Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk

Maria K. Wolters, Karl B. Isaac, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, United Kingdom
maria.wolters@ed.ac.uk, k.b.isaac@sms.ed.ac.uk, s.renals@ed.ac.uk

Abstract

Microtask platforms such as Amazon Mechanical Turk (AMT) are increasingly used to create speech and language resources. AMT in particular allows researchers to quickly recruit a large number of fairly demographically diverse participants. In this study, we investigated whether AMT can be used for comparing the intelligibility of speech synthesis systems. We conducted two experiments in the lab and via AMT, one comparing US English diphone to US English speaker-adaptive HTS synthesis and one comparing UK English unit selection to UK English speaker-dependent HTS synthesis. While AMT word error rates were worse than lab error rates, AMT results were more sensitive to relative differences between systems. This is mainly due to the larger number of listeners. Boxplots and multilevel modelling allowed us to identify listeners who performed particularly badly, while thresholding was sufficient to eliminate rogue workers. We conclude that AMT is a viable platform for synthetic speech intelligibility comparisons.

Index Terms: intelligibility, evaluation, semantically unpredictable sentences, diphone, unit selection, crowdsourcing, Mechanical Turk, HMM-based synthesis

1. Introduction

Crowdsourcing is increasingly used to create rich speech and language data sets [1]. Instead of relying on highly skilled annotators and transcribers, researchers ask anonymous Internet users to contribute annotations [2], transcriptions [3, 4], and ratings [5] for as little as a cent per task.

In this paper, we experiment with crowdsourcing the assessment of the intelligibility of synthetic speech. In intelligibility tests, we measure to what extent listeners can reproduce the content of one or more utterances produced by a given speech synthesis system. The degree to which listeners are successful depends on many factors apart from the synthesis system itself, such as the listener's hearing [6], the listening environment [7, 8], and the listener's familiarity with the languages and voices used. In laboratory settings, we can isolate the effect of system quality by controlling most of these confounding factors. When crowdsourcing, this is more difficult. We have no control over the circumstances under which listeners work, and we do not know to what extent they are honest about the information they share.

Our primary goal in this study is to establish whether the intelligibility of speech synthesis systems can be assessed using crowdsourcing, more specifically the microtask platform Amazon Mechanical Turk (AMT[1]). We hypothesised that the results of AMT workers (Turkers) would yield similar rankings of the intelligibility of speech synthesis systems compared to the results obtained in the laboratory with students (Lab Students) Our secondary goals were to investigate the effect of listener-specific conditions such as background noise and hearing on intelligibility, and to develop a method for screening out unreliable annotators.

Since our focus is on exploring and validating the methodology, we set up two experiments for which the outcome can be predicted relatively well from the literature, one comparing a US English diphone voice to the HTS 2007 [9] version of the same voice, and one comparing a UK English unit selection voice to its HTS 2007 variant. In both cases, we expected HTS 2007 to be significantly more intelligible than the other system.

The paper is structured as follows. In Section 2, we introduce the AMT crowdsourcing platform and discuss previous work on testing speech intelligibility over the Internet. Information about voices, stimuli, participants, and recruitment can be found in Section 3. The statistical analysis methodology is outlined in Section 4. The results of our comparison experiments are discussed in Section 5.1; inter-annotator variation is explored in Section 5.2. In Section 6, we discuss the implications of our results for Internet-based intelligibility testing and conclude with a plan of future work in Section 7.

2. Background

Intelligibility tests are often conducted in carefully controlled conditions, because we know that the presence and type of background noise affects how well people can understand speech [10]. Stimuli are typically presented in quiet or sound-proofed rooms over high quality headphones. When intelligibility tests are conducted over the Internet, we lose this control. Even if we ask listeners to use headphones, they may not use them, or they may use the low-quality ear buds that came with their cheap MP3 player. Listeners can take the tests in conditions that range from an office in a quiet rural cul-de-sac to a busy café next to major roadworks. This may well affect their performance [11]. In addition, we encounter classical issues with Internet experiments such as multiple submissions by the same person [12].

