Speech Intelligibility Prediction using Spectro-Temporal Modulation Analysis

Edraki, Amin; Chan, Wai Yip Geoffrey; Jensen, Jesper; Fogerty, Daniel

Published in:
IEEE/ACM Transactions on Audio Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2020.3039929

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Speech Intelligibility Prediction using Spectro-Temporal Modulation Analysis

Amin Edraki, Wai-Yip Chan, Jesper Jensen, and Daniel Fogerty

***Abstract*—Spectro-temporal modulations are believed to mediate the analysis of speech sounds in the human primary auditory cortex. Inspired by humans' robustness in comprehending speech in challenging acoustic environments, we propose an intrusive speech intelligibility prediction (SIP) algorithm, wSTMI, for normal-hearing listeners based on spectro-temporal modulation analysis (STMA) of the clean and degraded speech signals. In the STMA, each of 55 modulation frequency channels contributes an intermediate intelligibility measure. A sparse linear model with parameters optimized using Lasso regression results in combining the intermediate measures of 8 of the most salient channels for SIP. In comparison with a suite of 10 SIP algorithms, wSTMI performs consistently well across 13 datasets, which together cover degradation conditions including modulated noise, noise reduction processing, reverberation, near-end listening enhancement, and speech interruption. We show that the optimized parameters of wSTMI may be interpreted in terms of modulation transfer functions of the human auditory system. Thus, the proposed approach offers evidence affirming previous studies of the perceptual characteristics underlying speech signal intelligibility.**

***Index Terms*—speech intelligibility, speech quality model, spectro-temporal modulation**

I. Introduction

Expensive, time-consuming listening tests can be replaced by speech intelligibility prediction (SIP) methods during the development of speech processing systems. An ideal SIP algorithm would accurately estimate the intelligibility of a possibly degraded/processed signal as perceived by a group of normal-hearing listeners. SIP algorithms can be used repeatedly during system development to track the performance of the evolving and final system. In this study, we develop an intrusive or reference-based SIP algorithm that relies on inputting a clean reference signal. The algorithm is based on a spectral-temporal modulation analysis (STMA) of the clean and the degraded/processed speech signals.

Temporal modulation envelopes and modulation transfer functions (MTFs) have long been exploited for SIP, serving as the backbone for both intrusive and non-intrusive SIP. The Speech transmission index (STI) [1] is one of the first successful SIP algorithms to exploit the observation that reverberation tends to reduce the depth of speech temporal modulations. STI uses bandpass amplitude-modulated speech-shaped noise (SSN) as the input to the communication channel and measures the reduction in modulation depth in each frequency band. An intelligibility index is then calculated as a weighted sum of frequency bands. STI extends the range of degradations that is covered by the Articulation Index (AI) [2], [3] and Speech Intelligibility Index (SII) [4], to encompass convolutive distortions. AI and SII were designed for stationary additive noise distortions and bandwidth reduction [2]–[4].

Going beyond STI, Speech-to-Reverberation Modulation energy Ratio (SRMR) [5] performs spectral analysis on the bandpass modulation envelopes of the degraded speech signal. A non-intrusive or reference-free scheme, SRMR compares the amount of energy associated with low temporal modulation frequencies and the energy associated with high temporal modulation frequencies, exploiting the fact that the temporal modulation spectrum of natural speech exhibits a low-pass behavior [6], [7].

Even though STI, AI, and SII are suitable for a wide range of degradation conditions, the algorithms have several limitations. For example, the algorithms perform poorly for non-linearly processed speech signals, e.g., noisy speech processed by noise reduction algorithms [8]. Besides, SII does not make accurate predictions for fluctuating noise, as it relies on long-term power spectrum averages. Several extensions have been proposed to overcome these limitations [9]. For example, to extend SII to non-stationary noise, Rhebergen et al. proposed Extended SII (ESII) [9] which segments the signal into small time frames and computes the conventional SII within each frame. Then, the final intelligibility index is calculated as the temporal average of the intermediate SII values. Coherence SII (CSII) [10] was proposed to include broadband peak-clipping and center-clipping distortions. In a similar manner, Goldsworthy et al. proposed speech STI (sSTI) [11] which accounts for non-linear distortions by replacing the artificial speech-shaped noise in STI with actual speech signals [12]–[14].

In a somewhat similar manner to (E)SII, the glimpse proportion [15], [16] was defined as the proportion of spectro-temporal regions in which the local SNR is above a pre-defined threshold (*glimpses*), and was shown to be highly correlated with speech intelligibility [16], [17]. The glimpse proportion is similar to SII in that both assume that audibility determines intelligibility. However, while SII operates on a long-term average SNR of each acoustic frequency sub-band, glimpse proportion operates on local time-frequency regions. Even though glimpse proportion has shown a high correlation with speech intelligibility in the presence of modulated maskers, its application is limited to speech degraded by additive noise.

A. Edraki and W.-Y. Chan are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: a.edraki@queensu.ca, chan@queensu.ca).

J. Jensen is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark, and also with the Oticon A/S, 2765 Smørum, Denmark (e-mail: jesj@demant.com).

D. Fogerty is with the Department of Speech and Hearing Science, University of Illinois, Urbana-Champaign, USA (e-mail: dfogerty@illinois.edu).

Several studies suggested that normal-hearing listeners benefit from temporal fine structure (TFS) information in the presence of modulated maskers [18], [19]. TFS spectrum index (TFSS) [20] is a SIP algorithm that incorporates the Hilbert-derived TFS information into SIP. TFSS first decomposes the input speech signals into acoustic frequency sub-bands, using multiple bandpass filters. Next, the phase-modulated carrier of the Hilbert envelope is calculated for each bandpass signal. Then, within each frequency band, the magnitude-squared coherence between TFS signals is calculated. Finally, the overall speech intelligibility is computed as a weighted average of the coherence terms.

More recently, SIP methods have been proposed to cover a broader range of distortions. Taal et al. proposed Short-Time Objective Intelligibility (STOI) [21], which compares the temporal modulation envelopes of the clean and degraded speech samples in frequency sub-bands over short-time segments of speech to produce a similarity measure. STOI has shown high correlation with SI in many degradation conditions, including but not limited to noisy speech processed by single-channel noise reduction algorithms [21], speech processed by cochlear implants [22] and reverberant speech [23]. Although STOI performs well in many degradation conditions, it has limitations when additive noise with strong temporal modulation content is present in the signal under test [17]. Extended Short-Time Objective Intelligibility (eSTOI) [17] was proposed as an extension to STOI to cover a broader range of degradations. eSTOI is inspired by STOI and compares the spectro-temporal modulation envelopes of the clean and degraded speech signals over short-time segments. eSTOI has shown high correlation with SI in the presence of noise sources with highly modulated content but also shows high performance in situations in which other SIP algorithms work well [17], [23], [24].

Speech intelligibility in bits (SIIB) [25] is an information theoretic intelligibility metric that operates based on the hypothesis that the speech intelligibility is related to the mutual information between the modulation envelopes of the clean and degraded speech signals [26], [27]. SIIB uses a non-parametric mutual information estimator to estimate the information shared between the clean and degraded temporal envelopes.

Even though SIIB, STOI and eSTOI perform quite well in many degradation conditions, their fundamental processing steps are not strongly motivated by biological findings of the human auditory system. In another vein, more sophisticated models incorporating properties of the human auditory system into SIP have been proposed [15], [28]–[32]. For instance, HASPI [28] passes the clean speech signal and the signal under test through an auditory model and compares their envelopes and temporal fine structures to produce an intelligibility index [28]. HASPI has been shown to give accurate intelligibility estimates for a variety of degradation conditions, including speech processed using frequency compression and speech processed through noise-reduction algorithms. Furthermore, HASPI allows the auditory profile of the target listeners to be taken into account.

The Spectro-Temporal Modulation Index (STMI) [29], [30] uses a biologically-inspired modulation Gabor filter-bank [33] to decompose the clean and degraded speech spectrograms into a 4-D spectro-temporal modulation representation. The envelopes extracted from each spectro-temporal channel are then compared using normalized cross-correlation (NCC) to generate a SIP measure. STMI analyzes the effects of distortions on the *joint* spectro-temporal modulations (in contrast to commonly used temporal-only modulation analysis) in speech and is capable of predicting the speech intelligibility in the presence of several degradations where traditional SIP methods (such as STI) fail, e.g., phase jitter [29] and frequency-dependent phase shifts [29], [30]. However, STMI presumes a uniform contribution of different modulation frequencies to speech intelligibility, i.e. assigns equal weights to all the spectro-temporal modulation channels, which is not supported by behavioural findings of the human auditory system [34]. Although STMI uses an elaborate speech decomposition model, it is not capable of accurately predicting intelligibility in some degradation conditions, including speech processed by non-linear noise reduction algorithms, and speech degraded by temporally modulated noise [23]. In [23], OSTMI (denoted as ${}^{O}\text{STMI}^{O}$ in [23]) is proposed to improve STMI by using a modulation analysis filter-bank and a feature extraction scheme introduced in [35] and [36]. OSTMI uses a heuristic non-uniform weighting of different modulation channels for SIP, i.e., allowing larger contribution of spectral modulation frequencies that better predict speech intelligibility. It is shown [23] that the excellent performance of eSTOI is closely followed by OSTMI across many degradation conditions.

Motivated by the substantial performance improvement of OSTMI over STMI, here we propose an approach for SIP, based on STMA and assess the performance of an algorithm, wSTMI, designed to address the limitations of STMI and OSTMI. In contrast to the heuristic weighting of the modulation channels in OSTMI, here, we employ one dataset to train and another dataset to validate the weights in order to minimize the SIP error. The proposed approach features three distinctive characteristics: **i)** The proposed approach extends the concept of frequency-band importance function to spectro-temporal modulation channels and uses listening test data to optimize the weights assigned to different spectro-temporal modulation channels for SIP. **ii)** The resultant algorithm is more faithful to the processing in the human auditory system compared to other well-performing SIP algorithms such as eSTOI. Thus, the approach provides an opportunity to probe the perceptual components of the speech signal that affect speech intelligibility. **iii)** The approach provides a systematic way for further performance improvement with new degradation types.

The proposed approach first uses a voice activity detector to remove silent frames from the clean and degraded speech signals. Next, the clean and test speech spectrograms are calculated and passed through a spectro-temporal modulation filter-bank to produce modulation envelopes tuned to specific spectro-temporal modulation frequencies. The modulation envelopes of the clean and test signals are then compared using NCC, and combined using a simple linear regression model. The parameters of the regression model are determined through a regularized least squares method applied to a small training set. We show that the optimized set of parameters

may be interpreted in terms of a spectro-temporal modulation transfer function and is in line with previous biological and perceptual studies [34]. We also compare our findings of the relative importance of spectro-temporal modulation frequencies to previous studies on automatic speech recognition (ASR). We show that similar spectro-temporal modulation channels are crucial for both SIP and ASR. Finally, we show that the proposed algorithm performs well in all the degradation conditions investigated in this study, including conditions where established SIP algorithms perform less well.

