INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On the use of Band Importance Weighting in the Short-Time Objective Intelligibility Measure

Asger Heidemann Andersen[1,2], *Jan Mark de Haan*[2], *Zheng-Hua Tan*[1], *Jesper Jensen*[1,2]

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

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

Abstract

Speech intelligibility prediction methods are popular tools within the speech processing community for objective evaluation of speech intelligibility of e.g. enhanced speech. The Short-Time Objective Intelligibility (STOI) measure has become highly used due to its simplicity and high prediction accuracy. In this paper we investigate the use of Band Importance Functions (BIFs) in the STOI measure, i.e. of unequally weighting the contribution of speech information from each frequency band. We do so by fitting BIFs to several datasets of measured intelligibility, and cross evaluating the prediction performance. Our findings indicate that it is possible to improve prediction performance in specific situations. However, it has not been possible to find BIFs which systematically improve prediction performance beyond the data used for fitting. In other words, we find no evidence that the performance of the STOI measure can be improved considerably by extending it with a non-uniform BIF.
Index Terms: band importance function, speech intelligibility prediction, enhanced speech, speech in noise

1. Introduction

Speech Intelligibility Prediction (SIP) methods are increasingly being used by the speech processing community in lieu of time consuming and expensive listening experiments. Such methods can provide quick and inexpensive estimates of speech intelligibility in conditions where speech is subjected to e.g. additive noise, reverberation, distortion or speech enhancement. An early SIP method is the Articulation Index (AI) [1] which was proposed for the purpose of evaluating the intelligibility of speech transmitted via telephone. A more recent, improved and standardized, version of the AI is known as the Speech Intelligibility Index (SII) [2]. Further modifications of the SII have been proposed with aims of handling e.g. fluctuating masker signals [3, 4], non-linearly distorted speech [5], and binaural signals [6, 7]. More recently, the multi-resolution speech-based Envelope Power Spectrum Model (mr-sEPSM) has received attention for its physiological basis and its ability to predict intelligibility accurately across a wide range of conditions including reverberation, fluctuating maskers, and noise suppression [8]. The Short-Time Objective Intelligibility (STOI) [9] measure has recently gained popularity in the speech processing community. While the measure is simple and easy to use, it has also proven to predict intelligibility accurately in many conditions including e.g. additive noise, speech enhancement [9, 10], distortion from transmission via telephone [11], and hearing impairment [12]. Several variations of the STOI measure with various purposes and properties have recently been proposed [13, 14, 15, 16].

All of the above mentioned methods are roughly characterized by the same procedure: 1) split the involved speech signal into narrow frequency bands with a filterbank, thus mimicking the frequency selectivity of the basilar membrane, 2) estimate the amount of speech information conveyed in each frequency band, and 3) sum the information from all frequency bands, using some relative weighting that reflects how speech information is distributed across frequency. The frequency weighting function used in the third step is often termed a Band Importance Function (BIF). A BIF for the AI is determined in [1] by use of a graphical procedure, based on measured intelligibility of Highpass (HP) and Lowpass (LP) filtered noisy speech. Such BIFs are also used in the more recent SII [2]. The use of these has since spread to other SIP methods which are based on the SII [5, 3, 4, 6, 7]. The advent of modern computing has allowed fitting of BIFs, such as to maximize prediction accuracy for particular datsets of measured intelligibility [17]. Lastly, some authors have proposed SIP methods which use signal dependent BIFs, which are computed such as to reflect the instantaneous information distribution of speech across frequency [18, 14].

The STOI measure distances itself from other measures by being designed with a strong focus on simplicity, and therefore does not include any BIFs [9]. Instead, the STOI measure uniformly averages contributions from 15 one-third octave bands. The designers of the STOI measure [9] made this decision purely with the aim of simplicity, and do not report the effect of this decision (with the exception of noting that the resulting measure has a high performance, in spite of the uniform BIF). However, given the importance of BIFs assumed by other SIP methods, it appears likely that the performance of the STOI measure can be improved by extending it with a suitable BIF.

In this paper we investigate the effect of extending the STOI measure with fitted BIFs. In Sec. 2 we describe the STOI measure, including the modification of including BIFs, and following a similar approach given in [17], we describe how BIFs are fitted such as to minimize the prediction error for datasets of measured intelligibility. In Sec. 3 we describe the two datasets of measured intelligibility which we use for fitting BIFs. These datasets are further divided into different subsets. In Sec. 4 we investigate fitted BIFs for the different subsets of measured intelligibility. Sec. 5 concludes upon our findings.