Despite these shortcomings, more and more listening tests are conducted over the Internet, because there are important logistical advantages. Listeners can complete

[1] http://www.mturk.com

tests in their own time and the number of listeners that can do a test at the same time is only limited by the capacity of the web server that hosts the listening test. Many of the expert listeners for the Blizzard Challenges perform the tests over the Internet in their offices using headphones, and a sizeable contingent of naive listeners is recruited via email and social networking in addition to the extensive lab tests performed at the site that coordinates the challenge [13]. Listeners are usually asked to complete a short questionnaire to control for relevant factors such as use of headphones.

Recruiting participants using the Internet is not easy. Microtask platforms such as Amazon Mechanical Turk (AMT) provide a welcome link between experimenter and participant. People who are registered with AMT (Turkers) are paid small amounts of money to perform short, simple tasks. The demographics of Turkers are skewed compared to the general population; they are more likely to be not working or in part-time work, and they tend to be younger [14]. Since some Turkers are prone to gaming the system, Kittur et al. [5] recommend several measures for assessing the quality of a worker, such as externally verifiable questions, making it as hard to game the task as to complete it properly, and building in multiple ways of detecting suspect responses.

In speech and language technology, AMT has predominantly been used to create rich data sets [1]. Annotator and transcriber cost are one of the main reasons that good linguistic resources are so expensive to build. Although crowdworkers are cheaper, they are not always reliable [5, 2], and a lot of research has focused on minimising the impact of people whose work is of insufficient quality. Novotney and Callison-Burch [3] report that Turkers disagreed with gold-standard LDC transcriptions of the Switchboard corpus of conversational speech 20% of the time; while for data from a different domain, route instructions for robots, Marge et al. [4] found WERs of around 4–6%. Both Novotney and Callison-Burch and Marge et al. successfully improved WER by combining transcriptions from several different Turkers.

As in annotation and transcription, we have a ground truth in intelligibility testing—the words that were spoken by the speech synthesis system. However, there are two important differences. First, the more intelligible a system is, the smaller Turkers' WERs will be, so distance to the gold standard cannot easily be used to weed out bad workers. There are many legitimate sources of inter-listener variability, such as differences in listening environments. Therefore, we need to find a measure that allows us to separate representative listeners and outliers. Secondly, we are not only interested in absolute intelligibility scores, but we also want to rank voices in terms of significant differences in intelligibility [15]. So, whether AMT is a useful venue for administering intelligibility tests does not just depend on the absolute scores generated, but on whether these scores preserve the relative order of synthesis systems.

3. Method

We set up two experiments, one comparing two US English speech synthesis systems, and one comparing two UK English speech synthesis systems. Since we are mainly interested in the methodological issues that arise when conducting intelligibility tests via AMT, we selected pairs of systems where one system was known to be inferior to the other. In order to minimise variation due to speaker quality, we ensured that the same speaker had been used for both synthesis systems in each pair. Voice and synthesis method were between-subjects variables, so that each listener only heard one of the four system/voice pairs. The total duration of both experiments was around 20 minutes.

Participants were asked to provide demographic information, fill in a standard hearing screening questionnaire (ten item version of the Hearing Handicap Inventory for Adults; HHIA [16]), and transcribe 50 semantically unpredictable sentences (SUS, [17, 18]), having heard six practice stimuli first. Participants could only hear each sentence once. Finally, participants were asked to provide some information about the environment in which they completed the test and the type of headphones used. The experiment was administered using the software used for the Blizzard Challenge [15].

3.1. Voices and stimuli

For the US English experiment, we opted for the KAL diphone voice that is part of the standard Festival installation and the KAL HTS voice that was built at CSTR for demonstration purposes. The KAL HTS voice was built using speaker-adaptive HMM synthesis [19]. 523 sentences taken from the CMU KAL Communicator database were used to adapt a basic model, which was mixed-gender and had been trained on US English data. Since HTS tends to score very well in comparative tests, we hypothesised that KAL-HTS would be more intelligible than KAL-Diphone.

For the UK English experiment, we chose the unit selection (USel) and HTS version of the voice *Nick* as these were used for the speech-in-noise experiment reported in [8]. In that study, we found that Nick HTS was significantly more intelligible than the Nick USel for high signal-to-noise ratios. Therefore, we hypothesised that Nick HTS would be significantly more intelligible than Nick USel. Nick USel is based on a total of over 10 hours of recordings conducted over two years. Nick HTS was trained on around 7,000 sentences (9.5 hours) of speech taken from the original recordings. As enough training data is available, Nick HTS uses only speaker-dependent HMMs [9].

