Report on a binaural extension of the Speech Transmission Index method for nonlinear systems and narrowband interference

Anton Schlesinger[1], Juan-Pablo Ramirez[2], Jasper van Dorph Schuitman[1] and Marinus M. Boone[1]

1 *Acoustical Imaging and Sound Control, Delft University of Technology, 2600 GA Delft, The Netherlands*

2 *Quality and Usability Lab., Deutsche Telekom Laboratories, Berlin Institute of Technology, D-10587 Berlin, Germany*

A speech-based and binaural version of the Speech Transmission Index (STI) is presented. An efficient envelope regression method of the STI (Goldsworthy and Greenberg, 2004) provides the basis for the proposed method and offers the estimation of nonlinear distortion on speech intelligibility. The speech-based STI method is expanded by an auditory filterbank for the peripheral processing and a binaural processing stage. By this means, the influence of narrow- to broadband interferences and the binaural advantage on speech intelligibility in normal hearing and hearing impaired people can be predicted. The method has been primarily developed to assess nonlinear and binaural processors of speech enhancement, e.g. algorithms of the computational auditory scene analysis, but is generally qualified for other applications of binaural listening. To illustrate the binaural advantage in speech intelligibility, the binaural STI was contrasted with the monaural STI in room acoustical simulation. The performance of the proposed STI method was evaluated against a subjective test. Initial results show an appropriate operation of the proposed STI method. However, the binaural benefit is undervalued and nonlinearly processed speech is overestimated by the proposed STI with respect to subjective perception. For these conditions, further improvement of the objective measure is required.

INTRODUCTION

For reasons of efficiency and comparability, the objective assessment of the speech intelligibility is of interest in many fields of acoustics. An important example is the development of hearing aids. These devices have numerous algorithmic parameters as well as boundary conditions, such as the auditory threshold, that affect their performance and therefore the individual hearing rehabilitation. In particular, hearing aids that suppress noise often comprise a complex and binaural processing scheme based on the auditory scene analysis, blind source separation or beamforming. Current objective measures do generally not master the complexity of the auditory perception when optimizing these hearing aids. As a consequence, the audiological success often falls short of expectations. Therefore, an objective function of speech intelligibility that enables a comprehensive assessment should incorporate the fundamental aspects of the auditory processing.

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

Several methods have been developed to estimate the speech intelligibility at the output of a transmission system. Among these, the STI method has been proven to be a reliable monaural measure of for broadband distortion, like e.g. reverberation (Houtgast and Steeneken, 1973), by measuring the modulation depth to account for the unmasked portions of the speech-spectrum that contribute to speech intelligibility. The computation of the STI from the waveform of transmitted speech allows to estimate the influence of nonlinearities, e.g. as yielded by a nonlinear hearing aid (Goldsworthy and Greenberg, 2004), the speaking style (Payton and Shrestha, 2008) and phonemes on speech intelligibility. Furthermore, the STI method enables the reproduction of speech intelligibility in hearing impaired listeners. Holube and Kollmeier (1996) showed that the proficiency factor of the STI for hearing impaired people is rather dependent on the individual hearing threshold than on other psychoacoustical parameters. Their monaural STI implementation had a critical sampled filterbank that performed equally well as a detailed monaural psychoacoustical model.

