Bayesian Adaptive Method for Estimating Speech Intelligibility in Noise

Nikolay D. Gaubitch, Mike Brookes, Patrick A. Naylor, Dushyant Sharma

[1]*Imperial College London*

Correspondence should be addressed to Nikolay D. Gaubitch (ndg@imperial.ac.uk)

ABSTRACT

We present the Bayesian Adaptive Speech Intelligibility Estimation (BASIE) method – a tool for rapid estimation of a given speech reception threshold (SRT) and the slope at that threshold of multiple psychometric functions for speech intelligibility in noise. The core of this tool is an adaptive Bayesian procedure, which adjusts the signal-to-noise ratio at each subsequent stimulus such that the expected variance of the threshold and slope estimates are minimised. Simulation results show that the algorithm is able to achieve SRT estimates accurate to within ±1 dB in under 30 iterations. Furthermore, we discuss strategies for using BASIE to evaluate the effects of speech processing algorithms on intelligibility and we give two illustrative examples for different noise reduction methods with supporting listening experiments.

1. INTRODUCTION

The intelligibility of speech in background noise can be quantified in terms of a psychometric function (PF) which links the probability of a listener correctly understanding what is being said to the signal-to-noise ratio (SNR). Estimation of PFs is a widely researched topic in the field of psychophysics [1]. The PF is often modelled as a sigmoid function which can be parameterised in terms of the SNR corresponding to a defined intelligibility level, Ψ_0, and the slope of the PF at this SNR. In addition, there may be allowance made for guessing and lapses. This can be written as [1, 2]

$$\Psi(x) = \gamma + (1 - \gamma - \lambda)\Phi(x), \qquad (1)$$

where the guess rate is γ taken to be $1/N$ for a N-alternative forced choice experiment and the lapse rate is λ. Several sigmoid functions have been proposed for $\Phi(x)$; here we have used the cumulative normal distribution:

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt, \qquad (2)$$

where the slope, β, and the threshold, α, at $\Psi(x) = \Psi_0$ are governed by the variance, σ and mean μ respectively. It is straightforward to determine μ and σ as a function of $\gamma, \lambda, \alpha, \beta$ and Ψ_0. Thus, estimating the PF is reduced to estimating the two parameters σ and μ. An example PF generated with (1) and (2) is shown in Fig. 1.

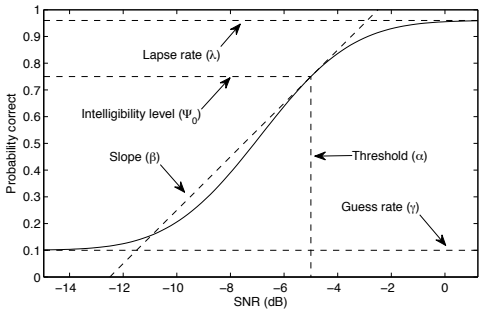

Fig. 1: An example psychometric function generated with the cumulative normal distribution. Indicated are the intelligibility level, threshold, slope, guess rate and lapse rate.

Measurements of subjective speech intelligibility scores often focus on speech in noise in order to determine the ability of a listener to understand speech under noisy conditions [2, 3]. Other than the spectral properties of the noise, several additional processes can affect the intelligibility of noisy speech; some processes are intentionally applied, such as speech enhancement software, while others unavoidable, such as codecs. Quantifying the effects on intelligibility of such processes is impor-

tant and can assist in the development of new algorithms and lead to a better understanding of existing ones. However, speech intelligibility testing is a very time consuming task which often limits the ability of researchers to examine the effects on intelligibility of such additional processes. It is, therefore, important to develop methods and tools for efficient and rapid subjective intelligibility evaluation.

