24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

AN ASSESSMENT OF OBJECTIVE INDICATORS OF SPEECH INTELLIGIBILITY IN NOISE AT LOW SIGNAL-TO-NOISE RATIOS

Simone Graetzer and Carl Hopkins

University of Liverpool, Acoustics Research Unit, School of Architecture, Liverpool, L69 7ZN, UK
email: s.graetzer@liverpool.ac.uk

This paper concerns how well objective indicators of speech intelligibility correlate with the low percentages of words that are correctly identified when speech is embedded in high levels of noise with signal-to-noise ratios (SNRs) down to -50 dB. The indicators under consideration are Short-Time Objective Intelligibility (STOI), the Normalised Covariance Metric (NCM) and the Coherence Speech Intelligibility Index (CSII). STOI is suitable for noisy or degraded speech, including non-linear processed or time-frequency weighted speech. Unlike the NCM, which is based on the normalised covariance between the entire original and degraded envelopes, STOI involves the correlation of the envelopes of clean and degraded (or processed) speech signals that have been divided into overlapping short-time (384 ms) segments. In this paper, speech is degraded by four types of additive noise: white noise, a 400 Hz sine wave, white noise with a 400 Hz sine wave and white noise with a 400 Hz sine wave and harmonics up to 3200 Hz. Listening tests involving normal-hearing human listeners have been carried out for male and female talkers using four SNRs per noise type, ranging from -10 dB to as low as -50 dB. The results characterise the relationship between the objective indicators and the performance-based measure from the subjective tests for speech communication in noisy conditions.
Keywords: speech intelligibility, noise, STOI, NCM

1. Introduction

The speech intelligibility performance of a communication channel, such as a room, an electronic communication system, or an electroacoustic system, can be quantified using subjective or objective means. While listening tests are the most reliable method of quantifying speech intelligibility, this method is not always feasible due to the time and expense, particularly if multiple processing schemes are to be tested. Hence, objective measures that can predict intelligibility are important. Various objective methods have been developed, some based on the signal-to-noise ratio (SNR), such as the Articulation Index (AI) [1], and the Speech Transmission Index (STI) [2,3], and others based on signal correlation or covariance, such as the Short-Time Objective Intelligibility (STOI) [4] metric and the Normalised Covariance Metric (NCM) [5]. In general, these objective methods are applied to signals with SNRs $\geq$ -10 dB. The focus in this study is the evaluation of signal intelligibility at very low SNRs (down to -50 dB) for normal hearing (NH) listeners.

Unlike the AI, which is based on the SNRs within frequency bands, the STI is based on the modulation transfer function (MTF) in those bands. Hence, the STI can be used to predict intelligibility scores for signals degraded by both additive noise and reverberation and echoes. However, like the AI and the Speech Intelligibility Index (SII), the traditional and indirect STI methods tend to over- or under-estimate the intelligibility of signals that have undergone noise reduction or some other non-linear operation, such as envelope thresholding (e.g. [6]). The application of the STI to non-linear operations may require speech, or speech-like, test signals [2]. Several speech-based versions of the

STI have been developed subsequently, including the NCM [5,7]. While most speech-based methods are unable to predict performance for both conventional acoustic degradation and non-linear processing, exceptions include the correlation or coherence-based, NCM, Coherence SII (CSII) [8], and STOI [4,9] measures.

NCM is based on apparent SNRs within frequency bands that are calculated on the basis of the squared normalised covariance between the test (or 'probe') signal and response signal envelopes [5]. Goldsworthy and Greenberg [5] found that of the STI-based parameters they considered, comprising magnitude and real cross-power spectrum methods, the envelope regression method and the NCM, only the NCM method produced (qualitatively) reasonable results for both conventional acoustic degradation and non-linear processing. With new Band Importance Functions (BIFs) [10], NCM performs well for Ideal Time Frequency Segregation (ITFS) processed speech [4]. However, Taal *et al.* [4,9] showed that even when NCM and CSII are implemented with new BIFs, STOI is more highly correlated with intelligibility scores for ITFS-processed speech.