The consideration of the binaural advantage by an objective speech measures has become a field of research only recently, mainly because of the computational complexity of this appendage. The spatial release from masking in speech intelligibility can be as high as 12 dB (Bronkhorst, 2000) when the direction of incidence of a single masker deviates from a target talker. In most of the natural listening conditions as well as in a set of speech processors, the principles of binaural hearing are effective and need to be considered. A first speech-based and binaural model that analyzes speech intelligibility in spatial configurations was developed by Beutelmann and Brand (2006) and is based on the equalization cancellation theory by Durlach (1972). In this model, the peripheral processing is performed by a Gammatone filterbank analysis of 30 bands. A hearing loss was simulated by the introduction of uncorrelated masking noise and the speech intelligibility was calculated through the Speech Intelligibility Index (SII). Their model performed well in single masker conditions and could also estimate the influence of room acoustics. However, the SII method is intrinsically not qualified to evaluate nonlinear speech processors for the reason that the target and noise signals have to be separated at the processors output prior to the SII computation. Another binaural model of speech intelligibility was implemented by Wijngaarden and Drullmann (2008). Their model utilizes a coincidence type or cross-correlation model of the binaural processing (Jefress, 1948). In addition to the binaural interaction processing, the model also computes the head shadow effect. Following the classical STI implementation, artificially modulated noise is used as a probe stimulus and analyzed in octave bands. Except for nonlinear conditions, the model consistently maps binaural speech intelligibility in spatial configurations.

Herein, a set of desirable methods is incorporated in a binaural model of speech intelligibility, which primarily aims to be adept at assessing nonlinear and binaural speech processors in a broad range of interferences. Therefore, the stochastic reformulation of the STI by Goldsworthy and Greenberg (2004) is complemented with a peripheral Gammatone filterbank analysis and the binaural processing stage, which is conceptually similar to the implementation of Wijngaarden and Drullman (2008).

In the following sections of this contribution, we sketch the processing of the proposed STI method, evaluate it in a listening test and illustrate how the binaural processing suppresses reverberation in a room acoustical simulation.

APPROACH

The proposed speech-based and binaural STI method was implemented in MATLAB. The general approach of the method is sketched in Fig. 1.

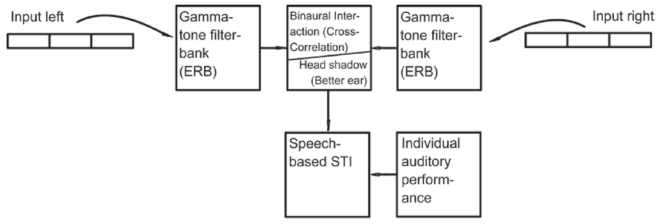

Fig. 1: The operation chart of the speech-based and binaural STI method.

Herein, all tokens of speech and speech shaped noise were originally filtered with the head related transfer functions (HRTFs), as measured using an artificial head, for the design of spatial sound scenes with the appropriate interaural time and level differences (ITDs and ILDs) of sound sources at different positions. These samples were then presented in a subjective test (see the following section) and processed by the described STI method. Since the envelope regression STI method compares the variances in the clean and the deteriorated signal, binaural recordings of the clean target and deteriorated signal form the input of the system. The overall method consists of following processing steps:

1) The speech signals are low pass filtered at 9.5 kHz and sampled at 22.05 kHz.
2) A Gammatone filterbank analysis (the implementation is taken from Wang (2006)) is performed using 30 equivalent rectangular bandwidth (ERB) bands with centre frequencies of 0.1 to 8 kHz. No middle ear filtering is modeled.
3) The output of the Gammatone filterbank is squared and lowpass filtered with a Hamming window of 10 ms to gain the intensity envelope of the signal with a cut off frequency of 50 Hz.
4) A resampling at 3150 Hz and a partitioning into windows of 30 ms is performed to calculate the binaural processing. The windows have an overlap of 0.75, which yields subsequently to the binaural processing stage an internal representation image with a sampling frequency of 132 Hz. At this sampling frequency, the R^2 measure indicates a fair predictability of the modulation transfer with speech samples of 0.5 s length throughout all bands.

5) The central binaural stage consists of the binaural interaction processing, computed between approx. 0.5 and 2 kHz using a cross-correlation, and the head shadow processing (below approx. 0.5 kHz and above approx. 2 kHz) by employing a "better ear" approach. The cross-correlation was calculated for ITDs between -0.8 and 0.8 ms at increments τ of 0.1 ms, which is on average in accordance with psychoacoustical data (Blauert, 1997, p. 342). The square root of the interaural time image is taken to maintain the required intensity representation of the signals for the calculation of the modulation transfer.

