# Non-intrusive codebook-based intelligibility prediction

Sørensen, Charlotte; Kavalekalam, Mathew Shaji; Xenaki, Angeliki; Boldt, Jesper Bünsow; Christensen, Mads Græsbøll

*Published in:*
Speech Communication

*DOI (link to publication from Publisher):*
10.1016/j.specom.2018.06.003



*Publication date:*
2018

*Document Version*
Accepted author manuscript, peer reviewed version


# Non-intrusive codebook-based intelligibility prediction

Charlotte Sørensen, Mathew Shaji Kavalekalam, Angeliki Xenaki, Jesper Bünsow Boldt, Mads Græsbøll Christensen

PII: S0167-6393(17)30362-X
DOI: 10.1016/j.specom.2018.06.003
Reference: SPECOM 2573

To appear in: *Speech Communication*

Received date: 25 September 2017
Revised date: 31 May 2018
Accepted date: 20 June 2018

Please cite this article as: Charlotte Sørensen, Mathew Shaji Kavalekalam, Angeliki Xenaki, Jesper Bünsow Boldt, Mads Græsbøll Christensen, Non-intrusive codebook-based intelligibility prediction, *Speech Communication* (2018), doi: 10.1016/j.specom.2018.06.003

# Non-intrusive codebook-based intelligibility prediction

Charlotte Sørensen[a,b,*], Mathew Shaji Kavalekalam[b], Angeliki Xenaki[a], Jesper Bünsow Boldt[a], Mads Græsbøll Christensen[b]

[a] *GN Hearing A/S, Denmark*
[b] *Audio Analysis Lab, CREATE, Aalborg University, Denmark*

**Abstract**

In recent years, there has been an increasing interest in objective measures of speech intelligibility in the speech processing community. Important progress has been made in intrusive measures of intelligibility, where the Short-Time Objective Intelligibility (STOI) method has become the de facto standard. Online adaptation of signal processing in, for example, hearing aids, in accordance with the listening conditions, requires a non-intrusive measure of intelligibility. Presently, however, no good non-intrusive measures exist for noisy, nonstationary conditions. In this paper, we propose a novel, non-intrusive method for intelligiblity prediction in noisy conditions. The proposed method is based on STOI, which measures long-term correlations in the clean and degraded speech. Here, we propose to estimate the clean speech using a codebook-based approach that jointly models the speech and noisy spectra, parametrized by auto-regressive parameters, using pre-trained codebooks of both speech and noise. In experiments, the proposed method is demonstrated to be capable of accurately predicting the intelligibility scores obtained with STOI from oracle information. Moreover, the results are validated in listening tests that confirm that the proposed method can estimate intelligibility from noisy speech over a range of signal-to-noise ratios.

*Keywords:* Hearing aids, non-intrusive, speech intelligibility prediction, STOI

## 1. Introduction

Human interaction depends on communication where speech has a central role. Inability to understand speech, e.g., due to hearing impairment, noisy background, or distortion in communication systems, can lead to ineffective communication and social isolation, and the development of speech enhancement methods[1, 2] is, therefore, a key concern in many applications. These include challenging applications such as hearing aids [3], telecommunication systems [4, 5], and architectural acoustics [6]. To assess the listening conditions in which speech processing would be beneficial, but also to evaluate the speech processing algorithms as such, a speech intelligibility measure is required [3, 5, 7].

A natural way of assessing the intelligibility of a degraded, i.e., processed, distorted or noisy speech signal is by performing subjective listening tests. Subjective speech intelligibility scores gives the percentage of correctly identified information from a degraded speech signal. However, subjective speech intelligibility experiments are time-consuming, expensive and cannot be used for real-time applications. Hence, there is a great interest in developing objective measures for speech intelligibility prediction. As opposed to subjective listening tests, objective intelligibility prediction algorithms are faster, cheaper and can be used for real-time processing.

The Articulation Index (AI) [8, 9] and the Speech Intelligibility Index (SII) [10] are some of the earliest metrics for prediction of speech intelligibility scores. The AI and SII use the signal-to-noise ratio (SNR) of speech excerpts in several frequency bands to estimate the intelligibility, hence they require that both the clean speech signal and the noise are available and uncorrelated as well as the noise to be stationary. The Extended SII (ESII) [11] and the Coherence SII (CSII) [12], are variants of SII which account for fluctuating

*Corresponding author.
*Email addresses:* csoerensen@gnresound.com, cs@create.aau.dk (Charlotte Sørensen), msk@create.aau.dk (Mathew Shaji Kavalekalam), axenaki@gnresound.com (Angeliki Xenaki), jboldt@gnresound.com (Jesper Bünsow Boldt), mgc@create.aau.dk (Mads Græsbøll Christensen)

[1]This work was supported by the Innovation Fund Denmark, Grant No. 99-2014-1.

noise and nonlinear distortions from clipping, respectively. The Speech Transmission Index (STI) [4] was introduced to predict the intelligibility of an amplitude modulated signal at the output of a transmission channel based on changes in the modulation depth across frequency of a probe signal. The STI, which requires a probe signal as reference, offers good prediction of speech intelligibility in reverberant and noisy conditions [4], but not for more adverse nonlinear distortions, such as those caused by spectral subtraction [13]. The Short-Time Objective Intelligibility (STOI) metric [14] predicts the intelligibility of a signal by its short-time correlation with its clean counterpart which is required as input. STOI estimates are accurate for time-frequency processed speech [15, 16]. The speech-based Envelope Power Spectrum Model (sEPSM) [17] estimates the SNR in the envelope-frequency domain and uses the noise signal alone as reference. The sEPSM accounts for the effects of additive noise and reverberation and some types of nonlinear processing such as spectral subtraction [17], but fails with other types of nonlinear processing such as ideal binary masks and phase jitter [16]. More recent work includes that of [18], which takes an information theoretical approach to the problem.

All the aforementioned methods are intrusive, i.e., they require either the clean speech signal or the noise interference as reference to estimate the intelligibility of the degraded signal. Access to the clean speech signal is impractical for many real-life applications or real-time processing systems. To overcome this limitation, a number of non-intrusive objective intelligibility measures have been proposed. The Speech to Reverberation Modulation energy Ratio (SRMR) [19] and the average Modulation-spectrum Area (ModA) [20] both provide intelligibility predictions based on the modulation spectrum of the degraded speech signal, i.e., in a non-intrusive manner. Other notable work includes the reduced dynamic range (rDR) based intelligibility measure [21], wherein the intelligibility is predicted directly from the dynamic range of the noisy speech, and the across-band envelope correlation (ABEC) metric [22], which is based on temporal envelope waveforms. Another approach to predict speech intelligibility non-intrusively is to first obtain an estimate of the clean speech signal which is thereafter used as reference to an intrusive method. Machine learning [23, 24], principal component analysis [7] or noise reduction [25, 26] methods have been proposed to reconstruct the clean signal from its degraded version and use it as input to the intrusive STOI metric for objective intelligibility prediction.

