Intelligibility Estimation in Law Enforcement Speech Processing

Patrick A. Naylor[1], Nikolay D. Gaubitch[1], Dushyant Sharma[1], Gaston Hilkhuysen[2], Mark Huckvale[2] and Mike Brookes[1]

Centre for Law Enforcement Audio Research (CLEAR),
[1]Department of Electrical and Electronic Engineering, Imperial College London, UK
[2]Division of Psychology and Language Sciences, University College London, UK
E-Mail: p.naylor@imperial.ac.uk
Web: www.ee.ic.ac.uk/naylor

Abstract

Speech recordings obtained in the context of law enforcement are often degraded in terms of quality and intelligibility. Several techniques for assessing the impact of speech enhancement algorithms on quality are available, both intrusive and nonintrusive, but the assessment of intelligibility is usually reliant on expensive and time consuming subjective listening scores. To address this issue, we describe some recent scoring experiments and an adaptive Bayesian procedure which efficiently estimates properties of the psychometric function from a small number of listening tests. A data-driven nonintrusive objective intelligibility estimation method is also described and tested on car and babble noise. It is shown to give intelligibility estimates that are well correlated with subjective scores. We aim to improve our understanding of the quality/intelligibility tradeoff and to study speech processing tools in the critical context of law enforcement.

1 Introduction

The work of Law Enforcement Agencies (LEAs) can be greatly assisted by audio recordings of, for example, police interviews or recordings made by the pubic on mobile phones. It is inevitable that these recordings will have widely varying sound quality and intelligibility since they are made in uncontrolled scenarios. Intelligibility is important so that a recording will be admissible as evidence in court and so that accurate transcripts can be made. Quality is important since the rate and accuracy of transcription is reduced for audio that is degraded by, for example, background noise or reverberation. The cost of transcription for noisy audio is about 50 % higher than for clean speech. The LEAs therefore require speech processing tools to improve both intelligibility and quality.

In the telecommunications sector, where speech has a Signal-to-Noise Ratio (SNR) typically between 40 dB and 0 dB, speech enhancement algorithms have been researched for many years [1, 2] with some notable successes. In law enforcement applications, the range of SNR is wider and extends to at least -20 dB. High levels of noise may be compounded with other degradations such as reverberation and nonlinear distortion. Speech intelligibility in law enforcement is therefore often much less than 100 %.

To develop and test speech enhancement technology for law enforcement, test data containing realistic types and levels of degradation is needed. Data from the telecommunications sector [3] is not usually adequate. It is also necessary to employ performance metrics for both quality and intelligibility. The effect of speech enhancement algorithms on speech quality can in principle be assessed using objective intrusive measures such as Perceptual Evaluation of Speech Quality (PESQ) [4], objective nonintrusive measures such as Low-Complexity, Nonintrusive Speech Quality Assessment (LCQA) [5] and subjective scoring techniques such as Mean Opinion Score (MOS) [6]. The effect of speech enhancement algorithms on intelligibility usually requires listening tests which are expensive and time consuming. It is therefore highly desirable either to employ subjective scoring techniques which are as efficient as possible and/or to employ objective measures.

In the remainder of this paper we review recent subjective scoring experiments targeting intelligibility. We then consider measurement of intelligibility and present a Bayesian Adaptive Speech Intelligibility Estimation (BASIE) method that efficiently estimates the Psychometric Function (PF) in a given noise condition. Lastly we review recent work on nonintrusive objective measures.

2 Scoring Experiments

A number of techniques have been presented in the literature for obtaining subjective quality and intelligibility scores for degraded speech [7].

Intelligibility can be measured at the phone level using nonsense syllables [8, 9], at the word level [10, 11] and the sentence level [12, 13] as well as using modified rhyme tests [14, 15] and diagnostic rhyme tests [16]. Experimental work to date within CLEAR has focused on measurements based on keywords.

