MEASURING THE EFFECT OF NOISE REDUCTION ON LISTENING EFFORT

MARK HUCKVALE, DEIZOM FRASI

Speech, Hearing and Phonetic Sciences, University College London, London, U.K.
m.huckvale@ucl.ac.uk, n.frasi@ucl.ac.uk

Noise reduction (NR) has become widely applied in the forensic audio domain to "improve" the quality of noisy speech recordings. In this paper we consider how such processing affects listener productivity in everyday speech communication. Two measures are presented: one based on reaction time to spoken digits, and one based on finding errors in a transcript of a spoken conversation. We explain why such tasks are a useful complement to measures based on intelligibility, then present the methodology and results for two evaluations of these measures using MMSE processing on speech corrupted by babble and car-noise. Finally we discuss the implications for the use of NR techniques and for our understanding of how signal quality affects speech communication.

1 INTRODUCTION

1.1 Background & Goals

Noise Reduction (NR) processing seeks to "improve" the quality of noisy speech signals, and is widely used in the forensic audio domain, either to make recordings "easier" to listen to, or to increase the amount of information that can be extracted from them. However, scientific evidence to support the supposed benefits of NR processing is contradictory. On the one hand studies of subjective opinion show that listeners do express a preference for some noise-reduced signals [1]. On the other hand, studies of objective intelligibility show that NR tends to decrease rather than increase human performance [2]. Why then would we consider NR processing at all? One answer is that there is an implicit assumption that any improvement in the physical SNR of the signal will lead to improved listening comfort, a decrease in listening effort and a reduction in fatigue. So that despite a possible negative effect on intelligibility, NR processing could still increase the productivity of listeners engaged in everyday speech communication tasks, such as monitoring or transcription of noisy signals. For example, it may be the case that NR processing improves the alertness of listeners or causes them to make fewer errors. This paper addresses the question whether NR *actually* confers these benefits by measuring the effect of NR on listener performance in two controlled communication tasks.

1.2 Performance-Measures of Speech Quality

To assess the impact of NR and other speech enhancement techniques on listening effort requires new testing methods. Firstly it is clear that subjective opinions by listeners are unreliable: the studies by Hu & Loizou [1,2] mentioned above show the essential contradiction that noise-reduced signals are preferred despite being less intelligible. Simply that listeners prefer some processing does not mean that it will lead to an increase in their productivity.

Secondly, if NR does decrease intelligibility, then the main areas for application will be where audio signals already have good intelligibility, so that any loss will not significantly impact understanding. However these are also the circumstances which are unsuited to intelligibility testing. At high intelligibility, the psychometric function of intelligibility against SNR is shallow, and large changes in signal quality are required to observe significant changes in intelligibility.

Thus we conclude that a third form of signal evaluation is required: one that measures objective speech communication performance even with signals of good intelligibility. We call such tests "performance-based measures of speech quality".

1.3 Previous Studies

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 communication performance for signals of high intelligibility. Two previous studies known to us are those of Durin et al [3], and Sarampalis et al [4].

Durin's study investigated the effect of telephone codec 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. The results showed that listener performance was affected by signal quality, in particular that reaction times increased as the codec bit rate was reduced. Their conclusion is that poorer quality signals make greater demands on a limited pool of cognitive resources available to the listener which caused decision times to increase.

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 NR speech processed by the MMSE algorithm [5]. Generally word intelligibility performance was reduced by NR processing, although recall performance was improved in one test condition.

These two studies suggest that the effect of signal quality on cognitive processing can be measured in laboratory tasks, although in both cases, the size of the effect was small, and there was considerable variability across listeners. In our own work we hope to build on these studies, to create more robust and sensitive tests, which would lead to larger effects of signal quality. These could be used to assess different signal enhancement strategies in the forensic audio domain.

2 METHODS

To investigate the impact NR has on listening effort directly, we have developed two simple speech communication tasks that can be used in the laboratory to measure the impact of signal quality on listener performance. These are designed to be applicable even when the signal is of good intelligibility, and where intelligibility tests are not appropriate.

The first task (the *Typometer*) is one of choice reaction time to isolated words: here the listeners must detect and identify a digit spoken in noise by pressing a key on a keypad. Listeners are encouraged to respond as quickly as possible, and it would be expected that increased listening effort would lead to an increase in the time taken to make a choice, even when the noise does not affect recognition accuracy.