In this paper, we present the Bayesian Adaptive Speech Intelligibility Estimation (BASIE) method. This tool was inspired by the Technique for Automatic Comparative Intelligibility Testing (TACIT) [4] and is a N-alternative forced choice test which enables rapid intelligibility testing of multiple *conditions* simultaneously. Here, a *condition* is defined as a particular setting of a speech enhancement algorithm. The simultaneous testing is achieved through interleaved estimation of several PFs. Thus, we use BASIE to evaluate the performance of a speech processing device by comparing the speech reception threshold (SRT) of processed speech with that of unprocessed speech. Using this interleaved approach is believed to reduce learning and fatigue effects with listeners [4]. The stimuli are digit triples and the user interface is a numeric keypad.

The remainder of the paper is organised as follows. In Section 2, we describe the components of BASIE including the adaptive PF estimation procedure, the data set and the user interface. In Section 3, we discuss how BASIE can be applied to the evaluation of speech enhancement algorithms. Simulation results are presented in Section 4 to verify the adaptive PF estimation procedure. The use of BASIE for evaluating speech enhancement algorithms is demonstrated in Section 5 with supporting results from a listening experiment. Finally, in Section 6 we draw conclusions from this work.

2. BASIE

An adaptive PF estimation algorithm is an iterative process. At each iteration the subject is presented with a noisy speech signal at a probe SNR and responds by indicating what he or she hears. Before the next iteration the probe SNR is increased or decreased according to some rule depending on the outcomes of previous iterations. Accordingly, there are three major components to be considered in the development of an intelligibility test: (i) the adaptive PF estimation, (ii) the data set and (iii) the user interface.

2.1. Adaptive PF Estimation

The objective of our adaptive PF estimation method is to select the probe SNR of the next trial such that we obtain as much information as possible about the PF. In this way, the number of trials required to estimate the threshold and the slope is minimised. The estimation procedure, based on that presented by Kontsevich and Tyler [5], is described below.

Let $\theta = (\alpha, \beta)^T$ be a two-dimensional vector containing the values of the threshold α and the slope β at $\Psi(x) = \Psi_0$. We can define a two dimensional probability density function (PDF), $p(\theta)$, which specifies the probability space of all possible psychometric functions functions and which we initialise to a non-informative prior distribution. At trial n the subject indicates a response, r_n, to a noisy speech sample at a probe SNR, x_n, such that $r_n = 1$ if the response was correct and $r_n = 0$ for an incorrect response. After the nth trial, the PDF is updated with the new result according to

$$p_n(\theta \mid x_n, r_n) = \frac{p_{n-1}(\theta) P(r_n \mid x_n, \theta)}{\sum_{\theta} p_{n-1}(\theta) P(r_n \mid x_n, \theta)}, \qquad (3)$$

where

$$P(r = 1 \mid x, \theta) = \Psi(x) \qquad (4)$$

$$P(r = 0 \mid x, \theta) = 1 - \Psi(x) \qquad (5)$$

and $\Psi(x)$ is given in (1). From $p_n(\theta)$, we can calculate the expected value and the covariance of the threshold and slope estimates according to

$$E_n(\theta) = \sum_{\theta} \theta \, p_n(\theta) \qquad (6)$$

and

$$Cov_n(\theta) = \sum_{\theta} \theta \theta^T p_n(\theta) - E_n(\theta) E_n(\theta)^T, \qquad (7)$$

where $E\{\cdot\}$ and $Cov\{\cdot\}$ denote expectation and covariance, respectively.

Next, we wish to find the SNR probe value for the following trial, x_{n+1}, such that the expected variances of the estimates of the threshold and the slope are minimised

$$x_{n+1} = \arg\min_{x} \big((1-\kappa) E\{Var_{n+1}\{\alpha \mid x\}\} + \kappa E\{Var_{n+1}\{\beta \mid x\}\} \big), \qquad (8)$$

where the variances, $Var\{\cdot\}$, are the diagonal elements of $Cov\{\theta\}$. For all the experiments in this paper the threshold/slope weighting constant is set to $\kappa = 0.5$.