CSII was proposed as an extension of the SII to predict speech intelligibility for non-linearly processed signals [8]. The calculation is based on the normalised cross-spectral density of the clean and degraded speech signals. SNR estimates are replaced with frequency band-dependent signal-to-distortion ratio (SDR) estimates. In later work, CSII was separated into three, separate indices, $CSII_{High}$, $CSII_{Mid}$ and $CSII_{Low}$, based on the root mean square (RMS) level [11]. The high index is associated with segments at or above the overall RMS level of the sentence; the mid index, at or up to 10 dB below the same level, and the low index, from 10 to 30 dB below the level. $CSII_{Low}$ and $CSII_{Mid}$ are combined linearly and transformed with a simple logistic function to derive a new measure, termed I_3 [8].

Taal *et al.* [4,9] introduced STOI primarily to assess speech intelligibility before and after ITFS-processing with a binary mask. A very high correlation between STOI and the intelligibility of noisy and ITFS-processed speech was reported (ρ = 0.95) [9]. STOI performed better than more complex metrics including SII and $CSII_{Mid}$, and performed comparably to an advanced version of the NCM (termed *NCM+*), in characterising the intelligibility of noise-suppressed speech [12]. Applying the Fourier transform to short segments allows STOI to reflect the non-stationarity of speech and capture low-frequency modulations. In this method, the clean signal, *x*, and the degraded signal, *y*, are down-sampled to 10 kHz and divided into 384 ms segments. Short-time discrete Fourier transform (STFT) coefficients are grouped into 15 third octave bands with centre frequencies between 150 and 3810 Hz. Signal *y* is normalised and clipped at a SDR β = -15 dB. Correlations of *x* and *y* in each band and frame are averaged. Taal *et al*. [4] claim that the clipping procedure ensures that 'sensitivity of the [STOI] model towards one T-F unit which is severely degraded is upper bounded. As a consequence, further degradation of a speech T-F-unit which is already completely degraded (*i.e.*, "unintelligible") does not lead to a lower intelligibility prediction' (p. 2127). However, Taal *et al*. [4] tested STOI with $\beta = -\infty$ without normalising *y*. For their range of SNRs, β = -15 dB was optimal.

In this paper, the aim is to determine whether at very low SNRs, STOI is more highly correlated with speech intelligibility when clipping is not performed (but when normalisation of the degraded signal is performed) than when it is performed. To this end, testing of the STOI-based, NCM, and CSII metrics with NH listeners was undertaken to answer the following questions:

1. At very low SNRs, how well do STOI-based parameters characterise the intelligibility of noisy speech?
2. Is STOI without clipping at -15 dB (STOI_WC) more consistent with intelligibility scores as measured by correlation coefficients and prediction errors than traditional STOI, NCM and CSII for noisy speech at very low SNRs?

2. Method

2.1 Subjects

40 untrained listeners were recruited as subjects (20 male and 20 female). Their ages ranged from 19 to 58 years with a median age of 23 years. All listeners used British English as a first language, and had self-reported good spelling ability. Before the experiment, each subject underwent an audiometric screening test according to ISO 8253-1 [13] to determine their threshold of hearing between 125 Hz and 8 kHz. The listeners' hearing thresholds did not exceed 20 dB HL.

2.2 Stimuli and signal presentation

Previous recordings of the IEEE [14] corpus from four talkers at a normal vocal effort were used [15]. Signals were down-sampled from 48 kHz (32 bit) to 16 kHz (16 bit), and energy below 50 Hz was removed. Due to this down-sampling, the direct form of the STI could not be used in this study.

