulting algorithm can be applied to any pair of time-aligned clean and degraded speech signals.

2. Glimpse Proportion

To motivate the proposed algorithm, we synopsize the GP introduced in [4, 5] and defined as the proportion of the regions in the TF representation of the speech signal where the local SNR exceeds a certain threshold:

$$\text{GP} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F} C[f,t], \tag{1}$$

where $1 \leq t \leq T$ indexes time frames, $1 \leq f \leq F$ indexes acoustic frequency channels, T and F are the number of time frames and acoustic frequency bins, respectively, and:

$$C[f,t] = \mathbb{1}_{+}(\text{SNR}[f,t] - \alpha) \tag{2}$$

where $\text{SNR}[f,t]$ denotes the local SNR, $\alpha \in \mathbb{R}$ denotes an SNR threshold, and $\mathbb{1}_{+} : \mathbb{R} \to \{0,1\}$ is the indicator function:

$$\mathbb{1}_{+}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases} \tag{3}$$

GP takes the time-aligned clean and AUN signals as input and computes the local SNR for each TF unit using the spectro-temporal excitation patterns (STEP) of the inputs, where STEP is a representation of the envelope of the basilar membrane response to a sound signal. GP has shown high correlation with the intelligibility of speech corrupted by AUN [4–7]. However, in many common scenarios, the difference between the clean and corrupted speech cannot be treated as AUN. E.g., for noise suppression processed speech, the difference signal is some non-linear mixture of the noise and speech signals.

http://dx.doi.org/10.21437/Interspeech.2021-605

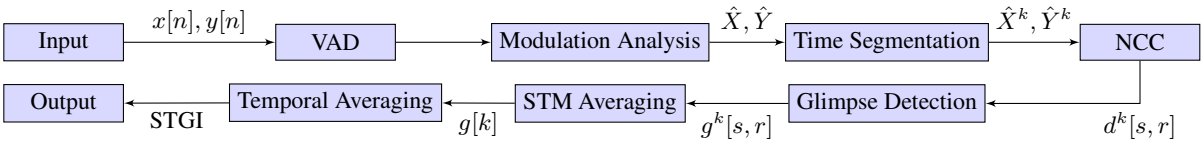

Figure 1: *A block diagram of the proposed STGI algorithm.*

3. Proposed Algorithm

To overcome the AUN limitation, we modify GP in two ways. First, we use a Gabor STM filter-bank [7, 11, 12] to decompose the input speech signals into STM frequency sub-bands. Perception of speech is crucially dependant on preserving the fidelity of salient STMs [9, 13–15], and STM analysis has previously shown promising results in SIP [6, 7, 10] and automatic speech recognition [11, 12]. Second, we modify the glimpse detection function $C(f,t)$ in Eq. 2 and employ NCC, as opposed to SNR, as an estimator of speech intelligibility [6–8, 10]. The NCC is computed between the STM envelopes of the clean and degraded speech signals and, thus, does not require an additive noise signal to be separately available.

Figure 1 shows a block diagram of the proposed algorithm. The algorithm takes the time-aligned clean $x[n]$ and degraded $y[n]$ speech signals as input. First, the STM envelopes of the input signals are extracted using a Gabor STM filter-bank. Next, the envelopes are divided into short-time segments, and each segment of the degraded signal is compared to that of the clean signal via NCC to produce a similarity measure. Next, similar to Eq. 2, a thresholding is applied to the similarity measures to detect STM glimpses. Finally, the overall intelligibility is estimated as the proportion of the glimpsed segments over the entire input. The building blocks of the algorithm are described in the following subsections.