The present paper, which is an extension of our prior work [27], proposes a non-intrusive intelligibility metric, which uses the STOI measure non-intrusively by estimating the features of the clean reference signal from its degraded version. The proposed method, however, estimates the reference signal by identifying the entries of pre-trained codebooks of speech and noise spectra which best fit the data, i.e., the noisy speech signal. The resulting new metric is dubbed Non-Intrusive Codebook-based STOI (NIC-STOI). The method is inspired by the work [28, 29] which demonstrates that codebook-based approaches offer effective speech enhancement, even under nonstationary noise such as babble noise. Moreover, the approaches of [28, 29] are based on low-dimensional parametrizations of both the noise and speech spectra, more specifically, via autoregressive (AR) models, something that engenders both effective training leading to small codebooks and computationally fast implementations. Furthermore, an AR process models the envelope of the signal's spectrum rather than its fine structure. Such models are suitable in this context since it is shown that the spectral envelope of speech is an important cue for intelligibility [30]. Compared to our previous work [26], which can be interpreted as sampling the speech spectrum at high-SNR frequencies based on the pitch, something that is consistent with the glimpsing model of speech perecetion [31], the new method is based on the complete speech spectrum. It should also be noted that we here address the problem of single-channel non-intrusive intelligibility prediction, which is a much more difficult task than the multichannel problem [25, 26], as the latter can use spatial information.

The rest of the paper is organized as follows. First, the principles of intelligibility prediction in the STOI method are described in Section 2. Then, the signal model that the proposed method is based on is detailed in Section 3, and the proposed non-intrusive method is described in in Section 4. The experimental details and results, which include both experiments with objective measures and a listening test, are first described in Section 5 and then discussed in Section 6. Finally, Section 7 concludes on the work.

## 2. Background

The STOI [14] metric predicts the speech intelligibility based on the correlation between the temporal envelopes of the clean and the degraded speech signal (see Fig. 1). First, the clean and degraded speech signals are decomposed in time-frequency representations using a discrete Fourier transform. Then, these time-frequency representations are grouped in one-third oc-

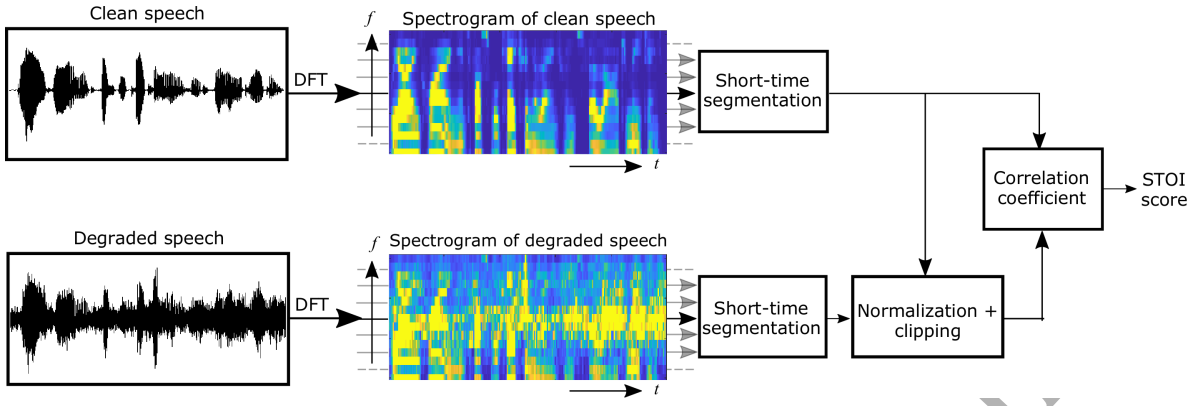


Figure 1: Block diagram of the STOI measure [14] that forms the basis for the proposed non-intrusive method. The STOI metric is based on the correlation between temporal envelopes of the clean and degraded speech in short time segments.

tave frequency bins and short-time segments (384 ms). The short-time segments are normalized in order to account for global level differences of the input signals. Furthermore, the short-time segments are clipped to prevent time-frequency units that are already completely degraded from excessively influencing the intelligibility score. Finally, the correlation of the signals is calculated over the short-time segments per frequency band. The STOI output is the average of the correlation coefficients across frequency bands and time-segments, i.e., a scalar value in the range 0-1 which relates monotonically to the average speech intelligibility scores.

## 3. Signal model

Assuming that a speech signal and a noise signal are generated by uncorrelated random processes, the corresponding noisy speech signal, $y(n)$, at time instance $n$ is $y(n) = s(n) + w(n)$. In the proposed method, both the speech and the noise are modeled as stochastic processes, namely AR processes [28, 29]. Using such a stochastic AR model, a segment of the speech signal is expressed as

$$s(n) = -\sum_{i=1}^{P} a_s(i)s(n-i) + u(n), \tag{1}$$

which can also be expressed in vector notation as

$$u(n) = \mathbf{a}_s^T \mathbf{s}(n) \tag{2}$$

where $P$ is the order of the AR process, $\mathbf{s}(n) = [s(n), s(n-1), \ldots, s(n-P)]^T$ is a vector collecting the $P$ past speech samples, $\mathbf{a}_s = [1, a_s(1), a_s(2), \ldots, a_s(P)]^T$ is a vector containing the speech auto-regressive parameters with $a_s(0) = 1$, and $u(n)$, which here models the excitation, is zero mean white Gaussian noise with excitation variance $\sigma_u^2$. Transforming the AR model into the frequency domain, $A_s(\omega)S(\omega) = U(\omega) \Leftrightarrow S(\omega) = U(\omega)/A_s(\omega)$, results in the following power spectrum:

$$P_s(\omega) = |S(\omega)|^2 = \frac{\sigma_u^2}{|A_s(\omega)|^2}, \tag{3}$$

where $A_s(\omega) = \sum_{k=0}^{P} a_s(k)e^{-j\omega k}$. Similarly to the speech signal, the noise signal can be modeled as

$$w(n) = -\sum_{i=1}^{Q} a_w(i)w(n-i) + v(n), \tag{4}$$

which can also be expressed as

$$v(n) = \mathbf{a}_w^T \mathbf{w}(n), \tag{5}$$

where $Q$ is the order of the AR process, $\mathbf{w}(n) = [w(n), w(n-1), \ldots, w(N-Q)]^T$ is a vector collecting the $Q$ past noise samples, $\mathbf{a}_w = [1, a_w(1), a_w(2) \ldots, a_w(Q)]^T$ where $a_w(0) = 1$, and $v(n)$ is zero mean white Gaussian noise with excitation variance $\sigma_v^2$. The noisy power spectrum is likewise given by

$$P_w(\omega) = |W(\omega)|^2 = \frac{\sigma_v^2}{|A_w(\omega)|^2}. \tag{6}$$

where $A_w(\omega) = \sum_{k=0}^{Q} a_w(k)e^{-j\omega k}$.

