A NON-INTRUSIVE SHORT-TIME OBJECTIVE INTELLIGIBILITY MEASURE

Asger Heidemann Andersen[⋆†], *Jan Mark de Haan*[†], *Zheng-Hua Tan*[⋆], *Jesper Jensen*[⋆†]

[⋆] Dept. of Electronic Systems, Aalborg University, 9220 Aalborg Øst, Denmark
[†] Oticon A/S, 2765 Smørum, Denmark

aha@es.aau.dk/aand@oticon.com, janh@oticon.com, zt@es.aau.dk, jje@es.aau.dk/jesj@oticon.com

ABSTRACT

We propose a non-intrusive intelligibility measure for noisy and non-linearly processed speech, i.e. a measure which can predict intelligibility from a degraded speech signal without requiring a clean reference signal. The proposed measure is based on the Short-Time Objective Intelligibility (STOI) measure. In particular, the non-intrusive STOI measure estimates clean signal amplitude envelopes from the degraded signal. Subsequently, the STOI measure is evaluated by use of the envelopes of the degraded signal and the estimated clean envelopes. The performance of the proposed measure is evaluated on a dataset including speech in different noise types, processed with binary masks. The measure is shown to predict intelligibility well in all tested conditions, with the exception of those including a single competing speaker. While the measure does not perform as well as the original (intrusive) STOI measure, it is shown to outperform existing non-intrusive measures.

Index Terms— non-intrusive speech intelligibility prediction, enhanced speech, speech in noise

1. INTRODUCTION

In recent years, Speech Intelligibility Prediction (SIP) has been investigated with great interest due to its potential as a tool in optimizing speech intelligibility across a wide range of applications, including e.g. architectural acoustics [1], telecommunications [2], and hearing aid signal processing [3, 4, 5]. Much of the recent work in the field is based on the classical methods: the Speech Intelligibility Index (SII) [6, 7] and the Speech Transmission Index (STI) [8, 9]. This has led to methods such as the Extended SII (ESII) [10, 11], the Coherence SII (CSII) [12] and the Binaural Speech Intelligibility Measure (BSIM) [13, 14]. Recently, the physiologically founded multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSM) [15] has received attention for its ability to predict intelligibility of speech in reverberation and modulated noise. Another recent method, the Short-Time Objective Intelligibility (STOI) measure [16], has become popular within the signal processing community because of its simplicity and proven ability to predict the impact of various speech processing algorithms [17, 2, 5]. Several variations of the STOI measure with specialized properties have been proposed [18, 3, 19].

The mentioned methods require access to a clean reference signal in addition to either the masker signal or the degraded speech signal. These are referred to as intrusive methods, because of their dependence on a clean reference signal. In some situations, intrusive methods cannot be applied because the clean reference signal is unknown or poorly defined, e.g. when attempting to predict the intelligibility of an unknown speech signal on a signal processing device in realtime.

The above concern has led to research into non-intrusive SIP. One such method is the Speech to Reverberation Modulation energy Ratio (SRMR) [20] which aims to predict the intelligibility of reverberated speech from the ratio between low and high modulation frequency energy. The SRMR has been shown to outperform a number of existing measures [20]. While originally formulated to predict the intelligibility of reverberant signals, the authors have later used the measure successfully to predict the intelligibility of noisy and processed signals [5]. A similar measure, the average modulation-spectrum area (ModA) [21], aims to predict the intelligibility of reverberated speech from the area of the modulation spectrum. This measure has been shown to compare favorably to other non-intrusive methods across a range of conditions spanning reverberation, additive noise, and distortion [5].

Another means to obtain non-intrusive SIP methods, is to estimate the output of existing intrusive methods, without using a clean reference signal. This can be done using machine learning, or by using noise reduction to estimate the clean signal from the degraded one. For instance, [22] uses a twin Hidden Markov Model (HMM) to estimate the STOI measure, while [23] uses tree based regression to predict both the STOI and PESQ [24] measures. A semi-non-intrusive method for hearing aids, using beamforming to estimate the clean signal, is proposed in [4] .

