TRANSIENT-BASED SPEECH TRANSMISSION INDEX FOR PREDICTING INTELLIGIBILITY IN NONLINEAR SPEECH ENHANCEMENT PROCESSORS

Anton Schlesinger

Institute of Communications Acoustics, Ruhr Universität Bochum, 44780 Bochum, Germany
anton.schlesinger@rub.de

ABSTRACT

A new speech intelligibility metric is proposed for the assessment of speech enhancement processors. These processors usually affect the fine structure in speech that is of fundamental importance to speech intelligibility. Classical metrics analyze the entire signal and thereby generally overestimate intelligibility. The measure presented here, therefore, isolates speech-transients by a cepstral smoothing technique and subsequently calculates speech intelligibility using an efficient version of the speech transmission index. By means of a genetic optimization of adjustable parameters, the proposed transition-based speech transmission index (TB STI) is adapted to the subjective data of linearly and nonlinearly processed speech. The method was assessed on untrained subjective data and showed a considerable improvement over other well-established measures.

Index Terms— Cepstrum, intelligibility, speech enhancement, speech perception, transients

1. INTRODUCTION

The algorithmic assessment of speech intelligibility of nonlinearly processed speech is a prerequisite for an efficient optimization of speech enhancement processors. The challenge persists in finding a measure that offers a functional relationship between linearly, i.e., unprocessed, noise-corrupted speech and nonlinearly processed noise-corrupted speech.

The envelope threshold distortion (also known as center clipping) characterizes the effect of nonlinearity of varying gain functions, which is a widely occurring distortion in speech enhancement processors [1, 2]. While there exists a set of qualified measures for linearly processed speech, e.g., the speech transmission index (STI) for the assessment of intelligibility in rooms [3], the development of an overall metric for linear and nonlinear distortions amounts to one of the major problems in speech processing. Although the research regarding a speech-based intelligibility assessment for noise-reduction algorithms has evolved for about three decades, we only recently observe a sudden proliferation of instrumental measures. This reflects the great demand for a comprehensive intelligibility metric. An overview on their applicability was recently given in [4]. Moreover, Taal et al. [5] developed the widely regarded short-time objective intelligibility measure (STOI) for time-frequency masked speech. For this nonlinear processing, which mimics speech enhancement, the STOI surpassed a selection of today's intelligibility measures [4, 5].

Even though most intelligibility metrics process the speech signal continuously, there are only a few measures that account for the time varying information content in speech. Kates and Arehart [1] proposed the coherence-based speech intelligibility index (SII) and extended this metric by an RMS-based intelligibility weighting method in short-time frames (I3). The rationale is that the faint transitional parts in speech are much more important to intelligibility than quasi steady state high energy vowel sections. Speech enhancement algorithms, however, alter mainly the amplitude in these low-level sections. The argument was further supported by Yoo et al. [6], who found that the isolated transient components in speech, i.e., consonants and consonant-vowel, and vowel-consonant and vowel-vowel "interfaces", comprise only 2 % of the original speech energy and are almost fully intelligible. Based on this understanding, we recently developed several coherence-based SII measures that use short-time subband intelligibility-based weights of the relative information content in speech, i.e., Shannon's entropy, to label transitional parts [7]. Although the approach was promising, the proposed metrics were consistently surpassed by the predictive power of the STOI and the I3 measure. The reason for this shortcoming is assumed to be in the very general classification paradigm of voiced and unvoiced speech components, knowing that the transitional components of speech are difficult to isolate [6].