The models of the the speech and noise in (2) and (5), respectively, can be motivated as follows. The AR model has a long history in speech processing, where one of its uses is in modeling the speech production system (see, e.g., [32]), where it corresponds to a cylinder model of the vocal tract which is excited by a noise signal generated by the lungs. The model is, though,

well-known not to be perfect. For example, it does not account for the nasal cavity and the Gaussian model is only a good model for unvoiced speech and less so for voiced speech [33]. Nevertheless, it remains useful for many purposes and here it is used as a low-dimensional representation of the speech spectrum. Regarding the noise, the model is good for many natural noise sources, but, in any case, it can be used for modeling arbitrary, smooth spectra of Gaussian signals [34].

## 4. The NIC-STOI measure

The proposed method provides an objective measure for speech intelligibility prediction given solely the degraded speech signal, i.e., non-intrusively.

The method is based on the speech and noise being additive and the AR models of the speech (2) and noise (5) signals. The speech and noise spectra are simultaneously estimated from the degraded speech signal using a Bayesian approach which uses the AR parameters as prior information for inference. The prior information is obtained from trained codebooks (dictionaries) of speech and noise AR parameters. The estimation is performed on short-time frames in order to account for non-stationary noise.

Figure 2 depicts a block diagram of the NIC-STOI algorithm. The methodology comprises three main steps: 1) estimation of the parameters for the speech and noise AR models, 2) computation of the time-frequency representations for the clean, $s$, and noisy speech, $y$, signals from the estimated parameters, 3) prediction of speech intelligibility of the noisy speech signal with the STOI framework from the estimated spectra.

### 4.1. Step 1: Parameter Estimation

Let the column vector $\boldsymbol{\theta} = [\mathbf{a}_s;\ \mathbf{a}_w;\ \sigma_u^2;\ \sigma_v^2]$ comprise all parameters to be estimated, i.e., the AR coefficients and the excitation variances of the models of both speech and noise.

Bayes' theorem facilitates the computation of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ of the model parameters $\boldsymbol{\theta}$ conditioned on the observation of $N$ noise samples, i.e., $\mathbf{y} = [\, y(0)\, y(1) \ldots\, y(N-1)\,]$, from the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, the prior distribution of the model parameters $p(\boldsymbol{\theta})$, and the marginal distribution of the data $p(\mathbf{y})$ [28, 29, 35]:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{7}$$

Based on the signal model introduced previously, the data likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, is a multi-variate zero-mean Gaussian distribution with covariance matrix, $\mathbf{R}_Y = \mathbf{R}_s + \mathbf{R}_w$, where $\mathbf{R}_s = \sigma_u^2(\mathbf{G}_s^T\mathbf{G}_s)^{-1}$ and $\mathbf{G}_s$ is a $N \times N$ lower triangular Toeplitz matrix defined by the AR parameters $\mathbf{a}_s$. More specifically, it is given by

$$\mathbf{G}_s = \begin{bmatrix} 1 & 0 & \ldots & & 0 \\ a_s(1) & 1 & & & \\ \vdots & a_s(1) & & & \\ a_s(P) & \vdots & \ddots & & \vdots \\ 0 & a_s(P) & \ddots & & \\ \vdots & \vdots & & 1 & \\ 0 & 0 & \ldots & a_s(1) & 1 \end{bmatrix} \tag{8}$$

while the noise covariance matrix can be expressed as $\mathbf{R}_w = \sigma_v^2(\mathbf{G}_w^T\mathbf{G}_w)^{-1}$ with $\mathbf{G}_w$ being defined in a similar manner as $\mathbf{G}_s$ but from $\mathbf{a}_w$. Then, the minimum mean square error (MMSE) estimate is given by [36]

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MMSE}} &= \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}\left[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})^2\right] = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) \\ &= \int_\Theta \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_\Theta \boldsymbol{\theta} \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} d\boldsymbol{\theta}, \end{aligned} \tag{9}$$

where $\Theta$ is the support space of the parameters to be estimated. Based on the independence of speech and noise signals, and further assuming that the AR process and excitation variances are independent, the prior distribution of the model parameters can be simplified as

$$p(\boldsymbol{\theta}) = p(\mathbf{a}_s, \sigma_u^2)p(\mathbf{a}_w, \sigma_v^2) \approx p(\mathbf{a}_s)p(\sigma_u^2)p(\mathbf{a}_w)p(\sigma_v^2).$$

Limiting the support of the AR parameter vectors $\mathbf{a}_s$ and $\mathbf{a}_w$ to predefined codebooks of size $N_s$ and $N_w$, respectively, the corresponding excitation variances are estimated through a maximum likelihood (ML) approach

$$\{\sigma_{u,ij}^{2,\text{ML}}, \sigma_{v,ij}^{2,\text{ML}}\} = \arg\max_{\sigma_u^2, \sigma_v^2} \log p(\mathbf{y}|\mathbf{a}_{s_i}^{\text{CB}};\ \mathbf{a}_{w_j}^{\text{CB}};\ \sigma_u^2;\ \sigma_v^2),$$

where $\mathbf{a}_{s_i}^{\text{CB}}$ is the $i^{th}$ entry of the speech codebook and $\mathbf{a}_{w_j}^{\text{CB}}$ is the $j^{th}$ entry of the noise codebook.

The Gaussian likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ can be expressed in the frequency domain in terms of the Itakura-Saito distortion measure between the observed, $P_y(\omega)$, and modeled, $\hat{P}_y^{ij}(\omega)$, noisy data power spectrum, i.e.,

$$p(\mathbf{y}|\mathbf{a}_{s_i}^{\text{CB}};\ \mathbf{a}_{w_j}^{\text{CB}};\ \sigma_{u,ij}^2;\ \sigma_{v,ij}^2) \propto e^{-d_{\text{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega))}, \tag{10}$$

where $d_{\text{IS}}(\cdot,\cdot)$ is the Itakura-Saito divergence, which is given by [29, 37]

$$\begin{aligned} &d_{\text{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) = \\ &\qquad \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)} - \ln\left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)}\right) - 1\right) d\omega. \end{aligned} \tag{11}$$

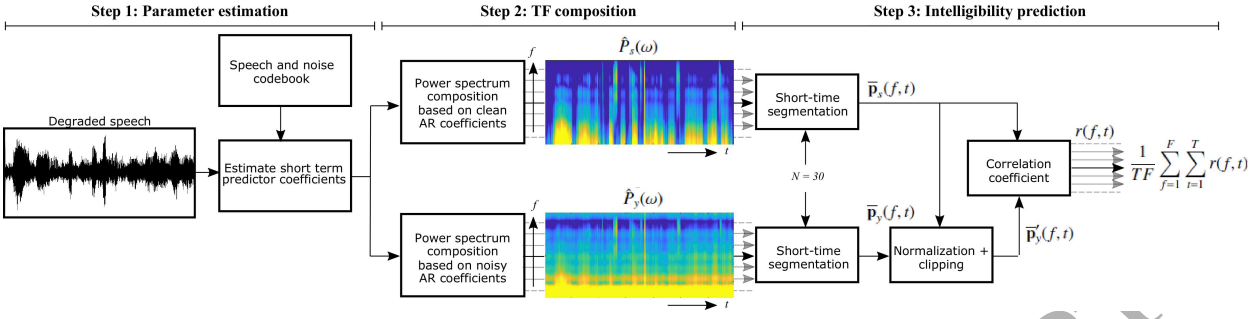


Figure 2: Block diagram depicting the processing scheme of the proposed non-intrusive codebook-based short-time objective intelligibility (NIC-STOI) metric. The relevant features of the clean and degraded speech signals are estimated using a codebook-based approach as time-frequency power spectra, which replace the estimates in the front-end of the STOI method.