In the present paper we propose a fully non-intrusive version of the STOI measure. The proposed measure estimates envelopes of the clean reference signal from the degraded signal by use of a statistical clean speech model. The measure requires training with clean speech, but does not require training with particular interferer- or processing types. The remainder of the paper progresses as follows: Sec. 2 describes the proposed measure, Sec. 3 evaluates the measure and compares it to a number of existing intelligibility measures, and Sec. 4 concludes upon the presented findings.

2. THE NI-STOI MEASURE

We describe the proposed Non-Intrusive STOI (NI-STOI) measure, which is similar to the original (intrusive) STOI measure [16]. The STOI measure assumes intelligibility to be related to the correlation between clean and degraded 1/3-octave band amplitude envelopes. However, as we do not assume a clean reference signal to be available, we estimate the clean speech envelopes from the degraded speech envelopes. This is done by use of a statistical model of clean speech. An overview of the NI-STOI measure is given in Fig. 1.

2.1. Generating a Clean Speech Model

To distinguish between speech and noise/distortion, we generate a modulation domain model of clean speech. This model is generated on the basis of a long clean speech signal, $x^{\mathrm{c}}(t)$, i.e. long enough to be considered representative of speech in general. Silent parts of the signal are removed with a Voice Activity Detector (VAD), and the signal is resampled to 10 kHz, as for the original STOI measure [16]. The resulting signal is Time Frequency (TF) decomposed with a short time Discrete Fourier Transformation (DFT) as specified in [16]. Let $\hat{x}^{\mathrm{c}}(k,m) \in \mathbb{C}$ denote the kth DFT-coefficient of the mth window.

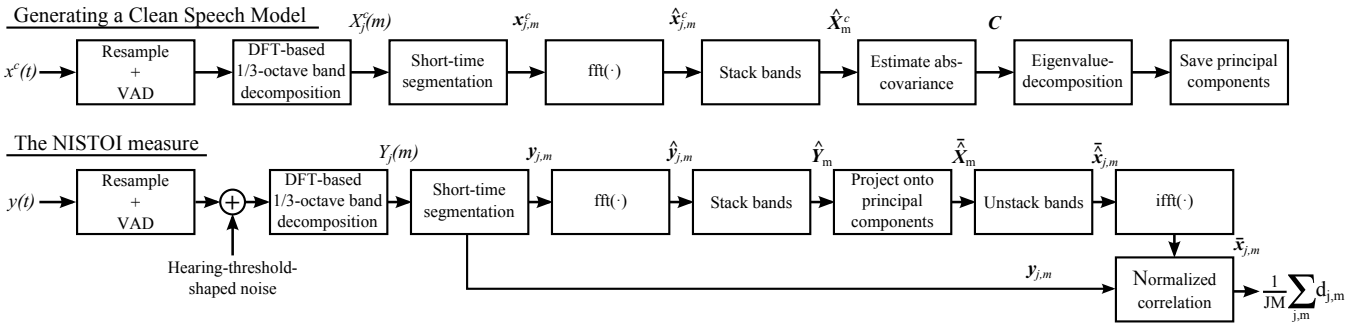

Fig. 1. A block diagram of the proposed non-intrusive intelligibility measure. The top illustrates how a clean speech model is generated from a sequence of clean speech, $x_c(t)$. The bottom illustrates how the intelligibility measure is evaluated for a degraded signal, $y(t)$.