The expected variance of the PF parameters for a given probe SNR, x, is given by

$$E\{Var_{n+1}(\theta \mid x)\} = \sum_{r=0}^{1} Var_{n+1}(\theta_i \mid x,r)\, P(r \mid x) \quad (9)$$

where $Var_{n+1}(\theta_i \mid x,r)$ is found from (3)-(7) and

$$P(r \mid x) = \sum_{\theta} p(\theta)\, P(r \mid x,\theta) \quad (10)$$

is the probability of obtaining response r in the next trial. Finally, we rescale and resample the PDF to cover ± 4 standard deviations around $E_{n+1}(\theta)$, which we found to be sufficient through experiments.

This process is repeated until satisfactory convergence is achieved. In our case, as in [5], the algorithm is executed for a fixed number of iterations.

As discussed in Section 1, our aim is to perform simultaneous estimation of multiple PFs in a manner similar to that of TACIT [4]. TACIT allows two tests to be run simultaneously and interleaves the estimation of a processed and a unprocessed signal by switching to the unprocessed signal on every 16th trial. Our approach extends that of TACIT in two ways: first, we accommodate the simultaneous estimate of any number of PFs and second, the next PF model to update is chosen to be that which provides the largest expected decrease of the cost function in (8). The reason for allowing more than two signals is related to the algorithm evaluation methodology and will be discussed further in Section 3.

2.2. User Interface and Data set

BASIE has been implemented in MATLAB. The user interface comprises a graphical numeric keypad where any number of digits can be entered. A sequence of digits corrupted by noise is played and the subject has to enter what he or she believes was said, even if this requires guessing.

We have used digit triples for our experiments where the intelligibility function was calculated for the triplet rather than for each digit [2]. The digit triplets give a theoretical guess rate of $\gamma = 0.001$ which almost eliminates the contribution of the guess rate in (1).

3. EVALUATION OF SPEECH PROCESSING ALGORITHMS

The option of interleaved estimation of several PFs facilitates simultaneous testing of speech processing algorithms or other speech corruption mechanisms such as speech codecs [4]. We are particularly interested in the processing gain defined here as the difference in SRT between the processed noisy speech and the unprocessed noisy speech

$$\text{Processing Gain} = SRT_{Noisy} - SRT_{Processed}. \quad (11)$$

Positive gain implies improvement in intelligibility while a negative gain implies degradation in intelligibility.

Furthermore, the ability of BASIE to estimate more than two PFs in one experiment, allows simultaneous evaluation of more than one process. Alternatively, this can be applied to one processing algorithm but with a varying set of parameters. This is the approach we shall consider for our listening experiments presented in Section 5. The default setting for BASIE is to search for the SRT at 75% intelligibility rather than the more conventional 50%. The reason for this is that the SRT search at an intelligibility of 75% is concentrated around higher SNRs, which provides a better operational environment for signal processing algorithms.

4. SIMULATION AND VALIDATION

We present simulation results to demonstrate the properties of the adaptive PF estimation component of BASIE described in Section 2.1. The objective of the experiment is twofold: (i) to validate that BASIE accurately identifies the threshold and the slope when the underlying PF model is known to be correct; (ii) to investigate the number of iterations that are required for the identification.

The outcome r of a subject was simulated by utilising the inverse of the scaled and shifted cumulative normal distribution of (1) and (2). The SRT at 75% intelligibility was arbitrarily set to $\alpha = -5$ dB. The slope, lapse rate and guess rate were set to, respectively, $\beta = 0.1$ dB^{-1}, $\lambda = 0.04$ and $\gamma = 0.001$. The values for slope and lapse rate were chosen based on the results in [2].

We ran 500 Monte Carlo simulations of 300 trials each. The SNR for each subsequent trial was calculated using either a fixed step-size of 0.5 dB, similar to the procedures in [4, 6] or with BASIE. The outcome was evaluated using two metrics: the *estimation bias*, calculated as the average difference between the estimates and the true SRT, and the *estimation variance*.