We synthesised 50 semantically unpredictable sentences for each of the four voices. These sentences are a subset of the 100 SUS that were used in the 2009 Blizzard Challenge [15]. The SUS consisted of 12 commands, 8 questions, 10 statements with a relative clause (complex statements), and 20 statements with no minor clauses (simple statements). One of the commands occurred twice, once in the middle of the sentence list and once towards the end. The sequence of SUS began and ended with a set of two commands, two questions, four simple statements, and two simple statements. The sequence of sentences was randomised once for all voices.

3.2. Participants and recruitment

Listeners were recruited from two sources, the University of Edinburgh student population (Lab Students) and AMT (Turkers). All listeners were required to be na-

tive speakers of US English. Lab Students were recruited through the Edinburgh University Student employment service and paid £5 for their participation. The experiment was conducted in a quiet meeting room; participants listened to stimuli over headphones. We recruited a total of 20 listeners, 5 per voice. 80% of listeners were female, and 90% were aged between 20 and 29. 19 listeners scored below 3 on the HHIA, with one female student scoring 14, above the cut-off for potential sensorineural hearing loss. This student did not report any problems with her hearing in the free comment field and had the second-best average WER out of the five students who listened to that particular voice.

Turkers were paid US$1 for the task, with the time for completion set to one hour. We restricted the experiment to US workers and required participants to wear headphones and be native speakers of US English. After initial slow recruitment, the task was re-released every day at a time that roughly corresponded to morning in the US. This led to an average of 20 new completed assignments per day. Out of a total of 229 Turkers, 73% completed the entire task, 11% completed the demographic questionnaire, but failed to transcribe all 50 sentences, and 16% did not complete any subset of the task.

In order to comply with the privacy policy of AMT, we allowed Turkers to opt out of almost all of the demographic questions except for current location within the US, country of birth, and dialect of English, which we used to filter out people who were not native speakers of US English and who were not born in the US. Only 6 of our participants were born outside of the US; of these, 1 was born in New Zealand, and the others were born in places not on our list of English-speaking countries. The New Zealand native also identified themselves as a native speaker of New Zealand English, while the others all stated that they were native speakers of US English. Finally, we excluded two additional participants with a mean WER above 0.9. One of these had trouble playing the stimuli, another failed to mention any problems with sound. This leaves us with a total of 159 participants. The WER criterion appears to be a good method for filtering out workers who did not complete the task conscientiously. The Turker with the highest WER in the remaining data set had a mean WER of 0.61 and scored 100% WER on only 5 sentences. The next worst scorers scored 100% WER on two out of a total of fifty sentences, and eight others had only one sentence with 100% WER.

All 159 Turkers specified their occupation. 1 chose not to state their age, and 1 did not reveal their gender. 8 did not specify their level of education. 52% of Turkers were female, 48% male. 58% were aged between 18 and 29, 33% between 30 and 49, and only 8% were aged 50 or older. Age groups and genders are distributed evenly across voices (Fisher's Exact Test, age: $p < 0.44$, gender: $p < 0.80$). 25% described themselves as computer scientists; they are distributed equally across all four voices (χ^2 test, $p < 0.60$). 17% listened to synthetic speech at least weekly.

40% of the Turkers worked full or part-time, 24% were students, 26% were homemakers, retired, or fell into a category not covered by our alternatives. 70% had a college education or a Bachelor's degree. While 10% had a postgraduate degree, 11% had only completed high school. 2 had not completed high school, but one of these appears to have been a high school student. We assessed background noise with three questions, level of background noise (quiet all of the time, quiet most of the time, equally quiet and noisy, noisy most of the time, noisy all of the time); type of noise (not applicable, radio/TV, chat, music, traffic, domestic); and character of noise (not applicable, constant, fluctuating, short bursts). 54% of all Turkers reported that their environment was quiet all of the time, for 38%, it was quiet most of the time. Only 2% reported a mostly noisy environment. The most frequent type of noise reported by the Turkers who heard a background noise was radio/television (40%), followed by traffic (22%) and chat (16%). 38% reported that the noise came in short, isolated bursts, for 25% the noise was constant, and for 26%, it fluctuated. Noise types, levels, and sources were evenly distributed across all four voices.