We then extract $J = 15$ 1/3-octave band envelopes from the TF-decomposed signal as follows [16]:

$$X_j^{\mathrm{c}}(m) = \sqrt{\sum_{k_1(j)}^{k_2(j)} |\hat{x}^{\mathrm{c}}(k,m)|^2}, \tag{1}$$

where $k_1(j)$ and $k_2(j)$ are the lower and upper bounds of the jth 1/3-octave band. The resulting envelope samples are arranged in vectors of $N = 30$ samples:

$$\mathbf{x}_{j,m}^{\mathrm{c}} = \left[X_j^{\mathrm{c}}(m-N+1),...,X_j^{\mathrm{c}}(m)\right]^T, \tag{2}$$

which are normalized to have zero mean and unit norm:

$$\tilde{\mathbf{x}}_{j,m}^{\mathrm{c}} = \frac{\mathbf{x}_{j,m}^{\mathrm{c}} - \mathbf{1}\mu_{\mathbf{x}_{j,m}^{\mathrm{c}}}}{||\mathbf{x}_{j,m}^{\mathrm{c}}||}, \tag{3}$$

where $\mu_{(\cdot)}$ denotes the mean of entries in a vector and $\mathbf{1}$ is a vector of ones. Let $\hat{\mathbf{x}}_{j,m}^{\mathrm{c}}$ denote the DFT of $\tilde{\mathbf{x}}_{j,m}^{\mathrm{c}}$, i.e. the modulation domain representation of the signal. We then stack modulation domain representations for all frequency bands into one vector:

$$\hat{\mathbf{X}}_m^{\mathrm{c}} = \left[\hat{\mathbf{x}}_{1,m}^{\mathrm{c}T},...,\hat{\mathbf{x}}_{J,m}^{\mathrm{c}T}\right]^T \in \mathbb{R}^{JN\times 1}. \tag{4}$$

The transition from (2) to (4) is illustrated on Fig. 1 by the blocks "fft(·)" and "Stack bands". We use the resulting vectors to estimate an amplitude covariance matrix for all modulation frequencies across all frequency bands:

$$\mathbf{C} = \frac{1}{M}\sum_{m=N}^{M+N-1} |\hat{\mathbf{X}}_m^{\mathrm{c}}||\hat{\mathbf{X}}_m^{\mathrm{c}}|^T, \tag{5}$$

where M is the number of frames, and $|\cdot|$ denotes the absolute value which is evaluated on an entry-wise basis for vectors. As we show in Sec 3, matrix $\mathbf{C}$ can be approximated well by a low rank matrix. This property can be used to distinguish between speech and non-speech components of degraded speech envelopes. We compute the eigenvalue decomposition of $\mathbf{C}$, resulting in a descending sequence of eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_{JN}$, and corresponding eigenvectors, $v_1, v_2, \ldots, v_{JN}$. In the following section we use the principal components, $v_1, \ldots, v_K$, with $1 \leq K \leq JN$, to estimate the envelopes of the unknown clean reference signal.

2.2. Computing the NI-STOI Measure

We now describe the computation of the proposed NI-STOI measure. This is done as for the original STOI measure [16], except that the clean envelope samples are estimated from the degraded ones, because only the degraded signal, $y(t)$, is assumed known. Silent regions of the signals are removed, by use of the same VAD as for the original STOI measure. While this *does* make use of the clean reference signal, $x(t)$ it ensures comparability with the original STOI measure. Then, the degraded signal is resampled to 10 kHz. At this stage we add a faint noise signal, shaped such that the energy in each 1/3-octave band corresponds to the average hearing threshold in quiet [25] (similar to what is done in e.g. [13]). This has shown necessary in conditions where aggressive speech processing renders the presented signal almost inaudible. The noise has little impact on predictions at normal speech levels. A TF decomposition, carried out as described in Sec. 2.1, results in DFT coefficients of the degraded signal, $\hat{y}(k,m)$. Using these, we define envelope samples, $Y_j(m)$, similar to (1), normalized envelope vectors, $\tilde{\mathbf{y}}_{j,m}$, similar to (3), and modulation domain vectors, $\hat{\mathbf{Y}}_m$, similar to (4). We then construct an estimate of the corresponding clean signal modulation vector, $\hat{\mathbf{X}}_m$, by assuming: 1) the phase of $\hat{\mathbf{X}}_m$ is the same as the phase of $\hat{\mathbf{Y}}_m$, and 2) the magnitude of $\hat{\mathbf{X}}_m$ can be approximated by projecting the magnitude of $\hat{\mathbf{Y}}_m$ into the space spanned by the K clean signal principal components, $v_1,...,v_K$, found in Sec. 2.1. These assumptions lead to the following estimate of $\hat{\mathbf{X}}_m$:

$$\hat{\bar{\mathbf{X}}}_m = e^{j\angle\hat{\mathbf{Y}}_m} \odot \sum_{k=1}^{K} v_k v_k^T |\hat{\mathbf{Y}}_m|, \tag{6}$$

where $e^{j\angle\hat{\mathbf{Y}}_m}$ is a vector in which all entries have the same phase as $\hat{\mathbf{Y}}_m$, but unit magnitude, while $\odot$ denotes entry-wise multiplication (Hadamard product). The resulting estimate is split into J vectors, $\hat{\bar{\mathbf{x}}}_{1,m},...,\hat{\bar{\mathbf{x}}}_{J,m}$, of length N, corresponding to the inverse of the operation described in (4). By computing the DFT of these vectors, we obtain estimates, $\bar{\mathbf{x}}_{1,m},...,\bar{\mathbf{x}}_{J,m}$, of the clean signal envelopes. From this point, we can compute the (intrusive) STOI measure, using the estimated clean speech envelopes in place of the true ones. To do this, we compute the correlation between the (estimated) clean and degraded envelopes [16][1]:

$$d_{j,m} = \frac{\left(\bar{\mathbf{x}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{x}}_{j,m}}\right)^T \left(\mathbf{y}_{j,m} - \mathbf{1}\mu_{\mathbf{y}_{j,m}}\right)}{||\bar{\mathbf{x}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{x}}_{j,m}}||\,||\mathbf{y}_{j,m} - \mathbf{1}\mu_{\mathbf{y}_{j,m}}||}. \tag{7}$$

[1]The original STOI measure includes a clipping stage which serves to limit the extent to which a single frame can be detrimental to overall predicted intelligibility, in cases where there is very little, or negative, correlation between the clean and degraded envelopes. We have chosen not to include this stage in the NI-STOI measure. The removal of the clipping mechanism has previously been shown not to decrease performance markedly [26, 3].

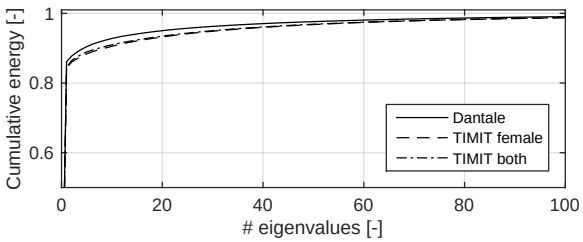

Fig. 2. The normalized cumulative sum of eigenvalues of $\mathbf{C}$ when generated with each of the three different clean speech corpora.

The NI-STOI measure is then computed, as described in [16], as the average normalized correlation between clean and degraded envelopes:

$$\text{NI-STOI} = \frac{1}{JM}\sum_{j,m} d_{j,m}. \tag{8}$$

In order to carry out direct predictions of intelligibility, in terms of a percentage of correctly answered words, the output of the NI-STOI measure is transformed with a logistic function [16]:

$$\bar{s}(x) = \frac{100\%}{1+e^{ax+b}}, \tag{9}$$

where x is the input NI-STOI measure and $\bar{s}$ is the estimated intelligibility in percent. The coefficients a and b are fitted to available data by maximum likelihood (as described in [27]).

3. RESULTS AND DISCUSSION

In this section we first show results from the training of a number of clean speech models. We then evaluate the proposed non-intrusive intelligibility measure by using it to predict the results of a listening experiment.

3.1. Clean Speech Models

To investigate the dependence on clean speech material, we train clean speech models as described in Sec. 2.1, using three different sources of clean speech: 1) all the sentences from the Dantale II corpus [28], 2) all sentences with female speakers from the TIMIT training corpus [29], 3) all sentences, male and female speakers, from the TIMIT training corpus.