This paper is structured as follows. In section II-A, the STMA scheme used in this study is described. STMI is briefly introduced in Section II-B, in order to background our work and show new performance results for STMI. The proposed SIP algorithm, wSTMI, is introduced in Section III. Section V covers the degradation types and datasets that we used for investigation in this study. The performance of the proposed SIP algorithm is evaluated and compared to other SIP algorithms in Section VI. In Section VII, we interpret the final model based on biological spectro-temporal MTFs. Lastly, Section VIII concludes the work.

II. Background

A. Spectro-Temporal Modulation Analysis

To establish a foundation for the proposed algorithm, we briefly discuss an STMA scheme proposed in [35], [36] that has shown promising results in ASR [35], [36] and SIP [23]. Figure 1 shows the STMA scheme introduced in [36] which comprises two stages of processing: auditory-spectral analysis, followed by modulation analysis. The two stages are discussed in the following subsections.

1) Auditory-Spectral Analysis: The STMA scheme operates at a sampling frequency of f_s. To mimic the early stages in the cochlea, first, a Mel frequency spectrogram is calculated. $O_{FFT}\%$ overlapping Hann-windowed frames of length W_{FFT} are zero padded to N_{FFT} samples and transformed using FFT. The absolute values of the FFT coefficients are processed using a Mel filter-bank. The frequency band from F_l to F_u is divided into F_M channels equally spaced on the Mel frequency scale. Each channel has a triangular shaped weighting window, and adjacent channels overlap by $O_M\%$. The output of each Mel-filter is the window-weighted sum of the FFT magnitude values in each band, calculated in accordance with the ETSI standard [35] [37]. The output of the Mel-filter is then compressed with the natural logarithm. The calculated spectrogram is denoted by $X[f,n]$ where $1 \leq f \leq F_M$ indexes acoustic frequency channels, and n indexes time frames.

2) Modulation Analysis: In this step, the Mel spectrogram $X[f,n]$ is transformed into a multi-resolution spectro-temporal decomposition using two Gabor modulation filter-banks [35]. The decomposition is performed in two steps: spectral-only, followed by temporal-only modulation filtering. This results in a collection of filtered spectrograms that exhibit modulation patterns characteristic of the modulation passbands of the associated filters. Equations 1 and 2 describe the one-dimensional filters used to perform the separate spectral- and temporal-modulation filtering:

$$h_b(x) = \begin{cases} 0.5 + 0.5\cos(\frac{2\pi x}{b}), & -\frac{b}{2} < x < \frac{b}{2}, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$g(x;\omega) = \begin{cases} h_{b_{max}}(x), & \omega = 0 \\ \cos(\omega x) \cdot h_{\nu\pi/\omega}(x), & \omega \neq 0 \end{cases} \tag{2}$$

where h_b is a Hann-envelope of width b, ν is the number of half-waves under the envelope, and $b_{max} < \infty$ denotes the maximum filter size. The spectral and temporal modulation filter-banks consist of S spectral and R temporal modulation filters [35], [36]:

$$\begin{aligned} G_S[f;s_i] &= g(f;s_i) \\ G_T[n;r_j] &= g(n;r_j) \end{aligned} \tag{3}$$

where s_i and r_j denote the spectral and temporal modulation center frequencies of the filters, respectively. Spectral and temporal modulation frequencies are also termed "scale" and "rate", and are represented by the notations s and r, respectively. Note that filter supports are inversely proportional to the center frequency, and are proportional to the number of half-waves under the envelope.

The decomposition can now be expressed as:

$$\tilde{X}[f,n;s_i,r_j] = X[f,n] * G_S[f;s_i] * G_T[n;r_j], \tag{4}$$

where $\tilde{X}$ is the resultant time-frequency representation, and $*$ denotes linear convolution. The spectrogram is zero-padded prior to the convolution so that the filtered spectrograms have the same size as the original spectrogram [35], [36]. The output of the STMA is a set of $S \times R$ filtered spectrograms. We let $\tilde{X}[f,n;s_i,r_j]$ denote the filtered spectrogram tuned to the spectral and temporal modulation frequencies s_i and r_j, respectively. For a detailed description of the filter banks, we refer the reader to [35] and [36]. The two-step modulation filtering process is illustrated in Figure 2, which shows a Mel-spectrogram before and after each modulation filtering step.

B. Spectro-Temporal Modulation Index (STMI)

STMI is an intrusive SIP algorithm introduced in [29], [30] based on STMA of the clean and degraded/processed input speech samples. STMI uses a biologically inspired STMA scheme [30], [38], [39] to produce multi-resolution spectro-temporal decompositions $\tilde{X}_M$ and $\tilde{Y}_M$ for the clean and degraded signals, respectively. While conceptually similar to the STMA scheme reviewed in Section II-A, the implementation of the spectro-temporal modulation filters used in STMI comprises more sophisticated biologically-inspired processing steps (e.g., using a lateral inhibitory network before modulation filtering). We emphasize this difference by denoting the resulting modulation filtered spectrograms as $\tilde{X}_M[f,n;s,r]$ and $\tilde{Y}_M[f,n;s,r]$. For a detailed description of STMI's STMA scheme, we refer the reader to [29], [30]. We emphasize that both STMA schemes produce a 4-D spectro-temporal decomposition of the input speech signal.

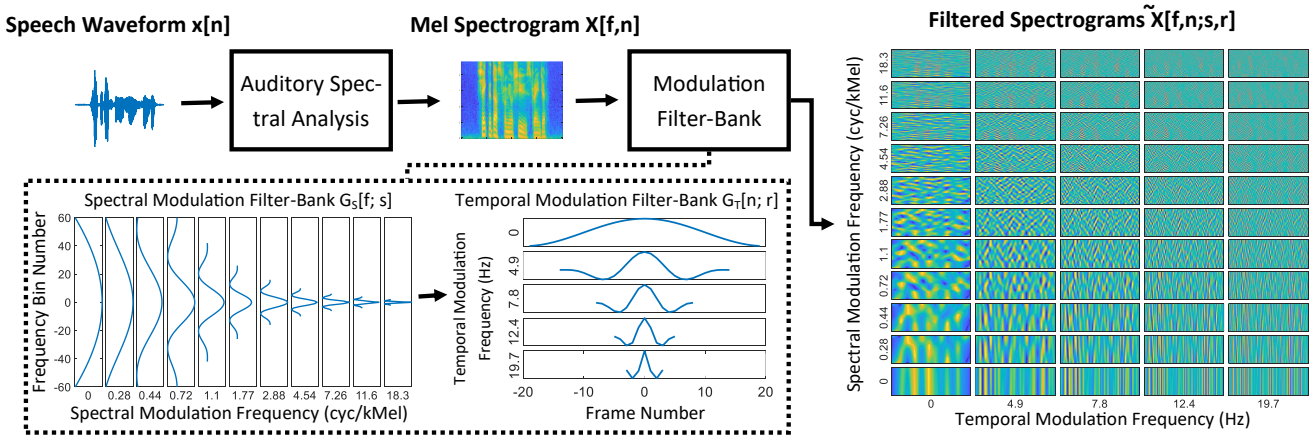

Fig. 1: A block diagram of the spectro-temporal modulation analysis (STMA) scheme described in Section II-A. The impulse responses of the modulation filters are plotted in the dashed box. Using S=11 spectral and R=5 temporal modulation filters result in R×S=55 spectro-temporally filtered spectrograms. Because the slow spectral modulation filters have very long supports, the diagram can only show part of the first four spectral impulse responses.

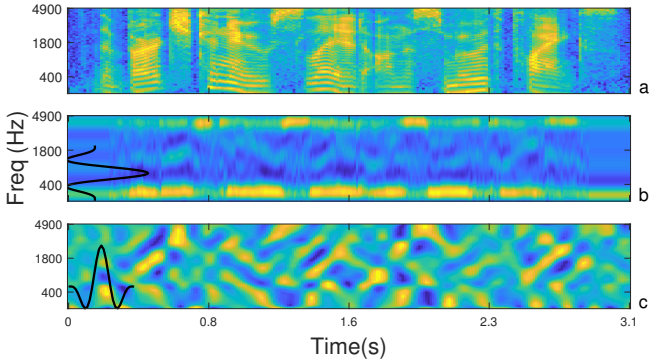

Fig. 2: An illustration of two-step spectro-temporal modulation filtering to produce a filtered spectrogram. (a) the Mel-spectrogram (b) the spectrally filtered spectrogram (c) the spectro-temporally filtered spectrogram. The impulse responses superimposed on the left are associated with spectral and temporal modulation center frequencies of 1.1 cyc/kMel and 4.9 Hz, respectively.

Next, the absolute values of the filtered spectrograms are summed along the frequency axis in order to capture the temporal variation of energy in each spectro-temporal modulation channel:

$$\begin{aligned}\bar{X}[n;s,r] &= \sum_f |\tilde{X}_M[f,n;s,r]| \\ \bar{Y}[n;s,r] &= \sum_f |\tilde{Y}_M[f,n;s,r]|.\end{aligned} \tag{5}$$

STMI then uses NCC to compare $\bar{X}$ and $\bar{Y}$ within each modulation channel. An intermediate intelligibility measure is defined as:

$$\bar{\rho}[s,r] = \frac{\langle \bar{X}[n;s,r] - \mu_{\bar{X}}, \bar{Y}[n;s,r] - \mu_{\bar{Y}} \rangle}{\|\bar{X}[n;s,r] - \mu_{\bar{X}}\| \; \|\bar{Y}[n;s,r] - \mu_{\bar{Y}}\|} \tag{6}$$

where the inner product and the induced norm are defined as:

$$\langle \bar{X}[n], \bar{Y}[n] \rangle = \sum_n \bar{X}[n] \cdot \bar{Y}[n] \tag{7}$$

$$\|\bar{X}[n]\| = \sqrt{\langle \bar{X}[n], \bar{X}[n] \rangle} \tag{8}$$

and $\mu_{\bar{X}}$ is defined as:

$$\mu_{\bar{X}} = \frac{1}{K} \sum_n \bar{X}[n] \tag{9}$$

where K is the total number of time frames. Note that the channel indexes s and r are dropped in Equations 7 to 9 for simplicity. Finally, an overall speech intelligibility estimate is defined as the average over all intermediate intelligibility measures:

$$\text{STMI} = \frac{1}{Z} \sum_{s,r} \bar{\rho}[s,r], \tag{10}$$

where Z denotes the total number of spectro-temporal modulation channels.