In this paper, we present a new approach of speech intelligibility assessment by evaluating only transitional parts in speech. The method combines a STFT-based cepstrum smoothing technique to identify transitional parts in speech and a speech-based STI version to calculate speech intelligibility, i.e., the TB STI. The cepstrum offers a very efficient means to separate the constitutive components of speech [8]. Subsequently, a binary mask (BM) is created as a stencil for the transitional parts in the clean and distorted speech signal, which are then analyzed in a second part of the algorithm proposed here. The analysis of speech intelligibility employs a speech-based envelope regression method of the STI. The choice of the STI is motivated by several advantages of the metric over purely spectral measures, e.g., the assessment of reverberation, non-stationary interference or the possibility for a binaural extension [2]. In order to optimize the proposed TB STI, a genetic algorithm was applied for the adjustment of a set of algorithmic parameters, whereby subjective scores of linearly and nonlinearly processed speech served as a target function. Finally, the generalizability of the optimized instrumental measure was tested on untrained subjective data.

In Sec. 2 the algorithmic approach is presented. Sec. 3 reports the optimization and assessment procedure. In conclusion, Sec. 4 gives a summary and an outlook.

2. ALGORITHM

In this section the algorithmic approach for predicting speech intelligibility of linearly and nonlinearly processed speech is given. Data

The author is much indebted to J.-P. Ramirez for conduction the listening tests at T-Labs Berlin.

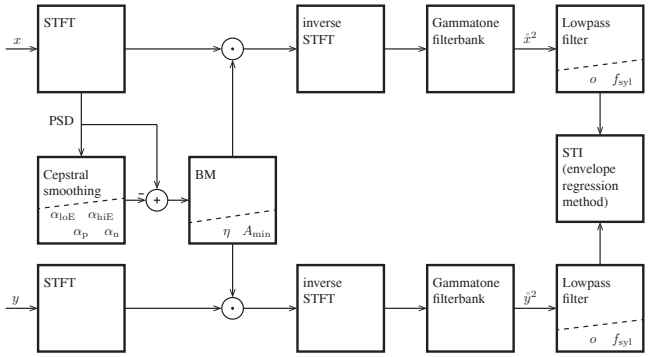

Fig. 1. Flow chart of the TB STI method. The parameters below the dashed lines are optimized via a genetic algorithm.

for the assessment of intelligibility is prepared by lowpass filtering the speech materials at 8 kHz, followed by (re-)sampling at 16 kHz. Silent passages, defined as the RMS level of -50 dB in frames of 10 ms with respect to the long term RMS of the sentence, are determined in the reference signal $x(n)$ with a voice activity detection procedure, and are subsequently discarded in the reference and the degraded signal $y(n)$ at equivalent time samples.

A flow chart of the algorithm is given in Fig. 1. The algorithm comprises two stages, i.e., the transient extraction stage and the speech intelligibility analysis stage. First, the transient detection stage is introduced. In the initial step, the signals $x(n)$ and $y(n)$ are transformed into the short-time Fourier transform (STFT) domain, using a 512-point DFT, considering an appropriate means for preventing circular convolution and spectral leakage in the transformation process. This results in the STFT representation $X(d,\iota)$ and $Y(d,\iota)$, whereby $d = 0, 1, \ldots, N-1$ and ι are the frequency bin and the frame index, respectively. The frame shift ΔT of the STFT is set to 8 ms, and the length of an analysis frame is 16 ms in the current implementation, so that the short-time stationarity of speech is preserved. The succeeding calculation of the BM is based on the power spectral density estimation (PSD) of the clean signal. Therefore, a first-order recursive smoothing operation is applied to the squared absolute magnitude spectrum:

$$\Psi_1(d,\iota) = \alpha\Psi_1(d,\iota-1) + (1-\alpha)|X(d,\iota)|^2. \qquad (2.1)$$

