3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

Performance-Based Measurement of Speech Quality with an Audio Proof-Reading Task

Mark Huckvale, Gaston Hilkhuysen, Deizom Frasi

Department of Speech, Hearing & Phonetic Sciences
University College London, London, U.K.
m.huckvale@ucl.ac.uk, g.hilkhuysen@ucl.ac.uk, n.frasi@ucl.ac.uk

Abstract

Existing measures of speech intelligibility and speech quality can be ineffective for evaluating new types of speech communication systems, such as wideband audio codecs, digital hearing aids and noise-reduction systems. We propose that new performance-based evaluation methods are required which tap into the cognitive effort listeners employ to understand speech through such systems. We present an example of such a method, based on the correction of transcripts of fluent spontaneous dialogues, and evaluate it for six different signal qualities, including telephone, added noise and noise-reduced conditions. We show that signal quality has a significant effect both in terms of transcript error detection accuracy and in terms of processing speed. We also show that in this test noise-reduction did not have any beneficial effect, despite the commonly recorded opinion that noise reduction improves signal quality.

Index Terms: speech quality, objective measures, proof-reading

1. Introduction

Speech signal processing technologies are exploited in the telecommunications, hearing aid, speech technology and forensic audio industries. Using these technologies speech signals are, among other things, encoded, decoded, compressed, companded, filtered, equalized, synthesized and noise-reduced. But with these processing techniques comes the necessity to assess their impact on the users of a speech communication system. Evaluation methods are required to check that systems are of adequate performance for their application. Evaluation can also be used to compare one system with another, to determine the optimal settings for processing parameters, or to aid in the development of new technologies. However, it is still unclear how best to evaluate a speech communication system. How should one assess a wide-band speech coder? An equalizer? A speech synthesis system? A noise-reduction system?

The "gold-standard" for the assessment of speech communication systems is intelligibility testing. Spoken utterances are processed through the system under test to listeners who immediately report back what was said. Intelligibility, measured in terms of % words correct can then be compared to the scores obtained from the utterances before processing to provide an objective estimate of the change in intelligibility caused by the system. Such testing has a long history in the telecommunications area, building on the work of Harvey Fletcher [1], and has been the subject of much empirical analysis. Modern testing protocols ensure reliable, sensitive and unbiased scores from a panel of human listeners [2]. Furthermore intelligibility testing has also given rise to predictive models of the change in intelligibility caused by processing. For example, the Speech Intelligibility Index (SII, [3]) provides a good account of changes to intelligibility caused by changes in gain or by linear filtering within a system.

When intelligibility testing is performed as a function of the signal-to-noise ratio (SNR) of the speech, a typical 'S'-shaped psychometric function is obtained. The slope of the curve is related to the redundancy in the speech materials. Meaningful sentences, where the context aids the recognition of ambiguous words, tend to have steep functions (small changes in SNR leading to big changes in intelligibility), while less-redundant materials, for example digits, tend to have shallower curves. The SNR value corresponding to the 50% intelligibility point on the curve is called the Speech reception Threshold (SRT). A common way to express the effect of a test system or channel on intelligibility is in terms of an "equivalent" change in the SRT caused by a change in SNR. Thus if some telecommunication system decreased intelligibility by some fraction, that change can be expressed as an equivalent change in SNR measured in dB.

From this description it should be easy to see that this kind of performance testing is of no use when the system under test makes very little change to intelligibility. Testing is particularly problematic when the intelligibility of the speech is high, the psychometric function is flat, and large changes in SNR provoke only small changes in intelligibility. Intelligibility testing then loses statistical power, since much larger experiments are required to estimate the changes in intelligibility scores with any degree of reliability.

Unfortunately, there is now an urgent need to assess the performance of systems that have very little impact on intelligibility. Presented with good quality speech on their input, modern telephones, VOIP systems or hearing aids output speech of high intelligibility. In speech synthesis, modern systems produce highly intelligible read speech from text [4]. Likewise contemporary noise reduction systems have been shown to make only minor changes to intelligibility [5]. With these systems, evaluation through conventional intelligibility testing makes no sense: it is neither effective (because of the difficulty of measuring small changes in intelligibility where the psychometric function is flat), nor relevant (since the goals of much contemporary speech signal processing is not to maximise intelligibility). This has become known as the challenge of measuring "speech quality" rather than intelligibility. What matters about such systems is not the intelligibility of the speech but factors such as "clarity", "comfort", "listening effort", "fatigue", or "pleasantness". Evaluation has changed from being about the primary goal of communicating word identity into being about secondary or "emergent" properties of communication related to the

10.21437/PQS.2010-14

cognitive processing performed by listeners to satisfy the primary goal.