Equation (11) makes use of the modeled noisy power spectrum, which is here given by

$$\hat{P}_y^{ij}(\omega) = \frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2}, \tag{12}$$

where $A_s^i(\omega) = \sum_{k=0}^{p} a_s^{i,\mathrm{CB}}(k) e^{-j\omega k}$ and $A_w^j(\omega) = \sum_{k=0}^{p} a_w^{j,\mathrm{CB}}(k) e^{-j\omega k}$ being the spectra of the $i^{th}$ and $j^{th}$ vector from the speech codebook and noise codebook, respectively.

Assuming that the modeling error between $P_y(\omega)$ and $\hat{P}_y^{ij}(\omega)$ is small and by using a second-order Taylor series approximation of $\ln(\cdot)$, the Itakura-Saito divergence can be approximated as [29]

$$d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) \approx \frac{1}{2} d_{\mathrm{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right), \tag{13}$$

where the log-spectral distortion between the observed and modeled noisy spectrum, $d_{\mathrm{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right)$, which is given by

$$d_{\mathrm{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right) = \tag{14}$$
$$\frac{1}{2\pi} \int_0^{2\pi} \left| \ln\left( \frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2} \right) - \ln\left(P_y(\omega)\right) \right|^2 d\omega$$

Finally, the ML estimates of the speech and noise excitation variances, $\sigma_{u,ij}^{2,\mathrm{ML}}$ and $\sigma_{v,ij}^{2,\mathrm{ML}}$ can be obtained by

$$\{\sigma_{u,ij}^{2,\mathrm{ML}}, \sigma_{v,ij}^{2,\mathrm{ML}}\} = \arg\min_{\sigma_u^2, \sigma_v^2} d_{\mathrm{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right), \tag{15}$$

which is solved by differentiating (14) with respect to $\sigma_u^2$ and $\sigma_v^2$ and setting the result equal to zero [28, 35]. This results in the following estimate of the excitation variance for the speech:

$$\sigma_{u,ij}^{2,\mathrm{ML}} = \frac{1}{\Psi_{ij}} \left( \sum_\omega \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4} \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2} - \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2} \right).$$

Similarly, the estimate of for excitation variance of the noise is given by

$$\sigma_{v,ij}^{2,\mathrm{ML}} = \frac{1}{\Psi_{ij}} \left( \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4} \sum_\omega \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^2} - \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2} \right).$$

The quantity $\Psi_{ij}$ is given by

$$\Psi_{ij} = \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4} \sum_\omega \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4} - \left( \sum_\omega \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \right)^2. \tag{16}$$

Finally, based on these estimates, the quantities in (9) are estimated from their discrete counterparts, which are given by

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} \boldsymbol{\theta}_{ij} \frac{p(\mathbf{y}|\boldsymbol{\theta}_{ij})}{p(\mathbf{y})} \tag{17}$$

and

$$p(\mathbf{y}) = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} p(\mathbf{y}|\boldsymbol{\theta}_{ij}), \tag{18}$$

where $\boldsymbol{\theta}_{ij} = [\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_{u,ij}^{2,\mathrm{ML}};\ \sigma_{v,ij}^{2,\mathrm{ML}}]$ is the resulting parameter vector for the $i^{th}$ entry of the speech codebook

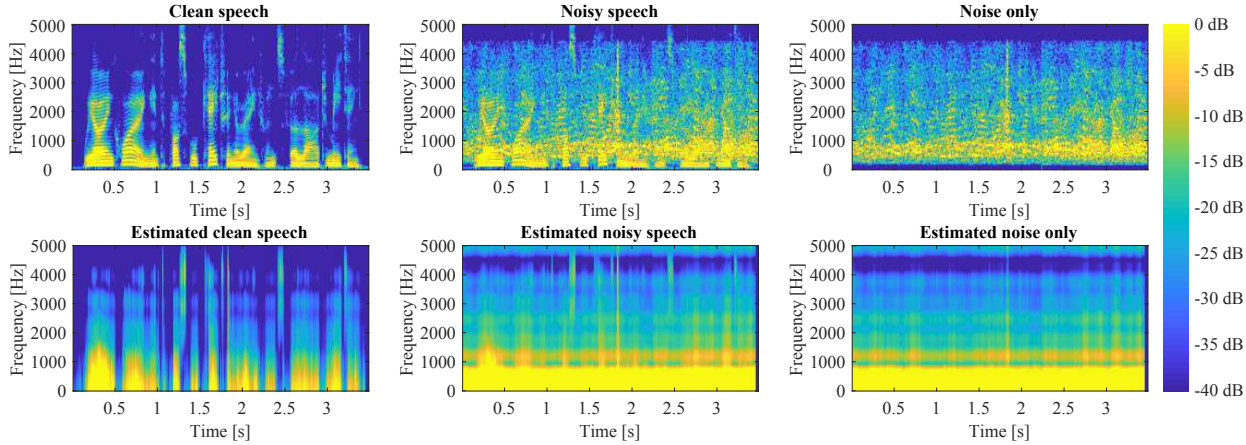


Figure 3: The top panel depicts from left to right, respectively, the spectra of the original clean speech signal, the degraded noisy speech signal at 0 dB SNR and noise only. In the bottom panel their corresponding estimated spectra using the codebook-based approach are depicted.

and the $j^{th}$ entry of the noise codebook and the final estimates are denoted as $\hat{\boldsymbol{\theta}} = [\hat{\mathbf{a}}_s;\ \hat{\mathbf{a}}_w;\ \hat{\sigma}_u^2;\ \hat{\sigma}_v^2]$. These estimates can be thought of as being obtained from an average over all possible models with each model being weighted by its posterior. We remark that codebook combinations that result in infeasible, negative values for either the speech or noise excitation variances should be neglected. Since all ML estimates of the excitation variances and the predefined codebook entries contribute with equal probability, the prior is non-informative and is omitted in (9). It should also be noted that the weighted summation of the AR parameters can be performed in the line spectral frequency (LSF) domain whereby a stable inverse filters is ensured, something that is not always the case when operatig directly on the AR parameters [28, 29].

### 4.2. Step 2: TF composition

The estimated parameters in $\hat{\boldsymbol{\theta}}$, obtained using (17), are then used to compute the time-frequency (TF) power spectra of the estimated speech and noise spectra as

$$\hat{P}_s(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2}, \tag{19}$$

where $\hat{A}_s(\omega) = \sum_{k=0}^{P} \hat{a}_s(k) e^{-j\omega k}$, and

$$\hat{P}_w(\omega) = \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}, \tag{20}$$

where $\hat{A}_w(\omega) = \sum_{k=0}^{Q} \hat{a}_w(k) e^{-j\omega k}$. The AR parameters, i.e., $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_w$, determine the shape of the envelope of the corresponding signals $\hat{S}(\omega)$ and $\hat{W}(\omega)$, respectively. The excitation variances, $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$, determine the overall signal power. Finally, the noisy spectrum is composed as the combined sum of the clean and the noise power spectra:

$$\hat{P}_y(\omega) = \hat{P}_s(\omega) + \hat{P}_w(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2} + \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}. \tag{21}$$

These time-frequency spectra replace the discrete Fourier transform of the clean reference signal and the noisy signal in the original STOI measure, respectively.