The second task (the *Proofometer*) is one of finding errors in a transcript of a spoken conversation. Here recordings of two speakers discussing the differences between two pictures are used as a source of natural spontaneous speech. Transcripts of the conversation are then corrupted with typical word insertions, deletions and substitutions. In the listening task, subjects must identify the location of transcript errors in real-time. Subjects are encouraged to find as many errors as possible, and it would be expected that increased listening effort would decrease the number of errors found, and increase the number of false alarms. We are also able to estimate the average processing delay between hearing the relevant word and clicking on the transcript. We expect that this delay would be increased by increased listening effort.

The effect of added noise and noise reduction on speech quality was evaluated using these two tasks in two listening experiments. Good quality recordings of speech in quiet were mixed with either babble noise at +6dB SNR, or car cabin noise at -3dB SNR to create two noise conditions. These were then processed with an MMSE noise reduction algorithm [5] as implemented in VOICEBOX [6] to create two further NR conditions. The noise levels were chosen to not significantly impair intelligibility of the speech.

3 TYPOMETER EXPERIMENT

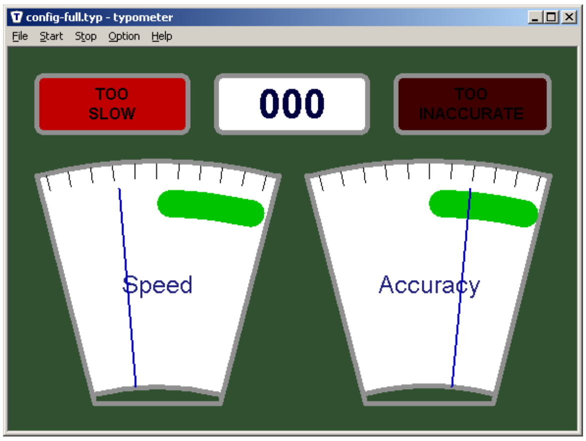


Figure 1: Typometer interface

3.1 Protocol

In this experiment, mean reaction time (RT) for spoken digit recognition was measured in the 5 audio conditions. Listeners were asked to key the correct digit 1-9 on a numeric keypad as quickly as possible after hearing the spoken digit. During the testing the noise was presented continuously, and not just during presentation of the speech. For the NR conditions, a recording of NR processed babble and NR processed car-noise was played in the background between the digits. Presentations of the digits followed each keyed response with a random delay between 0 and 2.5s. Subjects had up to 2 seconds to respond after the start of the digit, otherwise the response was 'timed out'. Each subject was tested across the five conditions in sequence within one session. The order of conditions was balanced across subjects using a double latin square which ensured that every condition and every condition dyad occurred in every position. Each condition was run until the subject had recorded a minimum of 10 correctly keyed repetitions of each digit. This took about 5-6 minutes per condition.

3.2 Results

There was no significant difference in digit recognition accuracy across the conditions. This confirms that the results are not caused by changes in signal intelligibility directly.

The mean RT for each subject for each condition for each digit was then calculated over the last 10 correct responses. The change in mean reaction time across audio conditions compared to the average reaction time for each subject is shown in Fig.2.

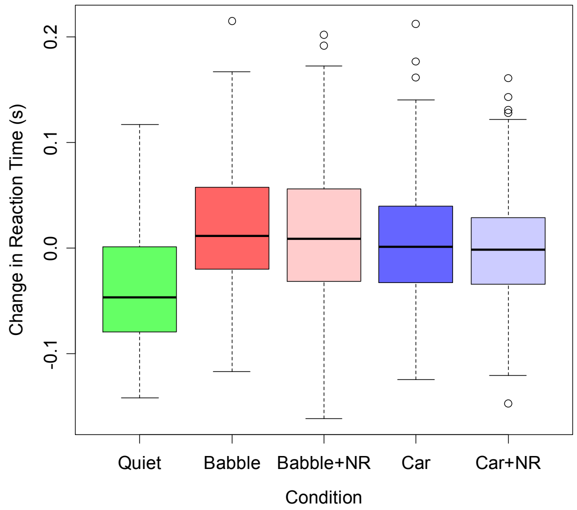


Figure 2: Typometer result. Change in reaction time to spoken digits in 5 audio conditions. Distributions show median, interquartile range, total range and outliers across listeners.