3.1. Auditory-Modulation Analysis

First, an ideal voice activity detector (VAD) based on the clean reference signal [6, 8] is used to remove silent frames from the clean and degraded signals. Next, a Mel-frequency spectrogram is calculated for both input signals following the ETSI standard [7, 11, 16]. The Mel-spectrograms for the clean and degraded speech signals are denoted by $X[f,t]$ and $Y[f,t]$, respectively, and consist of F acoustic frequency channels equally spaced on the Mel-frequency scale. Next, the Mel-spectrograms are decomposed using an STM filter bank [7, 11, 12]. The filter bank consists of S spectral and R temporal Gabor modulation filters covering the range of modulation frequencies that are known to be important for speech perception [11] and SIP [7]. The STM filter-bank is implemented as a two-step procedure: spectral modulation filtering followed by temporal modulation filtering. The filter bank's output consists of $S \times R$ filtered spectrograms that exhibit modulation patterns characteristic of the associated filters. We denote the resultant TF decompositions by $\hat{X}[f,t;s,r]$ and $\hat{Y}[f,t;s,r]$, where $1 \leq s \leq S$ and $1 \leq r \leq R$ index the spectral and temporal modulation center frequencies, respectively, of the associated filters. For a detailed description of the modulation filter-bank, we refer to [7, 12].

3.2. Envelope Comparison

As discussed in Sec. 2, GP uses the local SNR as a measure of speech degradation in each TF unit. Here, we develop a similarity measure between the clean and degraded speech signals by comparing their STM envelopes in short-time segments [6, 8]. We present the processing steps for a single STM channel, i.e., one filtered spectrogram; the steps are identical for all filtered spectrograms. Let $\hat{X}^k, 1 \leq k \leq T - N + 1$ denote the k-th short-time segment of length N:

$$\hat{X}^k = \begin{bmatrix} \hat{X}[1,k] & \hat{X}[1,k+1] & \dots & \hat{X}[1,k+N-1] \\ \hat{X}[2,k] & \hat{X}[2,k+1] & \dots & \hat{X}[2,k+N-1] \\ \vdots & \vdots & \ddots & \vdots \\ \hat{X}[F,k] & \hat{X}[F,k+1] & \dots & \hat{X}[F,k+N-1] \end{bmatrix}. \quad (4)$$

Note that the dependencies on s and r are dropped for simplicity. First, mean and variance normalization is applied to each row and each column of the short-time segments [6]. We denote the clean and degraded row- and column-normalized short-time segments by $\bar{X}^k = [\bar{x}^k_{ij}]$ and $\bar{Y}^k = [\bar{y}^k_{ij}]$, respectively. Next, the columns of $\bar{X}^k$ and $\bar{Y}^k$ are compared using NCC, and the NCCs are averaged across columns to produce the STM channel's similarity measure for the k-th short-time segment:

$$d^k[s,r] = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{F} \bar{x}_{ij} \bar{y}_{ij}, \quad (5)$$

Each of the STM channels $[s,r]$ contributes a similarity measure $-1 \leq d^k[s,r] \leq 1$ for the k-th time segment.

3.3. Glimpse Detection

As described in Eq. 2, GP compares the local SNR with a fixed threshold to determine the intelligibility of a TF unit. We follow a similar approach and compare the similarity measure $d^k[s,r]$ with a fixed threshold to determine each STM channel's glimpse estimate for the k-th short-time segment:

$$g^k[s,r] = \mathbb{1}_+(d^k[s,r] - \beta_{sr}), \quad (6)$$

where $-1 \leq \beta_{sr} \leq 1$ is the glimpsing threshold for the $[s,r]$ STM channel to be determined with the aid of a small dataset in Sec. 3.5. Note that each STM channel contributes a glimpse estimate for the k-th time segment. Lastly, the glimpsing index for the k-th short-time segment is computed as the average of glimpse estimates over all STM channels:

$$g[k] = \frac{1}{SR} \sum_{s=1}^{S} \sum_{r=1}^{R} g^k[s,r]. \quad (7)$$

Fig. 2 shows the glimpsing index $g[k]$ as a function of time for a speech signal sample degraded by AUN.