2. Methods

In this Section we outline the concepts we apply in investigating the use of BIFs together with the STOI measure.

2.1. The STOI Measure

The STOI measure estimates the intelligibility of a degraded speech signal, $y(t)$, by comparing it to a clean reference signal, $x(t)$. Both signals are resampled to 10 kHz and silent regions are removed by use of an ideal Voice Activity Detector (VAD) [9]. The signals are Time Frequency (TF) decomposed by use of a short time Discrete Fourier Transformation (DFT) (see details in [9]). Let the degraded signal DFT coefficient of the kth frequency bin and the mth frame be denoted $\hat{y}(k, m)$, and the corresponding clean signal DFT coefficient be denoted by $\hat{x}(k,m)$. Envelopes for each of $J = 15$

http://dx.doi.org/10.21437/Interspeech.2017-1043

one-third octave bands are extracted from the DFT coefficients [9]:

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

where $k_1(j)$ and $k_2(j)$ denotes, respectively, the lower and upper bounds of the jth one-third octave band. The one-third octave bands have center frequencies from 150 Hz and upwards in one-third octave steps. Corresponding envelope samples, $Y_j(m)$, are defined for the degraded signal. The resulting envelope samples are arranged in vectors of $N = 30$ samples [9]:

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

Corresponding vectors, $\mathbf{y}_{j,m}$ are defined for the degraded signal. We define a normalized and clipped version of $\mathbf{y}_{j,m}$, such as to minimize the sensitivity of the method to severely degraded TF-units [9]:

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{||\mathbf{x}_{j,m}||}{||\mathbf{y}_{j,m}||}\mathbf{y}_{j,m}(n), (1+10^{-\beta/20\text{dB}})\mathbf{x}_{j,m}(n)\right), \quad (3)$$

for $n = 1,...,N$, where $\beta = 15$ dB is a lower bound on signal-to-distortion-ratio [9]. The resulting short-time envelope vectors, $\mathbf{x}_{j,m}$ and $\bar{\mathbf{y}}_{j,m}$ are used to define intermediate correlation coefficients [9]:

$$d_{j,m} = \frac{\left(\mathbf{x}_{j,m} - \mathbf{1}\mu_{\mathbf{x}_{j,m}}\right)^T \left(\bar{\mathbf{y}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{y}}_{j,m}}\right)}{||\mathbf{x}_{j,m} - \mathbf{1}\mu_{\mathbf{x}_{j,m}}|| \, ||\bar{\mathbf{y}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{y}}_{j,m}}||}, \quad (4)$$

where $\mathbf{1}$ is a vector of ones, and $\mu_{(\cdot)}$ denotes the sample mean of a vector. The STOI measure is then obtained as the average of $d_{j,m}$ across all values of j and m [9]. This implies a uniform weighting (BIF) for all one-third octave bands j. In this paper, to allow for different BIFs, we instead define bandwise average correlations:

$$\tilde{d}_j = \frac{1}{M}\sum_m d_{j,m}, \quad (5)$$

where M is the number of time frames. These are averaged with the BIF $\mathbf{w} = [w_1, \ldots, w_J]^T$, to obtain the final frequency weighted STOI score:

$$s = \sum_{j=1}^{J} w_j \tilde{d}_j, \quad (6)$$

where $w_j \geq 0$ for $j = 1,...,J$ and $\sum_{j=1}^{J} w_j = 1$.

The resulting STOI score is a number in the range from 0 to 1, where a higher STOI score indicates higher intelligibility (e.g. percentage of words understood correctly). In order to transform the STOI score into a direct estimate of intelligibility in %, a logistic mapping is applied [9]:

$$f(s;a,b) = \frac{100\%}{1+\exp(as+b)}, \quad (7)$$

where a and b are fitted such as to maximize prediction accuracy on a well-defined dataset of measured intelligibility.