6) In order to refer the deteriorated to the clean signal in the STI calculation, the target is localized in a clean interaural time image by searching for maximum of the time averaged cross-correlation every 0.5 s.

7) If $x_i(t)$ denotes the clean and $y_i(t,\tau)$ the degraded samples of 0.5 s length, the modulation metric is calculated in each band i and for each τ through:

$$M_i(\tau) = \frac{\mu_{xi}}{\mu_{xi} + \mu_{zi}} \cdot \frac{E\left\{(x_i(t) - \mu_{xi})(y_i(t,\tau) - \mu_{yi})\right\}}{E\{(x_i(t) - \mu_{xi})^2\}}, \qquad \text{(Eq. 1)}$$

where μ_{xi}, μ_{yi} and μ_{zi} are the signal means and $z_i(t) = |y_i(t,\tau) - x_i(t)|$. The first term normalizes level differences between $x_i(t)$ and $y_i(t,\tau)$. The normalization also holds for nonlinear operations, if e.g. the modulation depth is abnormally increased. The second term is the quotient of the cross-covariance and the auto-covariance, an efficient version of the envelope regression method as only running averages have to be calculated (Goldsworthy and Greenberg, 2004). After the computation of the modulation transfer for each τ, a decision stage selects $M_i(\tau)$ that is maximal in each band i. The "better ear" approach is modeled in a similar way, by selecting the ear offering the highest modulation depth using the same window length. The analyzed windows overlap by half. Therefore, the threshold of the perceptibility of directional changes is 250 ms and about 50 – 150 ms higher than suggested from literature (Blauert, 1997, p. 323). The decision of this threshold value was taken for the two reasons that solely static scenes are analyzed and that the computational cost is decreased. The modulation metric is then, following the classical STI calculation scheme, related to the apparent SNR, clipped below –15 and above +15 dB and transformed to the transmission indices in bands. The transmission indeces are weighted with a band importance function, taken from Pavlovic (1987) for average speech (extended to 30 bands), and summed to the final binaural STI.

The influence of the absolute hearing threshold can be modeled in the STI method as masking noise within each band (Holube and Kollmeier, 1996). Since the model was evaluated with normal hearing subjects, a raised threshold is not included in the present implementation.

EVALUATION

To examine the quality of the proposed speech-based and binaural STI method, a listening test was performed. Two approaches were pursued. First, diotic and dichotic test conditions were analyzed by the proposed STI method. The distribution of these spatial conditions around a common word score – STI curve bears a quality measure for the applied listening test and the STI method (Wijngaarden and Drullman, 2008). Second, the binaural intelligibility level difference (BILD) was calculated using the standard psychometric curve of the applied speech corpus. In a similar manner Beutelmann and Brand (2006) analyzed their binaural model to show the binaural advantage as a function of the azimuth of the interfering noise source.

A detailed description of the listening test can be found in Ramirez *et al*. (2009a). In all test conditions, the consonant-vowel-consonant (CVC) corpus described in Ramirez *et al*. (2009b) was used. The speech files were recorded and digitized at 44.1 kHz. The clean and distorted wave forms were convolved with the HRTFs of an artificial head in a particular acoustical scene and 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. The recordings were stored for further analysis with the STI. The root mean square (RMS) level of the clean signal was set to -30 dB relative to the RMS level of the degraded signal. Silent gaps in the speech files, in here defined as the RMS level smaller than -50 dB relative to the overall RMS level, lead to an erroneous increase of the STI with the envelope regression method (Payton and Shrestha, 2008) and were therefore cut out with an automatic gap detection algorithm.