3.4. Intelligibility Prediction

The spectro-temporal glimpsing index (STGI) is defined as the temporal average of the short-time glimpsing index over the entire input:

$$\text{STGI} = \frac{1}{K} \sum_{k=1}^{K} g[k]. \quad (8)$$

3.5. Glimpsing Thresholds

Here, we describe the procedure to select the glimpsing threshold β_{sr} for each STM channel. For simplicity, we employ a greedy approach and aim to maximize each individual STM

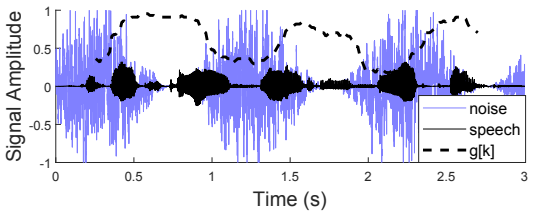

Figure 2: *Glimpse index $g[k]$ as a function of time for a speech sample degraded by amplitude modulated white Gaussian noise.*

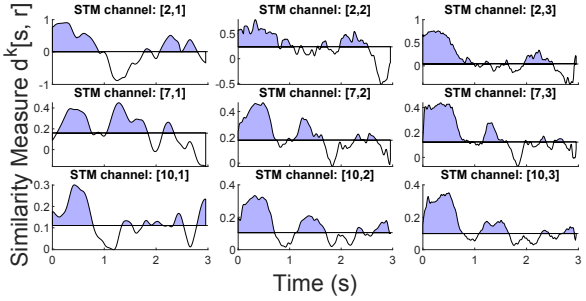

Figure 3: *An illustration of $u_l[s,r]$ for a number of STM channels computed for a signal pair. The ground truth intelligibility is 60%. Hence, for each STM channel, the threshold $u_l[s,r]$ is selected such that 60% of the short-time segments have a similarity measure $d_l^k[s,r]$ above the threshold $u_l[s,r]$. The shaded time-segments are detected as glimpses.*

channel's SIP performance in isolation, leading to solve a sequence of simple one-dimensional problems. Consider the l-th signal pair, $l = 1, ..., L$ of a training dataset comprising L clean and degraded speech signal pairs and a subjective intelligibility score I_l for each degraded signal. Let $u_l[s,r]$ denote the threshold that results in the correct intelligibility prediction for this signal pair if *only* the glimpsing index of the $[s,r]$ STM channel is used as the predictor of intelligibility. More specifically, for each STM channel and for a single signal pair, the threshold $u_l[s,r]$ is selected such that:

$$\frac{1}{K_l}\sum_{k=1}^{K_l} g_l^k[s,r] = I_l, \tag{9}$$

where K_l is the total number of time segments and:

$$g_l^k[s,r] = \mathbb{1}_+(d_l^k[s,r] - u_l[s,r]), \tag{10}$$

where $d_l^k[s,r]$ is the similarity measure for the k-th segment of the l-th sample of the training dataset. We note that such a threshold always exists, as long as a single STM channel and a single signal pair are considered. The glimpsing threshold β_{sr} for each STM channel is then calculated as the average of $u_l[s,r]$ over all the signal pairs in the training dataset:

$$\beta_{sr} = \frac{1}{L}\sum_{l=1}^{L} u_l[s,r]. \tag{11}$$

A subset of the ITFS-Kjems dataset (introduced in Sec. 4) consisting of 200 randomly chosen sentences were used as the training set. Fig. 3 shows the similarity measure $d_k^k[s,r]$ as a function of time for a number of STM channels of a speech signal sample, along with the thresholds $u_l[s,r]$. Note that the same threshold cannot be used for all the STM channels since both

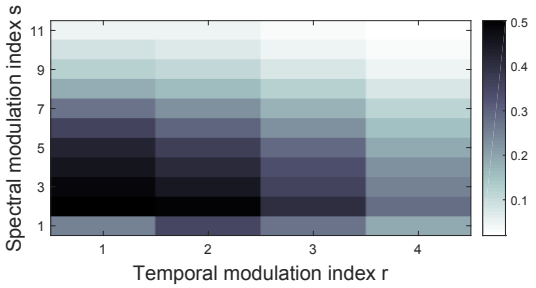

Figure 4: *Average glimpsing thresholds β_{sr}.*

clean and degraded speech signals have different statistics and result in different distributions of the NCCs across the channels. Fig. 4 shows the resulting thresholds for the STM channels.

4. Implementation and Evaluation

Tables 1 and 2 summarize the model parameters used in this study. The glimpsing thresholds β_{sr} were calculated following the procedure described in Sec. 3.5. We compare the SIP performance of STGI with a suite of state-of-the-art SIP algorithms in terms of i) Pearson correlation coefficient and ii) Spearman rank correlation coefficient between the ground truth intelligibility and the algorithm outputs.