III. Proposed SIP Algorithm

Although STMI uses a biologically inspired representation of the clean and degraded speech signals for SIP, it performs poorly in several degradation conditions [23]. We attribute this failure to two oversimplifications. Firstly, the filtered spectrograms $\tilde{X}_M$ and $\tilde{Y}_M$ are summed over acoustic frequency to calculate the intermediate intelligibility measures $\bar{\rho}[s,r]$ (Equations 5 and 6). Hence, distinct contributions of different acoustic frequencies to speech perception are ignored. Secondly, STMI is computed as the average of the intermediate intelligibility measure $\bar{\rho}[s,r]$ over all the spectro-temporal

modulation channels in Eq. (10). In other words, STMI presumes equal contribution of different spectro-temporal modulation frequencies to speech intelligibility. However, experimental measurement of the sensitivity of the human auditory system [30] suggests that distinct spectro-temporal modulation frequencies are not equally important for speech-perception by the human auditory system. In particular, Elliott et al. [34] measured the loss in intelligibility caused by eliminating specific modulation frequencies from speech. They showed that different modulation frequencies are not equally crucial for the comprehension of speech by human listeners. In [40], Kates et al. showed that the low temporal modulation rates provide the highest information for speech intelligibility using the normalized cross-covariance of the degraded signal envelope with that of a reference signal. In [41], Steinmetzger et al. showed that the temporal modulation frequencies are not equally important for SIP. Also, in [23], we showed that a non-uniform heuristic weighting of the spectro-temporal modulation frequencies remarkably improved the performance of STMI.

Motivated by the aforementioned shortcomings of STMI, here, we propose an improved algorithm we call weighted STMI (wSTMI). Figure 3 shows an overview of the proposed algorithm. First, a voice activity detector [17], [21] is used to remove silent frames from the clean and degraded speech signals. Next, the STMA scheme introduced in Section II-A is used to decompose the input speech signals. After that, an intermediate intelligibility measure is calculated for each spectro-temporal modulation channel. Finally, the intermediate intelligibility measures are combined using a spectro-temporal modulation importance function to estimate the intelligibility of the degraded speech. wSTMI differs from STMI in several ways. First, in [23], it was shown that STMI's SIP performance could be improved remarkably by replacing its STMA with the STMA scheme described in Section II-A. Below, the STMA scheme described in Section II-A is used to develop wSTMI. Secondly, we avoid the integration over the acoustic frequency axis. Instead, an intermediate intelligibility measure is calculated for each spectro-temporal modulation frequency channel. Thirdly, an optimization approach is proposed to combine the intermediate measures in order to account for the non-uniform contribution of the modulation frequencies to speech intelligibility. As in any other data-driven approach, the quality and generalizability of the fit is tied to the training data. Therefore, we evaluate the algorithm over multiple "unseen" datasets to ensure generalizability. Finally, we present some insight into the method and show that it is well in line with previous findings of the relative importance of different modulation frequencies for human and machine perception.

A. Intermediate Intelligibility Measure

Here, we propose to modify the intermediate intelligibility measure in Eq. (6). This modification is motivated by the fact that at all levels of the human auditory system, the speech signal is represented tonotopically, i.e., distinct acoustic frequencies are analyzed separately [42]. Hence, for each filtered spectrogram and each acoustic frequency bin, the reference and degraded signals are compared using NCC:

$$d[f; s_i, r_j] = \frac{\langle \tilde{X}[f, n; s_i, r_j] - \mu_{\tilde{X}}, \tilde{Y}[f, n; s_i, r_j] - \mu_{\tilde{Y}} \rangle}{\|\tilde{X}[f, n; s_i, r_j] - \mu_{\tilde{X}}\| \; \|\tilde{Y}[f, n; s_i, r_j] - \mu_{\tilde{Y}}\|}. \tag{11}$$

Here, in contrast to Eq. (6), the acoustic frequency bins are compared separately. Next, a new intermediate intelligibility measure $\rho[s_i, r_j]$ is defined as:

$$\rho[s_i, r_j] = \frac{1}{F_M} \sum_f d[f; s_i, r_j] \tag{12}$$

where F_M is the total number of Mel frequency bins.

Comparing speech spectrograms using NCC was beneficial for SIP in (e)STOI [17], [21]. However, (e)STOI compares the clean and degraded/processed speech spectrograms over short-time segments. In particular, it was shown [21] [17] that lengthening the segment in (e)STOI to above 384 ms decreases SIP performance, especially in the presence of non-stationary maskers. In contrast to (e)STOI, in the proposed approach, speech signals are not segmented into short-time windows, but, instead, integration is performed across the full duration of the speech signals in question (Eq. (11)). This "full duration" analysis was completed in the present experiment as a first investigation of the utility of the proposed algorithm. Future work will need to explore if a segmental implementation provides advantages to SIP beyond that implemented here, and under what conditions.

B. Intelligibility Estimation

To enable unequal contribution of distinct modulation frequencies to speech perception we propose to combine the set of intermediate intelligibility measures $\rho[s_i, r_j]$ using regression. In this study, a linear model is used to estimate the intelligibility of the degraded signal as a linear combination of the intermediate intelligibility measures:

$$\text{wSTMI} = \sum_{i=1}^{S} \sum_{j=1}^{R} w[s_i, r_j] \rho[s_i, r_j] \; + b \tag{13}$$

where $w[s_i, r_j] \in \mathbb{R}$ denotes the weights and $b \in \mathbb{R}$ is the intercept. Below, we show how the weights can be computed using least squares with L_1 regularization. Using a linear model offers some advantages over more complex models. Besides model simplicity, the number of parameters can be controlled via sparsification in order to reduce the risk of overfitting. Sparse linear models are easier to interpret and do not require a massive amount of data for training.

Eq. (13) can be written in a matrix form as:

$$\text{wSTMI} = \Phi^T W + b \tag{14}$$

where $W \in \mathbb{R}^{SR}$ and $\Phi \in \mathbb{R}^{SR}$ are constructed by stacking the weights $w[s_i, r_j]$ and intermediate measures $\rho[s_i, r_j]$ into a column vector, respectively.

IV. Implementation

The following subsections provide the implementation details for the proposed algorithm. Section IV-A covers the STMA parameters used in this study, and Section IV-B describes the optimization of the wSTMI's parameters.

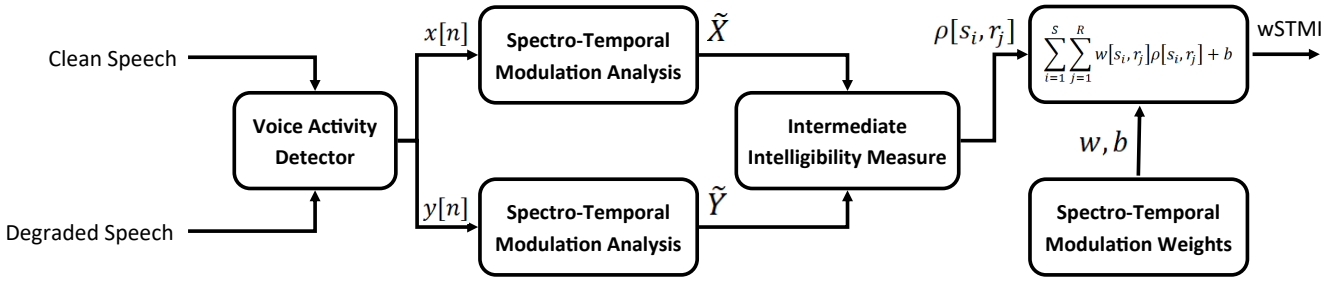

Fig. 3: A block diagram of the proposed SIP algorithm.

TABLE I: Auditory-Spectral Analysis Parameters

Parameter	Value	Parameter	Value
f_s	10 kHz	F_u	5000 Hz
O_{FFT}	50%	F_M	130
W_{FFT}	25.6 ms	b_s^m	390 channels
N_{FFT}	1024	b_t^m	40 frames
F_l	64 Hz		

TABLE II: Modulation Analysis Parameters

Parameter	Value
ν	3.5
S	11
R	5
Spectral Modulation Center Frequencies s_i (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 Center Frequencies r_j (Hz)	0, 4.9, 7.8, 12.4, 19.7

A. Spectro-Temporal Modulation Analysis

Tables I and II summarize the STMA parameters used in this study. To ensure that a sufficient range of high spectral modulation frequencies are covered, the calculation of the Mel-spectrogram consists of $F_M = 130$ frequency channels. The values of spectral and temporal modulation center frequencies are selected following the suggestions in [35] and [36]. Since a Mel-scaled frequency axis is used, the spectral modulation frequencies are specified in cycles per kMel (cyc/kMel). The selection of spectral and temporal modulation center frequencies ensures that adjacent filters exhibit a constant overlap in the modulation frequency domain. For the DC modulation filters, we adopt the suggestions in [36] and set $b_{max} = b_s^m = 3 \times F_M = 390$ channels for $s_i = 0$ and $b_{max} = b_t^m = 40$ time-frames for $r_j = 0$. For a more detailed description of the modulation filter bank, we refer the reader to [35], [36].

B. Parameter Optimization

We optimize W and b in Eq. (14) by minimizing the root mean square error (RMSE) between the wSTMI scores and the subjective intelligibility scores of a training data set. The minimization is augmented with a Lasso [43] sparsification constraint.

Consider a training dataset comprising L clean and degraded speech signal pairs and a subjective intelligibility score for each degraded signal. Let $\Phi_i \in \mathbb{R}^{SR}$ and $0 \leq I_i \leq 1, 1 \leq i \leq L$ denote the intermediate intelligibility vector and the subjective intelligibility score associated with the i-th signal, respectively. The optimization problem can be expressed as:

$$\min_{W,b} \left(\frac{1}{2L} \sum_{i=1}^{L} (I_i - b - \Phi_i^T W)^2 + \lambda \|W\|_1 \right), \qquad (15)$$

where $\lambda \geq 0$ is the regularization parameter, and $\|W\|_1$ is the L_1 norm of W [43]. Increasing λ increases our preference for a sparse model with fewer non-zero weights. Sparse linear models are easier to interpret and can generalize more accurately to unseen data [43]. Sparsification also reduces the number of selected modulation channels and hence the computational complexity of the SIP algorithm. Here, an algorithm based on cyclical coordinate descent is used to optimize W and b. For a detailed description of the algorithm, we refer the reader to [44]. We emphasize that the computational complexity of the optimization is of little concern, since it is performed once, and W and b are fixed afterward.

To optimize W and b, speech stimuli and the subjective intelligibility scores of the ITFS-Kjems and NELE-Taal datasets (described in Section V) are used for training and validation, respectively. The generalizability of the trained model is investigated by evaluating its performance over unseen data in Section VI. Figure 4 shows the RMSE and Pearson correlation between the model predictions and the subjective scores of the training and validation datasets as a function of λ. It is interesting to note that the RMSE curves for the training and validation datasets are almost flat for a wide range of λ, indicating that reducing the number of channels used has little impact on RMSE. For large values of λ, the performance over the validation set is better than the performance over the training set. We attribute this phenomenon to the fact that, unlike the common practice of partitioning one data set into a training and a validation set, two *different* datasets were used with one for training and the other for validation, resulting in drawing the training and validation samples from two different distributions.