A total of 16 conditions were tested: 4 SNRs x 4 maskers. (N.B. Four additional maskers are not discussed here.) Four maskers were added to the speech signals, with 1 s of noise at each end of the speech signal: Gaussian white noise at -40 dBW (termed *WN*), a 400 Hz tone at a sound pressure level (SPL) of 70 dB (SIN), a combination of white noise and a 400 Hz tone (WSIN), and a combination of white noise, the 400 Hz tone and its harmonics up to 3200 Hz (WSINS), where each sine was created at a level of 75 dB SPL. The 400 Hz tone was of interest as Stevens *et al.* noted that sine wave maskers are more effective at frequencies between 300 and 500 Hz than at other frequencies [16]. For the maskers containing white noise, a pseudo-randomly selected segment of white noise was added to the speech signal to obtain the required SNRs [17,18]. Preliminary trials were used to identify low SNRs for each masker that were likely to give percentages of correctly identified words between 0% and 10%. Three additional SNRs at +5, +10, and +15 dB relative to this base SNR were used to give percentages > 10%. The SNRs ranged between -10 and -25 dB for WN, -35 to -50 dB for SIN, -20 to -35 dB for WSIN, and -25 to -40 dB for WSINS. Each experimental condition was tested with two word lists (*i.e.*, 20 sentences). Each listener was exposed to 64 of the word lists. The order of the word lists and conditions was randomized across subjects. Each listener heard one talker only.

Diotic presentation of the stimuli used a playback system comprising Beyer Dynamic DT770 Pro headphones connected to a PC running MATLAB code with a custom graphical user interface. The audio output of the system was calibrated using a Brüel & Kjær type 4100 head-and-torso simulator (HATS) with type 4189 microphones in each ear canal. Participants were tested inside an audiometric booth with a background noise of 20 dB L_{Aeq}. The SPL at the entrance to the ear canal was 20 dB L_{Aeq} for the HATS wearing the headphones connected to the PC. Subjects chose their preferred listening level as 65, 70 or 75 dB L_{Aeq}. They were asked to identify as many words as possible in each sentence and were allowed to correct their spelling. Incorrect spelling was identified and assessed (see [15] for details) and words 'a' and 'the' were excluded from percentages reported here. The test took ≈ 2 hours, with breaks of up to five minutes taken every ≈ 30 minutes.

2.3 Implementations of metrics

STOI_WC is compared with STOI, NCM and CSII using three figures of merit: linear Pearson's product-moment (ρ) and Kendall's tau (τ) correlations between the metrics and intelligibility scores, and the standard deviation of the prediction error (σ). In testing, it was noted that STOI and STOI_WC varied at the second decimal place with pseudo-randomly selected segments of additive Gaussian white noise. Hence, correlations are computed on the basis of objective indices that are *averaged over sentences* within word lists. STOI calculation followed the reference implementation [4]. STOI and STOI_WC were calculated for each sentence. Any values below 0 were raised to 0 before averaging across sentences within word lists. STOI_WC was calculated as in the case of STOI including normalisation but without clipping at β = -15 dB. STOI-based values were converted to 'mapped' values via a logistic function to linearize the relationship between STOI or STOI_WC and intelligibility

scores and therefore report linear correlation coefficients and determine the distribution of prediction errors. The logistic function, $f(d) = 100/(1 + \exp(a \cdot d + b)$, is used to map a variable d (representing STOI or STOI_WC) with the free parameters, a (slope) and b (centre) as reported by Taal *et al*. [4,9]. Kendall's tau is rank-based and therefore independent of the mapping. The prediction error was computed as $\sigma = \sigma_d\sqrt{1 - \rho^2}$ where σ_d is the standard deviation of the percentage words correct. In the literature (e.g., [4]), one logistic mapping is applied across maskers for a single corpus. However, broadband maskers are used; sinusoidal maskers are not normally considered. Hence, SIN is mapped separately. Given that for maskers other than SIN, there is a clustering of intelligibility scores < 10% at the two or three lowest SNRs, psychometric functions (involving intelligibility scores by SNR) are not evaluated.