We evaluate the SIP performance of STGI across 15 listening test datasets, which together cover degradation conditions including modulated noise (ModN), checkerboard noise (CheckB), gated noise (GatedN), noise reduction processing (NR), reverberation (Rev), and near-end listening enhancement (NELE). Table 5 summarizes the data sets, while the accompanying references provide additional details. The first 6 datasets are for AUN, dataset 7 includes temporally filtered speech, datasets 8 and 9 include noisy speech processed by non-linear noise-reduction algorithms, dataset 10 covers ideal TF masking, datasets 11 and 12 consist of reverberated speech, and datasets 13-15 include NELE processed speech degraded by noise and reverberation.

Table 1: *Model parameters.*

Parameter	Value	Parameter	Value
f_s	10 kHz	F	130
S	11	R	4
N	40 (512 ms)		

Table 2: *Spectral and temporal modulation center frequencies.*

Parameter	Value
Spectral Modulation Filter Center Frequencies (cyc/kMel)	0, 0.28, 0.44, 0.72, 1.10, 1.77, 2.88, 4.54, 7.26, 11.6, 18.3
Temporal Modulation Filter Center Frequencies (Hz)	0, 4.9, 7.8, 12.4

5. Results and Discussion

Tables 3 and 4 show the Pearson and Spearman rank correlation coefficients between the ground truth intelligibility and each SIP algorithm's output over the datasets introduced in Sec. 4. The results indicate excellent performance of STGI over a wide range of degradations and datasets.

Comparing STGI and GP, a few observations can be made. Firstly, different from GP, the application of STGI is not limited

Table 3: *Performance of different SIP algorithms in terms of Pearson correlation coefficient.*

	CheckB1 Fogerty	CheckB2 Fogerty	GatedN Fogerty	ModN Jensen	ModN Fogerty	ModN Gibbs	ModFN Fogerty	NR Jensen	NR Hu	ITFS Kjems	Rev HINT	Rev IEEE	NELE Taal	NELE Cooke	NELE Chermaz	AUN Mean	Overall Mean
STGI	**0.80**	0.73	**0.94**	**0.93**	**0.94**	**0.94**	**0.93**	**0.99**	0.89	0.92	**0.95**	0.95	0.94	0.95	0.67	**0.88**	**0.90**
GP [4]	0.68	0.72	**0.94**	0.85	0.83	0.84	—-	—-	—-	—-	—-	—-	—-	—-	—-	0.81	—-
ESII [17]	0.67	0.67	0.81	0.82	0.82	0.81	—-	—-	—-	—-	—-	—-	—-	—-	—-	0.77	—-
HASPI [18]	**0.80**	**0.82**	0.59	0.61	0.61	0.80	0.83	0.92	0.68	0.88	0.84	**0.97**	0.44	0.86	**0.87**	0.71	0.77
wSTMI [7]	-0.41	0.46	0.89	0.92	0.93	0.92	0.91	0.94	**0.92**	0.93	0.91	0.94	0.96	**0.96**	0.86	0.62	0.80
OSTMI [10]	0.69	0.79	0.93	0.88	0.89	0.87	0.89	0.89	0.89	0.86	0.84	0.86	0.89	0.82	0.71	0.84	0.85
STOI [8]	-0.03	0.32	-0.50	0.45	-0.77	0.55	0.35	0.98	0.86	**0.94**	0.88	0.89	0.89	0.70	0.38	0.00	0.46
eSTOI [6]	0.48	0.54	0.74	0.85	0.76	0.87	0.88	0.97	0.90	0.93	0.88	0.89	0.88	0.92	0.72	0.71	0.81
SIIB [19]	0.60	0.57	0.42	0.83	0.78	0.80	0.82	0.94	**0.92**	0.81	0.78	0.92	**0.97**	0.90	0.62	0.67	0.78

Table 4: *Performance of different SIP algorithms in terms of Spearman rank correlation coefficient.*