The variable $\alpha = \exp(\Delta T/\tau)$ is the smoothing constant that depends on the filterbank frame-shift ΔT and the time constant τ. A copy of $\Psi_1(d,\iota)$, denoted $\Psi_2(d,\iota)$, is subjected to a cepstral smoothing operation. To independently modify the component signals of speech in $\Psi_2(d,\iota)$, the multiplicative signal in the spectral domain is first linearized by applying the logarithm and then transformed in the cepstral domain with an inverse DFT:

$$\psi(c,\iota) = \frac{1}{N}\sum_{d=0}^{N-1}\left\{\log_{\mathrm{n}}(\Psi_2(d,\iota))\right\} e^{j2\pi c\frac{d}{N}}, \qquad (2.2)$$

where $c = 0, 1, \ldots, N-1$ denotes the quefrency coefficient. The following signal modification is based on a first order recursive smoothing for each quefrency bin:

$$\overline{\psi(c,\iota)} = \alpha_{\mathrm{x}}\overline{\psi(c,\iota-1)} + (1-\alpha_{\mathrm{x}})\psi(c,\iota), \qquad (2.3)$$

where the overscore denotes the smoothed cepstral quantities. In the current implementation, four ranges are defined in order to cover the elemental parts of speech in the cepstral domain. Each cepstral range features a particular time constant α_{x} that accounts for the respective contribution to speech intelligibility:

$$\alpha_{\mathrm{x}} = \begin{cases} \alpha_{\mathrm{loE}} & \text{if} \quad c \in \{0,\ldots,c_{\mathrm{loE}}\} \\ \alpha_{\mathrm{hiE}} & \text{if} \quad c \in \{c_{\mathrm{loE}}+1,\ldots,c_{\mathrm{hiE}}\} \\ \alpha_{\mathrm{p}} & \text{if} \quad c \in \{c_{\mathrm{p}}\} \\ \alpha_{\mathrm{n}} & \text{if} \quad c \in \{c_{\mathrm{hiE}}+1,\ldots,N/2\} \setminus \{c_{\mathrm{p}}\} \end{cases}. \qquad (2.4)$$

The lowest range contains the quasi steady state broadband envelope of speech, i.e., the formants with maxima at resonances of the vocal filter. The second range comprises the fluctuating envelope components in the speech spectrum, i.e., the voiced fine structure of the speech spectra, that is largely dominated by dynamic articulators in speech sounds. The third broad range carries, with high probability, the high fluctuations in unvoiced speech. The pitch of speech also resides in this third range. Additionally, the cepstrum offers a robust way for estimating the pitch by taking the maximum value in the cepstral range of the first harmonic, e.g., $c_{\mathrm{p}} \in \{70\text{ Hz}\ldots 500\text{ Hz}\}$. With the relation $c_{\mathrm{p}} = f_{\mathrm{s}}/f_{\mathrm{p}}$, where f_{s} and f_{p} are the sampling frequency of the signal and the pitch frequency, respectively, the pitch quefrency is calculated via:

$$c_{\mathrm{p}} = \underset{c}{\operatorname{argmax}}\left\{\psi(c,\iota)|c_{\mathrm{p-low}} \le c \le c_{\mathrm{p-high}}\right\}. \qquad (2.5)$$

In the current implementation, the cepstral coefficients c_{loE}, c_{hiE}, $c_{\mathrm{p-low}}$ and $c_{\mathrm{p-high}}$ were set to 5, 10, 20 and 200, respectively. After the cepstral smoothing, the signal is transformed into the spectrum by calculating the DFT and element-wise exponentiation:

$$\Psi_2(d,\iota) = \exp\left\{\sum_{c=0}^{N-1}\overline{\psi(c,\iota)}e^{-j2\pi d\frac{c}{N}}\right\}. \qquad (2.6)$$

In the next step, a BM is computed from $\Psi_1(\mu,\lambda)$ and $\Psi_2(\mu,\lambda)$ in the following way:

$$\mathrm{BM}(d,\iota) = \begin{cases} 1 & \text{if} \quad \Psi_1(d,\iota) - \Psi_2(d,\iota) > \eta \\ A_{\min} & \text{otherwise,} \end{cases} \qquad (2.7)$$

where η is a fixed algorithmic criterion and $A_{\min}$ is the maximum attenuation allowed by the BM operation. By eventually multiplying $\mathrm{BM}(d,\iota)$ with $X(d,\iota)$ and $Y(d,\iota)$, the transitional parts in speech are isolated for further analysis in the second stage of the TB STI. As a last step in this transient extraction stage, the clean and the degraded stimuli are reconstructed by computing the inverse DFT of $X(d,\iota)$ and $Y(d,\iota)$, followed by an overlap-add reconstruction.