In this paper we set out an experimental methodology which seeks to investigate listener performance in a speech communication task even with speech of high intelligibility. Our aim is to work towards objective tests of speech quality that can be used to complement intelligibility testing. First we will discuss what is wrong with existing measures of speech quality, and the design goals for our new performance-based alternative.

2. Measuring Quality

Unfortunately, if you mention the term "speech quality" to a telecommunications engineer, he or she will think you are talking about the "mean-opinion-score" (MOS) rating scale of quality as defined by a number of industry standards (e.g. ITU P.800 [6]). The use of the MOS quality measure, in which listeners are asked to rate speech signals on a scale of 1-5 between "poor" and "good", is widespread and firmly established in the industry as synonymous with quality itself.

It is not hard to demonstrate the weaknesses of such a simple rating scale for assessing quality: (i) that listeners opinions of quality are based on many signal characteristics, and (ii) that listeners vary in the weight they give to different characteristics; (iii) that variability across and within listeners lead to experiments of low statistical power; (iv) that listeners are easily biased by the experimental conditions; (v) that experiments are short and untypical of everyday communication tasks; and (vi) that what listeners prefer is not the same as what is best for them. On the last point, it has been shown that listener scores vary even with the presentation level of the signals and that the best scores for a system can be obtained at levels other than the ones preferred by the listeners! [7].

The use of MOS to measure speech quality is, we propose, an admission of a failure to find an adequate operational definition of quality. The MOS scale is a way of obtaining an average opinion about the quality of a system without having to come to terms with what quality actually is.

So what makes a "quality" system? Harvey & Green [8] provide some possible definitions:

- a quality system is exceptional, that is, it stands out from amongst its peers, or
- a quality system is an expression of perfection, that is, it has no identifiable defects, or
- a quality system is fit for purpose, that is, it meets a set of goals set out for its application, or
- a quality system provides value for money, in that its performance and its cost are in balance, or
- a quality system is transformative, that is, it makes a step-change to the experience of its users.

From these, the definition that is perhaps most relevant to our current application and the easiest to operationalise is "fitness for purpose". If we could establish a set of criteria based on users' requirements and expectations, we may be able to create a quantitative scale of the extent to which a system is fit for its purpose.

Taking a system designed for communication of speech signals, what qualities would users expect a system to have aside from good intelligibility? Our suggestion is that the system should facilitate communication, in other words that it should make the job of working on speech tasks less effortful. This may not be the only possibility, users may like the signal to have a particular timbre, or sound "pleasant". But gauging quality in terms of listener effort - how much cognitive work the listener has to put in to complete their everyday tasks - is at least amenable to objective quantification. Also there are established cognitive models of effort, which make a link between the complexity of the task and the extent to which subjects make mistakes, lose attention, or become tired. A common assumption is that we draw cognitive resources from a limited pool. By increasing the complexity of the task, subjects run out of resources and so fail to fully evaluate the information required to make decisions [9]. Thus we propose to rebase the measurement of speech signal quality in terms of the cognitive effort required to conduct a speech communication task.

3. Performance-based Measures of Quality

The study of the effect of noise on human performance has a long history, and there are many psychoacoustic models that can be used to predict likely intelligibility performance from signal SNR. However few studies have investigated the impact of signal quality on cognitive load for signals of high intelligibility. Previous studies known to us are those of by Sarampalis *et al* [10], by Durin *et al* [11], and by Huckvale & Leak [12].

In the study by Sarampalis, subjects were asked to repeat and memorise words from sentences spoken in noise. Comparisons were made in task accuracy between noisy speech and noise-reduced (NR) speech processed by the MMSE algorithm [13]. Generally word intelligibility performance was reduced by NR processing, although recall performance was improved in one test condition. The dual-task design seems to have made the listening test much harder for the subjects, so causing them to make errors even for good quality speech. Our interpretation of this result is that by adding a memory task to the intelligibility test, Sarampalis has effectively shifted the psychometric function to the right, so as to obtain changes in performance with changes in signal quality for signals of otherwise similar intelligibility.