None of the Turkers and none of the Lab Students had been fitted with a hearing aid. Only 10 Turkers scored 10 or higher on the HHIA, which means that they possibly have a sensorineural hearing loss. In the free comments, however, an additional 6 who scored low on the HHIA mentioned hearing problems either due to hearing loss or in specific situations.

4. Statistical analysis

In the following, we will retain WER for descriptive statistics, but for statistical modelling, we converted it to the number of errors that were made on each sentence. WER itself is not normally distributed; 34.34% of all scores are 0 and 62.84% are below 0.2, and 84.45% are below 0.4. By replacing WER with the corresponding number of errors, we obtain an outcome variable that can be approximately characterised using the Poisson distribution, a well-understood type of distribution that is widely used for models of counts.

We modelled the effect of speech synthesis system type on the number of errors made using generalised linear mixed models [20, 21]. Individual-level effects were synthesis system (diphone vs. HTS for US English, unit selection vs. HTS for UK English), and type of sentence (command, question, statement, complex statement). We also included a term for the interaction between system and sentence type. We added two sets of group-level predictors, a sentence-level term and a listener-level term. The listener-level term only consisted of an intercept, which reflects individual differences in performance. We can use these intercepts to identify listeners who have particular problems with the material. The sentence-level term consisted of an intercept, which reflects differences in difficulty between sentences, and a slope for speech synthesis system. If the slope for a given sentence is negative, listeners are less likely to make errors on that particular sentence when a particular synthesis system is used; if the slope is positive, listeners are more likely to make errors. Models were fitted using the R [22] package `lme4` [23]; for Kruskal-Wallis, Wilcoxon, and Spearman tests, we used the package `coin` [24].

P-values are specified to an accuracy of two digits after the decimal point. Values of 0.00 mean that $p < 0.005$ or better. We introduced this restriction because many actual p-values are very small, even though the actual increase in model fit is relatively small.

Table 1: *Significance of individual-level predictors (ANOVA model comparison, χ^2 test).*

Predictor	US Lab		US Turk		UK Lab		UK Turk	
	AIC	P	AIC	P	AIC	P	AIC	P
Baseline	515	N/A	3784	N/A	401	N/A	3530	N/A
HTS	513	0.20	3798	0.00	416	0.00	3548	0.00
SUS Type	515	0.06	3796	0.00	405	0.01	3541	0.00
SUS×HTS	515	0.11	3794	0.00	407	0.01	3538	0.00

5. Results

5.1. Comparison of speech synthesis systems

The mean WER of Turkers across all four voices was 0.20 compared to 0.13 for the Lab Students—around 33.71% higher. If we consider only those Turkers who identified themselves as students, the difference between the two conditions is the same, with a mean WER of 0.19 for student Turkers. Both differences are statistically significant at the $p < 0.00$ level (Wilcoxon test). When comparing scores on the first and the last ten sentences, the Turkers show significant learning effects (p <0.00, Wilcoxon Test), but not the Lab Students (p <0.50). None of the groups managed to improve WERs on the command that is repeated.

While absolute WER scores from AMT are much higher, relative differences between systems are either preserved or enhanced. Figure 1 summarises WER for the two US English KAL voices, and Figure 2 shows WER for the UK English Nick voices. Both figures are boxplots, with the boxes representing the interquartile range and the whiskers 1.5 × the interquartile range. Dots are outliers; solid lines indicate means.