The calculation of the STI starts with an auditory-based peripheral frequency analysis by means of a Gammatone filter bank of 4th order. For reasons of efficiency, only 10 equivalent rectangular bandwidth (ERB) filters with center frequencies ranging approximately logarithmically from 0.5 to 7 kHz on the linear frequency axis, i.e., linearly on the ERB scale, are applied. No models of the middle ear filter or of the loudness adaption are included in the calculation. The length of the analysis window $w(n)$ of the STI method was set to 0.4 s, which offers a decent modulation transfer for the envelope regression method [9]. The Gammatone output signals $\mathring{x}(b,n)$ and $\mathring{y}(b,n)$, where b denotes the ERB channel, are squared and subsequently filtered with a FIR lowpass filter. This filter follows the syllabic rate of speech and is determined by the filter order o and the cutoff frequency f_{syl}, which are algorithmic parameters. Thereafter, the output of the lowpass stage is down-sampled by a factor of hundred to a sampling frequency of 160 Hz.

As a modulation metric, the stochastic reformulation of the envelope

regression method of Goldsworthy and Greenberg [10] was chosen. The modulation $\mathcal{M}(b)$ is calculated in each ERB band through the normalized covariance function:

$$\mathcal{M}(b) = \frac{\mu_{\mathring{x}}(b)}{\mu_{\mathring{x}}(b) + \mu_{\mathring{z}}(b)} \cdot \frac{E\{[\mathring{x}(b,n) - \mu_{\mathring{x}}(b)][\mathring{y}(b,n) - \mu_{\mathring{y}}(b)]\}}{E\{[\mathring{x}(b,n) - \mu_{\mathring{x}}(b)]^2\}}, \tag{2.8}$$

where $\mu_{\mathring{x}}$, $\mu_{\mathring{y}}$ and $\mu_{\mathring{z}}$ are the intensity means, and $\mathring{z}(n) = |\mathring{y}(n) - \mathring{x}(n)|$. The first factor in Eq. 2.8 was designed to account for nonlinear signal modifications that increase the modulation depth abnormally [10].

The remainder of the STI calculation is in accordance with the standard approach, as initially proposed in [3]. That is, the band-wise modulation metric is related to the apparent SNR per band:

$$\mathrm{aSNR}_w(b) = 10 \log_{10} \frac{\mathcal{M}(b)}{1 - \mathcal{M}(b)}, \tag{2.9}$$

which is subsequently clipped below -15 dB and above 15 dB, transformed to the transmission index and weighted with a band importance function of average speech. Subsequently, the weighted band transfer indices are summed to give the STI_w. The overall STI of a sentence is averaged across the analysis windows w.

By employing a Gammatone filter bank and a lowpass filtered hair cell approximation, the problem of assessing low level envelope regions with the envelope regression method, reported in [9], is circumvented through a leaky integration of the transitional parts in speech. As a result, the envelope regression method can be used to good effect.

3. SPEECH INTELLIGIBILITY MODEL ADAPTION

In order to adapt the proposed STI method to the perception of linearly and nonlinearly processed speech, an optimization of the algorithmic parameters shown in Fig. 1, was performed. For this reason, two speech intelligibility tests were conducted. The first test was used in the optimization of the STI method and the second test was used for the assessment of the algorithm on untrained data.

3.1. Optimization

We first describe the listening test used for the optimization of the proposed STI method. The speech material was taken from the semantically unpredictable sentence (SUS) corpus, which was developed by Ramirez et al. [11]. The speech files were recorded, lowpass filtered at 22.05 kHz and sampled at 44.1 kHz. The clean and distorted waveforms were corrected for the headphones that were used in the listening test. The masking signal was presented at a fixed level of 70 dB(A) SPL and the target level was changed to the respective SNRs used in the different test conditions. Four subjects of normal hearing ($<$ 15 dB hearing loss (HL) for both ears) participated in the first diotic listening test, which offered three trials per test condition. The recordings were stored for the optimization of the TB STI measure.