We select the regularization parameter $\lambda^* = 0.065$ to minimize the RMSE over the validation dataset. Figure 5 shows the selected set of weights W associated with λ^*. The horizontal and vertical axes indicate the temporal and spectral modulation center frequencies, respectively. The weights associated with the selected spectro-temporal modulation channels are also shown. Figure 5 shows that only 8 out of the given 55

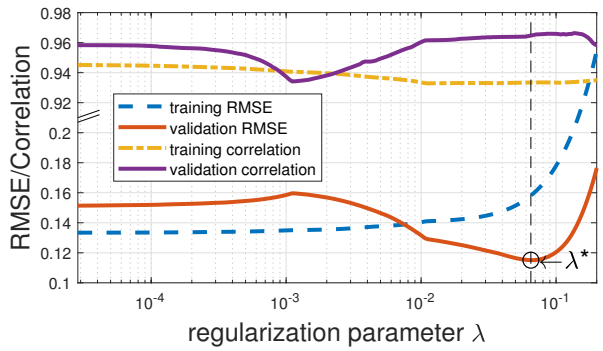

Fig. 4: The RMSE and the Pearson correlation between wSTMI predictions and subjective intelligibility scores of the training and validation datasets as a function of the regularization parameter λ. The highlighted value of $\lambda* = 0.065$ minimizes the RMSE over the validation dataset.

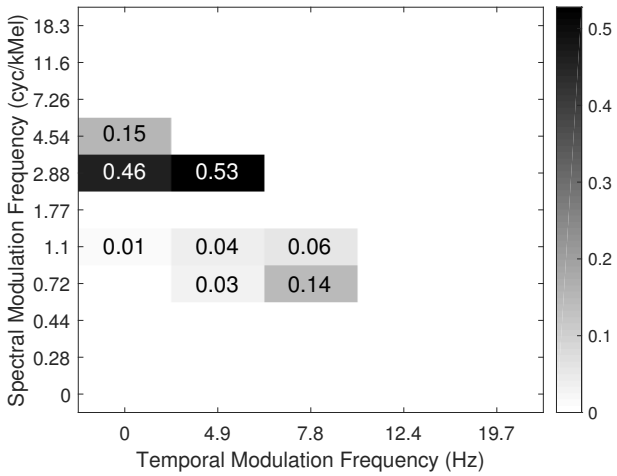

Fig. 5: Optimized weights W associated with $\lambda^* = 0.065$.

spectro-temporal modulation channels are selected for SIP. The optimized weights identify the importance of spectral modulation frequencies between 0.72 and 4.54 cyc/kMel and temporal modulation frequencies below 10 Hz.

V. Datasets

We evaluate wSTMI and other SIP algorithms using noise-corrupted/processed speech datasets collected previously from measuring the speech recognition performance of normal-hearing adults.

A. ITFS-Kjems

This dataset consists of speech subjected to ideal time-frequency segregation (ITFS) [45] processing. 150 Sentences from the Dantale II corpus [46] were degraded by four types of noise: unmodulated SSN, cafeteria noise, car interior noise, and noise from a bottling factory [47]. Noisy sentences were processed using an ideal binary mask (IBM) or a target binary mask (TBM). Noises were presented at three different SNRs: at 20% speech reception threshold (SRT), 50% SRT, and -60 dB SNR. Each binary mask was created using eight different relative-power criteria (RC). This resulted in a total of 7 (mask and noise type combinations) x 3 (SNRs) x 8 (RCs) = 168 conditions. N=15 subjects participated in the listening test. This dataset was used as the training set to optimize the parameters of wSTMI.

B. ModN-Jensen

This dataset consists of speech degraded by modulated noise [17]. Dantale II sentences were degraded by ten types of modulated noise, each presented at six SNRs selected to cover the full range of performance across the different noise types. Four of the maskers were selected from the ICRA noise corpus [48]: unmodulated SSN and 1/2/6-person babble. The ICRA signals are synthetic signals with spectral and temporal modulation properties similar to speech. Machine gun noise and destroyer operation room noise were selected from the Noisex corpus [49]. The rest of the maskers were sinusoidally amplitude-modulated SSN at 2, 4, 8, and 16 Hz. This resulted in a total of 10 (noise types) x 6 (SNRs) = 60 conditions. N=12 subjects participated in the experiment.

C. ModN-Fogerty

This dataset consists of speech degraded by modulated noise. In [50], IEEE sentences [51] spoken by a male talker [52] were degraded by six modulated maskers: unmodulated SSN and single-talker modulated SSN that was either time-compressed or expanded. Pitch-synchronous overlap-add time compression/expansion was employed to modify the modulation spectrum of the masker to run at 25%, 50%, 100%, 200%, and 400% of the duration; thereby, increasing or decreasing the rate of noise modulation. Maskers were presented at -7 dB SNR. This resulted in a total of 1 (unmodulated SSN) + 5 (modulated SSN) = 6 conditions. N=15 normal hearing subjects participated in the study.

D. ModN-Gibbs

In [53], IEEE sentences were corrupted by four types of noise: unmodulated SSN and three types of time-compressed/expanded speech-modulated noise. Unmodulated SSN and single-talker speech-modulated SSNs were created similar to those in Section V-C, and the noise modulation was time-compressed/expanded using pitch-synchronous overlap-add to run at 25%, 100%, or 400% of the original duration. Noises were presented at three SNR levels: -8, -4, and 0 dB. This resulted in a total of 4 (noise types) x 3 (SNRs) = 12 degradation conditions. Stimuli were presented to 5 listeners.

E. ModFN-Fogerty

This dataset [54] consists of temporally filtered speech, degraded by modulated noise. IEEE sentences, spoken by a male talker, and speech-modulated SSN were temporally filtered to retain modulations in either the low-pass (0-8 Hz) or high-pass (8-16 Hz) range. Moreover, the noise envelope was compressed (to 50% depth), left alone (100% depth), or expanded (to 200% depth) using an exponential function of

the instantaneous envelope amplitude (as in [55]), and noises were presented at two different SNRs (0 dB and -2 dB). This resulted in a total of 6 (temporal filtering configurations) x 3 (amplitude compressions) x 2 (SNRs) = 18 conditions. N=20 young normal hearing subjects participated in the listening test.

F. NR-Jensen

This dataset consists of speech processed by non-linear single-microphone noise reduction (NR) algorithms. In [56], the Dutch version of the noisy Hagerman sentences [57], [58] were degraded by unmodulated SSN at -8, -6, -4, -2, and 0 dB SNR. The degraded speech was processed by three NR algorithms aimed at finding binary or soft minimum MSE estimates of the short-time spectral amplitude. This resulted in a total of 4 (3 NR algorithms + 1 unprocessed) x 5 (SNRs) = 20 conditions. N=13 subjects participated in the listening test.

G. NR-Hu

In [52], IEEE sentences and isolated consonants were degraded by four types of noise: babble, car, street, and train at 0 and 5 dB SNRs. Degraded signals were then processed by eight non-linear NR algorithms, including spectral subtraction, sub-space, statistical model-based, and Wiener-type algorithms. This resulted in a total of 4 (noise types) x 9 (1 unprocessed + 8 algorithms) x 2 (SNRs) = 72 conditions. N=40 subjects participated in the listening test. This is the only dataset in this study that provides sentence instead of word recognition scores.

H. Reverberation datasets

In [59], HINT [60] and IEEE sentences were convolved with room impulse responses simulated using the image method. The sentences were processed subject to four reverberation times: T60 = 0.9, 1.2, 1.5, and 2.1 seconds, and three direct-to-reverberant ratios (DRR): 0, -10, -20 dB. This resulted in a total of 4 (T60) x 3 (DRR) = 12 reverberation conditions. The stimuli were presented to N=15 normal-hearing subjects. This is the only dataset explored here that did not use noise as a type of speech degradation.

I. NELE-Taal

In [61], sentences from the Dutch Matrix-test [57] were processed by two near-end listening enhancement (NELE) algorithms based on linear [62] and non-linear [61] approximations of the SII. The signals were then degraded by unmodulated SSN and babble noise at three different SNRs: [-20, -17, -14] dB and [-11 -8 -5] dB for the processed and unprocessed speech, respectively. The SNRs were chosen to produce roughly similar intelligibility for the processed and unprocessed conditions. This resulted in a total of 2 (noise types) x 3 (1 unprocessed + 2 algorithms) x 3 (SNRs) = 18 conditions. N=16 subjects participated in the listening test. This dataset was used as the validation set in optimizing the parameters of wSTMI.

J. NELE-Cooke

In [63], IEEE sentences were processed by 19 NELE algorithms and degraded by two types of noise. Unmodulated SSN and competing speaker (CS) noise were added to the processed speech at [1, -4, -9] dB and [-7, -14, -21] dB SNR, respectively. In [24], a subset of the dataset comprising 10 of the IEEE sentences for each degradation condition and nine of the NELE algorithms was used for SIP assessment. This resulted in a total of 2 (noise types) x 10 (1 unprocessed + 9 processed) x 3 (SNRs) = 60 conditions for our assessment.

K. NELE-Chermaz

In [64], a recording of the IEEE sentences [65] were processed by three NELE algorithms: SSDRC [66], AdaptDRC [67], [68], and AdaptDRC + OE [69]. The algorithms were chosen based on their performances in the Hurricane challenge [63]. The processed sentences were degraded in two realistic acoustic environments. The first environment was a small space with a short reverberation time (T60 = 300 ms) and CS noise. The second environment was a wide space with a long reverberation time (T60 = 1250 ms) and unmodulated SSN. Additive noises were presented at [-12.6, -7.6, -2.6] and [-1.6, 1.8, 5.6] dB SNRs for the small and large environments, respectively. This resulted in a total of 2 (environments) x 4 (1 unprocessed + 3 NELE algorithms) x 3 (SNRs) = 24 degradation conditions. Stimuli were presented binaurally to N=34 listeners. In our work, speech signals associated with the left channel (the better ear) were used for monaural SIP.

L. Int-Miller

In contrast to the previous datasets of degraded continuous speech, this dataset [70] presented temporally interrupted segments of clean speech alternating with noise. Interruption intervals for IEEE sentences were defined by calculating the running SNR between the target speech and a single-talker speech-modulated SSN at an average SNR of 0 dB. Two methods of interruption were employed based on the running SNR: positive (high-intensity speech) and negative (low-intensity speech) local SNR intervals which defined temporal intervals of speech either exceeding, or exceeded by, the competing noise level, respectively. The detected intervals were deleted and filled with one of the following noise types: (1) unmodulated SSN, or speech-modulated SSN based on (2) the missing portion of the speech, (3) the preceding speech, (4) a random segment of a different sentence, and (5) a time-compressed version of (4). The noise replacements were presented at two levels: according to the level of the replaced segment or according to the overall level of the sentence. This resulted in a total of 2 (interruption methods) x 5 (noise types) x 2 (noise levels) = 20 conditions. Stimuli were presented to 11 normal-hearing listeners.