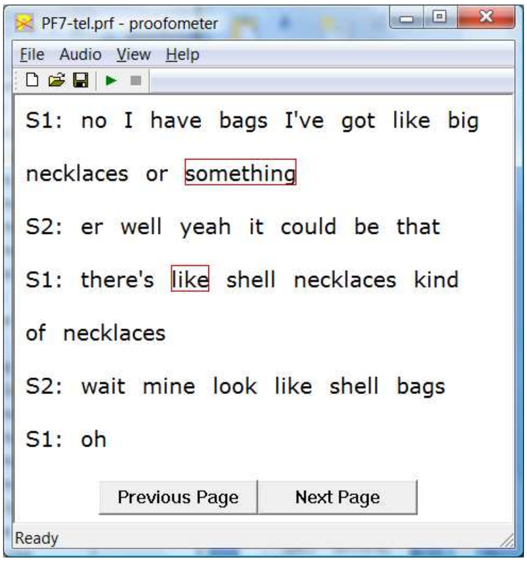
A repeated measures analysis of variance across conditions (5 levels) and digits (9 levels) with subject as a random factor (20 samples), shows a significant effect of condition (F[4,76]=25.7, p<0.001), and of digit (F[8,152]=50.7, p<0.001) on mean reaction time, and also an interaction between condition and digit (F[32,608]=4.5, p<0.001). A post-hoc analysis shows that none of the noise conditions, processed or unprocessed, are significantly different from one another, but all differ significantly from the quiet condition.

3.3 Discussion

In this experiment we investigated whether time taken to recognise digits in noise was improved by noise reduction processing. If we had seen an improvement this could have been taken as evidence that NR had reduced the cognitive effort required to process speech in noise. However we did not see any improvement in reaction time due to NR processing.

Closer inspection of the results suggest that there are two aspects of the increase in reaction time. Firstly digits such as "three", "four" and "five" are more strongly affected by the added noise, presumably by the effect of auditory masking of the initial fricative. On the other hand the increase in reaction time shown for digits such as "eight" in noise are maintained even after the noise and the effect of masking is reduced. Thus we conclude that there is an additional cognitive impact of the noise which delayed the digit recognition process itself. See [7] for more details.

4 PROOFOMETER EXPERIMENT



S1: no I have bags I've got like big necklaces or something

S2: er well yeah it could be that

S1: there's like shell necklaces kind of necklaces

S2: wait mine look like shell bags

S1: oh

Figure 3: Proofometer interface

4.1 Protocol

Four minute extracts from five different spontaneous conversations were used. These were amplitude equalised and processed down to a monophonic channel at 16000 samples/sec. Noise was added and NR processing performed as in the previous experiment to create five audio conditions. Transcripts of the spoken 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. The Proofometer program displays the transcripts on screen and the listener's task was to click on substituted or inserted words, or click on spaces where words had been deleted, see Fig.3. This error detection was done in real-time without pausing the audio. Each listener corrected a different transcript in each audio condition, with the transcripts, audio condition and processing order balanced across listeners.

4.2 Results

Performance was scored in terms of % accuracy = 100 × (number errors detected – number of false alarms) / (total number of errors). The change in % accuracy across conditions compared to each subject mean is shown in Fig.4.

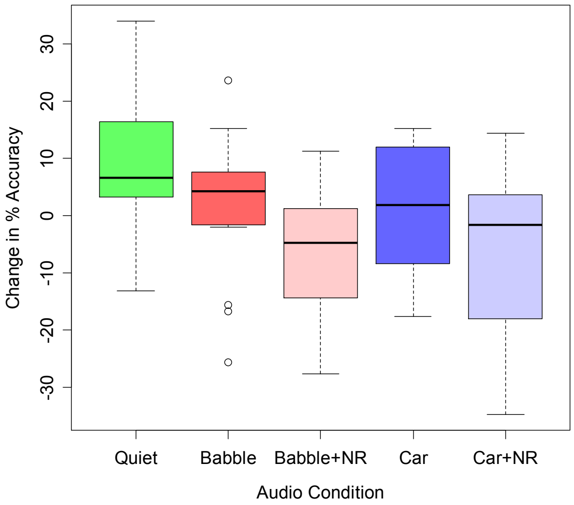


Figure 4: Proofometer result. Change in error detection accuracy in 5 audio conditions.

An analysis of variance across conditions (5 levels) with subject as random factor (18 samples) shows a significant effect of condition (F[4,68]=3.4, p=0.013). A post-hoc analysis shows that the noise reduced conditions are both significantly worse than the quiet condition, but neither are significantly different to the unprocessed noise condition.

To investigate the hypothesis that the change in listener performance is affected in part by an increase in listener effort, we also measured the average delay between when a listener clicked on an error in the transcript compared to the time that the respective word was played in the audio. The change in mean response lag for the unprocessed and processed audio conditions compared to the average for each subject is shown in Fig.5.