The NCM was calculated according to the ANSI 3.5 fixed-weight implementation [10, 19]. The clean and degraded signals were bandpass filtered into 20 frequency bands with centre frequencies ranging from 335 to 6910 Hz with eighth-order Butterworth filters. The envelopes were computed by means of the Hilbert transform and the signals were down-sampled to 32 Hz. Apparent SNRs were calculated in each frequency band on the basis of the normalised covariance of the clean and degraded envelopes [10]. Transmission indices were averaged over the weighted bands. Logistic mapping was performed after [9]. CSII calculations were performed on the basis of publicly available code [17]. Standard ANSI 3.5 [19] weights were used for the 16 frequency bands with centre frequencies ranging from 150 to 3400 Hz. Models used to derive I_3 followed Kates and Arehart [8]. More information concerning the NCM and CSII measures is given in references [5,10].

3. Results

The constants for the logistic function for mapping STOI, STOI_WC and NCM are given in Table 1. STOI slopes (a) are relatively steep, and NCM slopes, shallow. Merits ρ and σ are applied to the mapped scores. Table 2 reports the performance of metrics for each masker and set of talkers. A higher ρ and τ and a lower σ indicate better performance. In general, ρ was higher for SIN than other maskers. This is likely to be due to the greater spread of word intelligibility scores for this masker.

Table 1: Values for the free parameters of the non-linear mappings of STOI, STOI_WC and NCM (IEEE).

	Masker	STOI		STOI_WC		NCM	
		a	*b*	*a*	*b*	*a*	*b*
Male talkers	WN/WSIN/WSINS	-17.69	13.00	-13.05	6.60	-10.02	4.73
	SIN	-37.41	33.00	-22.31	16.85	-8.24	2.76
Female talkers	WN/WSIN/WSINS	-12.29	10.07	-16.31	8.40	-9.83	5.20
	SIN	-38.52	33.71	-22.48	17.32	-9.91	2.68

3.1 STOI and STOI without clipping

3.1.1 STOI

For maskers other than SIN, the STOI logistic parameter values are similar to those reported by Taal *et al*. [4] for a male talker and the IEEE sentences of a = -17.49 and b = 9.69. In comparison, SIN parameter values indicate a steeper slope and a rightwards shift. As shown in Table 2, for STOI, $\rho > 0.7$ with the exceptions of WN and WSINS for female talkers, while $\sigma < 10$ %, with the exception of the SIN condition. In Fig. 1, scatterplots show the relationship between STOI and intelligibility scores for male and female talkers with fitted lines deriving from logistic models. It is apparent that STOI ≥ 0.32, even when signals are unintelligible. For SIN, STOI > 7.5, which indicates over-estimation for sinusoidal maskers. Evidently, STOI does not form clusters associated with individual SNRs but is relatively continuous, and limited in range. Mapping the STOI outcomes using the logistic function and comparing these predicted intelligibility scores with measured intelligibility scores (not shown), mapped STOI tends to under-predict intelligibility at the highest SNR for WN and WSIN

maskers, and over-predict at -35 and -30 dB SNRs for the WSINS masker. Predictive power is very good for the SIN masker; however, in this case, the mapping has been masker-optimised.

3.1.2 STOI without clipping

For STOI_WC with maskers other than SIN, a = -13.05 and b = 6.60 for male talkers and a = -16.31 and b = 8.40 for female talkers (Table 1). Once again, parameter values are larger for SIN than the other maskers. STOI_WC $\rho > 0.7$ for the male talkers and for SIN and WSIN for the female talkers, and $\sigma < 10$ % with the exception of the SIN condition (Table 2). STOI_WC correlation coefficients tend to be higher and prediction errors lower than those for STOI. Of the four metrics, STOI_WC tends to perform best for these maskers at very low SNRs. Indicating significant differences between coefficients (excluding CSII_{Low} from comparison), 95% confidence intervals (CIs) for STOI_WC exceed the upper CI of (1) $\text{CSII}_{\text{High}}$ for SIN for male talkers, (2) CSI_{Mid} and (3) I_3 for WSINS for male talkers, and (4) STOI for WSINS for female talkers. Unintelligible signals are associated with STOI_WC ≈ 0 for maskers other than SIN (Fig. 2), *i.e.*, STOI_WC correctly estimates very poor signal intelligibility. However, for SIN, where there are zero or very low percentages of correct words, STOI_WC > 0.59, which suggests over-estimation. Compared to STOI, clusters with STOI_WC occur in four areas that correspond to the four SNRs, indicating better discrimination, and mapped STOI_WC is a more accurate predictor of measured intelligibility scores. For male talkers, there is only a very slight over-estimation for the highest SNR for WN and for -30 and -35 dB SNR for WSINS.