Intelligibility may be characterized using a PF plotting intelligibility as a function of distortion level, usually SNR. An example is shown in Fig. 2. The impact of a speech enhancement algorithm on intelligibility can be seen by comparing the PFs before and after processing.

The study by Hu and Loizou [17] shows the effect of speech enhancement as a reduction in the intelligibility of the signal. In contrast, the same processing has been shown to improve SNR and increase perceived quality [18].

A study of the manner in which experienced operators use a commercial speech enhancement system was presented in [19]. The operators worked on three concatenated IEEE sentences [20], which were corrupted by babble noise at five SNRs, ranging from 0 to -12 dB SNR in 3 dB steps. The operators were asked to find the settings of the controls which gave what they perceived to be "maximum intelligibility". Each operator adjusted the controls for a particular SNR five times. Using an Analysis of Variance, it was found that there was no correlation amongst the choice of settings so that each operator had a different opinion on which settings gave best intelligibility. Subsequently, the intelligibility of the enhanced and unprocessed speech was measured by normal hearing naive listeners and reported as the $\log_2$ of the ratio between the number of correct and number of incorrect keywords

in the listeners' responses, known as the Berkson scale. It was found that the effect of the speech enhancement was to reduce intelligibility by about 1 Berkson, meaning that for a fixed number of correct words, the number of incorrect words doubled due to speech enhancement.

The effects of noise suppression on intelligibility were further studied using 200 IEEE sentences [20] from a male speaker. These were corrupted by car and babble noise at 5 SNRs chosen in each case to correspond to Speech Intelligibility Index (SII) [21] values of 0.1, 0.3, 0.5, 0.7 and 0.9. The noisy speech was then processed by spectral subtraction [22] (SS), minimum mean square error spectral estimation [23] (MMSE) and subspace enhancement [24, 25] (SBS). The noise spectrum was estimated using the minimum statistics algorithm [26, 27]. The resulting database contained 20 conditions per enhancement type. Listening tests were performed with 20 subjects for SS and MMSE experiments, where it was found that sufficient statistical power was achieved with 10 subjects. The experiment design was based on a Latin square design, each subject performing the scoring task in 10 sessions (with a randomization of presentation order). The task was to give a verbal response to the stimuli. The results for the SS experiment are shown in Fig. 1. A reduction in intelligibility is observed for both car and babble noise when the speech processed by spectral subtraction is presented to the listeners. Similar results are obtained for the MMSE and SBS techniques, highlighting and quantifying the intelligibility loss due to application of speech enhancement algorithms.

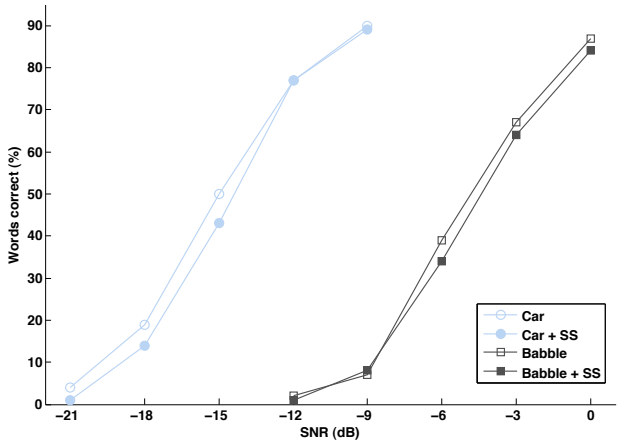

Figure 1: Subjective intelligibility scores for car, babble noise and enhancement of the same by spectral subtraction.