Durin's study investigated the effect of telephone codecs on performance in a letter recognition task and a digit memory task. In the letter task subjects hear a spoken description of a letter and have to respond quickly whether the description matches a displayed letter. In the digit memory task, five spoken digits are played to the subject who must subsequently indicate whether a displayed digit was one of the set. Interestingly, results show changes in both recognition accuracy and in reaction time with changes in codec bit rate. While changes in accuracy could be attributed to a shift in the intelligibility function caused by the dual-task, the shift in reaction times adds a different dimension to the experiment - tapping into effects of the signal on the cognitive processing required by the task. Durin *et al* suggest poorer quality signals make greater demands on cognitive resources which makes the words harder to remember, so causing increases in decision times and a reduction in accuracy.

In the study by Huckvale & Leak, reaction times to the identification of spoken digits were measured directly under 5 different signal conditions. The conditions involved the addition of car noise or babble to clean speech, and the subsequent processing by an MMSE noise-reduction algorithm [13]. Results showed that changes in signal quality made significant changes to the reaction times of listeners to the spoken digits. Importantly, however, these changes in reaction time occurred even though the accuracy of the listeners in the task of identifying digits did not change across conditions.

From these studies we can now contrast two explanations of a speech communication error: a signal is of *poor intelligibility* if errors are made even if the signal is given the listener's full attention and unlimited time to respond, while a signal is of *poor quality* if errors are made because signal does not allow the listener to give it their full attention or enough processing time. From this distinction between intelligibility and quality we can begin to identify the requirements for an evaluation methodology that would provide us with a means of assessing the effect of signal quality on speech communication. A method should:

(i) be based on objective measurements, that is, measurements of human performance not human opinion,

(ii) exploit increases in cognitive load caused by a complex task to shift the psychometric function of intelligibility so that subjects make errors even for otherwise highly intelligible signals,

(iii) include measures of reaction time or other physiologically-based signals to add a dimension of measurement directly related to cognitive load.

There is another requirement, however, not met by the previously mentioned studies:

(iv) be based on a speech task relevant to the situation in which the communication systems is used.

The previous studies used only isolated words rather than realistic speech materials, a fact that makes it harder to justify the value of the experimental results to the users of a speech communication system.

Lastly, the quantitative analyses in these studies are just on the edge of statistical significance. Thus overall, the goals of our work are to extend current research in two-ways: to improve the relevance and the reliability of performance-based measures of speech signal quality.

4. Audio Proof-Reading

The experimental task we have designed is based on audio proof-reading. This task was chosen because of its clear similarity with a number of everyday speech activities, such as dictation, interview transcription, or general business transactions over the telephone. Transcripts of a spontaneous conversation are corrupted with typical word insertion, word deletion and word substitution errors. In the listening task, subjects listen to the conversation in real-time and must identify the location of transcript errors from a displayed transcript. Subjects are encouraged to find as many errors as possible, with the expectation that increased listening effort would decrease the number of errors detected and increase the number of false alarms. As well as primary accuracy on the task, the idea is also to measure the average processing delay between the listener hearing the relevant word and marking the transcript. It would be expected that increased listening effort would lead to increased response times.

The design of the task allows us to control complexity through the number and type of errors introduced into the transcript. Measurements of accuracy and response speed provide two means to assess cognitive load. The use of spontaneous speech gives the method ecological validity. We now describe the particular implementation used in our experiment.