3.2 NCM and CSII

Logistic mapping parameter values for the NCM indices are reported in Table 1. Indices range between 0 and 0.57 for these maskers and SNRs. Maskers SIN and WSIN are associated with similar index maxima, despite markedly different maximal intelligibility scores. In Table 2, figures of merit are identified for the mapped NCM and CSII measures. For male talkers, NCM values $\rho > 0.8$. For female talkers and WN and WSINS, values are low at $\rho = 0.57$ and $\rho = 0.62$, respectively. NCM performs comparably to STOI-based measures, except in the case of WSINS for male talkers, where performance exceeds STOI (*i.e.*, CIs do not overlap). With the exception of SIN -40 dB and WSINS -25 dB SNRs, there is predominantly an over-estimation of measured intelligibility scores. NCM ρ values tend to be higher than those of CSII measures. $\text{CSII}_{\text{High}}$ is limited to the range 0 to 0.50, CSII_{Mid}, 0 to 0.37, CSII_{Low}, 0 to 0.28, and I_3, 0.10 to 0.66, which indicates under-estimation for SIN, at least for metrics other than I_3. I_3, CSII_{Mid} and $\text{CSII}_{\text{High}}$ are associated with ρ values that are similar to those for STOI. CSII_{Mid} and I_3 are associated with slightly higher ρ values than $\text{CSII}_{\text{High}}$. CSII_{Low} is associated with $\rho < 0.3$ for all maskers other than SIN, suggesting unsuitability for signals with very poor intelligibility.

4. Discussion

For the SNRs selected, WN acted as an effective masker, while even at SNRs between -25 and -50 dB, SIN was a relatively ineffective masker (but *c.f.* [16]). For SIN, STOI and STOI_WC values ≥ 0.59 occurred even when signals were unintelligible, which suggests that these metrics are sub-optimal for sinusoidal maskers. This may be due to STOI's assumption of frequency band independence. Overall, higher linear correlation coefficients and lower prediction errors occurred with STOI_WC compared to STOI, which suggests that when signals are highly degraded, STOI_WC is more reliable for intelligibility prediction. Even for unintelligible signals, STOI values ≥ 0.32, while they were approximately zero for STOI_WC, *i.e.*, STOI tends to over-predict speech intelligibility [20]. Moreover, mapped STOI_WC better predicts intelligibility scores than STOI. This can be attributed to STOI inflating correlations by reducing degraded signal (y) magnitudes to magnitudes similar to those of the clean signal (x) where speech energy is weak or absent, whereas STOI_WC retains real degraded signal magnitudes. The STOI clipping procedure is intended to ensure that the sensitivity of the metric to signal degradation is upper bounded and is typically effective in noise only

regions. Taal *et al.* [9] stated that clipping ensures that intelligibility scores are not under-estimated by STOI for noisy signals. However, especially at very low SNRs, where there may be very large differences between *x* and *y* magnitudes, clipping leads to over-estimation.

Table 2: Correlations between speech intelligibility scores (%) and mapped STOI, STOI_WC and NCM, and CSII metrics, where $p < 0.01$ with the exception of $CSII_{Low}$ where $\rho < 0.20$. Best performers for ρ and σ figures of merit are in bold.