### 4.3. Step 3: Intelligibility Prediction

The STOI measure is used for intelligibility prediction with the estimated spectra $\hat{P}_s(\omega)$ (19) and $\hat{P}_y(\omega)$ (21) as inputs. First, the frequency bins of $\hat{P}_s(\omega)$ and $\hat{P}_y(\omega)$ are grouped into 15 one-third octave bands denoted by $\overline{P}_s(f,t)$ and $\overline{P}_y(f,t)$, respectively, with the lowest center frequency set to 150 Hz and the highest set to 4.3 kHz. The short-time region of the temporal envelopes of the clean speech is defined as $\overline{\mathbf{p}}_s(f,t) = [\overline{P}_s(f,t-N+1), \overline{P}_s(f,t-N+2), \ldots, \overline{P}_s(f,t)]^T$, where $N$ is the length of the short-time regions and is set to 30, resulting in a short-time region of 384 ms as in the original STOI implementation [14]. In the same manner, the short-time region of the degraded speech is given by $\overline{\mathbf{p}}_y(f,t)$. The short-time regions of the degraded speech, $\overline{\mathbf{p}}_y(f,t)$, are further clipped by a normalization procedure in order to de-emphasize the impact of region in which noise dominates the spectrum:

$$\overline{\mathbf{p}}'_y(f,t) = \min\left(\frac{\|\overline{\mathbf{p}}_s(f,t)\|_2}{\|\overline{\mathbf{p}}_y(f,t)\|_2}\overline{\mathbf{p}}_y(f,t), (1+10^{-\beta/20})\overline{\mathbf{p}}_s(f,t)\right)$$

where $\|\cdot\|_2$ denotes the $l_2$ norm and $\beta = -15$ dB is the lower signal-to-distortion ratio. The local correlation coefficient, $r(f,t)$, between $\overline{\mathbf{p}}'_y(f,t)$ and $\overline{\mathbf{p}}_s(f,t)$ at frequency $f$ and time $t$, is defined as

$$r(f,t) = \frac{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^T (\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})}{\sqrt{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^2}\sqrt{(\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})^2}},$$

where $\mu(\cdot)$ denotes the sample average of the corresponding vector. Given the local correlation coefficient, the NIC-STOI prediction is given by averaging across all bands and frames as

$$d_{NS} = \frac{1}{TF}\sum_{f=1}^{F}\sum_{t=1}^{T} r(f,t). \tag{22}$$

## 5. Experimental Details and Results

### 5.1. Performance Measures

The non-intrusive intelligibility prediction is given by $d_{NS}$, for the different conditions to be evaluated. Whereas the ground truth, denoted by $d_S$, for these conditions are given by the intrusive STOI scores. Similarly to the approach in [24], the original true STOI score is expected to be well-correlated with the subjective intelligibility. Thus, the purpose is to predict the intrusive STOI score of a given condition using a non-intrusive method. The performance of the objective intelligibility predictions are evaluated using three performance metrics often used for assessing objective intelligibility predictions [3, 14, 39]:

- The Pearson correlation coefficient ($\rho$) quantifies the linear relationship between the predicted non-intrusive intelligibility scores and true STOI scores or subjective intelligibility scores, where a higher $\rho$ indicates higher correlation.
- Kendall's Tau ($\tau$) characterizes the ranking capability by describing the monotonic relationship between the predicted intelligibility scores and true STOI scores or subjective intelligibility scores, where a higher $\tau$ represents better performance [40]. It is defined as $\tau = 2(n_c - n_d)/N(N-1)$, where $n_c$ is the number of concordant pairs, i.e. ordered in the same way, and $n_d$ is the number of discordant pairs, i.e. ordered differently.
- The standard deviation of the prediction error ($\sigma$) is given as a measure of the estimation accuracy of the predicted non-intrusive intelligibility scores, where a lower $\sigma$ implies better results.

### 5.2. Experimental Details

The results reported in this paper are based on both objective measurements and subjective listening tests. For the results based on the objective measures, the proposed metric, NIC-STOI, is evaluated on a test set of 100 speech utterances (full sentences), 50 male and 50 female, randomly selected from the EUROM_1 database of the English corpus [41]. The interfering additive noise signal is babble noise from the AURORA database. The babble noise contains many speakers in a reverberant acoustical environment. The sentences and interfering additive noise signal are both resampled to 10 kHz. Segments randomly selected from the additive noise signal are added to the EUROM_1 sentences at different SNR levels in the range of -30 to 30 dB SNR in steps of 10 dB SNR.

For further evaluation of the proposed metric, a subjective listening test has also been carried out to provide a data set for comparing NIC-STOI and SRMR. Stimuli were the fixed-syntax sentences from the GRID corpus database [38] mixed with the babble signal from the AURORA database with an SNR range -8 to 0 dB. The grid corpus consists of sentences spoken by 34 talkers (16 female and 18 male). The sentences are simple, syntactically identical phrases, e.g. place blue in A 4 again, and the listeners are asked to identify the color, letter, and digit after listening to the stimuli using a user-controlled MATLAB interface. The syntax and words of the GRID corpus are shown in Table 1. A total of nine subjects were used for the experiment which took around 30 minutes per subject. Intelligibility was defined as the number of keywords correctly identified per stimulus resulting in a fraction of either 0, 1/3, 2/3, or 1 being correct. A total of 220 stimuli were used, approximately 2 s in duration each, with the same stimuli being used for both NIC-STOI and SRMR: 5 SNR levels times 44 different sentences. We remark that to reduce intra- and intersubject variability the condition-averaged results are used for comparison and mapping of the objective results to subjective performance [3, 42]. Measuring intelligibility on a short time-scale (i.e., from short stimuli less than 2 s in duration each) with non-stationary noise types implies a high variance for both subjective and objective evaluations, i.e., precise estimation of intelligibility requires multiple sentences and not only a few words. However, it is difficult to execute subjective listening tests using long sentences or phrases as stimulus for which reason the average of many shorter sentences is here used instead.