Four subjects of normal hearing (with a hearing loss lower than 15 dB for both ears) participated in the diotic test (3 trials per condition). The conditions ranged from no deterioration to several forms of linear deteriorations using echo, reverberation and a single masker of speech shaped stationary noise. Situations 24 to 27 indicate nonlinear envelope threshold distortions from soft to severe, respectively, which increase the modulation depth (Goldsworthy and Greenberg, 2004).

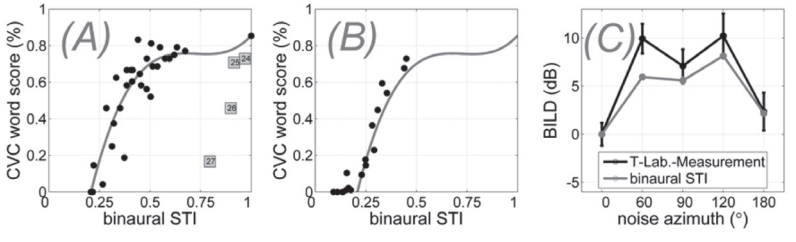

Fig. 2: The CVC-STI curve for the diotic (A) and the dichotic (B) test conditions. The BILD (C) for a SRT test and the STI computation, including the confidence levels.

The results of the diotic test are depicted in Fig. 2A. A third order polynomial fit has been laid through the data points (excluding nonlinear conditions) with a standard deviation of 11.9%. When comparing the CVC-STI curve of the proposed binaural STI method with a classical monaural CVC-STI curve (e.g. Steeneken and Houtgast, 2002), a steeper slope and a higher underestimation of the STI at around 0.5 are apparent. One explanation is that the speech-based STI saturates at low SNRs towards a higher STI than classical STI implementations (the same is observed for positive SNR values, when the speech-based STI saturates earlier, the classical STI-SNR curve is thus compressed). A second reason is that short speech segments, as used in the proposed STI method, expose a lower window-wise SNR than found for whole CVC sentences. The STI values of the classical method with artificial modulated noise are thus not comparable with the STI values of the here proposed method. These features are however of negligible consequence for the competence of the proposed STI method. Consequential for the assessment of nonlinearly processed speech is the observation that the tested nonlinear conditions (number 24 to 27, envelope threshold distortion) fall off the CVC-STI curve. Under these conditions, important parts of speech are suppressed while the modulation in the non-suppressed parts is maintained and analyzed. Therefore, the STI overestimates the speech intelligibility. In spite of the increase of the modulation depth, the envelope regression method of Goldsworthy and Greenberg (2004) is however capable to indentify the trend correctly.

The results of the dichotic test are shown in Fig. 2B. The dichotic conditions comprise free-field presentations of the target in frontal position and a single masker at angles ranging from 60° to 180°. Eight normal hearing subjects participated in the test (three trials per condition). The proposed STI method slightly underestimates the binaural conditions for a word score higher than 0.4 (and for word scores lower than 0.1). Two reasons might be accountable. First, one normal hearing subject out of four in the diotic test performed considerably bad and shifted slightly the diotic fitting CVC-STI curve. Second, the proposed STI method does not fully reproduce the binaural advantage and underestimates therefore the speech intelligibility.

The second reason is supported by the findings of the subjective test and the computation of speech intelligibility in different target-masker conditions. The results are displayed in Fig. 2C. Eight normal hearing listeners participated in a speech reception threshold (SRT) test to determine the BILD. For the computation of the BILD it was assumed that the maximum likelihood fit of the psychometric function of the CVC set (Ramirez *et al*., 2009b),

$$p(SNR) = \left(\frac{1}{1 + e^{(M - SNR)/S}} \right), \qquad \text{(Eq. 2)}$$

with M is the SRT, S the steepness (with respect to the applied corpus, S is 2.1 for a slope of 12%/dB for speech shaped stationary noise) and p the probability of a correct response, changes only by M for all free-field dichotic conditions. This assumption