3 Databases for Validation

The C-Qual database has been presented in [28] and contains degraded speech with mean opining scores. The clean speech used is the same as in ITU-T P.23 [3] allowing comparisons. The degradations are intended to represent law enforcement situtations and include additive car, babble and hum noise, reverberation, colloration and some nonlinear effects. As a first experiment, this database has been used to validate the performance of the PESQ quality measure in the law enforcement context. These tests showed that PESQ in its current form is more suitable for use in additive noise conditions (correlation of 0.94 for car, babble and hum in -30 to 30 dB SNR range) and correlates poorly with subjective quality scores for non-linear distortions such as peak clipping and drop-outs (correlation of 0.17).

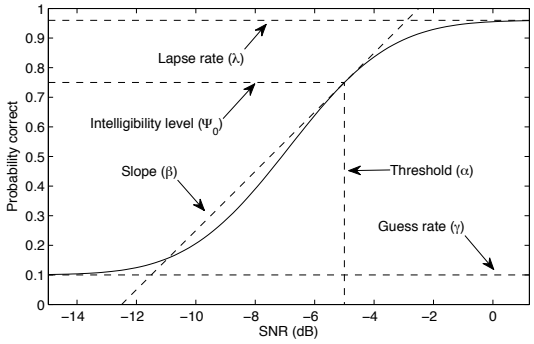

Figure 2: Psychometric funciton for speech intelligibility in noise.

4 BASIE

The PF for speech intelligibility in noise is often modelled as a sigmoid function parametrized in terms of the SNR corresponding to a chosen intelligibility level, Ψ_0, and the slope of the PF at this SNR. This can be written as [29, 30]

$$\Psi(x) = \gamma + (1-\gamma-\lambda)\Phi(x), \tag{1}$$

where x is the SNR, Ψ is the probability of a correct response, λ is the lapse rate and γ is the guess rate. An example is shown in Fig. 2.

In BASIE we model $\Phi(x)$ as a cumulative normal distribution so that the slope, β, and the threshold, α, at a chosen $\Psi(x) = \Psi_0$ are governed by the distribution's variance, σ, and mean, μ. BASIE estimates these terms using a technique similar to [31]. At each iteration n, the listening 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 $r_n = 0$ otherwise. For the next iteration, the probe SNR is adjusted according to some rule depending on the outcomes of previous iterations. The objective of BASIE is to select the probe SNR of the next trial such that we obtain as much information as possible about the PF in order to estimate the Speech Reception Threshold (SRT) and the PF slope at the SRT within the minimum number of iterations. One approach [32] is to choose the SNR probe value for iteration x_{n+1} such that a weighted sum of the expected variances of the estimates of the threshold, α, and the slope, β, are minimised by

$$\begin{aligned} x_{n+1} = \arg\min_x \big(&(1-\kappa) E\left\{ Var_{n+1}\left\{ \alpha \mid x \right\} \right\} \\ &+ \kappa E\left\{ Var_{n+1}\left\{ \beta \mid x \right\} \right\} \big), \end{aligned} \tag{2}$$

where here we use a 75% SRT threshold, $\kappa = 0.5$ and $E\{\cdot\}$ is the expectation operator.

We define a two dimensional Probability Density Function (pdf), $p(\theta)$ that specifies the probability space of all possible PFs, where $\theta = (\alpha, \beta)^T$ is a two-dimensional vector containing the values of the threshold α and the slope β at $\Psi(x) = \Psi_0$. At the nth iteration, the pdf is updated with the new result according to

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

Maximum Noise Attenuation (dB)	Processing Gain (dB)
-1	0.5 ± 2.2
-5	0.67 ± 1.57
-10	-0.15 ± 1.9
-20	-3.69 ± 2.42
-30	-4.74 ± 1.9
-40	-8.49 ± 2.71

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

where $P(r = 1 \mid x, \theta) = \Psi(x)$, $P(r = 0 \mid x, \theta) = 1 - \Psi(x)$ 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. This process is repeated until satisfactory convergence is achieved. The algorithm is executed for a fixed number of iterations, normally determined empirically.