The AR parameters and excitation variances of both the speech and noise signal are estimated on frames with

Table 1: Sentence syntax of the GRID database [38] which is used in the subjective listening test. Each sentence is constructed from (in order) a combination of a command, color, preposition, letter digit, and adverb.

| Command | Color | Preposition | Letter | Digit | Adverb |
|---|---|---|---|---|---|
| bin | blue | at | A-Z | 0-9 | again |
| lay | green | by | (no W) | | now |
| place | red | in | | | please |
| set | white | with | | | soon |

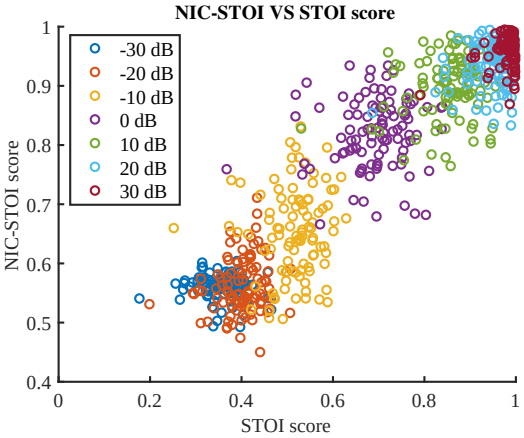


Figure 4: Scatter plot of the predicted STOI scores using the non-intrusive codebook-based STOI, NIC-STOI, metric.

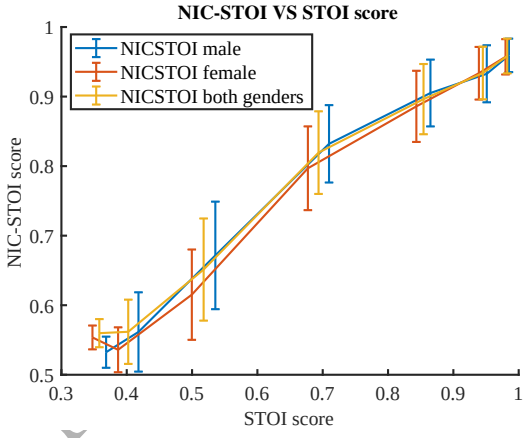


Figure 5: Averaged NIC-STOI scores (± standard deviation) against the intrusively computed STOI score.

a length of 256 samples. The speech and, thus, the estimated parameters are assumed to be stationary over these very short 25.6 ms frames. The frames are windowed using a Hann window with 50 % overlap between adjacent frames. The AR model orders $P$ and $Q$ of the speech and noise, respectively, are set to 14 in accordance with the literature [28, 29, 35]. The speech codebook is trained using the generalized Lloyd algorithm (GLA) on 10 minutes of speech from multiple speakers in the EUROM_1 database in order to ensure a sufficiently general speech model [28, 43]. We stress that the speakers included in the test set are not used for the training of the speech codebook. The noise codebook is trained on 2 minutes of babble talk. The sizes of the speech and noise codebooks are $N_s = 64$ and $N_w = 8$, respectively.

### 5.3. *Experimental Results*

An example of the spectrum of a speech signal from the test set is shown in Fig. 3 . The spectra of the original clean speech signal, the degraded noisy signal at 0 dB SNR and the noisy only are depicted in the top panel from left to right, respectively. The corresponding estimated spectra of the relevant signal features are shown in the bottom panel. The spectra are generated using trained codebooks of speech and noise spectral shapes. The estimated clean spectrum (bottom left panel) and estimated noisy spectrum (bottom middle panel) are used as input to the intrusive STOI framework.

The performance of the NIC-STOI metric is evaluated against the intrusively computed scores of the original STOI metric as ground truth. In Fig. 4, the estimated NIC-STOI scores have been plotted against the intrusive STOI scores. The plot shows good performance by means of a strong monotonic relationship between NIC-STOI and STOI, such that a higher NIC-STOI score also corresponds to a higher STOI score. Furthermore, a strong linear correlation can be observed between the two measures. This observation is also supported by the performance criteria, where NIC-STOI achieves a Pearson's correlation of $\rho = 0.94$, Kendall's Tau of $\tau = 0.70$ and a standard deviation of the prediction error $\sigma = 0.14$ for STOI, implying a high correlation. This indicates that the proposed non-intrusive version of STOI can offer a comparable performance to the original intrusive STOI.

Fig. 5 depicts the averaged predictions (± standard deviation) of the NIC-STOI scores in the scatter plot in

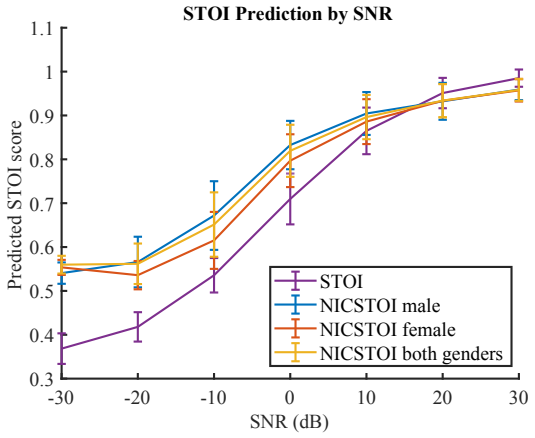


Figure 6: Averaged NIC-STOI and STOI scores (± standard deviation) per SNR condition.

Fig. 4 for male (blue line), female (red line) and both genders (yellow line), where the performance measures are given in Tab. 2. As it can be observed, the measure performs equally well whether the method is tested using either a gender specific clean speech codebook or a generic clean speech codebook. This suggests that the method generalizes well and does not capture gender specific effects due to the very generic and smooth structure of the spectra of the auto-regressive processes.

In Fig. 6 the STOI measure (purple line) and the NIC-STOI measure (male: blue line; female: red line; both genders: yellow line) are depicted as function of SNR. There is a clear monotonic correspondence between NIC-STOI and STOI, such that a higher STOI measure results in a higher NIC-STOI score. Furthermore, the NIC-STOI scores also increase with increasing SNR.

Subjective results, in terms of intelligibility as a function of SNR, are shown in Fig. 7 together with objective results obtained using the proposed NIC-STOI and SRMR. The error bars in the Figure are 95 % confidence intervals computed using a normal distribution for the SRMR and NIC-STOI methods and the normal approximation for the binomial confidence interval of the subjective results from the listening test. Note that to map the objective results to subjective intelligibility, a sigmoid function has been fitted to the average data as described in Section 5.2. As can be seen, the proposed method performs well and is capable of predicting the speech intelligibility with similar variance over a range of SNRs. The results do not, however, enable the conclusion that NIC-STOI is superior to SRMR although NIC-STOI has a better alignment with the subjective data, as both metrics have a good performance, even at

Table 2: Performance of the proposed metric in terms of Pearson's correlation ($\rho$), and Kendall's tau ($\tau$) and the standard deviation of the prediction error ($\sigma$) between NIC-STOI and STOI.

| Condition | $\rho$ | $\tau$ | $\sigma$ |
|---|---|---|---|
| Male | 0.93 | 0.70 | 0.14 |
| Female | 0.94 | 0.71 | 0.13 |
| Both genders | 0.94 | 0.70 | 0.14 |

low SNRs, and the confidence intervals overlap. Concerning the probability intervals, the intervals for both NIC-STOI and SRMR are large, as is to be expected, due to the short sentences in the GRID corpus and the limited number of stimuli for each SNR level. One thing to note is that the variance for SRMR increases as the SNR decreases, whereas NIC-STOI exhibits a similar variance across SNRs.