is supported by our measurement at fixed SNRs. Therefore p is identified with the diotic fitting CVC-STI curve (the influence of the afore mentioned expected shift of the diotic fitting CVC-STI curve by using a greater population in the diotic test would have a small effect on the standard deviation). Setting M to zero, the SNR indicates the difference between the STI calculation and the subjective measurement of the BILD. The results reveal a lower estimated BILD by the proposed STI method in each target-masker combination than found in the listening test. The difference is in the order of 4 dB in the $N_{60}S_0$ situation and diminishes to approx. 0 dB in the $N_{180}S_0$ situation. The observed tapering might be answered by psychoacoustical data that shows a higher density of multiplier coincidence cells around the median plane (Blauert, 1997, p. 342) instead of a constant distribution as implemented in the here proposed STI method.

APPLICATION OF THE PROPOSED STI METHOD IN ROOM SIMULATION

To evaluate differences between the monaural and binaural STI as proposed in this paper on various positions in an acoustic environment, a virtual room was simulated. The room is shoebox-shaped with dimensions W x L x H = 15 x 20 x 5 m (V = 1500 m^3). A virtual source (omnidirectional) was positioned in front of the room, off centre and at a height of 2 m. A total of 225 receivers were located in the room at the same height as the source, spaced 1 meter apart in the x- and y-directions. For each receiver location r two impulse responses were simulated:

1) A monaural impulse response $p_{r,m}(t)$, as would have been measured in a real room using an omnidirectional microphone.

2) A binaural impulse response $p_{r,b}(t)$, as would have been measured using an artificial head.

The impulse responses were generated using custom software which is designed for simulating shoebox-shaped rooms. The configuration was such that image source modeling was used for reflections up to the second order. For the simulation of higher order reflections statistical modeling was performed. The binaural room impulse responses $p_{r,b}(t)$ were generated by convolving the sound field with HRTFs from an artificial head as measured in an anechoic room. It was chosen to set a uniform absorption coefficient for the room boundaries, leading to a reverberation time of 1.25 s in low to mid frequencies. Above mid frequencies, a roll-off at higher frequency bands is modeled due to sound absorption through air. For each receiver position r both the monaural and binaural STI values were determined using the simulated room impulse responses $p_{r,m}(t)$ and $p_{r,b}(t)$ respectively. The results are shown in Fig. 3.

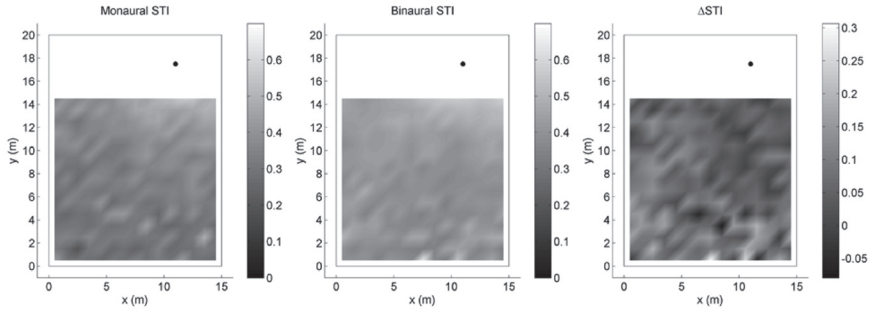

Fig. 3: The monaural (left) and binaural (centre) STI values for all receiver positions in the virtual room and the difference between binaural and monaural STI values, relative to the monaural STI (right). The source is denoted by a closed dot.