	CheckB1 Fogerty	CheckB2 Fogerty	GatedN Fogerty	ModN Jensen	ModN Fogerty	ModN Gibbs	ModFN Fogerty	NR Jensen	NR Hu	ITFS Kjems	Rev HINT	Rev IEEE	NELE Taal	NELE Cooke	NELE Chermaz	AUN Mean	Overall Mean
STGI	**0.89**	0.82	**0.87**	**0.94**	**1.00**	**0.92**	**0.93**	**0.98**	0.84	0.92	0.89	0.93	0.97	0.94	0.67	**0.91**	**0.90**
GP [4]	0.86	0.79	0.59	0.87	0.43	0.77	—-	—-	—-	—-	—-	—-	—-	—-	—-	0.72	—-
ESII [17]	0.84	0.71	0.59	0.84	0.37	0.78	—-	—-	—-	—-	—-	—-	—-	—-	—-	0.69	—-
HASPI [18]	0.80	**0.88**	0.60	0.76	0.26	0.76	0.82	0.95	0.64	0.92	0.88	0.94	0.39	0.91	**0.91**	0.68	0.76
wSTMI [7]	-0.37	0.23	0.59	**0.94**	0.71	0.91	**0.93**	0.96	0.90	**0.96**	0.90	**0.95**	0.97	**0.95**	0.89	0.50	0.76
OSTMI [10]	0.80	0.87	0.60	0.91	0.49	0.87	**0.93**	0.90	0.84	0.89	**0.97**	0.91	0.85	0.85	0.69	0.76	0.82
STOI [8]	-0.15	0.38	-0.74	0.48	-0.94	0.54	0.67	0.96	0.80	**0.96**	**0.97**	0.94	0.89	0.72	0.31	-0.07	0.45
eSTOI [6]	-0.10	0.68	0.60	0.92	0.26	0.82	0.91	**0.98**	0.87	**0.96**	0.96	0.94	0.89	**0.95**	0.68	0.53	0.75
SIIB [19]	0.69	0.59	0.60	0.84	0.26	0.80	0.87	**0.98**	**0.92**	0.85	0.93	0.90	**0.98**	0.89	0.66	0.63	0.78

Table 5: *Summary of listening test datasets.*

Name	Degradation
1 CheckB1-Fogerty [20]	Checkerboard noise
2 CheckB2-Fogerty [21]	Checkerboard noise
3 GatedN-Fogerty [20]	Gated speech shaped noise (SSN)
4 ModN-Jensen [6]	Modulated noise
5 ModN-Fogerty [22]	Time compressed/expanded modulated noise
6 ModN-Gibbs [23]	Same as ModN-Fogerty
7 ModFN-Fogerty [24]	Temporally filtered speech + speech-modulated noise
8 NR-Jensen [25]	Unmodulated SSN
9 NR-Hu [26]	Babble, car, street, and train noise
10 ITFS-Kjems [27]	Unmodulated SSN, cafeteria , car and bottling factory noise
11/12 Rev-HINT/IEEE [28]	Reverberation
13 NELE-Taal [29]	Speech shaped and babble noise
14 NELE-Cooke [30]	Unmodulated SSN and competing speaker
15 NELE-Chermaz [31]	Unmodulated SSN and competing speaker + reverberation

to AUN and it can be used for SIP in acoustic conditions where the speech and noise are not separately available. Secondly, STGI consistently outperforms GP in the presence of AUN.

Another interesting observation is the failure of several state-of-the-art SIP algorithms in the presence of checkerboard and gated noises. These conditions consist of artificial maskers that provide a controlled environment for studying the influence of the spectro-temporal properties of the glimpses on speech perception. A SIP algorithm's ability to successfully capture and model such properties can lead to performance gains in other degradation conditions. We note that several successful SIP algorithms (including wSTMI, eSTOI, and SIIB) that have excellent performance across many acoustic conditions, struggle to predict intelligibility in the presence of checkerboard and gated noise. On the other hand, STGI consistently shows high correlation with speech intelligibility in such degradation conditions (CheckB1-Fogerty, CheckB2-Fogerty, and GatedN-Fogerty datasets). Like STGI, wSTMI also decomposes the input speech signals using a Gabor STM filter bank and compares the filtered spectrograms using NCC. However, different from STGI, wSTMI uses a sparse linear model to combine information from only the most salient 8 STM channels. While the approach has shown excellent performance in a wide range of degradations, it fails to predict intelligibility in the presence of checkerboard noise. Unlike conventional distortions, the maskers' energy in the checkerboard noise datasets is distributed over a small number of STM channels, thus preserving much of the speech information in the other STM channels. While the human auditory system can take advantage of the information from less affected STM channels, the sparse representation of wSTMI cannot capture this opportunistic behaviour. STGI emulates this behaviour by pooling the information from all the STM channels in the range that is usable for speech perception.