Fig. 2 shows the cumulative sum of descending eigenvalues for the three models. For all three models, the majority of the energy in the modulation magnitude spectra can be accounted for by a single principal component. This can be seen as an indication that the modulation domain representation is a strong starting point for low dimensional representations of speech. Contrarily, it should be noted that this representation includes neither the phase of the speech signal nor the phase of the envelopes, and therefore cannot be used to reconstruct the original speech signal. Fig. 3 shows the first six principal components, obtained by training with the Dantale II speech material. Here, the first component, v_1, is most important, as it codes for the majority of the modulation energy. The shape of this component indicates that most of the modulation energy is contained at low modulation frequencies (i.e. less than 10 Hz). This fits well with the rationale of the SRMR measure which considers low frequency modulations to be carriers of speech information.

3.2. Experimental Data

To evaluate the performance of the proposed measure, we use it on a set of data [30] which was also used for evaluating the original STOI measure [16]. Intelligibility was measured for 15 normal hearing subjects, using the Dantale II sentence material [28]. Four types of noise was used: bottling factory hall noise, café noise, car noise, and Speech Shaped Noise (SSN), each presented at three different Signal to Noise Ratios (SNRs). The noisy signals were processed with Ideal Binary Masks (IBMs) and Target Binary Masks (TBMs) at eight different Relative Criterion (RC) values[2] [30]. Two sentences were presented with each combination of the above for a total of 15 subjects $\times$ 7 noise/mask combinations $\times$ 8 RC values $\times$ 3 SNRs $\times$ 2 repetitions $=$ 5040 sentences. The dataset is described in detail in [30].

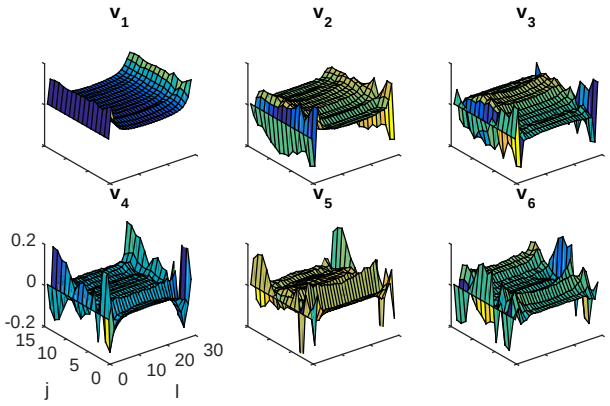

Fig. 3. The first six principal components, obtained by training with clean speech from the Dantale II corpus. The axes, as shown on the lower left plot, are identical for all the plots. The j-axis denotes the 15 1/3-octave bands, while the l-axis denotes the 30 modulation frequency bins resulting from the DFT of an envelope vector. Note that the plots are symmetric on the l-axis.

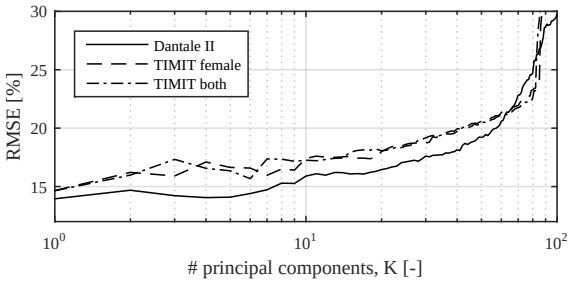

Fig. 4. Prediction performance in terms of RMSE, of NI-STOI, vs., K, for each of the three clean speech models. The conditions with café noise were excluded in this analysis. See the text for details.