As expected the STI values are highest for receiver positions close to the source. Further away from the source, the late reverberation of the room can mask the direct sound, leading to a decrease in speech intelligibility. However, the human auditory system is capable of suppressing sound coming from directions other than that of the source, as discussed earlier. Therefore the monaural STI will underestimate the perceived intelligibility. As can be seen in Fig. 3 the binaural STI is generally higher, because it takes this suppression effect into account. The difference between the monaural and binaural STI values is most apparent at locations far from the source, where the room reverberation is considerably higher than the direct sound level. To demonstrate this, the difference $\Delta STI = STI_{binaural} - STI_{monaural}$ is shown in Fig. 3 (right). In the diffuse field, the difference can be as high as 50% or sometimes even higher. The monaural and the binaural STI were calculated with the envelope regression method. No mapping was performed to relate the here computed STI values to the STI values of the classical method using artificially modulated noise. Therefore, the STI values of the proposed method are only comparable with other STI methods to a limited extend.

CONCLUSION

A speech-based and binaural STI method has been proposed that aims to assess speech intelligibility in spatial configurations as well as the influence of nonlinearities and narrow- to broadband interferences. The performance of the STI has been evaluated with a listening test.

The results show a good predictability for linear narrow- and broadband distortions. The correct assessment of narrowband distortion, a crucial issue in the optimization of speech processors (e.g. the assessment of musical noise due to artifacts in the processing), is attributed to the application of an auditory peripheral filterbank.

However, the analyzed nonlinear distortions are not correctly mapped by the proposed STI. Although the envelope regression STI method by Goldsworthy and Greenberg

(2004) is capable to rate correctly the severity of nonlinear distortions, the method will likely not be sufficient for optimization tasks of nonlinear processors. For these applications, the proposed STI method has to be expanded by a more accurate model of the auditory processing, e.g. by including a multi-scale analysis of spectro-temporal modulations as suggested by Elhilali *et al*. (2003).

The proposed STI method shows the ability to model the binaural advantage. At present, the BILD of the STI computation is lower than found in the subjective test. An alteration and non-uniform distribution of the time-increments (coincidence cells) in the binaural stage might yield a better consistency with the subjective results. In conclusion, the proposed method is conceptually suitable for the assessment of speech intelligibility in a wide range of spatial configurations and distortions. Refinement of the method is essentially needed for the assessment of the binaural advantage and nonlinear distortions.

ACKNOWLEDGEMENT

The authors wish to express their sincere thanks to Sander J. van Wijngaarden and Alexander Raake for their support.

REFERENCES

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331-342.

Blauert, J. (**1996**). *Spatial hearing* (MIT Publication).

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions," Acustica **86**, 117-128.

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Communication **41**, 331-348.

Durlach, N. I. (**1972**). *Binaural Signal Detection: Equalization and cancellation theory of binaural masking-level differences* (Academic New York), Vol. II, pp. 371-462.

Goldsworthy, R. L., and Greenberg, J. E. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679-3689.

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

Houtgast, T., and Steeneken, H. J. M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66-73.

Jefress, L. A. (**1948**). "A place theory of sound localization," J. Comp. Physiol. Psychol. **41**, 35-39.

Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413-422.

Payton, K., and Shrestha, M. (**2008**). "Analysis of short-time speech transmission index algorithms," Acoustics 08 Proceedings, Paris, France, 633-638.

Ramirez, J. P., Raake, A., and Reusch, D. (**2009a**). "Intelligibility assessment method for semantically unpredictable sentences in german," in *ISAAR 2009: Binaural Processing and Spatial Hearing* J. Buchholz, J. C. Dalsgaard, T. Dau, and T. Poulsen (The Danavox Jubilee Foundation).

Ramirez, J. P., Raake, A., and Reusch, D. (**2009b**). "Intelligibility assessment method for semantically unpredictable sentences in german" NAG-DAGA Proceedings, Rotterdam.

Steeneken, H. J. M., and Houtgast, T. (**2002**). "Basics of the STI measuring method," in *Past, present and future of the Speech Transmission Index*, edited by Sander J. van Wijngaarden, (TNO Human Factors).

van Wijngaarden, S. J., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J. Acoust. Soc. Am. **123**, 4514-4523.

Wang, L. (**2006**) *Computational Auditory Scene Analysis* (John Wiley and Sohns, Inc. Publication).