VI. Results Analysis

We evaluate the performance of the proposed algorithm under several degradation conditions. We emphasize that only normal-hearing listeners were considered in this study. Figures of merit are introduced in Section VI-A. A comparison to other SIP algorithms is presented in Section VI-B.

TABLE III: SIP methods for comparison. SIP methods marked with (*) require the speech and additive noise realizations to be available separately.

Method	Description
wSTMI	Weighted Spectro-Temporal Modulation Index.
STMI	Spectro-Temporal Modulation Index [29], [30].
OSTMI	Modified Spectro-Temporal Modulation Index [23].
STOI	Short-Time Objective Intelligibility [21].
eSTOI	Extended Short-Time Objective Intelligibility [17].
HASPI	Hearing-Aid Speech Perception Index [28].
SIIB	Speech Intelligibility In Bits [25].
CSII-high	The high-level Coherence Speech Intelligibility Index [10].
CSII-I3	A linear combination of high/mid/low-level CSII [10].
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio [71].
TFSS	The Temporal Fine-Structure Spectrum index [20].
ESII*	Implementation of Extended SII [9].
Glimpse*	Implementation of Cooke's glimpse method [15].

A. Figures of Merit

SIP algorithms are conventionally evaluated using the Pearson and Spearman correlation coefficients or variants. The reason is that the subjective intelligibility score scale generally depends on listening test parameters not known to the algorithm, such as the corpus, subject response protocol, and scoring method, e.g., word scoring versus sentence scoring. Thus, a SIP algorithm can be made more widely applicable to different listening test protocols and scoring scales by requiring the algorithm's output ("objective" scores) to track the *trend* of the subjective scores instead of the latter's absolute values. To reconcile the objective score and the subjective score scales, a common practice is to use a monotonic mapping with few parameters, such as a sigmoid or a third-degree polynomial, to map the objective scores to the subjective scale. In [21], a logistic mapping is used to map model outputs to subjective score estimates:

$$\bar{I} = \frac{1}{1 + \exp(c\tilde{I} + d)} \tag{16}$$

where $\tilde{I}$ is the SIP algorithm output, and $c < 0$ and d are constants whose values are chosen with least square fitting. The parameters of the logistic mapping are calculated separately for each dataset and each algorithm.

We use five figures of merit to evaluate the performance of SIP algorithms: i) the Pearson correlation coefficient between the subjective I and objective $\tilde{I}$ intelligibility scores, ii) the Pearson correlation coefficient between the subjective I and logistic mapped $\bar{I}$ intelligibility scores, iii) the Spearman rank correlation coefficient, iv) the RMSE between I and $\tilde{I}$, and v) the RMSE between intelligibility score I and the logistic mapped predicted intelligibility $\bar{I}$.

B. Performance Evaluation

We compare the performance of wSTMI to the other SIP algorithms listed in Table III. Included are the best-performing SIP algorithms known to-date. Different from other algorithms, SIIB needs to input speech samples with at least 20 s length. To meet this requirement, SIIB-1 uses sample repetition, and SIIB-2 entails concatenation of several distinct short samples.

Figure 6 shows scatter plots for each dataset and wSTMI's output before applying a logistic mapping. Each point in the plot represents a different condition in the dataset. Tables IV and V show the Pearson correlation before and after applying the logistic mapping in Equation 16, respectively. To determine statistically significant differences between correlation values (Tables IV and V), pairwise comparisons using the Williams's test [72] were performed for each dataset between the best performing SIP method and the others. Methods which did not perform significantly worse than the best performing algorithm ($p < 0.05$) are marked with (*) in Table V. Table VI displays the Spearman rank correlation. Tables VII and VIII show the RMSE before and after applying the logistic mapping in Equation 16, respectively.

Not surprisingly, wSTMI shows excellent performance for the training and validation datasets, for which the model parameters W and b were optimized (ITFS-Kjems and NELE-Taal). However, importantly, wSTMI also performs well for datasets not used in the training phase. Also interesting to note that wSTMI performs well with strongly modulated noise maskers, despite the fact that such signals were not present in the training dataset. The top performance of wSTMI is closely followed by eSTOI, SIIB-2, and OSTMI. Focusing on the modulated noise conditions, i.e., ModN-Jensen, ModN-Fogerty, ModN-Gibbs, and ModFN-Fogerty, it is clear that wSTMI outperforms existing SIP methods.

STMI, OSTMI, and wSTMI all compare the spectro-temporal modulation envelopes of the clean and degraded speech signals to estimate the intelligibility of the latter. STMI integrates the filtered spectrograms across the acoustic frequency axis in order to capture the temporal fluctuation of energy in each modulation channel, while OSTMI and wSTMI avoid this integration by making individual comparisons between acoustic frequency bins within each modulation channel. Moreover, while STMI assigns equal weights to different spectro-temporal modulation frequencies, OSTMI uses a heuristic feature-selection scheme which assigns larger weights to intermediate spectral modulation frequencies. wSTMI further improves the performance of OSTMI by optimizing the weights assigned to different spectro-temporal modulation channels to maximize the SIP performance across a training dataset. The top SIP performance of wSTMI signifies the importance of taking into account different contributions of distinct spectro-temporal modulation channels to speech intelligibility.

Recall that STOI, eSTOI, OSTMI, and wSTMI estimate the intelligibility by comparing the modulation envelopes of the clean and degraded input signals. While STOI compares the temporal modulation envelopes of the clean and degraded signals, the other three algorithms use the joint spectro-temporal modulation envelopes. It is interesting to note that eSTOI, OSTMI, and wSTMI outperform STOI in the presence of modulated noise (ModN-Jensen, ModN-Fogerty, ModN-Gibbs, NELE-Cooke, and NELE-Chermaz). This suggests the significance of spectral modulation analysis in modulated

noise.

TFSS uses a Hilbert-derived temporal fine structure (TFS) waveform for SIP. The TFS waveform in [20] is the phase-modulated carrier of the Hilbert envelope. This contrasts with SIP methods that employ primarily envelope information, such as STOI, eSTOI, and wSTMI. Though TFSS was tested only on the NR-Hu dataset in [20], TFSS performs quite well for a wide range of distortions, including modulated noise and reverberation. The decent performance of TFSS for modulated noise conditions may be explained by the significant contribution of TFS information to speech perception in the presence of modulated noise [18], [73], [74].

We note that HASPI is computed as a linear combination of cepstral correlation and auditory coherence terms (low, mid, and high-level coherence). Here, we used a fixed set of parameters proposed in [28] to evaluate the performance of HASPI. In [24], the parameters of HASPI were optimized for each dataset to maximize performance. By comparing the performance of HASPI with and without parameter tuning, a few observations can be made. Even though HASPI achieves top SIP performance in many degradation conditions in both setups, it performs less well for NR-Hu and ModN-Jensen when used without parameter tuning. Moreover, when used without adaptation, HASPI performs poorly for NELE-Taal while delivering high performance for similar distortions in NELE-Cooke and NELE-Chermaz. These observations indicate that a careful tuning of HASPI's parameters is crucial for achieving robust SIP performance.

Many SIP algorithms performed poorly on ModN-Fogerty. Recall that ModN-Fogerty consists of speech degraded by time-compressed/expanded speech shaped modulated noise presented at a fixed SNR. Therefore, the time-compression/expansion rate of noise is the only factor governing the intelligibility in this dataset. None of the SIP methods investigated in this study were optimized for such distortions. STOI and SI-SDR show negative correlation with intelligibility across this dataset. In general, we desire a SIP algorithm whose outputs show strong positive correlation across many types of speech distortions.

Focusing on the interrupted speech conditions, i.e., Int-Miller dataset, it is evident that wSTMI, CSII-high, SI-SDR, and HASPI outperform other SIP methods. Interestingly, OSTMI, SIIB, CSII-I3, and TFSS are negatively correlated with intelligibility for this dataset. Figure 7 shows the scatter plots for each algorithm and the Int-Miller dataset before applying a logistic mapping. Recall that Int-Miller consisted of temporally interrupted speech. Also, the +SNR conditions (high-intensity speech) were more intelligible than their -SNR (low-intensity speech) counterparts. As the scatter plots in Figure 7 imply, many SIP algorithms predicted lower intelligibility for the +SNR conditions, resulting in a negative correlation. In addition, even though CSII-high and SI-SDR achieved high overall correlation for this dataset, they might not be suitable for this type of degradation. The scatter plots show that the predictions within each cluster are flat or negatively correlated with intelligibility for CSII-high and SI-SDR.

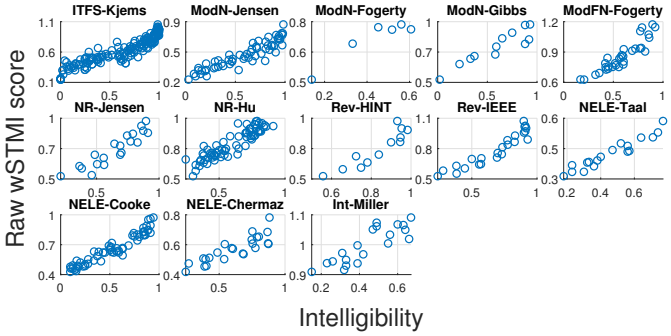

Fig. 6: Scatter plots of listening test intelligibility scores against wSTMI's output, before applying a logistic mapping. The horizontal axis shows the intelligibility and the vertical axis shows wSTMI's output.

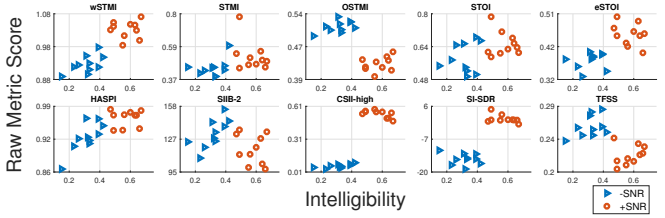

Fig. 7: Scatter plots of listening test scores for the Int-Miller dataset versus different algorithms' outputs, before applying a logistic mapping. The horizontal axis shows the intelligibility and the vertical axis shows the predicted intelligibility.

VII. Discussion

In Section VI, we pointed out that the accuracy of HASPI predictions depends on the degree of parameter tuning. In general, the accuracy of a SIP method is tied to the data used to develop the algorithm [24]. For example, a SIP algorithm designed with additive noise in mind might not work well on speech processed by non-linear noise reduction algorithms. In order to verify that the proposed algorithm is not narrowly tuned to a specific noise/processing condition- but works broadly, we evaluated it using 11 datasets that were not used for algorithm parameter tuning.