5. Methodology

Recordings of two speakers discussing the differences between two pictures were used as a source of natural spontaneous speech. Four minute extracts from six different spontaneous conversations were used. These were downsampled to 16000 samples/sec, and amplitude equalised using a moving window of size 10 seconds. The two speaker signals were then added to create a monophonic signal which was presented binaurally. Two noise conditions were created from the original speech by adding babble noise at +6dB SNR, or car cabin noise at -3dB SNR. The noise levels were chosen on the basis of other studies we have performed on the effect of these two noise types on word intelligibility. The levels chosen give SII performance indices greater than 0.9 and informal listening tests showed that the materials were still highly intelligible. The noise conditions were then processed with an MMSE noise reduction algorithm [13] as implemented in VOICEBOX [14] to create two further noise-reduced conditions. Lastly, a telephone signal quality condition was produced by band-passed filtering the original audio between 300 and 3200Hz, downsampling to 8000 samples/sec, applying a mu-law encoding and decoding, then upsampling back to 16000 samples/sec; this simulates the G711 telephone codec. Thus there were six audio conditions: Quiet, Telephone, Babble Noise, Babble Noise+NR, Car Noise, Car Noise+NR.

Transcripts of the spoken dialogue extracts were randomly corrupted with 50 errors: 30 word substitutions, 10 word insertions and 10 word deletions. To disguise the corruptions, so that they could not be guessed from the transcript alone, word edits were chosen from equivalent contexts found in other transcripts of different speakers describing the same pictures. See Table 1 for examples.

Error type	**Original**		**Error**
Substitution	just says push	⇨	just saying push
	got brown hair	⇨	got blonde hair
	just says peaches	⇨	just have peaches
Insertion	in <S2:>	⇨	in it <S2:>
	yeah and	⇨	yeah peaches and
	I think	⇨	I don't think
Deletion	and there's a	⇨	and a
	the top right	⇨	the right
	then just below	⇨	then below

Table 1: Examples of errors introduced into the transcripts. (<S2:> = change of speaker)

To run the experiment, a program called the *Proofometer* was created to replay the audio, display the transcripts and collect the listener's responses, see Figure 1. Using the Proofometer program, the listener's task was to listen to the replayed conversation and click on substituted or inserted words, or click on spaces where words had been deleted. The audio, once started, could not be paused.

Twenty-five subjects attempted the task, although results from seven had to be discarded because of the low number of errors detected or because of a high number of false-alarms across all conditions. Of the remaining 18 subjects, each listener corrected a different transcript in each audio condition, with the transcripts, audio condition and processing order balanced across listeners. All listeners were British English speakers with no known hearing impairments. The experiment was conducted over headphones in everyday listening environments. Subjects chose their preferred presentation level

within a practice session which was then kept constant for all conditions.

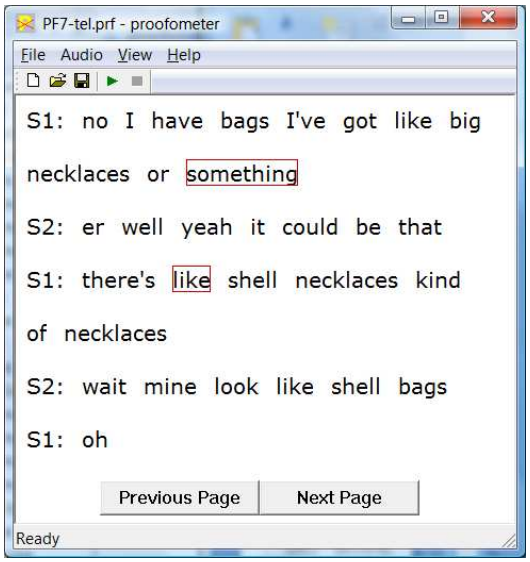

Figure 1: Proofometer interface. Users click on words or spaces to indicate errors.

6. Experimental Results

6.1. Transcript error detection

There are a number of ways to record listener performance on the transcript error detection task. Figure 2 shows the raw % correct errors detected for all listeners across all conditions.

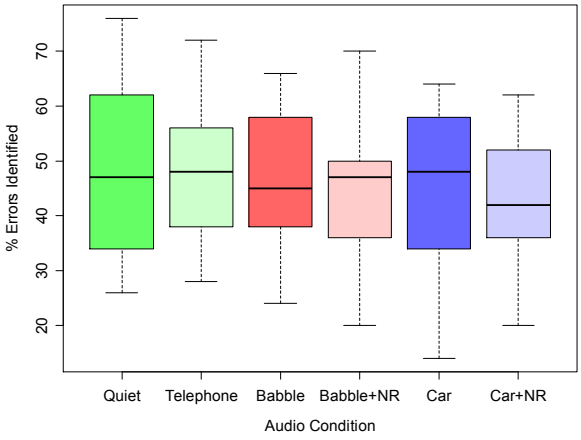

Figure 2: % error detection rate across audio conditions (N=18).