In order to examine the effect of speech synthesis system (HTS), type of SUS (SUS), and the interaction between SUS type and system (HTS×SUS), we removed all variables that contained the predictor to be tested from the fitted model described in Section 4. So, when testing for the effect of HTS and SUS, we removed both the variable itself and the interaction term, because otherwise part of the effect of the removed variable would have been captured in that term. We then assessed the difference between full and reduced models using ANOVA and the χ^2 test for establishing significance. The results are summarised in Table 1. For each combination of experiment and participant group, we give the AIC of the baseline model with all predictors, followed by the AIC of the models that result when one of these predictors is removed, and the probability that the difference between the original and the reduced model is significant. While the results from the Lab Students did not differentiate between the two US voices, clear differences emerge in the data collected from AMT. Looking at the boxplots in Figure 1, we see that this is due to a few of the laboratory students who had particular problems understanding the KAL HTS voice (upper end of the boxplot).

There are also significant effects of sentence type and interactions between sentence type and system. This is illustrated by Table 2, which gives mean WERs by SUS Type and voice, calculated from Turkers' responses. Commands are easiest for people who heard KAL Diphone; for all other SUS types, Nick HTS gives the best results.

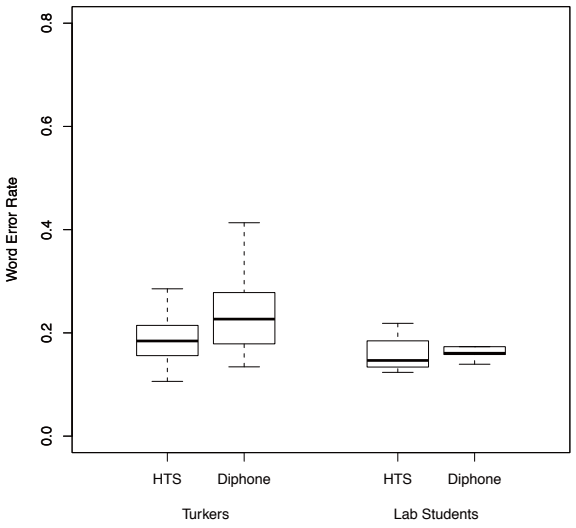

Figure 1: *Word Error Rates for KAL (US).*

5.2. Characterising individual annotators

Figures 1 and 2 reflect substantial variation in Turkers' WER scores. In order to find out whether listener and environmental characteristics affected scores, we quantified the effect of exposure to synthetic speech, age group, gender, background noise levels, and background noise type on listeners' mean scores using a linear model. (Since the mean scores follow a log-normal distribution, we predicted the logarithm of this variable.) An ANOVA showed no significant effects. Nevertheless, looking at the residual plots, we were able to identify three individuals who performed particularly well and one person who performed much worse than expected—this is the Turker with the highest overall mean score.

We can do better than modelling mean scores, however. The statistical model described in Section 4 contains a term where separate intercepts are fitted for each listener. These intercepts describe the overall WER trends for each listener after global effects of speech synthesis system and sentence type have been taken into account. The histograms for US English (Figure 3) and UK English (Figure 4) show several clear outliers with very large intercepts. These correspond to Turkers whose individual mean score is more than two interquartile ranges above the mean for their respective voice. And so we come full circle to our original graphs of mean scores—