2.2. Fitting of Band Importance Functions

We now turn to the determination of the BIF, $\mathbf{w}$. We determine this, such as to minimize the prediction error in terms of Root-Mean-Square Error (RMSE). This is heavily inspired by the approach taken in [17] (which fits RMSE optimal weights for the SII). Specifically, we assume that speech intelligibility has been measured in L conditions (e.g. different types of reverberation, distortion or processing at different Signal to Noise Ratios (SNRs)), and is given by $p(l)$, $l = 1\ldots,L$, where $0\% \leq p(l) \leq 100\%$ is the average fraction of correctly repeated words. We furthermore assume that samples of clean and degraded speech are available for each condition, such that we may compute bandwise average correlations, $\tilde{d}_j(1),...,\tilde{d}_j(L)$, with $j = 1,...,J$, for each condition, using (5). For a given BIF, $\mathbf{w}$, we can compute a weighted STOI score for each condition, by (6). We can further transform this score into a direct prediction of intelligibility by (7). The RMSE of this prediction can be written as:

$$\text{RMSE}(\mathbf{w},a,b) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(p(l) - f\left(\sum_{j=1}^{J} w_j \tilde{d}_j(l);a,b\right)\right)^2}. \quad (8)$$

We jointly determine a, b and $\mathbf{w}$ such as to minimize the RMSE, as given by (8):

$$\begin{aligned} &\underset{a,b,\mathbf{w}}{\text{minimize}} \quad \text{RMSE}(\mathbf{w},a,b) \\ &\text{subject to} \quad \sum_{j=1}^{J} w_j = 1 \quad \text{and} \quad w_j > 0, \quad j = 1,...,J. \end{aligned} \quad (9)$$

This optimization problem is non-convex and we are not aware of a method to solve it analytically. Instead, we apply the MATLAB Optimization Toolbox to numerically find solutions which are locally optimal.

3. Experimental Data

We use two datasets of measured intelligibility to investigate the fitting of BIFs according to (9), and to compare the resulting prediction performance with that of the original STOI measure.

3.1. The "Kjm" dataset [19]

The first dataset was used in the initial evaluation of the STOI measure [9] and is described in detail in [19]. For this dataset, intelligibility was measured for 15 normal hearing Danish subjects using the Dantale II corpus [20]. Measurements were carried out for 1) four noise types: Speech Shaped Noise (SSN), café noise, bottling factory noise and car noise 2) processing by two types of binary masks, Ideal Binary Masks (IBMs) and Target Binary Masks (TBMs), 3) eight different threshold values for binary mask generation and 4) three different SNRs. Since IBMs and TBMs are identical for SSN, there are only seven combinations of noise types and binary masks. The three SNRs were chosen individually for each noise type. Intelligibility was measured for a total of: 15 subjects $\times$ 7 noise/mask combinations $\times$ 8 RC values $\times$ 3 SNRs $\times$ 2 repetitions $\times$ 5 words/sentence $=$ 25200 words. By averaging performance across subjects, repetitions and words, we obtain measured intelligibility for 168 conditions. The authors of [19] have kindly supplied both clean and degraded audio files for the conditions.

For this study, the data is divided into eight subsets such as to investigate the BIFs arising from fitting to different types of data. Firstly, the dataset is divided into four subsets depending on noise type. Secondly, the dataset is divided according to the three SNR conditions (low, medium and high). Lastly, one subset is defined to include all the data. We refer to these subsets with the label "Kjm".

3.2. The "S&S" dataset [21]

The second dataset [21] was collected in an effort to derive BIFs for the AI. Speech intelligibility was measured for 8 normal hearing subjects using a recording of the CID W-22 word lists. Measurements

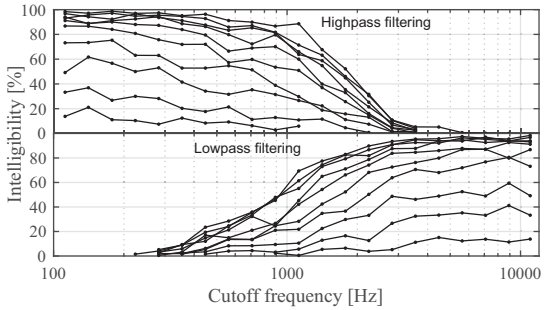

Figure 1: *Replotted experimental results, as reported in tables 2–3 of [21]. The top plot shows measured intelligibility of HP filtered noisy speech versus cutoff frequency. Each line represents measurements at a particular SNR. The bottom plot shows the same type of results for LP filtering.*