WN	Merit	STOI	STOI_WC	NCM	$CSII_{High}$	$CSII_{Mid}$	$CSII_{Low}$	I_3
Male talkers	ρ	0.78	0.80	**0.81**	0.73	0.77	0.06	0.80
	τ	0.64	0.64	0.65	0.65	0.63	0.21	0.63
	σ	4.49	4.29	**4.14**	4.85	4.57	7.11	4.32
Female talkers	ρ	0.56	**0.65**	0.57	0.58	0.59	0.25	0.60
	τ	0.49	0.54	0.51	0.54	0.54	0.15	0.54
	σ	3.25	**2.99**	3.23	3.2	3.16	3.8	3.14
SIN	Merit	STOI	STOI_WC	NCM	$CSII_{High}$	$CSII_{Mid}$	$CSII_{Low}$	I_3
Male talkers	ρ	0.84	**0.85**	0.82	0.71	0.84	0.76	0.84
	τ	0.63	0.65	0.65	0.53	0.64	0.58	0.64
	σ	15.3	**14.79**	16.02	19.61	15.14	18.02	15.25
Female talkers	ρ	**0.9**	0.85	0.89	0.82	0.84	0.87	0.85
	τ	0.71	0.65	0.72	0.62	0.65	0.65	0.66
	σ	**12.36**	14.43	12.81	15.94	15.01	13.76	14.66
WSIN	Merit	STOI	STOI_WC	NCM	$CSII_{High}$	$CSII_{Mid}$	$CSII_{Low}$	I_3
Male talkers	ρ	0.82	**0.84**	0.83	0.79	0.81	0.03	0.78
	τ	0.72	0.72	0.72	0.65	0.66	0.14	0.66
	σ	7.78	**7.31**	7.45	8.21	7.82	13.47	8.4
Female talkers	ρ	0.72	**0.84**	0.76	0.73	0.79	0.17	0.81
	τ	0.58	0.67	0.62	0.63	0.65	0.17	0.65
	σ	4.85	**3.75**	4.57	4.82	4.28	6.89	4.10
WSINS	Merit	STOI	STOI_WC	NCM	$CSII_{High}$	$CSII_{Mid}$	$CSII_{Low}$	I_3
Male talkers	ρ	0.62	0.78	**0.84**	0.74	0.59	0.08	0.56
	τ	0.57	0.64	0.68	0.63	0.51	0.16	0.50
	σ	8.11	6.46	**5.54**	6.88	8.32	10.28	8.57
Female talkers	ρ	0.40	**0.66**	0.62	0.53	0.56	0	0.57
	τ	0.41	0.58	0.57	0.53	0.50	0.02	0.52
	σ	5.81	**4.77**	4.98	5.4	5.26	6.34	5.22

