INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Validation of the Non-Intrusive Codebook-based Short Time Objective Intelligibility Metric for Processed Speech

Charlotte Sørensen[1,2], *Jesper B. Boldt*[2], *Mads G. Christensen*[1]

[1]Audio Analysis Lab, CREATE, Aalborg University, Denmark
[2]GN Hearing A/S, Lautrupbjerg 7, DK-2750, Ballerup, Denmark
{csoerensen,jboldt}@gnresound.com, mgc@create.aau.dk

Abstract

In recent years, objective measures of speech intelligibility have gained increasing interest. However, most speech intelligibility metrics require a clean reference signal, which is often not available in real-life applications. In a recent publication, we proposed a method, the Non-Intrusive Codebook-based Short-Time Objective Intelligibility (NIC-STOI) metric, which allows using an intrusive method without requiring access to the clean signal. The statistics of the reference signal is estimated as a combination of predefined codebooks that best fit the degraded signal by modeling the speech and noisy spectra. In this paper, we perform additional validation of the NIC-STOI in more diverse noise condition as well as for speech processed non-linearly with binary masks, where it is shown to outperform existing non-intrusive metrics.

Index Terms: Hearing aids, non-intrusive, speech intelligibility prediction, codebooks, STOI

1. Introduction

In recent years, objective measures of speech intelligibility have gained increasing interest as a tool for automatically adapting and optimizing speech enhancement algorithms in, e.g., hearing aids [1]. The articulation index (AI) [2] and the speech transmission index (STI) [3] are some of the earliest metrics that predict the intelligibility for a limited type of degradations, like linear filtering and additive noise. Recently, the speech-based envelope power spectrum model (sEPSM) [4] and the short-time objective (STOI) metric [5] were developed for more complex distortion types and are reported to have high prediction accuracy [1].

However, these metrics are all intrusive, i.e., they require a clean reference in order to predict the speech intelligibility of a degraded signal. In some scenarios, e.g., real-time processing, it is impractical to use intrusive metrics for predicting speech intelligibility. To overcome this limitation, a number of non-intrusive intelligibility prediction methods have been introduced. The Speech to Reverberation Modulation energy Ratio (SRMR) [6] and the average Modulation-spectrum Area (ModA) [7] both provide a non-intrusive estimate of the speech intelligibility based on the modulation spectrum of the degraded speech signal. Another way to predict speech intelligibility non-intrusively is to first obtain an estimate of the clean signal from its degraded version and then use this as reference to an intrusive metric. For instance, machine learning [8, 9], noise reduction [10, 11], principal component analysis [12] and neural network [13] methods have been proposed as approaches to obtain a reference signal to use inside the STOI framework from the degraded speech signal. Another non-intrusive version of the STOI metric, the non-intrusive codebook-based STOI (NIC-STOI), is proposed in [14, 15]. This is based on estimating the spectrum of the reference signal from its degraded version by identifying combinations of pre-trained codebook entries of speech and noise spectra, parametrized by Auto-Regressive (AR) parameters, which best fit the degraded speech signal. The evaluation of the NIC-STOI metric in [15] is shown to be highly correlated with STOI and subjective listening scores for additive babble noise interference. However, since methods for predicting speech intelligibility are often used to evaluate the effects of non-linear processing, a method that is also suitable for such types of processing is desirable [1]. Therefore, in this paper, the NIC-STOI metric is further validated on speech in different noise conditions, which has been non-linearly processed with Ideal Binary Masks (IBMs) [16].

2. The NIC-STOI metric

The NIC-STOI metric, proposed in [14, 15], is based on STOI but does not require access to a clean reference signal. Figure 1 depicts an overview of the NIC-STOI algorithm. The algorithm consists of three main steps: 1) estimation of the AR speech and noise model parameters 2) computation of the clean and noisy time-frequency spectra 3) prediction of intelligibility within STOI. In the following, a condensed description of the NIC-STOI metric is presented. A more thorough description is available in [15].