As described by Eq. (11), the sum in the NCC is computed over the entire utterance. This is different from the short-time NCC used by other SIP methods such as (e)STOI [17], [21], which computes the NCC across short signal segments, e.g., of duration 384 ms [17], [21], and then averages the results. STOI's approach can be regarded as a time-varying normalization of the input envelopes, while wSTMI normalizes the entire envelopes at once. While the non-segmental approach of wSTMI has shown good SIP performance over a broad range of degradations and datasets, it might fail in certain acoustic conditions. For example, consider a sentence with five unrelated words degraded by a burst of noise that aligns with one of the words in the sentence. As the noise level increases, the intelligibility will drop to 80%, because only one of the words is affected by the noise. However, as calculated in Eq. (11), the NCC will be dominated by the high-energy portion of the noise and hence possibly decrease the predicted

TABLE IV: Performance of Different SIP Algorithms in Terms of Pearson Correlation Coefficient *Before* Applying a Logistic Mapping. SIP algorithms that did not perform significantly worse than the best performing algorithm ($p < 0.05$) are marked with (*).

	ITFS-Kjems	ModN-Jensen	ModN-Fogerty	ModN-Gibbs	ModFN-Fogerty	NR-Jensen	NR-Hu	Rev-HINT	Rev-IEEE	NELE-Taal	NELE-Cooke	NELE-Chermaz	Int-Miller	Mean
wSTMI	0.93*	**0.92**	**0.93**	**0.92**	**0.91**	0.94	0.92*	**0.91**	0.94	0.96*	**0.96**	0.86*	**0.86**	**0.92**
STMI	0.76	0.50	0.54	0.79	0.76	0.25	0.64	0.89*	0.90	0.36	0.76	0.70	0.35	0.62
OSTMI	0.86	0.88	0.89*	0.87	0.89*	0.89	0.89*	0.84	0.86	0.89	0.82	0.71	-0.76	0.71
STOI	**0.94**	0.45	-0.77	0.55	0.35	**0.98**	0.86	0.88*	0.89	0.89	0.70	0.38	0.44	0.61
eSTOI	0.93*	0.85	0.76	0.87	0.88	0.97*	0.90*	0.88*	0.89	0.88	0.92	0.72	0.68	0.86
HASPI	0.88	0.61	0.61	0.80	0.83	0.92	0.68	0.84	**0.97**	0.44	0.86	**0.87**	0.76*	0.79
SIIB-1	0.84	0.68	0.03	0.67	0.82	0.95*	**0.93**	0.78	0.82	**0.97**	0.87	0.63	-0.61	0.64
SIIB-2	0.81	0.83	0.78	0.80	0.82	0.94	0.92*	0.78	0.92	**0.97**	0.90	0.62	-0.31	0.75
CSII-high	0.50	0.60	0.19	0.74	0.67	0.96*	0.87	0.85	0.85	0.73	0.70	0.16	**0.86**	0.67
CSII-I3	0.65	0.50	0.00	0.78	0.67	0.86	0.92*	0.86	0.90	0.92	0.75	0.44	-0.19	0.62
SI-SDR	0.43	0.53	-0.80	0.76	0.56	0.95*	0.42	0.62	0.72	0.31	0.42	0.15	0.81*	0.45
TFSS	0.47	0.82	0.85*	0.79	0.77	0.92	0.91*	0.77	0.80	0.96*	0.64	0.29	-0.62	0.64
ESII	—	0.82	0.82*	0.81	—	—	—	—	—	—	—	—	—	—
Glimpse	—	0.85	0.83*	0.84	—	—	—	—	—	—	—	—	—	—

TABLE V: Performance of Different SIP Algorithms in Terms of Pearson Correlation Coefficient *After* Applying a Logistic Mapping. SIP algorithms that did not perform significantly worse than the best performing algorithm ($p < 0.05$) are marked with (*).

	ITFS-Kjems	ModN-Jensen	ModN-Fogerty	ModN-Gibbs	ModFN-Fogerty	NR-Jensen	NR-Hu	Rev-HINT	Rev-IEEE	NELE-Taal	NELE-Cooke	NELE-Chermaz	Int-Miller	Mean
wSTMI	0.95	**0.94**	**0.94**	**0.94**	**0.93**	0.95	0.92*	**0.96**	**0.97**	**0.97**	**0.97**	**0.90**	**0.86**	**0.93**
STMI	0.78	0.54	0.54	0.81	0.76	0.25	0.64	**0.96**	0.92	0.36	0.76	0.70	0.35	0.63
OSTMI	0.89	0.89	0.90*	0.89	0.92*	0.90	0.90	0.95*	0.90	0.89	0.85	0.73	-0.76	0.74
STOI	**0.96**	0.45	-0.77	0.56	0.38	0.98*	0.86	**0.96**	0.94*	0.90	0.71	0.39	0.44	0.62
eSTOI	0.95	0.92*	0.75	0.90	0.91*	**0.99**	0.90	**0.96**	0.94*	0.90	0.95*	0.73	0.69	0.88
HASPI	0.91	0.75	0.61	0.80	0.86	0.93	0.68	0.94*	**0.97**	0.44	0.91	**0.90**	0.76*	0.82
SIIB-1	0.87	0.83	0.03	0.71	0.82	0.98*	**0.93**	**0.96**	0.93	**0.97**	0.89	0.66	-0.61	0.69
SIIB-2	0.89	0.89	0.79	0.85	0.88	0.98	0.92	0.96	0.93	0.97	0.94	0.65	-0.31	0.79
CSII-high	0.51	0.64	0.19	0.75	0.67	0.97*	0.87	0.95*	0.90	0.73	0.70	0.16	**0.86**	0.68
CSII-I3	0.70	0.59	0.01	0.79	0.78	0.87	**0.93**	**0.96**	0.91	0.92	0.77	0.44	-0.19	0.65
SI-SDR	0.42	0.51	-0.80	0.76	0.57	0.96	0.41	0.61	0.74	0.31	0.41	0.15	0.81*	0.42
TFSS	0.51	0.90	0.85*	0.83	0.78	0.95	0.91	0.93	0.87	0.96*	0.67	0.28	-0.62	0.68
ESII	—	0.85	0.81	0.82	—	—	—	—	—	—	—	—	—	—
Glimpse	—	0.87	0.83	0.86	—	—	—	—	—	—	—	—	—	—

TABLE VI: Performance of Different SIP Algorithms in Terms of Spearman Rank Correlation Coefficient.

	ITFS-Kjems	ModN-Jensen	ModN-Fogerty	ModN-Gibbs	ModFN-Fogerty	NR-Jensen	NR-Hu	Rev-HINT	Rev-IEEE	NELE-Taal	NELE-Cooke	NELE-Chermaz	Int-Miller	Mean
wSTMI	**0.96**	**0.94**	**0.71**	**0.91**	**0.93**	0.96	0.90	0.90	0.95	0.97	**0.95**	0.89	**0.82**	**0.90**
STMI	0.78	0.51	0.03	0.76	0.72	0.30	0.65	0.90	0.94	0.24	0.76	0.70	0.59	0.60
OSTMI	0.89	0.91	0.49	0.87	**0.93**	0.90	0.84	**0.97**	0.91	0.85	0.85	0.69	-0.68	0.72
STOI	**0.96**	0.48	-0.94	0.54	0.67	0.96	0.80	**0.97**	0.94	0.89	0.72	0.31	0.41	0.59
eSTOI	**0.96**	0.92	0.26	0.82	0.91	**0.98**	0.87	0.96	0.94	0.89	**0.95**	0.68	0.68	0.82
HASPI	0.92	0.76	0.26	0.76	0.82	0.95	0.64	0.88	0.94	0.39	0.91	**0.91**	0.72	0.76
SIIB-1	0.84	0.82	-0.03	0.73	0.81	**0.98**	**0.92**	0.90	0.92	0.97	0.88	0.64	-0.64	0.67
SIIB-2	0.85	0.84	0.26	0.80	0.87	**0.98**	**0.92**	0.93	0.90	**0.98**	0.89	0.66	-0.26	0.74
CSII-high	0.53	0.69	0.26	0.71	0.66	0.95	0.80	0.94	0.92	0.72	0.70	0.33	0.74	0.69
CSII-I3	0.70	0.56	0.03	0.80	0.66	0.88	0.89	0.90	**0.97**	0.90	0.77	0.45	-0.10	0.65
SI-SDR	0.44	0.56	-0.94	0.77	0.51	0.96	0.70	0.51	0.80	0.32	0.43	0.19	0.71	0.46
TFSS	0.58	0.90	**0.71**	0.81	0.78	0.96	0.87	0.90	0.89	0.96	0.65	0.32	-0.65	0.67
ESII	—	0.84	0.37	0.78	—	—	—	—	—	—	—	—	—	—
Glimpse	—	0.87	0.43	0.77	—	—	—	—	—	—	—	—	—	—

intelligibility to much below 80%. Calculating the NCC sum over short segments could be beneficial in such scenarios, because the segmental approach limits the time extent of the influence of the noise burst. Future work will need to consider extending wSTMI to using short-time NCC for SIP. However, it is encouraging in this first investigation of the wSTMI as

TABLE VII: Performance of Different SIP Algorithms in Terms of RMSE Before Applying a Logistic Mapping. SIIB estimates the amount of information shared between a talker and a listener in bits per second. SI-SDR calculates the scale-invariant speech to distortion ratio in dB. A non-linear mapping is necessary to map the output of these two algorithms to intelligibility predictions. Therefore, RMSE can not be calculated for these algorithms before applying a mapping.

	ITFS-Kjems	ModN-Jensen	ModN-Fogerty	ModN-Gibbs	ModFN-Fogerty	NR-Jensen	NR-Hu	Rev-HINT	Rev-IEEE	NELE-Taal	NELE-Cooke	NELE-Chermaz	Int-Miller	Mean
wSTMI	**0.15**	**0.30**	**0.13**	**0.17**	**0.11**	0.13	0.10	**0.25**	0.09	**0.11**	**0.14**	0.26	0.55	**0.20**
STMI	0.29	0.51	0.22	0.34	0.24	0.30	0.18	0.34	0.23	0.35	0.34	0.23	0.15	0.29
OSTMI	0.55	0.60	0.31	0.47	0.40	0.52	0.41	0.67	0.55	0.42	0.45	0.58	0.19	0.47
STOI	0.23	0.36	0.21	0.25	0.17	**0.12**	0.15	0.27	0.14	0.14	0.21	**0.20**	0.22	**0.20**
eSTOI	0.37	0.53	0.20	0.32	0.29	0.37	0.17	0.50	0.38	0.36	0.35	0.46	**0.13**	0.34
HASPI	0.25	0.59	0.24	0.27	0.22	0.18	0.18	0.33	**0.06**	0.44	0.30	0.35	0.52	0.30
CSII-high	0.50	0.54	0.16	0.24	0.19	0.16	**0.09**	0.43	0.29	0.29	0.30	0.50	0.21	0.30
CSII-I3	0.47	0.59	0.34	0.37	0.30	0.34	**0.09**	0.49	0.36	0.41	0.46	0.57	0.22	0.39
TFSS	0.68	0.64	0.40	0.57	0.48	0.63	0.51	0.76	0.63	0.46	0.56	0.64	0.26	0.56
ESII	—	0.53	0.28	0.43	—	—	—	—	—	—	—	—	—	—
Glimpse	—	0.43	0.15	0.29	—	—	—	—	—	—	—	—	—	—

TABLE VIII: Performance of Different SIP Algorithms in Terms of RMSE After Applying a Logistic Mapping.