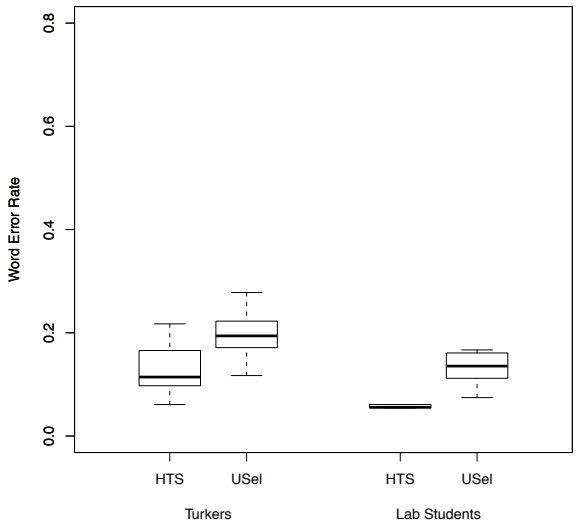

Figure 2: *Word Error Rates for Nick (UK).*

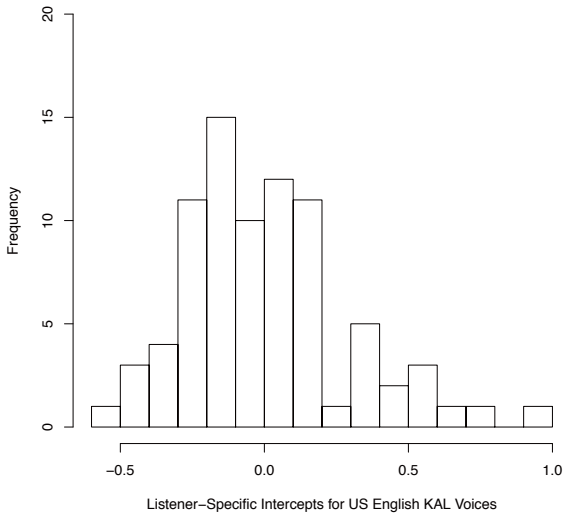

Figure 3: *Listener-level intercepts, US English.*

such Turkers will also be represented as empty circles in a typical boxplot.

Table 2: *WER by sentence type for all voices.*

SUS Type	US (KAL)		UK (Nick)		Avg.
	Diphone	HTS	USel	HTS	Overall
Command	0.17	0.25	0.20	0.20	0.21
Question	0.25	0.17	0.17	0.07	0.16
Statement	0.29	0.17	0.21	0.12	0.20
Complex Stat.	0.25	0.21	0.22	0.13	0.20
Average	0.25	0.20	0.20	0.13	0.20

6. Discussion

To the best of our knowledge, this is the first paper where speech intelligibility is measured using AMT to recruit listeners. Although absolute WERs are much worse than in the laboratory situation, the AMT results reflect relative differences in intelligibility fairly well. For many applications, for example when a new speech synthesis approach needs to be compared to older technology or to a human voice, this will be sufficient, but not if we want to know whether it is possible to understand synthetic speech perfectly. Listeners who perform less well than expected can be identified through simple boxplots; these results agree well with the outcomes of multilevel models where separate group-level intercepts are fitted for each listener.

The result of our test experiment using UK English voices is as expected, while the outcome of the comparison of the US English voices is somewhat surprising, given that HTS generally yields very high-quality voices. The poor performance of KAL HTS may be due to the lack of material available for adapting the average speaker model. The results on the KAL voices illustrate the particular advantage of AMT. For small effect sizes, many listeners are needed to obtain sufficient power, especially if the comparison is between subject. Thus, the AMT data were able to reveal a clear difference between the diphone and HTS voices, which had not emerged from the laboratory study due to the small number of listeners.

All Turkers who completed the task fit the recruitment criteria, except for one who misread the instructions and five native speakers of US English who had not been born in the US. The percentage of sentences that are completely mistranscribed (100% WER) appears to be a good criterion for identifying people who enter random words instead of carefully listening to the sentences. We are somewhat concerned by the high drop-out rate. Around 15% of all Turkers failed to complete the task. We did not ask non-completers what their main problem was, but from previous experience with Internet listening experiments, we suspect that sound is the main culprit.

While the sample of Turkers we recruited is not representative of the population of the US, it is far more diverse than the student and expert samples which are typically recruited for listening experiments. Overall word error rates are in line with WERs reported by Novotney and Callison-Burch [3] for transcription of spontaneous speech. It remains to be seen how WERs on human speech compare with those on synthetic speech. The WERs obtained by the Turkers in our experiments also vary far more than the WERs of the Lab Students. Although we asked all participants to fill in an extensive questionnaire about their listening situation and their hearing, none of these variables was able to cover a sizeable amount of this variation. Indeed, quite a few Turkers who scored well on the HHIA mentioned specific hearing problems in comments. We also suspect that the questions we used to assess environmental noise levels during the test may not have adequately reflected true noise levels and sources.