We next consider the use of BASIE to measure the effect of speech enhancement on intelligibility. In this context, we evaluate a speech enhancement algorithm by the processing gain defined as the difference in SRT between the processed noisy speech and the unprocessed noisy speech such that

$$\text{Processing Gain} = SRT_{Noisy} - SRT_{Processed}. \qquad (4)$$

The ability of BASIE to estimate more than two PFs in one experiment allows simultaneous evaluation of the processing algorithm with various parameter settings.

We carried out a listening experiment with six subjects using a Noise Reduction Module (NRM) from a commercial audio workstation to enhance speech degraded by car cabin noise. The original speech data was anechoic digit triplets from the TIDigits database, which were normalized to have the same activity level [33] before noise was added with an intensity adjusted to the required SNR. The samples were presented to the listeners through an 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$=0$ dB; these were excluded from the results but enabled the subjects to familiarise themselves with procedure.

The NRM has several adjustable parameters. We considered here only the parameter with the greatest apparent perceptual effect: the 'maximum noise attenuation' setting. This was varied as: Maximum Noise Attenuation (dB)$=\{-1,-5,-10,-20,-30,-40\}$. The remaining parameters were set to the default values prescribed by the algorithm implementation. The subjects were asked to perform the experiment twice under identical conditions. The two sets of experiments were undertaken on two consecutive days. For each experiment, BASIE was run for 150 iterations taking approximately 10 minutes per subject per experiment.

The results shown in Table 1 are given in terms of mean and standard deviation of the processing gain calculated 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 as shown in Fig. 3. These results are over a relatively small number of subjects such that their statistical significance is not confirmed. However, a valuable observation is the small improvement of intelligibility seen for the attenuation settings in the range of -5 dB to -1 dB. This type of result could serve as an indicator of 'safe operational regions' for a speech enhancement algorithm over which speech quality may be improved without degrading intelligibility.

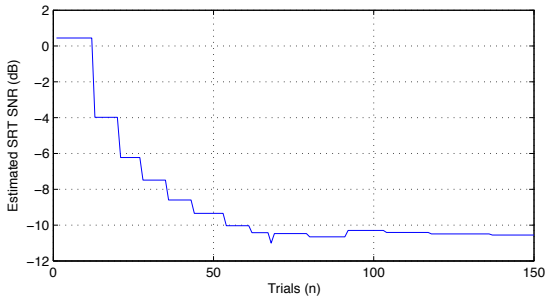

Figure 3: Example of BASIE convergence for unprocessed noisy speech.

5 Data-driven Objective Intelligibility Estimation

Regardless of the efficiency of a subject-based intelligibility test, it is still nevertheless desirable to be able to estimate intelligibility of a segment of audio automatically. This permits offline processing to seek out intelligence in long audio recordings such that, for example, transcription can be attempted on only those segments with intelligibility estimates above a chosen threshold. Feature based methods for quality estimation have already been presented, for example [5]. We now describe a non-intrusive objective intelligibility estimation method and present some initial results with car and babble noise and for speech processed by spectral subtraction.

Low-Complexity, Nonintrusive Speech Intelligibility Assessment (LCIA) [34] employs a feature set [5] derived from Linear Predictive Coding (LPC) supplemented by the Importance Weighted Signal-to-Noise Ratio (iSNR) resulting in 11 features per frame. An utterance, or segment of audio, is represented by 44 statistical per-frame features. We apply a two-step dimensionality reduction scheme using feature subset selection based on feature correlations followed by a feature extraction step based on PCA using a development data-set. It was found that selecting 8 features from the 44 global features and extracting 7 linear combinations after the feature extraction gave good results. A joint Gaussian Mixture Model (GMM) is trained on the 7 resulting features and the intelligibility score for each speech utterance in the training data.