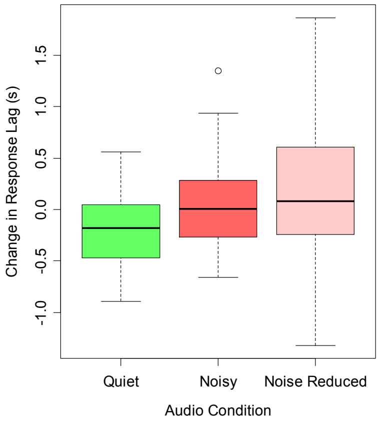


Figure 5: Proofometer result. Change in error detection response lag in unprocessed and processed audio conditions.

Analysis of variance across conditions (3 levels) with subject as a random factor (18 samples) shows a significant effect of condition (F[2,88]=3.4, p=0.036). A post-hoc analysis shows that the noise reduced condition has a significantly longer delay than the quiet condition, but is not significantly different to the unprocessed noise condition.

4.3 Discussion

In this experiment we investigated whether the accuracy or the speed with which errors were identified in an audio transcript were improved by NR processing. If we had seen an improvement this could have been taken as evidence that NR has reduced the cognitive effort required to process speech in noise. However we did not see any improvement in accuracy or speed due to NR processing.

In the results we could not partial out any changes in intelligibility caused by NR processing. We did not find intelligibility differences in the Typometer experiment that used similar audio conditions, but here very different speech materials were used, and these may have been more strongly affected by the noise and NR. Thus it is possible that some of the significant performance decrease found in the NR condition over the quiet condition could have been due to a decrease in intelligibility.

5 CONCLUSIONS

5.1 Methodology

The two experiments achieve our primary goal to demonstrate that signal quality can lead to measurable changes in performance on some laboratory tests of speech communication. However results from both trials show that variability across and within subjects is large. This makes the tests less sensitive to small changes in signal quality than we would like. For example, the Typometer experiment showed that the addition of noise causes a significant increase in reaction time, but any changes due to noise reduction were too small to assess. In the Proofometer experiment, although listeners performed significantly worse in the noise-reduced conditions compared to the quiet condition, we were not able to demonstrate a difference between the added noise conditions and the quiet condition.

To pursue this type of testing further we need to increase the sensitivity of the tests. 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.

5.2 Noise Reduction

Perhaps the most significant result from these tests was the absence of any demonstrated performance increase due to noise-reduction processing. In neither the Typometer nor Proofometer tests did NR lead to a significant change in performance over the unprocessed noisy condition. To the contrary, in the Proofometer test there was an indication that NR made matters worse.

In the Typometer test we saw a clear auditory-level masking effect of added noise in that reaction time increase of added noise was greater for digits that started with quiet fricatives. However we also saw cognitive level effects in both experiments: either due to the fact that release from masking in the Typometer case did not cause a reduction in reaction time, or from the increasing response lag in the Proofometer test. The implication is that for noise reduction to have performance benefits we need to consider its impact at both auditory and cognitive levels. For example, it may well be the case that NR processing not only removes noise, but also distorts the speech signal left behind. Or alternatively, that NR makes the residual noise left in the signal more "speech like" so that it interferes with phonetic level processing of the speech information [8].

5.3 Forensic Audio

In the experiments reported here we tested two noise types at one signal-to-noise ratio with a single NR algorithm. It is not possible to generalise from these results to all the circumstances and algorithms found within forensic audio. Forensic audio experts will frequently choose the nature and degree of processing to match the materials and types of noise or distortion present in them. Even given the materials we used in our tests, they might not have chosen to exploit the MMSE algorithm with its default settings as we did.

The importance of these results are that we have shown in principle how speech enhancement could be objectively evaluated. There is still some way to go to make tests that are robust and reliable enough to discriminate between different enhancement approaches. Ultimately, however, it will be easier to explain that forensic audio processing is worthwhile if we can show that it leads to objectively better speech communication.

6 ACKNOWLEDGEMENTS

Thanks to Gaston Hilkhuysen for help with stimulus preparation and for statistical advice. Thanks to Jayne Leak for running subjects with the Typometer. The picture description recordings were made by Rachel Baker as part of the project "Speaker-controlled variability in connected discourse", funded by the ESRC. This work is supported in part by the Centre for Law-Enforcement Audio Research, funded by the U.K. Home Office.

7 REFERENCES

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

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

[3] 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.

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

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

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

[7] Huckvale, M. and Leak, J. (2009) "Effect of noise reduction on reaction time to speech in noise", Interspeech 2009, Brighton.

[8] Hilkhuysen, G. & Huckvale, M., “Signal properties reducing intelligibility of speech after noise reduction”, European Conference on Signal Processing (EUSIPCO), Denmark 2010.