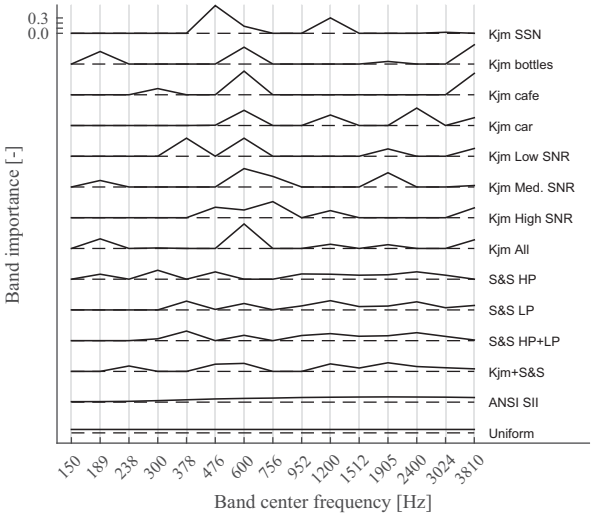

Figure 2: *Fitted BIFs for eight subsets of the "Kjm' data, three subsets of the "S&S" data, one set including data for both experiments, as well as two non-fitted standard BIFs. The scaling of the vertical axes is the same for all BIFs.*

were carried out for 1) HP and LP filtered speech masked by SSN, 2) 21 filter cutoff frequencies and 3) 10 different SNRs. SNRs were uniformly spaced in 2 dB intervals between -10 and +8 dB. In total, this amounts to 2 filter types (HP/LP) $\times$ 21 cutoff frequencies $\times$ 10 SNRs $=420$ conditions. However, some conditions were skipped because intelligibility was almost zero, and therefore only 308 conditions were measured [21]. The results are shown in Figure 1.

It has not been possible to obtain either clean or degraded speech for the conditions of this experiment. Nor has it been possible to obtain recordings of the CID W-22 word lists. We therefore recreated similar stimuli as accurately as possible, in order to allow for computing STOI scores. To this end, 150 random sentences were selected from the TIMIT database [22] and concatenated. Both HP and LP filtering was carried out using 512th order linear phase Finite Impulse Response (FIR) filters, designed using the windowing method. SSN was generated by filtering white noise such as to have the same long time spectrum as the TIMIT sentences. The concatenated, non-filtered, TIMIT sentences were used as a clean reference signal, $(x(t))$, while filtered speech, mixed with SSN, was used as degraded speech $(y(t))$. The SNR is defined to be the energy ratio of speech and noise *before* filtering the speech (as in [21]).

We define three divisions of this dataset: 1) the conditions with HP filtering, 2) the conditions with LP filtering, and 3) all the data. We refer to these subsets with the label "S&S". We also define one set of data, "Kjm+S&S", which includes all data from both experiments.

3.3. ANSI SII- and Uniform BIFs

In addition to BIFs fitted with (9), we include two additional BIFs: 1) The BIF specified for use with the SII in Table 3 of [2]. Linear interpolation was used to determine a BIF for the exact center frequencies of the one-third octave bands of the STOI measure. This BIF, shown in Figure 2, places increased weight on the higher frequency bands, as compared to the uniform BIF. 2) A uniform BIF, as used in the original STOI measure [9], i.e. $w_j = 1/J, \quad j = 1,...,J$.

4. Results and Discussion

BIFs were fitted to the defined subsets of data by finding local minima for (9), using the `fminsearch`-solver in the MATLAB Optimization Toolbox[1]. The resulting BIFs are shown in Figure 2. Most strikingly, all BIFs fitted to subsets of the "Kjm"-data place the majority of the weight on few frequency bands. The heavily weighted bands are not the same across the BIFs (except for band 7, which is consistently weighted strongly by all "Kjm"-BIFs except the one fitted to the SSN conditions). Such solutions could indicate some degree of overfitting, and it should be remarked that the smaller subsets of the "Kjm"-data involve only 24, 48 or 56 data points, to which 17 parameters are fitted (i.e. a, b and $\mathbf{w} \in \mathbb{R}^{15\times 1}$). However, the full set of all 168 data points of the "Kjm"-data results in a BIF with similar properties. It should also be noted that while the BIFs place most weight on a few bands, these few bands are generally spread out across the entire frequency range. Another explanation of the sparse BIFs could therefore be that the values of $\tilde{d}_j$ are highly correlated for adjacent bands, and thus supply redundant information. It is possible that smoother BIFs can be obtained by adding some form of regularization to (9).