Tests were performed using spectral subtraction applied to 20 different noise conditions in 200 IEEE sentences as employed in Section 2. A 50% cross-validation training scheme was used in which the data is equally divided into test and training sets, with the training set containing all the

conditions present in the test set. However, the test speech material is not available in the training set. The test and training sets are swapped and the performance is the average over the two sets. The results we have obtained for LCIA have a Spearman correlation coefficient [35] of 0.96 with SII and 0.92 with subjective intelligibility scores. The statistical properties of the spectral dynamics were found to be the most important feature (with an individual correlation of 0.90 with intelligibility) suggesting that the rate of change of the spectrum provides important information for intelligibility. This finding is consistent with the use of modulation domain features.

6 Conclusion

Subjective scoring experiments indicate that some speech enhancement processes may improve perceived quality at the cost of reduced intelligibility. An efficient technique known as BASIE has been reviewed which can estimate the SRT on the PF in a few minutes of a listeners time. Nevertheless, it is still advantageous to investigate fully automatic objective measures of intelligibility and one such measure, LCIA, has been shown to be well correlated with human intelligibility scores. These types of measures are essential tools for appropriate deployment of speech enhancement algorithms into critical applications of law enforcement.

References

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, 2006.

[2] P. C. Loizou, *Speech Enhancement Theory and Practice*. Taylor & Francis, 2007.

[3] *ITU-T coded-speech database*, International Telecommunications Union (ITU-T) Supplement P.Sup23, Feb. 1998.

[4] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[5] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[6] *Methods for subjective determination of transmission quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.800, Aug. 1996. [Online]. Available: http://www.itu.int/rec/T-REC-P.800/en

[7] M. Brookes, N. D. Gaubitch, M. Huckvale, and P. A. Naylor, "Speech cleaning literature review," 2009. [Online]. Available: www.clear-labs.com/Tutorial-LitReview/index.html

[8] H. Fletcher and J. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, vol. 8, pp. 806–854, 1929.

[9] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Am.*, vol. 31, no. 6, pp. 667–677, Jun. 1959.

[10] J. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58(9), pp. 955–991, 1948.

[11] I. Lehiste and G. E. Peterson, "Linguistic considerations in the study of speech intelligibility." *J. Acoust. Soc. Am.*, vol. 31, no. 3, pp. 280–286, 1959.

[12] D. Kalikow, K. Stevens, and L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1337–1351, 1977.

[13] M. Nilsson, S. Soli, and J. Sullivan, "Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[14] G. Fairbanks, "Test of phonemic differentiation: the rhyme test," *J. Acoust. Soc. Am.*, vol. 30, no. 7, pp. 596–600, 1958.

[15] A. House, C. Williams, M. Hecker, and K. Kryter, "Articulation testing methods: consonant differentiation with a closed response set," *J. Acoust. Soc. Am.*, vol. 37, no. 1, pp. 158–166, 1965.

[16] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, 1983.

[17] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, pp. 1777–1786, 2007.

[18] ——, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, Jul. 2007.

[19] G. Hilkhuysen and M. Huckvale, "Adjusting a commercial speech enhancement system to optimize intelligibility," in *Proc AES Conf on Audio Forensics*, Hillerød, Jun. 2010.

[20] M. W. Smith and A. Faulkner, "Perceptual adaptation by normally hearing listeners to a simulated "hole" in hearing," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4019–4030, 2006.

[21] ANSI, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, ANSI Standard S3.5–1997 (R2007), 1997.

[22] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[24] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[25] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[26] 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.

[27] ——, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, Jun. 2006.

[28] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, M. Brookes, and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in *Proc AES Conf on Audio Forensics*, Hillerød, Jun. 2010.

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

[30] 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.

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

[32] N. D. Gaubitch, M. Brookes, P. A. Naylor, and D. Sharma, "Bayesian adaptive method for estimating speech intelligibility in noise," in *Proc AES Conf on Audio Forensics*, Hillerød, Jun. 2010.

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

[34] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Denmark, Aug. 2010.

[35] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*. Englewood Cliffs, NJ: Prentice-Hall, 1998.