2.1. Step 1: Estimate parameters

It is assumed that a speech and noise signal are random uncorrelated processes such that the noisy speech signal is given by $y(n) = s(n) + w(n)$ [17, 18]. The speech and noise are modeled as stochastic AR processes expressed as $u(n) = \mathbf{a}_s^T \mathbf{s}(n)$ and $v(n) = \mathbf{a}_w^T \mathbf{w}(n)$, respectively, where $\mathbf{s}(n) = [s(n), s(n-1), \ldots, s(n-P)]^T$ and $\mathbf{w}(n) = [w(n), w(n-1), \ldots, w(N-Q)]^T$ are vectors collecting the P and Q past samples, $\mathbf{a}_s = [1, a_s(1), a_s(2), \ldots, a_s(P)]^T$ and $\mathbf{a}_w = [1, a_w(1), a_w(2) \ldots, a_w(Q)]^T$ are vectors containing the AR parameters with $a_s(0) = 1$ and $a_w(0) = 1$. Finally, $u(n)$ and $v(n)$ models the speech and noise excitations as zero mean white Gaussian noise with excitation variance σ_u^2 and σ_v^2, respectively.

The parameters to be estimated, i.e., the speech and noise AR coefficients and excitation variances are given by the vector $\theta = [\mathbf{a}_s;\ \mathbf{a}_w;\ \sigma_u^2(n);\ \sigma_v^2(n)]$. Using Bayes' theorem, the minimum mean square error (MMSE) estimate given N noisy samples, i.e., $\mathbf{y} = [\ y(0)\ y(1)\ \ldots\ y(N-1)\ ]$ can be given by [17, 19, 18]:

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) = \int_{\Theta} \theta \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} d\theta, \tag{1}$$

where Θ denotes the support space to be estimated.

http://dx.doi.org/10.21437/Interspeech.2019-1625

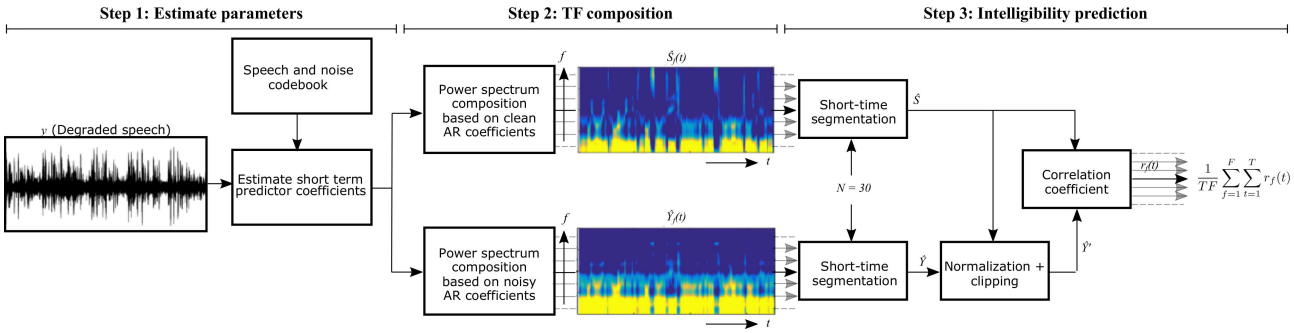

Figure 1: *A block diagram of the NIC-STOI metric adapted from [14, 15]. Using a codebook-based approach the speech and noisy spectra are jointly modeled from pre-trained codebooks of both speech and noise and are then used within STOI.*

The excitation variances are estimated through a maximum likelihood (ML) approach by limiting the AR parameters $\mathbf{a}_s$ and $\mathbf{a}_w$ to predefined codebooks of size N_s and N_w:

$$\{\sigma^{2,\mathrm{ML}}_{u,ij}, \sigma^{2,\mathrm{ML}}_{v,ij}\} = \arg\max_{\sigma^2_u,\sigma^2_v} \log p(\mathbf{y}|\mathbf{a}^{\mathrm{CB}}_{s_i};\ \mathbf{a}^{\mathrm{CB}}_{w_j};\ \sigma^2_u;\ \sigma^2_v),$$

where $\mathbf{a}^{\mathrm{CB}}_{s_i}$ and $\mathbf{a}^{\mathrm{CB}}_{w_j}$ are the i^{th} and j^{th} entry of the speech and noise codebook, respectively. The Gaussian likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is given by:

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

where $d_{\mathrm{IS}}(\cdot,\cdot)$ is the Itakura-Saito divergence between the observed, $P_y(\omega)$, and modeled, $\hat{P}^{ij}_y(\omega)$, noisy spectrum expressed as [18, 20]:

$$d_{\mathrm{IS}}(P_y(\omega), \hat{P}^{ij}_y(\omega)) = \frac{1}{2\pi}\int_{\Psi}\left(\frac{P_y(\omega)}{\hat{P}^{ij}_y(\omega)} - \ln\left(\frac{P_y(\omega)}{\hat{P}^{ij}_y(\omega)}\right) - 1\right)d\omega, \quad (3)$$

where $\hat{P}^{ij}_y(\omega) = \frac{\sigma^2_u}{|A^i_s(\omega)|^2} + \frac{\sigma^2_v}{|A^j_w(\omega)|^2}$, and A^i_s and A^j_w are the i^{th} and j^{th} entry from the speech codebook and noise codebook, respectively. The support space, Ψ, excludes values below a threshold in order to disregard time-frequency units with low energy or where the binary masks renders the presented signal inaudible. This threshold is here set to 40 dB below peak energy.