The linear degradations used in the optimization comprised five conditions of additive noise, using the long term spectrum of a male speaker and a global RMS-based mixing SNR of -8, -5, -2, 1, 4 and ∞ dB. Furthermore, a bandpass filter for wideband speech, i.e., a frequency transfer between 0.05 to 7 kHz, was applied to these conditions. The nonlinear degradations comprised four envelope threshold conditions of percentages of 50, 60, 75 and 90 % of the cumulative distributions of the sentence waveform. A description of the generation of envelope thresholding is given in [1].

In order to optimize the TB STI method, a 3rd order polynomial was fitted to the data points subsequent to each solution increment of a genetic optimization procedure. In the course of the optimization, the r-squared measure r^2 served as an objective function of the model fit. Table 1 gives the resulting parameters after the optimization. The outcomes of the classical STI (using the speech-based envelope regression version of [10] with a length of $w(n)$ of 0.4 s) and the optimized TB STI method are given in Fig. 2 (A) and (B), respectively. The outcomes show that r^2 is considerably increased as a result of the assessment of only the transitions in speech and the optimization. Furthermore, Kendall's τ is given. It indicates a substantial improvement in favor of the proposed method. An inspection of the optimized parameters highlights the importance of the fluctuating envelope fine structure in speech. Consequently, α_{hiE} is strongly smoothed in the cepstral domain and, hence, maintained after the subtraction operation in the STFT domain. Furthermore, α_{p} and α_{n} are moderately smoothed, thereby accounting for their relative contribution to intelligibility. On the contrary, α_{loE} is hardly smoothed, showing the negligible importance of the quasi steady state low-frequent spectral envelope toward intelligibility.

3.2. Assessment

In order to assess the proposed TB STI method, the measure was evaluated on untrained data and compared to well-established metrics, which are the previously introduced STOI and I3. The applied listening test set comprised five conditions of additive speech shaped noise of a male speaker, i.e., a global mixing SNR of -5, -3, -1, 1 and ∞ dB, and five conditions of envelope thresholding, i.e., at a threshold of 60, 75, 80, 85 and 90 % of the cumulative distribution of a sentence's waveform. In both kinds of distortions, the signal was band-limited to a lower and an upper cutoff frequency of 0.05 and 7 kHz, respectively. Eight people with a normal hearing ($<$ 15 dB HL) participated in a percent correct score test that randomly accessed the 288 sentences of the SUS corpus of Ramirez et al. [11]. Thus, the chances of reproducing the combinations of sentences and distortions of the listening test that was used in the optimization was negligible. The subjects had to respond to three versions of each condition, which were presented to the right ear. The entire test set was then used in the metrics calculation.

In order to facilitate the comparison with other measures and to derive an absolute speech intelligibility score, we linearized the objective results s of the STOI and the I3 measure with $f(s) = 100/(1 + \exp(ms + n))$ and the TB STI with $f(s) = 100/(1 + (ms + n)^q)$. The parameters m, n and q are tuning parameters that were computed in the minimization procedure of the RMS error σ. The choice for the respective linearization function $f(s)$ was based on the minimum σ value. The results of the comparison and the linearization parameters are given in Figure 2 (C), (D) and (E). Additionally, the sample correlation r and Kendall's τ are given. A thorough explanation of the linearization method and the applied statistics can be found in [4, 5].

As the results show, the proposed method offers the highest correlation r of 0.96 and the lowest error σ of 16 %. Only the $\tau = 0.82$ remains below the $\tau = 0.87$ value of the I3 metric. Overall, the TB STI performs better than other measures in predicting speech intelligibility of the considered speech data.

4. CONCLUSION

In this study, a new speech intelligibility assessment method is presented that combines the transient detection stage with the speech-

Table 1. TB STI parameter ranges and results of the genetic optimization.