	ITFS-Kjems	ModN-Jensen	ModN-Fogerty	ModN-Gibbs	ModFN-Fogerty	NR-Jensen	NR-Hu	Rev-HINT	Rev-IEEE	NELE-Taal	NELE-Cooke	NELE-Chermaz	Int-Miller	Mean
wSTMI	0.11	**0.11**	**0.05**	**0.10**	**0.07**	0.07	0.07	0.04	0.09	0.05	**0.07**	0.09	**0.08**	**0.07**
STMI	0.21	0.24	0.14	0.17	0.12	0.19	0.13	0.04	0.14	0.17	0.18	0.14	0.14	0.15
OSTMI	0.15	0.14	0.06	0.14	0.10	0.13	0.09	0.04	0.09	0.08	0.15	0.14	0.15	0.11
STOI	0.09	0.26	0.16	0.24	0.14	0.04	0.09	**0.03**	0.07	0.08	0.20	0.18	0.14	0.13
eSTOI	**0.10**	0.12	0.10	0.12	0.08	**0.03**	0.07	0.04	0.07	0.08	0.10	0.14	0.11	0.09
HASPI	0.14	0.20	0.11	0.16	0.10	0.07	0.13	0.04	**0.05**	0.16	0.12	**0.06**	0.10	0.11
SIIB-1	0.16	0.17	0.16	0.21	0.11	0.04	**0.06**	0.04	0.09	**0.04**	0.12	0.15	0.15	0.11
SIIB-2	0.14	0.15	0.10	0.15	0.09	0.04	**0.06**	0.04	0.09	**0.04**	0.11	0.15	0.15	0.10
CSII-high	0.28	0.22	0.16	0.19	0.14	0.05	0.07	0.04	0.10	0.12	0.20	0.20	**0.08**	0.14
CSII-I3	0.25	0.23	0.16	0.18	0.12	0.10	0.07	0.04	0.09	0.07	0.18	0.18	0.15	0.14
SI-SDR	0.30	0.25	0.16	0.19	0.16	0.05	0.16	0.10	0.07	0.17	0.25	0.43	0.09	0.19
TFSS	0.28	0.13	0.08	0.16	0.12	0.06	0.07	0.05	0.12	0.05	0.21	0.19	0.15	0.13
ESII	—	0.15	0.09	0.17	—	—	—	—	—	—	—	—	—	—
Glimpse	—	0.14	0.09	0.15	—	—	—	—	—	—	—	—	—	—

implemented over the entire signal, that it performs similarly or better than the other SIP methods, including those using segmental analysis. As mentioned, the comparison of segmental and non-segmental methods, particularly in transient noise scenarios, will need to be considered.

In Section IV-B, we employed two datasets (ITFS-Kjems and NELE-Taal) to optimize the linear combination of intermediate intelligibility measures for SIP. Here, we present some insights into the optimized set of parameters. In order to interpret the optimized weights W, first, we seek to determine whether the same spectro-temporal modulation channels will be selected (assigned non-zeros weights) if other datasets were used for training. Figure 8 shows the set of weights optimized for different training datasets as a function of the regularization parameter λ. We see that as λ increases, W tends to a sparse selection and almost the same cluster of spectro-temporal modulation channels are selected, irrespective of the training stimuli. Notably, increasing λ prunes off the "redundant" channels for the linear model. Increasing sparsification tends to assign larger weights to the spectral modulation frequencies between 0.72 and 4.54 cyc/kMel and temporal modulation frequencies below 10 Hz, essentially without affecting SIP performance. The importance of spectro-temporal modulation frequencies, as determined here, is tied to the type of degradations that were used to tune the parameters of the algorithm. Hence, tuning the model parameters using degradation conditions that were not considered in this study might lead to different results. For instance, Steinmetzger et al. [41] demonstrated that measuring temporal modulation frequencies in the human-pitch frequency range (roughly 60-400 Hz) can improve SIP accuracy when a masker is periodic and/or slowly amplitude modulated. In any case, the selected modulation channels in Figure 5 appear to work well over a wide range of degradations. The following subsections present some insights into the optimized set of parameters.

A. Comparison to MTFs of Human Auditory System

The optimization in Section IV-B assigns larger weights to the modulation channels that are important for SIP. As MTFs represent the relative importance of the spectro-temporal modulation frequencies for human auditory perception, we expect to observe similarities between the MTFs of the human auditory system and the optimized weights W.

In [29] and [30], spectro-temporal MTFs were estimated by measuring the detection thresholds of different modulation frequencies for the human auditory system. The MTFs in [30]

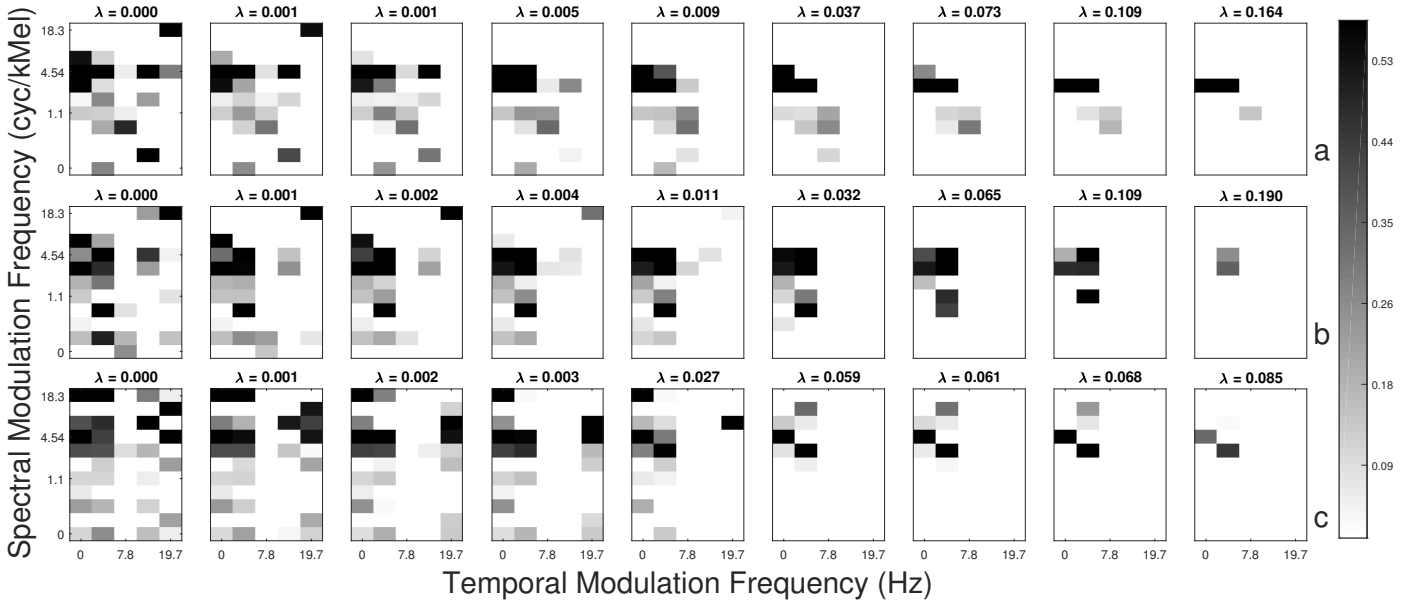

Fig. 8: Optimized weights W as a function of regularization parameter λ using different training datasets (a) ITFS-Kjems (b) ModN-Jensen (c) NELE-Taal.

show a lowpass behavior in both the spectral and temporal modulation directions with cut-off frequencies of about 2 cyc/octave and 16 Hz, respectively. Moreover, the MTF is relatively flat in the bandpass region. In [34], MTFs were measured by evaluating the relative importance of the modulation frequencies for speech perception (in contrast to discrimination thresholds in [30]). A modulation filtering approach was used to restrict the modulation patterns in the speech signal and a joint spectro-temporal MTF was derived. In [40], the envelope modulation fidelity was calculated using the normalized cross-covariance of the degraded and reference signal envelopes. Even though Kates et al. used temporal modulation frequencies up to 325 Hz, the results indicated that the low temporal modulation frequencies provided the highest amount of information for speech intelligibility. The findings in [29], [30], [34], [40] are consistent with the optimized W depicted in Figure 5, but with one key difference. While the cut-off spectral and temporal modulation frequencies agree, Elliot et al. [34] found that the lowest spectral modulation frequencies (near DC) are important for speech perception. In contrast, our optimization method tends to favor the intermediate spectral modulation frequencies.

From Figure 8 it follows that for larger values of λ (larger sparsity), the optimization scheme favours a lowpass behaviour with respect to the temporal modulations. Specifically, temporal modulation frequencies below 10 Hz tend to be selected. This is in slight contrast to previous studies of the relative importance of modulation frequencies for speech perception. For example, Drullman et al. [13] examined the effect of varying the amount of temporal modulation frequencies available to human subjects for speech recognition and found that they needed temporal modulation frequencies up to 16 Hz. The weights selected for wSTMI are an outcome of predicting speech intelligibility using a referenced-based STMA framework. One may argue that the task of SIP is different than that of speech recognition, and hence, one might not expect the important spectro-temporal modulations for the two tasks to be identical.

B. Comparison to Automatic Speech Recognition

In (Schadler et al. [35] and [36]), the relative importance of different spectro-temporal modulation frequencies for automatic speech recognition (ASR) was determined using the STMA front-end (Section II-A) and hidden Markov models with Gaussian mixture emissions as the back-end. Schadler et al. evaluated the performance sensitivity of their ASR algorithm, when features from a single spectro-temporal modulation channel were dropped. The results showed that the intermediate spectral modulation frequencies (between 1.34 and 2.8 cyc/kMel) and low temporal modulation frequencies were more important for ASR. This is consistent with our findings and implies that the relatively important spectral modulation frequencies exhibit a bandpass behavior.

C. Selected Spectral Modulation Frequencies

We present a filtering example to help further understanding the role of spectral modulation analysis in SIP. Figure 9 shows a series of spectral modulation filtered speech spectrograms for the utterance "should we chase?" spoken by a male (no temporal modulation filtering is employed). We can see that the filtered spectrograms with intermediate spectral modulation frequencies between 0.72 and 4.54 cyc/kMel best preserve the clarity of the formants and their transition patterns. A similar observation was made for several other speech signals. The importance of formant transitions in the perception of natural speech has been shown in several studies, including [75]–[77]. Thus, based on the above observation, we argue that

the intermediate spectral modulation frequencies are crucial for SIP.