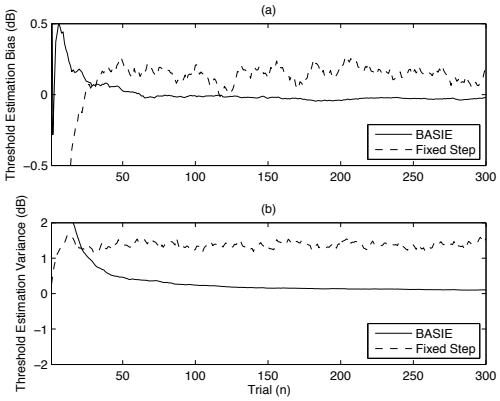

Fig. 2: (a) Bias and (b) variance of the speech reception threshold estimates from BASIE compared to a fixed step-size of 0.5 dB as an average of 500 Monte Carlo simulations.

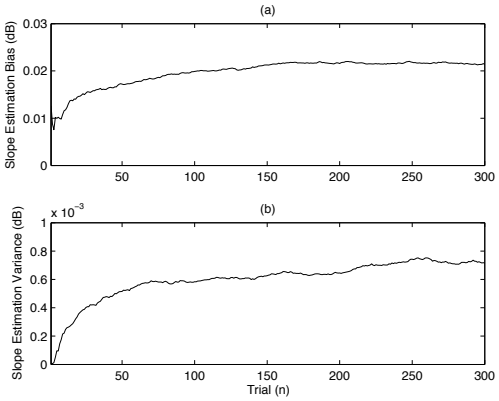

Fig. 3: (a) Bias and (b) variance of the slope estimates from BASIE as an average of 500 Monte Carlo simulations.

The results for the estimation of the speech reception threshold are shown in Fig. 2 and the results for the slope estimates of BASIE are shown in Fig. 3. It can be seen that BASIE converges to the correct SRT values within about 30 iterations and achieves greater accuracy compared to the fixed step-size approach. The average of the values between 30 and 40 iterations give a bias of 0.07 dB with a variance of 0.7 dB for BASIE. This reduces to a bias of 0.01 dB with a variance of 0.2 dB for the average of the values between 100 and 110 iterations. In contrast, the fixed step-size approach results in a bias of approximately 0.1 dB and a variance of 1.3 dB in both cases. An accurate estimate of the SRT is important when this method is to be used for algorithm evaluation. The slope estimate converges after about 300 iterations and the result is biased. These results are in agreement with the results presented in [5].

5. EXPERIMENT USING BASIE

In this Section, we present results from a pilot study of subjective SRT evaluation with the aim to demonstrate the use of BASIE in a more realistic scenario.

A listening experiment was performed with six subjects, all fluent in the English language. None of the subjects were aware of any significant hearing loss. The speech data comprised anechoic recordings of digit triplets drawn from the TIDigits database [7]. All speech samples were normalised to have the same active level according to the procedure described in the ITU-T P.56 standard [8] before noise was added with an intensity adjusted to the required SNR. The noise considered for the listening experiments was a car noise recording from the inside of a car. The samples were played through a RME Fireface 800 and Sennheiser HD650 headphones. At the beginning of the experiment, each subject was asked to adjust the audio to a comfortable listening level which was then kept fixed throughout the experiment. The first five samples were unprocessed noisy speech presented at SNR= 0dB; these were excluded from the recorded results and were used as training samples to give the subject a chance to familiarise themselves with the user interface.

Two separate experiments were performed with two different algorithms:

- a noise reduction module from a commercial audio workstation;
- a MATLAB implementation of spectral subtraction [9] with the noise estimated using minimum statistics [10]. The implementation (`specsub.m`) can be found in the MATLAB toolbox VOICEBOX [11].

Both algorithms have several adjustable parameters but the parameter that has the most apparent perceptual effect was found to be the control of maximum noise attenuation. Consequently, the algorithms were executed with varying maximum attenuation according to the following levels: Maximum Noise Attenuation (dB) $= \{-1, -5, -10, -20, -30, -40\}$. The remaining parameters for each algorithm were set to the default values as prescribed by the algorithm implementation. The subjects were asked to perform the experiment twice under identical conditions. The two sets of experiments were undertaken in two consecutive days. For each experiment, BASIE was run for 150 iterations taking approximately 10 minutes per subject per experiment.