## 6. Discussion

Since the framework of NIC-STOI is based on an AR model, it only captures the overall envelope structure and not the fine structure of the speech signal as illustrated in Fig. 3 [29, 37]. The envelope of the speech has been shown to be a good predictor for speech intelligibility in previous intrusive intelligibility frameworks, i.e. STI and EPSM [4, 17, 30]. Extensive vocoder simulations also support these findings, where a high speech intelligibility can be obtained in quiet solely from the envelope content in only four spectral bands [30]. As such, only modeling the envelope structure of the clean speech as the essential features in NIC-STOI is assumed to be an appropriate predictor for speech intelligibility. Moreover, the promising results in [28], which show improvements of STOI scores for single channel enhancement over the noise signal, also support that the proposed model captures the essential features of the speech, as the estimated AR parameters and excitation variances are used in a speech production model in [28] to enhance the noisy speech with a Kalman filter.

Both the reported objective and subjective results show that the proposed method works well. The subjective results show that the proposed method can predict the intelligibility of a listening experiment over a range of 10 dB. Although the predicted values exhibit a high variance, as is to be expected of this type of experiment, this variance is similar to the one obtained with SRMR. The objective results indicate that NIC-STOI performs very well for a broad range of SNRs, even down to -30 dB SNR where the noisy speech is expected to be unintelligible. It should be noted that while NIC-STOI appears to deviate from STOI for very

low SNRs, this is less important as, according to [3], a STOI score of 0.6 approximately corresponds to zero intelligibility. Even though the absolute value of STOI depends highly on the specific speech material and listening environment, the broad working range of NIC-STOI should cover the range of intelligibility. Hence, any score below this threshold can be simply assumed unintelligible. Here, it is also important to stress that the overall aim of NIC-STOI is to have a monotonic relation with the intrusively computed STOI scores, and not necessarily to predict the absolute STOI scores. However, the offset observed between the predicted NIC-STOI scores and STOI scores in Fig. 6 can easily be accounted for by the observed linear trend between the two measures depicted in Fig. 4, such that the absolute STOI score can be predicted by means of the estimated NIC-STOI score.

It should be noted that STOI was among the first intrusive intelligibility metrics with very good performance, but since it was first introduced other intrusive metrics have also been proposed that show good performance. The front-end of NIC-STOI, that forms the basis of the present work, could also quite possibly be used for other intrusive frameworks, provided that they are also based on spectral features of the noisy and clean speech. Regarding this, it is interesting to note that the estimation of the parameters in short-time segments based on the current observation makes the front-end suitable for non-stationary noise conditions. However, STOI does not work well for highly non-stationary interferers due to the analysis window length. Therefore, it could be interesting to investigate using the Extended STOI (ESTOI) as a back-end to NIC-STOI instead, since this method has been developed to work well for highly modulated noise sources [44].

Correlation-based metrics including STOI are generally not suitable for predicting the intelligibility of reverberant speech and, thus, it is likely that NIC-STOI will fail in such conditions [14, 45]. Furthermore, the short time frames used in STOI might also have a negative impact on the application of NIC-STOI to reverberant speech, as short time frames cannot capture all the effects of reverberation, such as temporal smearing [14]. Currently, SRMR and ModA are the most well-studied non-intrusive intelligibility metrics. They have both been proposed for predicting the intelligibility of reverberant speech, where they both show good performance [3, 19, 20]. Even though these metrics are aimed for reverberant speech, they have also been tested for noisy and processed speech [3], where they perform reasonably well. However, it would seem that SRMR and ModA are a more suitable choice for reverberant speech, while our proposed method, NIC-STOI, which takes into account the presence of noise, is a more suitable choice for additive degradations, such as background noise and interferences. In this connection, it should also be mentioned that the proposed method is computationally much more demanding than SRMR and ModA, mainly due to the codebook search, although approximate methods for implementation of this exist [46].

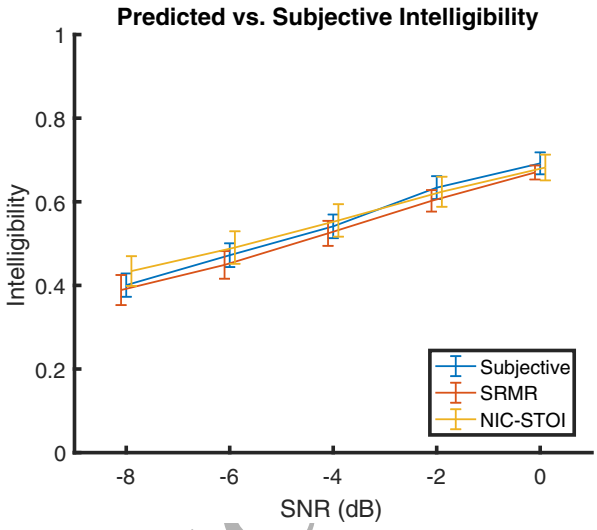


Figure 7: Intelligibility as a function of SNR for subjective listening experiments and as predicted by the proposed NIC-STOI and SRMR. Shown are the means and their 95 % confidence intervals.

In closing, we remark that the proposed method is not expected to account well for non-linear signal processing, since it is based on an additive noise model as well as the codebooks being trained on clean speech signals and noise signals. However, testing the method on the Ideal Time-Frequency Segregation (IFTS) data set from [47], which was used for evaluating the original STOI measure [14], results in a Pearson correlation of 0.70, which is surprisingly good. For comparison, NIC-STOI outperforms the non-intrusive intelligibility metric, SRMR [3, 19], which achieves a Pearson correlation of 0.24 [7], although it should be noted that SRMR, as already mentioned, was designed for reverberant speech. However, the newly proposed Non-Intrusive STOI (NI-STOI) measure [7] achieves a Pearson correlation of 0.71 for the data set [47], which is on par with the results obtained for NIC-STOI. We remark that NI-STOI is not completely non-intrusive, as it is based on the ideal voice activity detector used in the intrusive STOI metric [7].

## 7. Conclusion

In this paper, a non-intrusive codebook-based short-time objective intelligibility metric, called NIC-STOI, has been proposed. It is based on an intrusive intelligibility metric, STOI, but, unlike STOI, it does not require access to the clean speech signal. Instead, the proposed method estimates the spectrum of the reference signal by identifying the entries of pre-trained spectral codebooks of speech and noise spectra, parametrized by auto-regressive parameters, which best fit the observed signal, i.e., the noisy speech signal. This is done in a statistical framework wherein parameters are estimated by minimizing the Itakura-Saito divergence for combinations of speech and noise models. This is equivalent to maximum likelihood estimation for Gaussian distributed signals. The proposed NIC-STOI metric is shown, in experiments, to be highly correlated with STOI (with a Pearson correlation of 0.94 and a standard deviation of the prediction error of 0.14) and is also validated in a listening experiment assessing speech intelligibility. Hence, it can be used for the assessment of speech intelligibility when a clean reference signal is not available. This could be used, for example, for online optimization of hearing aids.

## References


[1] P. C. Loizou, Speech Enhancement: Theory and Practice, Signal processing and communications, Taylor & Francis, 2007.

[2] Y. Hu, P. C. Loizou, Subjective comparison and evaluation of speech enhancement algorithms, Speech Commun. 49 (78) (2007) 588 – 601.