VIII. Conclusion

We have presented a monaural intrusive speech intelligibility prediction algorithm based on spectro-temporal modulation analysis of the input speech samples. The proposed algorithm, which we call wSTMI, combines intermediate intelligibility measures from different modulation channels using a sparse linear model and extends the concept of frequency-band importance function to spectro-temporal modulation channels. The linear model parameters were optimized using a Lasso regression approach. We showed that the optimized parameters can be interpreted in terms of MTFs and are consistent with other findings of the human auditory system. We evaluated the performance of wSTMI and other state-of-the-art SIP algorithms across several datasets and distortions. Compared to other SIP methods, wSTMI performs well across all the investigated acoustic conditions. Notably, wSTMI outperforms other SIP methods in the presence of highly non-stationary distortions, e.g., single and multi-speaker speech modulated noise. The proposed SIP approach also provides a systematic way for performance-improvement with hitherto untested acoustic conditions.

Acknowledgment

This work was partly (A.E. & W.-Y.C.) supported by the Natural Sciences and Engineering Research Council of Canada, the Demant Foundation, and the Vector Institute. A portion of this work (D.F.) was also supported by the National Institutes of Health, National Institute on Deafness and Other Communication Disorders, Grant No. R01-DC015465 (D.F.). The authors would like to thank the following researchers for providing intelligibility data and implementations of their intelligibility metrics: Fei Chen, Carol Chermaz, James Kates, Cees Taal, and Steven Van Kuyk. Finally, the authors would like to thank the three anonymous reviewers for their insightful comments.

References

[1] *Sound system equipment — Part 16: Objective rating of speech intelligibility by speech transmission index*, 60268–16-2003, International Electrotechnical Commission, Geneva, 2003.

[2] *American National Standard: Methods for the Calculation and Use of the Articulation Index*, S3.5-1969 (R1986), American National Standards Institute, New York, 1969.

[3] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[4] *Methods for calculation of the speech intelligibility index*, ANSI S3.5, American National Standard Institute, New York, 1997.

[5] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[6] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *J. Acoust. Soc. Amer.*, vol. 66, no. 5, pp. 1364–1380, 1979.

[7] S. P. Bacon and N. F. Viemeister, "Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners," *Int. J. Audiology*, vol. 24, no. 2, pp. 117–134, 1985.

[8] C. Ludvigsen, C. Elberling, G. Keidser, and T. Poulsen, "Prediction of intelligibility of non-linearly processed speech," *Acta Oto-Laryngologica, Supplement*, vol. 109, no. 469, pp. 190–195, 1990.

[9] 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.

[10] J. Kates and K. Arehart, "Coherence and the Speech Intelligibility Index," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 2604–2604, 2004.

[11] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[12] V. Hohmann and B. Kollmeier, "The effect of multichannel dynamic compression on speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 97, no. 2, pp. 1191–1195, 1995.

[13] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[14] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 585–592, 1995.

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

[16] 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.

[17] 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.

[18] K. Hopkins and B. C. J. Moore, "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *J. Acoust. Soc. Amer.*, vol. 125, no. 1, pp. 442–446, 2009.

[19] S. Shamma and C. Lorenzi, "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2818–2833, 2013.

[20] F. Chen, L. L. Wong, and Y. Hu, "A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech," *Speech Commun.*, vol. 55, no. 10, pp. 1011–1020, 2013.

[21] 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.

[22] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[23] 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.

[24] 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.

[25] ——, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, 2018.

[26] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, 2014.

[27] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, 2014.

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

[29] 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.

[30] 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.

[31] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Amer.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[32] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 134, no. 1, pp. 436–446, 2013.

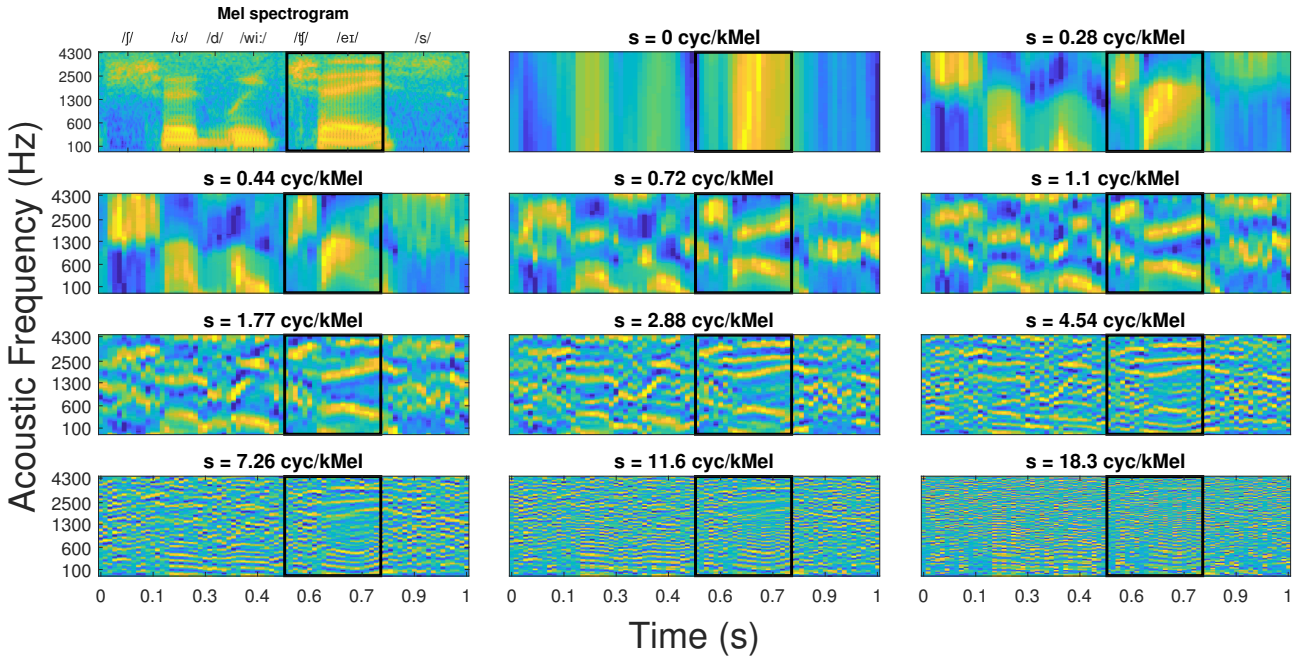

Fig. 9: An illustration of spectral modulation filtered spectrograms for different spectral modulation frequencies. The Mel spectrogram in the top left panel is filtered with the spectral modulation filter-bank $G_S[f; s_i]$ presented in Section II-A. No temporal modulation filtering is performed. The formant trajectories for the word "**chase**" are highlighted. The first four formants of the vowel change from 380, 1820, 2580, and 3490 Hz to 330, 2140, 2880, and 3720 Hz.

[33] 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.

[34] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, 2009.

[35] 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.

[36] 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.

[37] *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.

[38] S. Shamma, "Methods of neuronal modeling," *Spatial Temporal Process. Auditory Syst.*, pp. 411–460, 1998.

[39] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Speech Audio Process.*, vol. 2, no. 3, pp. 421–435, 1994.

[40] J. M. Kates and K. H. Arehart, "Comparing the information conveyed by envelope modulation for speech intelligibility, speech quality, and music quality," *J. Acoust. Soc. Amer.*, vol. 138, no. 4, pp. 2470–2482, 2015.

[41] K. Steinmetzger, J. Zaar, H. Relaño-Iborra, S. Rosen, and T. Dau, "Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations," *J. Acoust. Soc. Amer.*, vol. 146, no. 4, pp. 2562–2576, 2019.

[42] W. L. Gulick, G. A. Gescheider, and R. D. Frisina, *Hearing: Physiological acoustics, neural coding, and psychoacoustics.* Oxford University Press, 1989.

[43] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statistical Soc.: Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[44] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statistical Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[45] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[46] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[47] 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.

[48] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Int. J. Audiology*, vol. 40, no. 3, pp. 148–157, 2001.

[49] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[50] 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.

[51] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[52] 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.

[53] 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.

[54] 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.

[55] F. Apoux, N. Tribut, X. Debruille, and C. Lorenzi, "Identification of envelope-expanded sentences in normal-hearing and hearing-impaired listeners," *Hearing Res.*, vol. 189, no. 1-2, pp. 13–24, 2004.

[56] 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.

[57] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in noise test (matrix)," *Proc. 8th EFAS Congr., 10th DGA Congr.*, 2007.

[58] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian Audiology*, vol. 11, no. 2, pp. 79–87, 1982.

[59] 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.

[60] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[61] 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.

[62] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," *ITG-Fachtagung Sprachkommunikation*, vol. 8, no. 4, pp. 1919–1923, 2010.

[63] 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.

[64] 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.

[65] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Commun.*, vol. 55, no. 4, pp. 572–585, 2013.

[66] T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vol. 1, 2012, pp. 634–637.

[67] H. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 3577–3581.

[68] ——, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 138, no. 5, pp. 2692–2706, 2015.

[69] J. Grosse and S. Van De Par, "A speech preprocessing method based on overlap-masking reduction to increase intelligibility in reverberant environments," *J. Audio Eng. Soc.*, vol. 65, no. 1-2, pp. 31–41, 2017.

[70] R. E. Miller, B. E. Gibbs, and D. Fogerty, "Glimpsing speech interrupted by speech-modulated noise," *J. Acoust. Soc. Amer.*, vol. 143, no. 5, pp. 3058–3067, 2018.

[71] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2019, pp. 626–630.

[72] E. J. Williams, "The comparison of regression variables," *J. Roy. Statistical Soc.: Ser. B (Methodological)*, vol. 21, no. 2, pp. 396–399, 1959.

[73] S. Sheft, M. Ardoint, and C. Lorenzi, "Speech identification based on temporal fine structure cues," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 562–575, 2008.

[74] G. Gilbert and C. Lorenzi, "The ability of listeners to use recovered envelope cues from speech fine structure," *J. Acoust. Soc. Amer.*, vol. 119, no. 4, pp. 2438–2444, 2006.

[75] D. Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," *J. Acoust. Soc. Amer.*, vol. 72, no. 2, pp. 379–389, 1982.

[76] W. Strange, "Information for vowels in formant transitions," *J. Memory Lang.*, vol. 26, no. 5, pp. 550–557, 1987.

[77] K. R. Kluender, J. A. Coady, and M. Kiefte, "Sensitivity to change in perception of speech," *Speech Commun.*, vol. 41, no. 1, pp. 59–69, 2003.