The BIFs fitted to the subsets of the "S&S"-data appear much smoother than those fitted to the "Kjm"-data. At the same time, the "S&S"-BIFs are similar to one-another. Especially the BIFs fitted to the "S&S LP"- and the "S&S LP+HP"-subsets show some similarity to the SII BIF, by weighting the higher frequency bands slightly higher than the lower ones. The joint set of data from both experiments, "Kjm+S&S", leads to a BIF which is quite similar to the one fitted to the "S&S HP+LP"-data. This could indicate that the RMSE of the "S&S"-data is more sensitive to differences in BIFs, and that this dataset therefore ends up having the most influence on the optimal BIF. This is not surprising, as the "S&S"-data is designed specifically with the purpose of containing as much information as possible about which frequency bands are important to speech intelligibility (i.e. to facilitate the derivation of BIFs).

We evaluate the performance of all 14 BIFs on all 12 subsets of data, using two different performance metrics: 1) RMSE, and 2) Kendall's tau. The results are shown in color-coded tables in Figure 3.

[1]The default solver was initialized 100 times with random starting values, and the best solution across these was used.

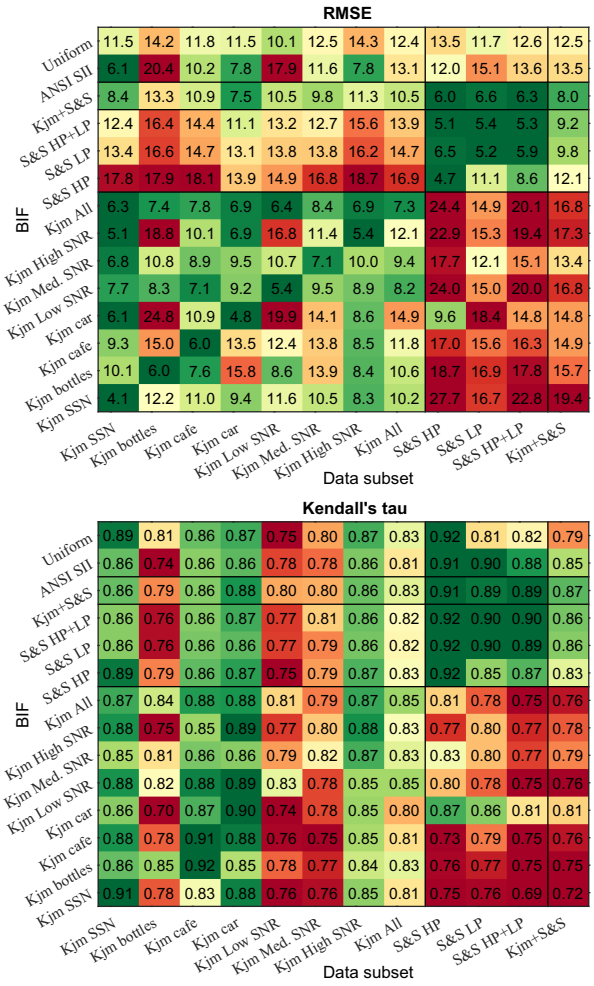

Figure 3: *Cross evaluation of all the BIFs with the 12 defined subsets of data. Each row shows the performance for one BIF, when evaluated on the different subsets of data. Each column shows the performance of the different BIFs when evaluated for one particular data subset. The top plot shows RMSE in % and the bottom plot shows Kendall's tau. Red colors indicate poorer than average performance and green colors indicate better than average performance.*