3.3. Predictions

We first consider the overall performance of the NI-STOI measure, and its dependency on the number of principal components, K, and the clean speech material used for training. The conditions with café noise are not included in this analysis, for reasons which are discussed later. Fig. 4 shows the Root-Mean-Square Error (RMSE) of predictions versus K. The best performance is obtained with the Dantale II speech model. This indicates that, in spite of the simplicity of the applied speech model, some

[2]The RC value is an algorithm parameter which determines the density of the computed binary mask. See [30] for details.

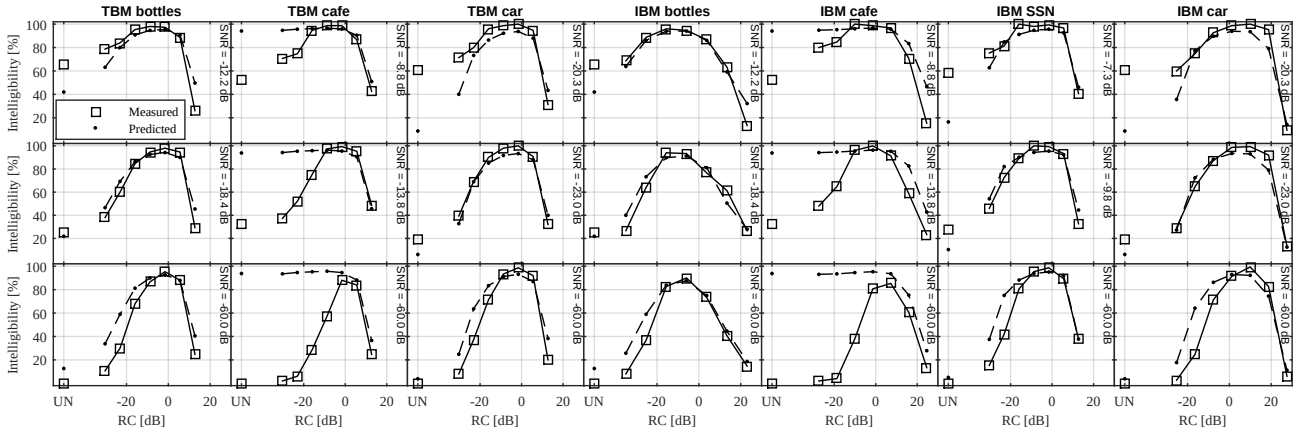

Fig. 5. NI-STOI predictions with $K = 1$, compared with measured results. The clean speech model based on Dantale II sentences was used, and the logistic mapping function has been fitted to all measured data except the conditions with café noise. Columns correspond to different noise- and processing types, while rows correspond to different input SNRs, decreasing downwards. The horizontal axis shows the RC-value of the processing algorithm. A low RC corresponds to mild processing while high RC corresponds to heavy processing. Unprocessed conditions are denoted "UN".

degree of talker specific modeling can occur. Contrarily, the TIMIT-based speech models perform about equally well. This suggests that the models do not capture gender specific effects. This may be due to the absence of pitch information in the speech representation. With regards to the number of principal components, K, there is a clear tendency that more components lead to poorer performance. While performance is relatively constant for K between one and ten, the best performance is obtained for $K = 1$. This corresponds well with the observation, supported by Fig. 2, that most of the modulation energy is captured by a single principal component. Adding more components adds rather little clean speech information, but may let more noise and distortion into the clean envelope estimate.

Fig. 5 shows NI-STOI predictions for the individual conditions. Again, the logistic mapping function was fitted without the café noise conditions. The figure indicates a good overall fit between predictions and measurements. Large deviations are seen for the conditions with café noise at low RC-values (corresponding to little or no processing), where intelligibility is predicted to be very high, while in fact it is rather low. It should be noted that the café noise consists mainly of a single interfering female talker. Since the NI-STOI measure is non-intrusive, it has no means of determining which speaker is the target one (assuming that the clean speech model is not talker specific). Therefore, one can argue that any non-intrusive SIP method is bound to fail in such a condition, unless supplied with additional information about which speaker is the target one. This also explains why the overall quality of the logistic mapping can be increased by excluding the café noise conditions during fitting.