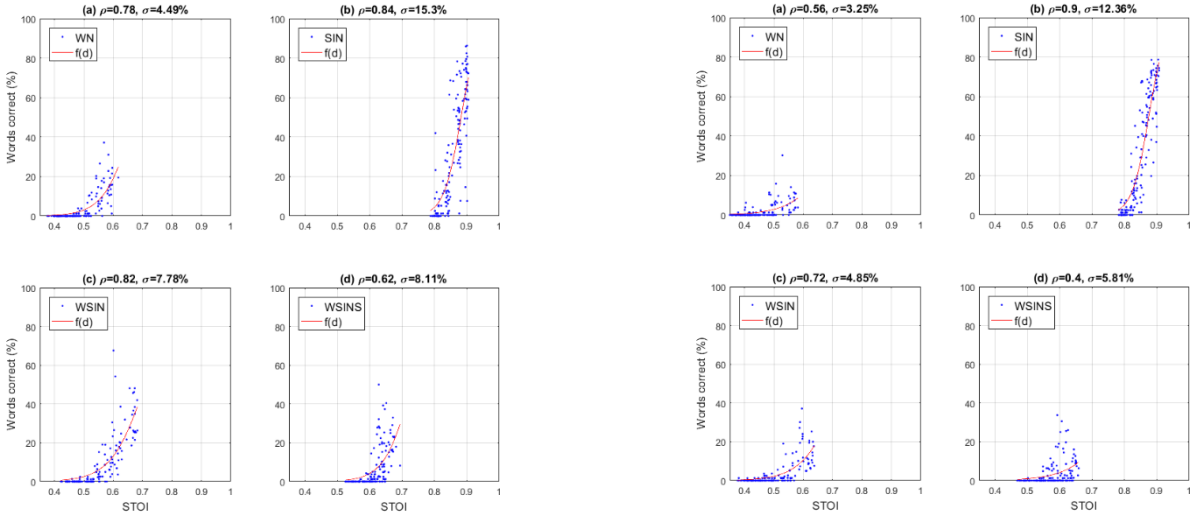

Figure 1: Scatterplots of STOI and speech intelligibility scores (%) for male (Left) and female (Right) talkers. Signals degraded by four maskers: (a) WN, (b) SIN, (c) WSIN, (d) WSINS.

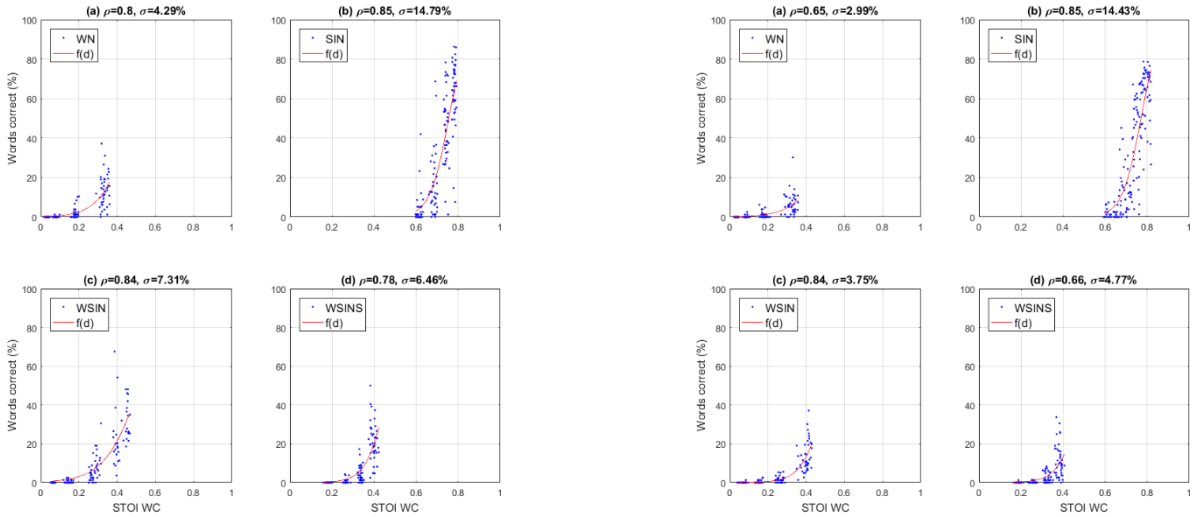

Figure 2: Scatterplots of STOI_WC and speech intelligibility scores (%) for male (Left) and female (Right) talkers. Signals are degraded by four maskers: (a) WN, (b) SIN, (c) WSIN, (d) WSINS.

NCM is a reasonable estimator of the intelligibility of noisy speech (see [4]). However, only in the case of male talkers and the WSINS masker was NCM more highly correlated with intelligibility scores than STOI-based metrics. The performance of CSII I_3 tended to be poorer than STOI_WC. Among the three-level CSII measures, CSII_{Mid} tended to yield the highest correlations. This is consistent with previous findings [10,11]. I_3 often, but not always performed better than CSII_{Mid}.

While STOI-based measures were primarily designed for ITFS-processed speech, and STOI has been shown to outperform both NCM and CSII in that context [4], it has been demonstrated that for highly degraded noisy signals STOI-based measures tend to perform as well as STI- and SII-based measures, NCM and CSII, if not better. Of these measures, STOI_WC tends to be most highly correlated with intelligibility scores for signals highly degraded by noise.