[3] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, S. Scollie, Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools, IEEE Signal Process. Mag. 32 (2) (2015) 114–124.

[4] H. J. M. Steeneken, T. Houtgast, A physical method for measuring speechtransmission quality, J. Acoust. Soc. Am. 67 (1) (1980) 318–326.

[5] S. Jørgensen, J. Cubick, T. Dau, Speech intelligibility evaluation for mobile phones., Acustica United with Acta Acustica 101 (2015) 1016–1025.

[6] T. Houtgast, H. J. M. Steeneken, A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria, J. Acoust. Soc. Am. 77 (3) (1985) 1069–1077.

[7] A. Heidemann Andersen, J. de Haan, Z.-H. Tan, J. Jensen, A non-intrusive short-time objective intelligibility measure, in: ICASSP, 2017, pp. 5085–5089.

[8] J. B. Allen, Harvey Fletcher's role in the creation of communication acoustics, J. Acoust. Soc. Am. 99 (4) (1996) 1825–1839.

[9] N. R. French, J. C. Steinberg, Factors governing the intelligibility of speech sounds, J. Acoust. Soc. Am. 19 (1) (1947) 90–119.

[10] ANSI S3.5, 1997, Methods for the calculation of the Speech Intelligibility Index, American National Standards Institute, New York, USA (1997).

[11] K. S. Rhebergen, N. J. Versfeld, W. A. Dreschler, Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise, J. Acoust. Soc. Am. 120 (6) (2006) 3988–3997.

[12] J. M. Kates, K. H. Arehart, Coherence and the speech intelligibility index, J. Acoust. Soc. Am. 117 (4) (2005) 2224–2237.

[13] C. Ludvigsen, C. Elberling, G. Keidser, Evaluation of a noise reduction method comparison between observed scores and scores predicted from STI, Scand. Audiol. Suppl. 38 (1993) 50–55.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech, IEEE Trans. Audio, Speech, and Language Process. 19 (7) (2011) 2125–2136.

[15] D. Wang, On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis, Springer US, Boston, MA, 2005, pp. 181–197.

[16] H. Relao-Iborra, T. May, J. Zaar, C. Scheidiger, T. Dau, Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain, J. Acoust. Soc. Am. 140 (4) (2016) 26702679.

[17] S. Jørgensen, T. Dau, Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing, J. Acoust. Soc. Am. 130 (3) (2011) 1475–1487.

[18] S. van Kuyk, W. B. Kleijn, R. Hendriks, An instrumental intelligibility metric based on information theory, in: ICASSP, 2018.

[19] T. H. Falk, C. Zheng, W.-Y. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech, IEEE Trans. Audio, Speech, and Language Process. 18 (7) (2010) 1766–1774.

[20] F. Chen, O. Hazrati, P. C. Loizou, Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure, Biomedical Signal Processing and Control 8 (3) (2013) 311 – 314.

[21] F. Chen, Modeling noise influence in speech intelligibility non-intrusively by reduced speech dynamic range, in: Interspeech, 2016.

[22] F. Chen, Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation, Biomedical Signal Processing and Control 24 (2016) 109 – 113.

[23] M. Karbasi, A. H. Abdelaziz, D. Kolossa, Twin-hmm-based non-intrusive speech intelligibility prediction, in: ICASSP, 2016, pp. 624–628.

[24] D. Sharma, Y. Wang, P. A. Naylor, M. Brookes, A data-driven non-intrusive measure of speech quality and intelligibility, Speech Commun. 80 (2016) 84–94.

[25] C. Sørensen, J. B. Boldt, F. Gran, M. G. Christensen, Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids, in: EUSIPCO, 2016, pp. 1358–1362.

[26] C. Sørensen, A. Xenaki, J. Boldt, M. Christensen, Pitch-based non-intrusive objective intelligibility prediction, in: ICASSP, 2017, pp. 386–390.

[27] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, M. G. Christensen, Non-intrusive intelligibility prediction using a codebook-based approach, in: EUSIPCO, 2017, pp. 226–230.

[28] M. Kavalekalam, M. Christensen, F. Gran, J. Boldt, Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach, in: ICASSP, 2016, pp. 191–195.

[29] S. Srinivasan, J. Samuelsson, W. Kleijn, Codebook-based Bayesian speech enhancement for nonstationary environments, IEEE Trans. Audio, Speech, and Language Process. 15 (2) (2007) 441–452.

[30] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech recognition with primarily temporal cues, Science 270 (5234) (1995) 303–304.

[31] M. Cooke, A glimpsing model of speech perception in noise, J. Acoust. Soc. Am. 119 (3) (2006) 1562–1573.
[32] J. H. L. Hansen, J. G. Proakis, J. R. Deller, Jr., Discrete-Time Processing of Speech Signals, Prentice-Hall, 1987.
[33] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, Sparse linear prediction and its applications to speech processing, IEEE Trans. Audio, Speech, and Language Process. 20(5) (2012) 1644–1657.
[34] J. Makhoul, Linear Prediction: A Tutorial Review, Proceedings of the IEEE 63(4) (1975) 561–580.
[35] S. Srinivasan, J. Samuelsson, W. Kleijn, Codebook driven short-term predictor parameter estimation for speech enhancement, IEEE Trans. Audio, Speech, and Language Process. 14 (1) (2006) 163–176.
[36] S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
[37] K. Paliwal, W. Kleijn, Quantization of LPC parameters, in: Speech Coding and Synthesis, Elsevier Science, 1995, pp. 433–468.
[38] M. Cooke, J. Barker, An Audio-visual corpus for speech perception and automatic speech recognition (L), J. Acoust. Soc. Am. 120(5) (2006) 2421–2424.
[39] A. H. Andersen, J. M. de Hann, Z.-H. Tan, J. Jensen, Predicting the intelligibility of noisy and nonlinearly processed binaural speech, IEEE Tran. Audio, Speech, Lang. Process. 24 (11) (2016) 1908–1920.
[40] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[41] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, J. Zeiliger, EUROM - a spoken language resource for the EU, in: EUROSPEECH, Vol. 1, 1995, pp. 867–870.
[42] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, M. Wältermann, Speech quality estimation: Models and trends, IEEE Signal Processing Magazine 28 (6) (2011) 18–28.
[43] Y. Linde, A. Buzo, R. Gray, An algorithm for vector quantizer design, IEEE Trans. Communications 28 (1) (1980) 84–95.
[44] J. Jensen, C. H. Taal, An algorithm for predicting the intelligibility of speech masked by modulated noise maskers, IEEE Trans. Audio, Speech, and Language Process. 24 (11) (2016) 2009–2022.
[45] R. L. Goldsworthy, J. E. Greenberg, Analysis of speech-based speech transmission index methods with implications for nonlinear operations, J. Acoust. Soc. Am. 116 (6) (2004) 3679–3689.
[46] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, 1993.
[47] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, D. Wang, Role of mask pattern in intelligibility of ideal binary-masked noisy speech, J. Acoust. Soc. Am. 126 (3) (2009) 1415–1426.