	α_{loE}	α_{hiE}	α_{p}	α_{n}	η	A_{min}	o	f_{syl} [Hz]
lower bound	0.0001	0.0001	0.0001	0.0001	$1e-8$	0.0001	2	3
upper bound	0.99	0.99	0.99	0.99	0.5	0.5	500	40
results	0.19	0.83	0.47	0.42	0.06	0.0037	491	4

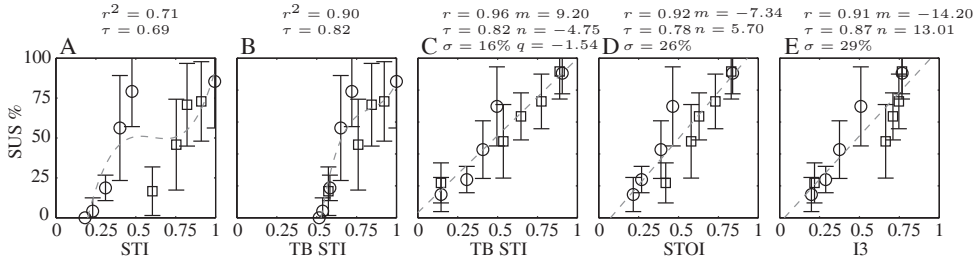

Fig. 2. Subjective versus objective results of linear (○) and nonlinear (□) speech conditions. The Plots A and B show the non-linearized results of the classical envelope regression-based STI and the TB STI, respectively. Plots C, D and E show the linearized results of a second listening test for the assessment of the TB STI, the STOI and I3, respectively. The errorbars denote the subjective standard deviations.

based STI in order to assess nonlinear speech enhancement processors. The method was adapted to the subjective perception of linearly and nonlinearly processed speech. In a subsequent assessment on untrained speech data, the proposed measure showed a high predictive power and outperformed well-established and dedicated measures of nonlinear speech enhancement.

The method draws its strengths from the intelligibility assessment of merely the spectrally fluctuating speech-transients. The cepstrum of the frequency resolution of the DFT offers an excellent tool for isolating transients of different spectral scales. Furthermore, by a genetic model adaptation to subjective date, we verified the importance of mainly voiced and fluctuating spectral components for speech intelligibility. These results corroborate recent findings in speech perception, primarily stating that not high frequency transients, but "kinematic vowel-like sounds" are of foremost importance to speech intelligibility [12]. Consequently, the subjectively optimized TB STI method appears equally useful to the fields of phonetics and linguistics in order to study the basic elements of speech intelligibility.

5. REFERENCES

[1] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[2] A. Schlesinger, *Binaural Model-Based Speech Intelligibility Enhancement and Assessment in Hearing Aids*, Ph.D. thesis, Delft University of Technology, The Netherlands, 2012.

[3] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *The Journal of the Acoustical Society of America (in press)*, 2011.

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

[6] S. D Yoo, R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1138–1149, 2007.

[7] A. Schlesinger and M. M. Boone, "The characterization of the relative information content by spectral features for the objective intelligibility assessment of nonlinearly processed speech," in *Proceedings of the Interspeech Conference*, Makuhari, Japan, 2010.

[8] C. Breithaupt and R. Martin, *Advances in Digital Speech Transmission*, chapter Noise Reduction-Statistical Analysis and Control of Musical Noise, John Wiley & Sons Ltd., 2008.

[9] K. L. Payton and M. Shrestha, "Analysis of short-time speech transmission index algorithms," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3071–3071, 2008.

[10] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[11] J.-P. Ramirez, A. Raake, and D. Reusch, "Intelligibility assessment method for semantically unpredictable sentences in German," in *Fortschritte der Akustik–DAGA*, Rotterdam, The Netherlands, 2009, pp. 1013–1015.

[12] C. E. Stilp and K. R. Kluender, "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 27, pp. 12387–12392, 2010.