5. Conclusions

In the present study, the performance of STOI without clipping (STOI_WC) was compared with STOI, NCM and CSII performance. It was shown that a high correlation (up to $\rho = 0.9$) can be obtained for signals with additive noise at very low SNRs using STOI-based measures. For speech degraded by noise at very low SNRs and low signal intelligibility, STOI without clipping tends to perform better than STOI, NCM and CSII for the four maskers considered.

Acknowledgement: This research was supported by Her Majesty's Government.

REFERENCES

1 French, N. R., & Steinberg, J. C. Factors governing the intelligibility of speech sounds, *J. Acoust. Soc. Am.*, **1**, 90-119, (1947).

2 Steeneken, H. J., and Houtgast, T. A physical method for measuring speech-transmission quality, *J. Acoust. Soc. Am.*, **67** (1), 318-326, (1980).

3 IEC 60268-16, Sound system equipment–Part 16: Objective rating of speech intelligibility by speech transmission index. International Electrotechnical Commission IEC, (2013).

4 Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech, *IEEE Transactions on Audio, Speech, and Language Processing*, **19** (7), 2125-2136, (2011).

5 Goldsworthy, R.L., and Greenberg, J. E. Analysis of speech-based speech transmission index methods with implications for non-linear operations, *J. Acoust. Soc. Am.*, **116** (6), 3679-89, (2004).

6 Drullman, R., Temporal envelope and fine structure cues for speech intelligibility, *J. Acoust. Soc. Am.*, **97** (1), 585-592, (1995).

7 Holube, I., and Kollmeier, B., Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model, *J. Acoust. Soc. Am.*, **100** (3), 1703-1716, (1996).

8 Kates, J. M., and Arehart, K. H., Coherence and the speech intelligibility index, *J. Acoust. Soc. Am.*, **117** (4), 2224-2237, (2005).

9 Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J., On predicting the difference in intelligibility before and after single-channel noise reduction, *Proceedings of the International Workshop on Acoustic Echo Noise Control*, Tel-Aviv, Israel, September, (2010).

10 Ma, J., Hu, Y. and Loizou, P., Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions, *J. Acoust. Soc. Am.*, **125** (5), 3387-3405, (2009).

11 Arehart, K. H., Kates, J. M., Anderson, M. C., and Harvey Jr, L. O., Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners, *J. Acoust. Soc. Am.*, **122** (2), 1150-1164, (2007).

12 Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M., Effects of noise suppression on intelligibility. II: An attempt to validate physical metrics, *J. Acoust. Soc. Am.*, **135** (1), 439-450, (2014).

13 ISO 8253–1:2010, Acoustics: Audiometric test methods part 1: Basic pure tone air and bone conduction threshold audiometry, Geneva: International Organization for Standardization, (2010).

14 IEEE Recommended practice for speech quality measurements, *IEEE Transactions on Audio and Electroacoustics*, **17** (3), 227-246, (1969).

15 Robinson, M., Hopkins, C., Worrall, K., and Jackson, T., Thresholds of information leakage for speech security outside meeting rooms, *J. Acoust. Soc. Am.*, **136** (3), 1149-1159, (2014).

16 Stevens, S. S., Miller, J., and Truscott, I., The masking of speech by sine waves, square waves, and regular and modulated pulses, *J. Acoust. Soc. Am.*, **18** (2), 418-424, (1946).

17 Loizou, P. C., *Speech enhancement: theory and practice*. CRC Press, Boca Raton, FL (2013).

18 ITU-T P.56, Objective measurement of active speech level. ITU-T Recommendation P.56 (1993).

19 ANSI S3.5-1997, Methods for Calculation of the Speech Intelligibility Index, Acoustical Society of America, American National Standards Institute, New York, (1997).

20 Tang, Y., Cooke, M., and Valentini-Botinhao, C. Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech, *Computer Speech & Language*, **35**, 73-92, (2016).