The spectral subtraction algorithm was executed from MATLAB and the probed SNRs were used as suggested by the adaptive procedure in Section 2.1. The commercial system does not allow control of the algorithm parameters through software. Instead, the speech signals were preprocessed for all settings and at SNRs ranging between -20 dB and 5 dB in steps of 1 dB. Consequently, the probe SNRs suggested by the adaptive procedure were rounded to the nearest available integer SNR.

The results are presented in Table 1 and are given in terms of the mean and the standard deviation (SD) of the processing gain, defined in (11), calculated over both tests and over all six subjects. The SRT for each condition was calculated as an average of the last ten trials when it was assumed that the PF estimation has converged. These results are over a relatively small number of subjects and the absolute values are not the main point of interest. However, two valuable observations can be made:

(i) For the commercial system, there is a small improvement of intelligibility for the attenuation settings in the range of -5 dB to -1dB. This could serve as an indicator of "safe operational regions" for this type of noise when intelligibility is important.

(ii) The method of spectral subtraction appears much less sensitive to the noise reduction level in terms of intelligibility and intelligibility is always compromised by about 2 dB.

Another interpretation of these results is that, if a larger scale listening experiment was to be performed for the commercial system, it would be useful to consider the region of noise suppression above -10 dB. Thus, BASIE could serve as a tool of pilot studies prior to larger scale experiments.

Table 1: Results from the listening experiments in terms of mean and standard deviation (SD) of the processing gain (positive gain indicated in bold) calculated over six subjects and two tests.

Maximum Noise Attenuation (dB)	Processing Gain (dB) Mean/SD	
	Commercial System	Spectral Subtraction
-1	**0.50** / 2.2	-2.56 / 1.32
-5	**0.67** / 1.57	-1.39 / 2.65
-10	-0.15 / 1.9	-1.74 / 2.91
-20	-3.69 / 2.42	-0.72 / 1.07
-30	-4.74 / 1.9	-1.72 / 2.00
-40	-8.48 / 2.71	-1.77 / 1.56

6. CONCLUSIONS

We have presented a tool for rapid intelligibility estimation of speech in noise. The core of the tool is a Bayesian approach for selecting the SNR of the next trial so as to deduce as much information as possible about the underlying psychometric function. The method was shown to be able to estimate the SRT with accuracy to within ± 1 dB in under 30 trials. It is also able to estimate the slope, although, this requires over 300 trials and the accuracy of the estimates can still be improved. This is not a major drawback for the current application since the key interest is the relative SRTs between processed and unprocessed speech. It was demonstrated using results from a listening experiment how BASIE can be used to evaluate speech processing algorithms and to identify the existence of "safe operational regions" where noise suppression may be applied with little risk of compromising intelligibility.

7. REFERENCES

[1] S. A. Klein, "Measuring, estimating, and understanding the psychometric function: A commentary," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1421–1455, 2001.

[2] C. Smits and T. Houtgast, "Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests," *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1608–1621, Sep 2006.

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

[4] K. Worrall, R. Fellows, J. Causer, and L. Craigie, "Inteligibility testing at HM Government Communications Centre," *Proc. Institute of Acoustics*, vol. 28, no. 6, p. 12, 2006.

[5] L. L. Kontsevich and C. W. Tyler, "Bayesian adaptive estimation of psychometric slope and threshold," *Vision Research*, vol. 39, pp. 2729–2737, 1999.

[6] R. Plomp and A. M. Mimpen, "Improving the reliability of testing the speech reception threshold for sentences," *Audiology*, vol. 18, pp. 43–52, 1979.

[7] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Mar. 1984, pp. 328 – 331.

[8] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[9] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[11] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html