Finally, (1) is computed from its discrete counterpart:

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_s N_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w}\boldsymbol{\theta}_{ij}\frac{p(\mathbf{y}|\boldsymbol{\theta}_{ij})}{p(\mathbf{y})} \quad (4)$$

and

$$p(\mathbf{y}) = \frac{1}{N_s N_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w} p(\mathbf{y}|\boldsymbol{\theta}_{ij}), \quad (5)$$

where $\boldsymbol{\theta}_{ij} = [\mathbf{a}^{\mathrm{CB}}_{s_i};\ \mathbf{a}^{\mathrm{CB}}_{w_j};\ \sigma^{2,\mathrm{ML}}_{u,ij};\ \sigma^{2,\mathrm{ML}}_{v,ij}]$. The priors in (1) are non-informative, since the codebook entries and the ML excitation variance estimates contribute with equal probability and are, thus, omitted.

2.2. Step 2: TF composition

Using the estimated parameters, $\hat{\boldsymbol{\theta}}$, from (4) the Time-Frequency (TF) spectrum of the estimated speech and noise signal are given by:

$$\hat{P}_s(\omega) = \frac{\hat{\sigma}^2_u}{|\hat{A}_s(\omega)|^2}, \quad (6)$$

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

$$\hat{P}_w(\omega) = \frac{\hat{\sigma}^2_v}{|\hat{A}_w(\omega)|^2}, \quad (7)$$

where $\hat{A}_w(\omega) = \sum_{k=0}^{Q}\hat{a}_w(k)e^{-j\omega k}$. The shape of the envelope of the estimated signals are given by the AR parameters, i.e., $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_w$, while the overall signal power is given by the excitation variances, i.e., $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_v$. Then, the noisy spectrum is given by the sum of the speech and noise power spectra:

$$\hat{P}_y(\omega) = \hat{P}_s(\omega) + \hat{P}_w(\omega) = \frac{\hat{\sigma}^2_u}{|\hat{A}_s(\omega)|^2} + \frac{\hat{\sigma}^2_v}{|\hat{A}_w(\omega)|^2}. \quad (8)$$

2.3. Step 3: Intelligibility Prediction

The estimated speech and noise TF spectra, i.e., $\hat{P}_s(\omega)$ (6) and $\hat{P}_y(\omega)$ (8), are then used as inputs in the original STOI metric as replacement for the discrete Fourier transform of the clean and noisy signal, respectively.

The TF spectra $\hat{P}_s(\omega)$ and $\hat{P}_y(\omega)$ are grouped into 15 one-third octave bands and short-time regions of 384 ms denoted by $\overline{\mathbf{p}}_s(f,t)$ and $\overline{\mathbf{p}}_y(f,t)$ as in the original STOI implementation [5]. In order to de-emphasize the impact of noise dominated regions, the entries in $\overline{\mathbf{p}}_y(f,t)$ are clipped by an element-wise normalization procedure:

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

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

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

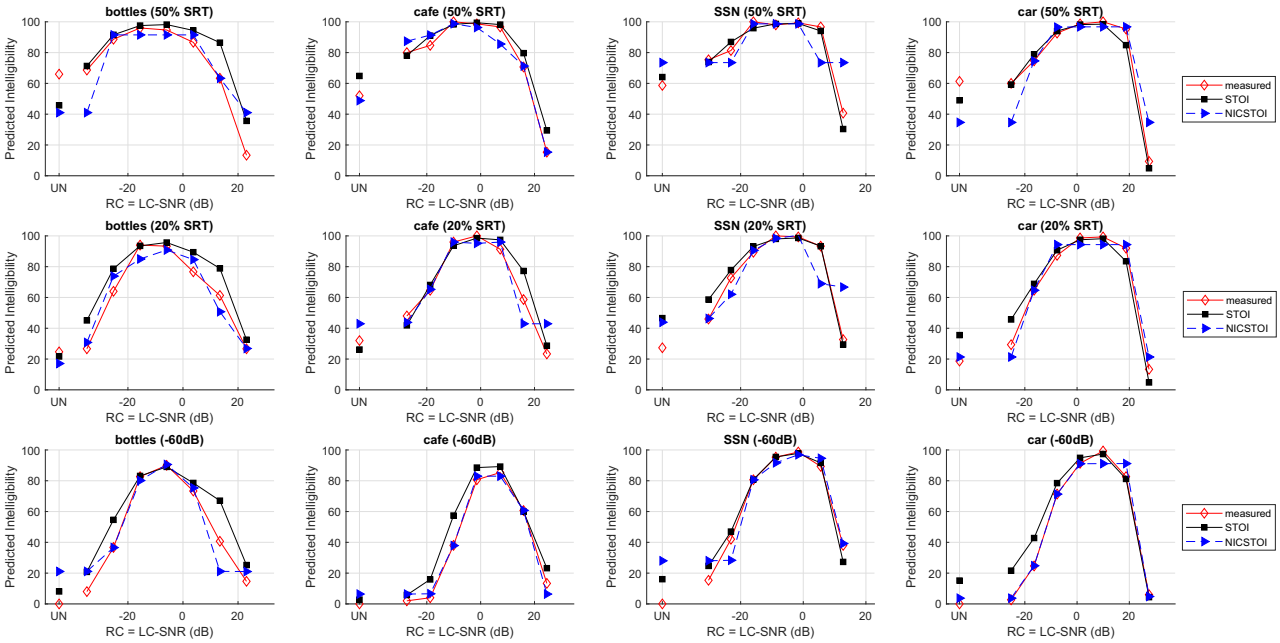

Figure 2: *NIC-STOI (blue) predictions evaluated against STOI (black) predictions and subjective (red) intelligibility scores. Each row refers to the three different SNRs used in the data set (50 % SRT, 20 % SRT and -60 dB) with the the first row corresponding to the highest SNR and the last row the lowest SNR. The columns refer to the four different noise types (bottling factory hall, cafe, Speech Shaped Noise (SSN) and car noise). The x-axis refers to the Relative Criterion (RC) values, which determines the density of the computed Ideal Binary Masks (IBM). "UN" refers to unprocessed conditions.*

where $\mu(\cdot)$ is the mean of the vector. Finally, the NIC-STOI intelligibility prediction is given by averaging the correlation coefficient, i.e. $r(f,t)$, across all bands and frames as

$$d_{NS} = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} r(f,t). \tag{9}$$

3. Experimental Details

In order to further validate the performance of the NIC-STOI metric presented in [14, 15] we here evaluate it on the same data as in the original paper on STOI [16, 5]. Subjective intelligibility scores have been obtained from 15 normal hearing subjects. Stimuli were the Dantale II sentence material [21] mixed with four different noise types: bottling factory hall noise, cafe noise, Speech Shaped Noise (SSN) and car noise at three different Signal to Noise Ratios (SNRs). The noisy signals were processed with IBMs at eight different Relative Criterion (RC) values, which determines the density of the computed binary mask [16]. Materials from 5 subjects is used to train the codebooks, whereas the results from the remaining 10 subjects are used for testing. The data and experimental details are described in detail in [16].

The speech and noise AR parameters and variances are estimated from 25.6 ms frames windowed using a Hann window with 50% overlap. Over these short time frames the estimated parameters are assumed to be stationary. The signals were resampled to 10 kHz as in the original STOI metric. The speech and noise AR model order P and Q, respectively, are set to 14 according to literature [17, 19, 18]. The speech codebook is trained on 50 clean speech sentences from the Dantale II data set not included in the training set using the generalized Lloyd algorithm (GLA) [17, 22]. The noise codebook is trained on 50 sentences of each noise type condition without IBM processing concatenated into a single vector. The sizes of the speech and noise codebooks are $N_s = 64$ and $N_w = 8$, respectively. The support space of the Itakura-Saito divergence, Ψ, is computed by taking the Fourier transform of the input signal and limiting the dynamic range to 40 dB from the highest value such that TF units below this threshold are not included in the calculation. In order to reduce intra- and intersubject variability the results are condition-averaged per noise and SNR combination and are then mapped to subjective performance across all conditions [1]. The performance of the metric is evaluated using Pearson's correlation (ρ) which gives the linear relationship, Kendall's tau (τ) which gives the ranking capability and the root mean square error (RMSE).