From the raw results it can be seen that there is great variability within a condition, and much overlap across conditions. The difficulty of the task seems to have been set about right, since subjects were not obtaining 100% correct scores even in the best conditions.

Figure 3 shows the % accuracy across conditions, calculated from the formula

100 × (# errors found – # false alarms) / 50

There were 50 transcript errors to be detected. Figure 4 shows the d-prime value across conditions, calculated from the formula:

InvNorm(#errors found/50)-InvNorm(#false alarms/50)

Where *InvNorm* means the inverse normal distribution. The false-alarm rate is calculated out of 50, since listeners knew that 50 errors needed to be detected, so would not be expected to make more than 50 button clicks in one trial. In fact figures 3 and 4 are very similar.

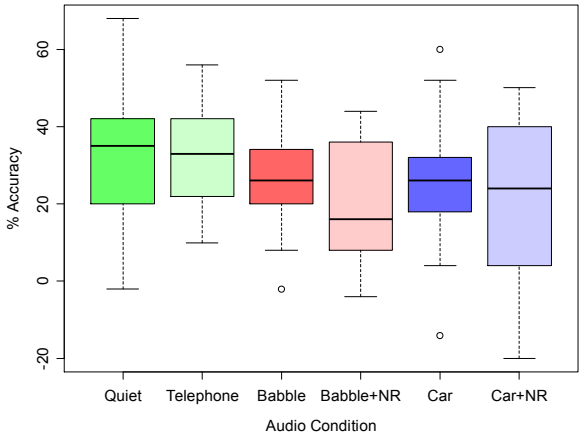

Figure 3. % accuracy across audio conditions (N=18).

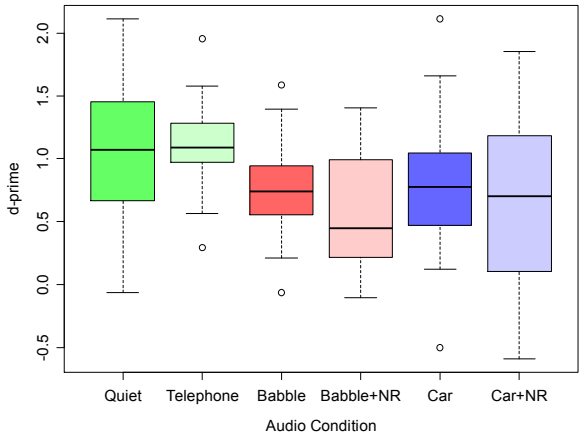

Figure 4. d-prime across audio conditions (N=18).

The difference between Figure 2 and Figures 3 & 4 is the inclusion of the false alarms. Because more false alarms occurred in the noisy and noise-reduced conditions, the effect of condition is now emphasized. However the graphs still show high variability within conditions, because of the wide variation in average performance on the task across listeners. The picture is clearer if we look in the change in performance *within* subjects across conditions. The change in % accuracy across conditions compared to each subject mean is shown in Figure 5. The figure shows how the addition of noise and the addition of noise-reduction both seem to degrade performance. An analysis of variance of % accuracy as a function of both condition (6 levels) and subject (18 levels), shows a significant effect of condition ($F(5,85)=4.6$, $p< 0.001$) and of subject ($F(17,85)=5.4$, $p< 0.001$). A Tukey post-hoc analysis shows significant differences between the noise-reduced

conditions and each of the quiet and telephone conditions. Similar results were obtained from the d-prime measure.

The results show no measurable effect of the telephone quality processing compared to the full-bandwidth condition.

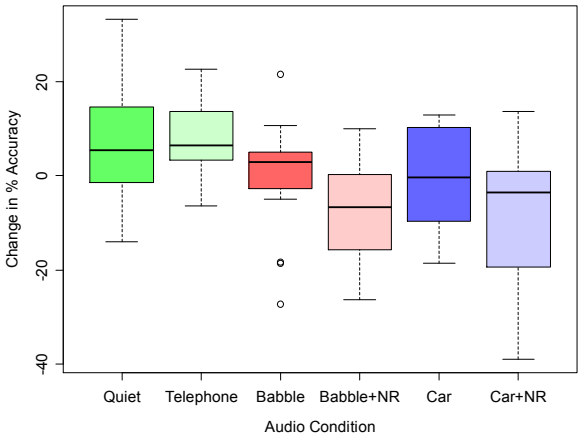

Figure 5. Relative % accuracy change per speaker across audio conditions (N=18).