6. Conclusion

We presented an intrusive speech intelligibility prediction (SIP) algorithm based on detecting glimpses in a spectro-temporal modulation decomposition of the input speech signal. The algorithm -STGI- was developed by overhauling a pre-existing speech intelligibility prediction scheme -GP- and thereby expanding its narrow applicability to a wide range of distortions, including speech processed by non-linear noise suppression algorithms and reverberated speech. The results suggested that the proposed algorithm outperforms established SIP methods in the presence of temporally and spectro-temporally gated maskers, while also consistently provides high performance for the other tested degradation conditions.

7. References

[1] M. Cooke and D. P. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Commun.*, vol. 35, no. 3-4, pp. 141–177, 2001.

[2] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[3] A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 1990, pp. 845–848.

[4] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[5] Y. Tang and M. Cooke, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2016, pp. 2488–2492.

[6] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[7] A. Edraki, W. Y. Chan, J. Jensen, and D. Fogerty, "Speech intelligibility prediction using spectro-temporal modulation analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 210–225, 2021.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2719–2732, 1999.

[10] A. Edraki, W. Y. Chan, J. Jensen, and D. Fogerty, "Improvement and assessment of spectro-temporal modulation analysis for speech intelligibility estimation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1378–1382.

[11] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 131, no. 5, pp. 4134–4151, 2012.

[12] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 137, no. 4, pp. 2047–2059, 2015.

[13] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, no. 2-3, pp. 331–348, 2003.

[14] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, 2005.

[15] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," *J. Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[16] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, ETSI ES 201 108 V1.1.3, European Telecommunications Standards, 2003.

[17] K. S. Rhebergen and N. J. Versfeld, "An SII-based approach to predict the speech intelligibility in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 2394–2394, 2004.

[18] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, 2020.

[19] S. Van Kuyk, W. Bastiaan Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[20] D. Fogerty, B. L. Carter, and E. W. Healy, "Glimpsing speech in temporally and spectro-temporally modulated noise," *J. Acoust. Soc. Amer.*, vol. 143, no. 5, pp. 3047–3057, 2018.

[21] D. Fogerty, V. A. Sevich, and E. W. Healy, "Spectro-temporal glimpsing of speech in noise: Regularity and coherence of masking patterns reduces uncertainty and increases intelligibility," *J. Acoust. Soc. Amer.*, vol. 148, no. 3, pp. 1552–1566, 2020.

[22] D. Fogerty, J. Xu, and B. E. Gibbs, "Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum," *J. Acoust. Soc. Amer.*, vol. 140, no. 3, pp. 1800–1816, 2016.

[23] B. E. Gibbs and D. Fogerty, "Explaining intelligibility in speech-modulated maskers using acoustic glimpse analysis," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. EL449–EL455, 2018.

[24] D. Fogerty, R. E. Miller, J. B. Ahlstrom, and J. R. Dubno, "Effects of age, modulation rate, and modulation depth on sentence recognition in speech-modulated noise," *J. Acoust. Soc. Amer.*, vol. 145, no. 3, pp. 1718–1718, 2019.

[25] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 92–102, 2012.

[26] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[27] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[28] D. Fogerty, A. Alghamdi, and W.-Y. Chan, "The effect of simulated room acoustic parameters on the intelligibility and perceived reverberation of monosyllabic words and sentences," *J. Acoust. Soc. Amer.*, vol. 147, no. 5, pp. EL396–EL402, 2020.

[29] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, 2013.

[30] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge." in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 3552–3556.

[31] C. Chermaz, C. Valentini-Botinhao, H. Schepker, and S. King, "Evaluating near end listening enhancement algorithms in realistic environments," *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, pp. 1373–1377, 2019.