We first consider performance in terms of RMSE, as given by the top plot of Figure 3. Each fitted BIF is optimized to minimize the RMSE on one particular dataset. This is seen in Figure 3 as a diagonal with high performance, projecting from the lower left corner. It can be noted that BIFs fitted on one subset of the "Kjm"-data often leads to a low RMSE when used on another subset of the "Kjm"-data, with some exceptions. This contradicts the notion of overfitting being a major problem with the small subsets of the "Kjm"-data. A similar observation holds for the "S&S"-data, where rather good performance is obtained regardless of which BIF is evaluated for what subset of data. In general it appears that lower RMSE can be obtained on the"S&S"-data, which suggests that this dataset contains either less statistical variation or less varied combinations of noise and processing. When using BIFs fitted to the "Kjm"-data for predictions of the "S&S"-data, and vice versa, performance is mostly low. This suggests some fundamental difference between the two datasets, caused e.g. by differences in target speech material. However, the combined "Kjm+S&S"-BIF manages to obtain good performance across all subsets of both sets of data. The uniform- and SII BIFs also obtain decent performance across most conditions, especially when considering that these are not fitted to any of the available data. With the exception of the "Kjm+S&S"-BIF, the uniform BIF, as used in the original STOI measure, has the smallest maximum RMSE (i.e. the highest number of the row: 14.3%). However, RMSE measured on all the available data combined, as shown in the rightmost column, is lowest for the "Kjm+S&S"-BIF, by a considerable margin. All BIFs fitted on the "Kjm"-data lead to quite poor performance when evaluated for the combined data, while the "S&S"-BIFs lead to much better performance. This should be viewed in light of the fact that the "S&S"-dataset is almost twice as big as the "Kjm"-dataset and therefore weighs more in the combined performance evaluation.

One can argue that it is unfair to fit BIFs to data from one listening experiment and validate it on data from another, because the speech material may have different degrees of complexity and the different groups of subjects may not perform equally well. These factors are, to a large extent, modeled by the parameters a and b, which control the mapping from STOI measure to predicted intelligibility in percent. The bottom plot in Figure 3 shows performance in terms of Kendall's tau. This statistic is interesting because it depends only on the extent to which predictions are correctly ordered, and is therefore independent of a and b. Here, we also see that fitting and testing with the same set of data gives improved performance, but to a somewhat smaller extent than what is the case with the RMSE which is directly optimized in (9). It is also seen that poor performance results when fitting BIFs on the "Kjm"-data and evaluating on the "S&S"-data, as was also the case when measuring performance in terms of RMSE. However, the opposite is not the case: fitting BIFs on the "S&S"-data and evaluating on the "Kjm"-data leads to performance which is almost as good as what is obtained when fitting with the "Kjm All"-set. This contrasts the results seen when evaluating with RMSE, and may indicate that a and b are important for fitting details about the specific experiment, and are not transferable from one experiment to another. On the other hand, this result also indicates that the BIF, $\mathbf{w}$, fitted on the "S&S"-data actually generalizes well to the "Kjm"-data. Overall, the BIF fitted to the "Kjm+S&S"-set performs better than the uniform BIF, in terms of Kendall's tau, when evaluated on the "Kjm+S&S"-set. However, this difference seems to stem mainly from the "S&S LP"-conditions. The other conditions do not indicate that performance is improved considerably above that of the uniform BIF.

5. Conclusions

We have investigated the use of Band Importance Functions (BIFs) in the Short-Time Objective Intelligibility (STOI) measure. BIFs were fitted to several different datasets of measured intelligibility, such as to minimize the Root-Mean-Square Error (RMSE). This can decrease prediction RMSE substantially in comparison with the uniform weighting of frequency bands normally used in the STOI measure. However, when cross evaluation was carried out between different sets of data, or when performance was measured using Kendall's tau, the use of BIFs appeared to result in neither a large or a consistent improvement in performance across the evaluated conditions. It is therefore not possible to say from this limited study, whether the improved average performance generalizes to other conditions. Across most of the evaluated conditions, it appears that the uniform BIF, as applied in the original STOI measure, is nearly optimal.

6. Acknowledgements

This work was funded by the Oticon Foundation and the Danish Innovation Foundation.

7. References

[1] 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.

[2] A. S. S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index*, ANSI Std. S3.5-1997, 1997.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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, Jul. 2013.

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

[10] 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, Sep. 2014.

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

[12] 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.

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

[14] 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: IEEE, Mar. 2016, pp. 5365–5369.

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

[16] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, US: IEEE, Mar. 2017, pp. 5085–5089.

[17] J. M. Kates, "Improved estimation of frequency importance functions," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. EL459–EL464, Nov. 2013.

[18] J. Ma, Y. H. Philipos, and C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[19] 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, Sep. 2009.

[20] 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.

[21] G. A. Studebaker and R. L. Sherbecoe, "Frequency-importance and transfer functions for recorded CID W-22 word lists," *Journal of Speech and Hearing Research*, vol. 34, pp. 427–438, Apr. 1991.

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