In Table 1, we evaluate the proposed measure against a number of existing measures. Being intrusive, the STOI measure outperforms the NI-STOI measure in all cases. However, when excluding the café noise conditions, the NI-STOI performance comes somewhat close to that of the STOI measure. There are no major differences between the results for the three different clean speech models. We also compare to two variations of the SRMR measure which the authors have kindly made available to the public: 1) the original SRMR measure [20], and 2) a later version of the measure which has been improved with the aim of lowering the output variability [31]. While these measures are mainly aimed at predicting intelligibility of reverberated speech, they have been successfully applied for noisy and processed speech [5]. As shown in Table 1, the improved SRMR measure outperforms the original

	Pearson correlation	RMSE	Kendall's τ	
STOI [16]	0.958	9.5%	0.824	÷ café
NI-STOI (Dantale II)	0.907	13.9%	0.777	
NI-STOI (TIMIT F)	0.897	14.6%	0.764	
NI-STOI (TIMIT M+F)	0.897	14.6%	0.768	
SRMR [20]	0.311	42.0%	0.207	
SRMR-norm [31]	0.550	31.6%	0.388	
STOI [16]	0.959	9.4%	0.822	All conds.
NI-STOI (Dantale II)	0.711	25.2%	0.529	
NI-STOI (TIMIT F)	0.704	25.4%	0.516	
NI-STOI (TIMIT M+F)	0.702	25.5%	0.513	
SRMR [20]	0.237	45.2%	0.036	
SRMR-norm [31]	0.394	38.6%	0.156	

Table 1. Comparison of intelligibility measures with and without café noise. In both cases, the logistic mapping function was fitted without café noise. The NI-STOI measure was used with $K = 1$.

one, especially when the café noise conditions are excluded. However, the NI-STOI measure also consistently outperforms both measures. The higher performance of the NI-STOI measure is especially pronounced in the absence of the café noise conditions. This result should, however, be viewed in the light of the fact that the NI-STOI measure is trained with clean speech material, while the SRMR measure is not trained or fitted in any manner (except for the logistic mapping, (9)).

4. CONCLUSIONS

We have proposed a non-intrusive intelligibility measure based on the Short-Time Objective Intelligibility (STOI) measure. Similar to the original STOI measure, the proposed measure is aimed at predicting the intelligibility of noisy and non-linearly processed speech. The model estimates unknown clean speech envelopes from degraded envelopes, by use of a clean speech model. The performance of the proposed measure was evaluated with a dataset consisting of speech in different types of noise processed with binary masks. This indicated that the proposed measure performs better than an existing non-intrusive measure, but not as good as the original (intrusive) STOI measure.

5. REFERENCES

[1] S. van Wijngaarden, "The speech transmission index after four decades of development," *Acoustics Australia*, vol. 40, no. 2, pp. 134–138, Aug. 2012.

[2] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[3] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 1908–1920, 2016.

[4] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *The European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, Aug. 2016, pp. 1358–1362, EURASIP.

[5] 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," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

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

[7] ANSI Std. S3.5-1997, "Methods for calculation of the speech intelligibility index," 1997.

[8] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

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

[10] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[11] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[12] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[13] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[14] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

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

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

[17] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363–1374, Sept. 2014.

[18] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5365–5369, IEEE.

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

[20] T. H. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[21] 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, pp. 311–314, Dec. 2013.

[22] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin HMM-based non-intrusive speech intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Sept. 2016, pp. 624–628, IEEE.

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

[24] "P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[25] B. C. Moore, *An introduction to the psychology of hearing*, Brill, sixth edition, 2013.

[26] Cees H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in coclear implants based on an intelligibility metric," in *The European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 504–508, EURASIP.

[27] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, June 2002.

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

[29] DARPA, "TIMIT, acoustic-phonetic continuous speech corpus," .

[30] 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, pp. 1415–1426, Sept. 2009.

[31] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan les Pins, France, Sept. 2014, pp. 55–59, IEEE.