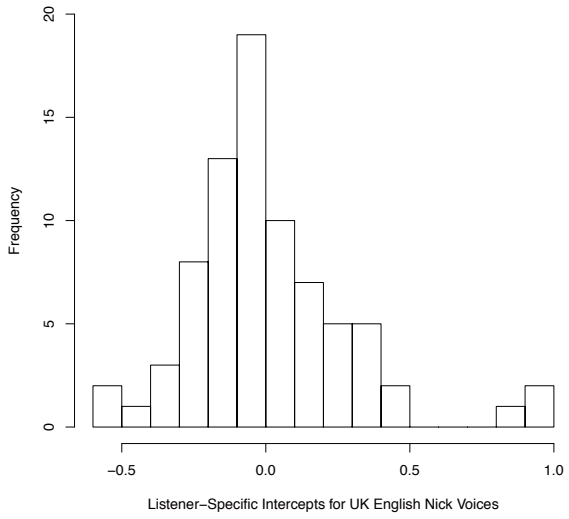

Figure 4: *Listener-level intercepts, UK English.*

7. Future work

Now that we have established that AMT is a viable platform for conducting speech intelligibility tests, we intend to use AMT for investigating methodological issues in intelligibility testing that typically require a large number of listeners, such as the choice of sentence material, and the effect of learning. We believe that this year's Blizzard Challenge may contain a sizeable sample of judgements collected via AMT [13]. A reanalysis of this data set could be used as a further test of our finding that AMT is good at reproducing relative differences in intelligibility, while also providing baseline data from a human speaker. In future experiments, we may replace the HHIA with a few simple questions that ask people directly whether they have ever experienced problems with their hearing. We also plan to replace the three questions regarding ambient noise levels with a larger set of easier and more detailed questions where Turkers are asked about the type of environment (home/café/train...) and the presence of noise sources that are specific to the environment (e.g., whether the café was busy, whether there was a lot of traffic noise, whether the radio or TV was on).

8. Acknowledgements

This research was funded by EPSRC grant no. EP/G060614/1 (MultiMemoHome). We thank Vasilis Karaiskos and Junichi Yamagishi for useful discussion and comments.

9. References

[1] C. Callison-Burch and M. Dredze, "Creating speech and language data with amazon's mechanical turk," in *Proc. NAACL*, 2010.

[2] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proc. EMNLP*, 2008.

[3] S. Novotney and C. Callison-Burch, "Automatic speech recognition with non-expert transcription," in *Proc. NAACL*, 2010.

[4] M. Marge, S. Banerjee, and A. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proc. ICASSP*, 2010.

[5] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proc. CHI*, 2008.

[6] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "Making synthetic speech accessible to older people," in *Proc. SSW-6*, Aug. 2007, pp. 288–293.

[7] H. S. Venkatagiri, "Segmental intelligibility of four currently used text-to-speech synthesis methods." *J Acoust Soc Am*, vol. 113, 2003.

[8] K. Isaac, M. Wolters, and S. Renals, "How intelligible is synthetic speech in noise?" in *Interspeech*, subm.

[9] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW-6*, 2007.

[10] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J Acoust Soc Am*, vol. 117, no. 4, 2005.

[11] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Prediction of the intelligibility for speech in real-life background noises for subjects with normal hearing." *Ear Hear*, vol. 29, no. 2, 2008.

[12] U. D. Reips, "Standards for Internet-based experimenting," *Exp Psychol*, vol. 49, 2002.

[13] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, 2010.

[14] P. G. Ipeirotis, "Demographics of Mechanical Turk," New York University, Tech. Rep., 2010.

[15] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge Workshop*, 2009.

[16] C. W. Newman, B. E. Weinstein, G. P. Jacobson, and G. A. Hug, "The Hearing Handicap Inventory for Adults: psychometric adequacy and audiometric correlates." *Ear Hear*, vol. 11, no. 6, 1990.

[17] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, 1996.

[18] H. T. Bunnell and J. Lilley, "Analysis methods for assessing TTS intelligibility," in *Proc. SSW-6*, 2007.

[19] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans ASLP*, vol. 17, no. 6, 2009.

[20] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press, 2007.

[21] H. Baayen, *Analyzing Linguistic Data*. Cambridge University Press, 2008.

[22] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009, ISBN 3-900051-07-0.

[23] D. Bates and M. Maechler, *lme4: Linear mixed-effects models using S4 classes*, 2009, R package version 0.999375-31.

[24] T. Hothorn, K. Hornik, M. van de Wiel, and A. Zeileis, "Implementing a class of permutation tests: The coin package," *Journal of Statistical Software*, vol. 28, no. 8, 2008.