6.2. Response Delay

Having established that changes in signal processing affected listener performance on finding transcript errors, we now turn to how quickly they responded. Reaction times were measured by estimating the time of each word in each recording and, for each correctly identified error, measuring the delay between the time at which the word was played and the time at which the mouse was clicked. The mean delay was then calculated for each subject and for each condition. The changes in response delay within each subject across audio conditions are shown in Figure 6.

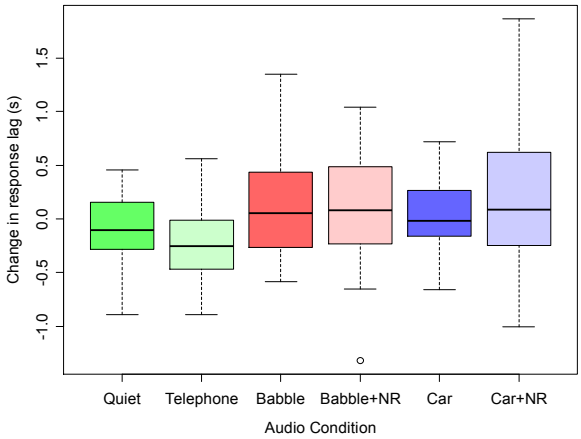

Figure 6: Change in response lag across audio conditions (N=18).

A weakness of our current method is that measurements of reaction time lack precision, with considerable random error in recording the replay time and the response time. This may be a contributing factor to the variability observed in Figure 6. Although the noisy and noise-reduced conditions seem to have larger response delays, no significant effects are observed. The situation can be clarified somewhat by pooling the conditions into the categories Quiet (Quiet+Telephone), Noisy (Babble+Car) and Noise-reduced (BabbleNR+CarNR), see Figure 7.

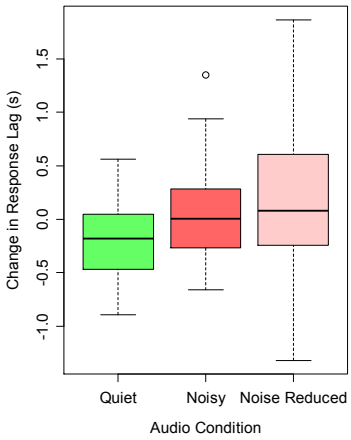

Figure 7: Change in response lag in unprocessed and processed audio conditions.

Analysis of variance of response delay across combined conditions (3 levels) and subjects (18 levels) shows a significant effect of condition ($F[2,88]=3.4$, $p=0.037$). A Tukey post-hoc analysis shows that the noise-reduced conditions together have a significantly longer delay than the quiet conditions, but are not significantly different to the unprocessed noise conditions.

7. Discussion

7.1. Implications for measures of speech quality

In this experiment we investigated whether the accuracy or the speed with which errors were identified in an audio transcript were affected by signal quality for 6 different audio conditions all of good speech intelligibility. We have been able to show that signal quality did affect listener performance on both the primary task of finding errors and on secondary effects of response delay. We interpret these results as meaning that changes in signal quality lead to measurable changes in cognitive effort required to process speech.

This work has replicated some of the results of the study by Sarampalis *et al* [10], in that we have shown how a complex communication task can cause changes in error-rate with signal quality even with materials of good intelligibility. However, in our experiment, the noise-reduction process made matters worse rather than better. Our results are more similar to those of Durin *et al* [11], in that we also observed increased reaction times as signal quality was degraded, although we did not observe a difference due to telephone quality processing. However, our task was based on normal, spontaneous dialogue materials rather than on digits.