4. Results and Discussion

The performance of NIC-STOI is depicted in Fig 2 (blue) against measured subjective scores (red) and the original intrusive STOI metric (black). It can be observed that NIC-STOI is highly correlated with the subjective scores across all noise conditions. Furthermore, NIC-STOI is also highly correlated with STOI, which supports the earlier findings in [14, 15]. The highest deviation can be observed for the SSN noise condition, which can perhaps be explained by the noise codebook weighting this condition less when being trained on all the noise conditions concatenated.

In Table 1, NIC-STOI is evaluated against existing state of the art intelligibility metrics. The best performance is obtained by the intrusive metric STOI. However, even though NIC-STOI is non-intrusive, it comes close to being on par with the performance of STOI. NIC-STOI is compared to three other non-intrusive metrics: NI-STOI [12], SRMR [6] and SRMR-

Table 1: *Performance of the intelligibility metrics in terms of Pearson's correlation (ρ), Kendall's tau (τ) and the standard deviation of the prediction error (RMSE).* [1]*The results for NI-STOI are obtained from [12], since it was not possible to obtain an implementation of this metric. The results are from the same data set but without the cafe condition included in the logistic mapping.*

Condition	ρ	τ	RMSE
STOI [5]	0.955	0.821	8.9 %
NIC-STOI [15]	0.940	0.791	11.4 %
NI-STOI[1] [12]	0.711	0.529	25.2 %
SRMR-norm [23]	0.392	0.155	38.4 %
SRMR [6]	0.235	0.034	45.0 %

norm [23]. The results from NI-STOI are obtained from [12], since it was not possible to obtain an implementation, while implementations of the latter two are publicly available. The NI-STOI metric is aimed to predict the intelligibility of non-linearly processed speech, while the SRMR metric and the improved SRMR-norm are aimed to predict the intelligibility of reverberated speech, but have successfully been applied for noisy and processed speech [1]. As shown in Table 1, NIC-STOI outperforms all three existing non-intrusive intelligibility metrics. It should, however, be noted that NI-STOI is only trained using clean speech material [12] and SRMR and SRMR-norm is not trained at all. The cafe noise condition is primarily composed by a single interfering speaker such that additional information is needed in order to determine, which speaker is the target. NIC-STOI is trained with both clean speech and noise material, which makes it able to account for the cafe condition. Excluding the cafe noise condition in NI-STOI, NIC-STOI still has the best performance ($\rho = 0.940$, $\tau = 0.791$, RMSE $= 11.4\%$) even though the performance of NI-STOI comes close to that of NIC-STOI ($\rho = 0.907$, $\tau = 0.777$, RMSE $= 13.9\%$) [12].

5. Conclusion

In this paper, the Non-Intrusive Codebook-based Short-Time Objective Intelligibility metric, NIC-STOI, has been investigated more thoroughly on a large data set with subjective scores in diverse noise conditions. NIC-STOI non-intrusively estimates the spectrum of a reference signal from its degraded version and uses this as input to an intrusive intelligibility metric, STOI. The reference signal is estimated as combinations of entries from pre-trained speech and noise spectral codebooks, parametrized by auto-regressive parameters, which best fit the degraded signal by minimizing the Itakura-Saito divergence. In order to account for binary mask processing a small adjustment of NIC-STOI is implemented in which only time-frequency units above a certain threshold is included in the Itakura-Saito divergence. The NIC-STOI metric is highly correlated with subjective intelligibility scores on the non-linearly processed speech data set and outperforms existing non-intrusive metrics.

6. Acknowledgements

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

7. References

[1] 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.

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

[3] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[4] 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. Am.*, vol. 130, no. 3, pp. 1475–1487, 1980.

[5] 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, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[6] 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, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[7] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 311 – 314, 2013.

[8] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[9] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, pp. 84–94, 2016.

[10] C. Soerensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *EUSIPCO*, August 2016, pp. 1358–1362.

[11] C. Sørensen, A. X. J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *ICASSP*, March 2017, pp. 386–390.

[12] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *ICASSP*, March 2017, pp. 5085–5089.

[13] ——, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1925–1939, 2018.

[14] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Non-intrusive intelligibility prediction using a codebook-based approach," in *EUSIPCO*, 2017, pp. 216–220.

[15] ——, "Non-intrusive codebook-based intelligibility prediction," *Speech Communication*, vol. 101, pp. 85 – 93, 2018.

[16] 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. Am.*, vol. 126, no. 3, p. 14151426, 2009.

[17] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *ICASSP*, March 2016, pp. 191–195.

[18] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[19] ——, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[20] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*. Elsevier Science, 1995, pp. 433–468.

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

[22] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[23] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *IWAENC*, Sept. 2014, p. 5559.