Overall, although significant effects of signal processing condition were observed, there was also a great deal of variation across subjects and conditions which reduced the sensitivity of the testing. Thus to pursue this type of testing further we need to increase the power of the experimental design. We can do this in a number of ways: by reducing variability within the test materials (e.g. by making all the transcripts in the Proofometer test equally hard), by improving the training given to listeners (to reduce a small learning effect), by motivating subjects better (to reduce effects of

attention loss), or in the worse case, running larger numbers of subjects. The precision of our measurements of reaction time needs to be improved

7.2. Implications for noise reduction

Previous studies of the effects of noise reduction have produced the contradictory result that while noise reduction generally decreases intelligibility [5], listeners seem to prefer noise-reduced speech when asked their opinion about the quality of the signal [15]. Our original hope for this experiment was that noise-reduction might have a beneficial effect on cognitive load even if it had little impact on intelligibility. This is what Sarampalis *et al* [10] found in at least one experimental condition. In fact, we found no evidence that the particular noise-reduction technique we applied had any benefit over listening to the noisy speech directly.

The fact that our results match previous intelligibility test results rather than previous MOS test results perhaps confirms the unreliability of MOS testing. Just because listeners prefer a signal quality does not mean that it leads to better speech communication.

Why doesn't noise reduction lead to a reduction in cognitive load? It may be that the noise reduction processing leaves behind a distorted speech signal that in itself is as hard to process as the noisy signal. The average number of false alarms in the noisy and noise-reduced conditions were significantly higher than in the quiet conditions (independent samples t-test, $t(102)=3.7$, $p < 0.001$). This is perhaps an indication that degraded signals affect listening effort by creating attentional distractions.

If noise reduction is going to be useful in a speech communication application, we need to show that it has some benefits over listening to the noisy signal directly, either in terms of intelligibility or listening effort. We need objective measures of signal quality like those presented in this paper to demonstrate and assess the benefits of noise reduction.

8. Acknowledgements

The picture description recordings were made by Rachel Baker as part of the project "Speaker-controlled variability in connected discourse", funded by the ESRC. Thanks to Mark Wibrow for tracking down references. This work is supported in part by the Centre for Law-Enforcement Audio Research, funded by the U.K. Home Office.

9. References

[1] Allen, J., "Harvey Fletcher's role in the creation of communication acoustics", J.Acoust.Soc.Am 99 (1996) 1825-1839.

[2] ISO/TR 4870, "The construction and calibration of speech intelligibility tests", International Standards Organisation, 1991.

[3] ANSI S3.5, "Methods for the calculation of the speech intelligibility index", American National Standards Institute, 1997.

[4] Bennett, C., "Large scale evaluation of corpus-based synthesizers: results and lessons from the Blizzard challenge 2005", in Proc. Interspeech 2005, Lisbon.

[5] Hu, Y. and Loizou, P. "A comparative intelligibility study of single-microphone noise reduction algorithms", J.Acoust.Soc.Am 122 (2007) 1777-1786.

[6] ITU-T P.800, "Methods for subjective determination of transmission quality", ITU-T Recommendations, 1996.

[7] ITU-T P.10, "Vocabulary of terms on telephone transmission quality and telephone sets, ITU-T Recommendations, 1998. See paragraph 17.31.

[8] Harvey, L., & Green, D., "Defining Quality", Assessment & Evaluation in Higher Education, 18 (1993) 9-34.

[9] Fraser, S., Gagne, J-P., Alepins, M., & Dubois, P., "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues", J. Speech, Language and Hearing Research, 53 (2010) 18-33.

[10] Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E., "Objective measures of listening effort: Effects of background noise and noise reduction", J. Speech, Language and Hearing Research, 52 (2009) 1230-1240.

[11] Durin, V., Gros, L., & Hericher, G., "Reaction times and performances in recognition tasks to assess speech quality", Audio Engineering Society Convention, May 2008, Amsterdam.

[12] Huckvale, M. & Leak, J. "Effect of noise reduction on reaction time to speech in noise", Interspeech 2009, Brighton.

[13] Ephraim, Y. & Malah, D. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", IEEE Trans. Acoustics Speech and Signal Processing, 32 (1984) 1109-1121.

[14] Brooks, M., "VOICEBOX: Speech Processing Toolbox for MATLAB", Department of Electrical & Electronic Engineering, Imperial College, London UK, 2008. Available from: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[15] Hu, Y., Loizou, P., "Subjective comparison of speech enhancement algorithms", Proc. ICASSP 2006, 153-156.