



Native and Non-native Speaker Judgements on the Quality of Synthesized Speech

Anna C. Janska¹, Robert A. J. Clark²

¹IMPRS NeuroCom, University of Leipzig, Germany

²CSTR, The University of Edinburgh, U.K.

janska@rz.uni-leipzig.de, robert@cstr.ed.ac.uk

Abstract

The difference between native speakers' and non-native speakers' naturalness judgements of synthetic speech is investigated. Similar/difference judgements are analysed via a multidimensional scaling analysis and compared to Mean opinion scores. It is shown that although the two groups generally behave in a similar manner the variance of non-native speaker judgements is generally higher. While both groups of subject can clearly distinguish natural speech from the best synthetic examples, the groups' responses to different artefacts present in the synthetic speech can vary.

Index Terms: speech synthesis, evaluation, non-native

1. Introduction

Text-to-Speech Synthesis (TTS) systems are employed increasingly in a variety of contexts, be it in automated services, such as travel information; assistive technologies, such as screen readers; interactive dialogue systems or entertainment applications.[11] What is often neglected, is that the language being spoken may in fact be a language other than the native language (L1) of a great number of potential users. In particular TTS services are often provided in English for an inhomogenous group of target users who have L1s other than English. This fact makes it a worthwhile endeavour to investigate how non-native speakers (NNSs) of English perceive output of a TTS system in English. Despite this, most research on the subjective evaluation of TTS has been conducted with native speakers (NSs) even if only to reduce the number of uncontrolled factors in the evaluation. In previous research, for example [12, 1], it has been found that the perception of NSs of a language differs qualitatively from that of non-native speakers, as their mental representations of the target language are not the same. This will surface, particularly when noise is introduced into the speech signal: even proficient NNSs can be expected to show a strong decrease in perceptual performance. This becomes particularly apparent in intelligibility tasks. [2] reports that in the course of the evaluation of TTS systems in the *Blizzard Challenge 2005*, NNSs were observed to encounter considerable difficulties in Modified Rhyme Tasks and the Semantically Unpredictable Sentences task. A large extent of NNSs' test results had to be excluded from the evaluation procedure, as participants either did not give answers to all the questions within a task, or gave up on a task altogether. The specific nature of these problems NNSs encounter has not been further investigated, so it is unknown whether these problems also extend to the evaluation of speech quality as well [2].

The main problem in predicting how NNSs judge the quality of synthesized speech is that NNSs are actually a truly di-

verse group. Unless a specific subgroup with a shared L1 is selected, hardly any specific predictions can be made, as the groups are too inhomogenous: a listener's perception of English is not only dependent on the listener's language competence, but it is also mediated to some extent by their L1.

With this in mind this study further investigates the following two questions:

1. Does the fluent NNS behave in the same way as the NS when evaluating the quality of synthetic speech?
2. Can the quality judgements of NSs be expected to find agreement in NNSs when evaluating synthetic speech?

2. Method

Two perceptual evaluations are carried out to compare how NSs and NNSs evaluate a number of synthetic speech stimuli. The first is a Multidimensional scaling (MDS) design where subjects are asked to make a pairwise comparison of stimuli and the second a Mean Opinion Score (MOS) [8, 13] design where subjects are asked to rate stimuli. Part of the reasoning for this design is to determine whether the richer output obtained by MDS incorporates a result similar to what would be achieved by a MOS design.

2.1. Stimuli

To ensure a range of different quality in individual stimuli but at the same time to ensure that all stimuli were representative of current state-of-the-art speech synthesis, 10 speech stimuli were chosen from the Blizzard challenge 2008 test set A. These stimuli consisted of pairs of stimuli from 4 different system entries and a natural speech control. To select the stimuli, first one high ranking sentence was chosen from each of four high ranking systems from the 2008 test set. For each system, another sentence that was perceived as perceptually more distant from natural recorded speech was then added. Finally two natural speech sentences were added to the set of stimuli. Natural speech recordings were included to "anchor the scale". [11, p. 537] and we would expect these stimuli to be highly ranked by both natives and non-natives alike in the experiments below. The length of the sentences ranges from 1.38 to 3.4 seconds, and from 8 to 15 syllables. ¹

The text for the stimuli are given in Table 1. The numbers represent the system in question, where system 1 is natural speech and T and B are pairs of sentences for each system. The

¹The stimuli used in the experiment can be accessed at http://homepages.inf.ed.ac.uk/s0674876/listening_test_july_2009_wavfiles/

label	type	sentence	syllables	duration
T1	natural	For good measure, he offered an unreserved apology.	15	2.9s
T2	synthesized	Billy could help Saxon little in her trouble.	12	2.2s
T3	synthesized	UCA based air traffic controllers are also unsettled.	15	3.4s
T4	synthesized	We are pulling on in the morning to circle city.	14	2.2s
T5	synthesized	I believe the two years suspension are harsh.	11	2.4s
B1	natural	Power cuts affect refrigerated medicines and food stuffs.	15	3.4s
B2	synthesized	But they can live in a pigsty.	8	1.38s
B3	synthesized	He was puzzled by the slowness of its progress.	12	2.2s
B4	synthesized	Thus he waited, keeping perfectly quiet.	11	2.4s
B5	synthesized	The bloodshed was not confined to Copenhagen.	12	2.5s

Table 1: The text spoken, number of syllables and duration of the stimuli used.

T sentences are the first chosen, highest scoring sentences from the original Blizzard 2008 challenge.

2.2. Participants

Altogether 32 participants from different vocational backgrounds were tested, 16 of which were NSs of some variety of English, while the remaining 16 were NNSs of English. The NNSs were fluent in English (they were mostly students being taught in English), and their first languages were from different language families. The pool of participants was self-selecting: Participants were chosen on a first-come, first-serve basis in their response to an advertisement. The majority of participants was paid seven pounds sterling to take the experiment, which took none of the participants longer than 40 minutes to complete.

2.3. Procedure

The experiment was conducted in a quiet computer lab. Instructions, as well as stimuli were presented via a web browser. Answers were given by clicking the respective radio-button on the screen. The subjects listened to the stimuli with closed-back Sennheiser headphones and at a volume level they could adjust themselves.

2.3.1. Part 1 – Multidimensional Scaling (MDS) design

To facilitate producing a coordinate space of the stimuli in terms of how similarly they are perceived by each group of listeners through a multidimensional scaling protocol, each stimulus was paired with every other stimuli, so that paired comparisons between all stimuli were made, in both orders. These stimulus pairs were then presented in a random order. Participants had to decide whether both items of each pair were equal or different in their degree of naturalness in line with the instructions used by [10].

2.3.2. Part2 – Mean opinion scores (MOS)

Part 2 was devised to rank the systems according to their naturalness. The listeners' scores were conditional similarity data, which means that values cannot be compared directly between subjects [4, p. 14], but they provided a valid basis for generating ranks for the systems. Each stimulus was presented three times in random order. Participants had to rate on a scale from 1 to 10 (1 being the lowest, and 10 the highest), how natural a sentence

sounded.² Part 2 of the experiment was presented after part one to ensure that subjects were familiar with the overall variation in quality of the stimuli by this point, again to try to encourage a wider use of the scale provided.

3. Results

The data from the participants for part 1 resulted in 32 responses for each of the 100 stimuli pairs, producing a proximity matrix comprising of 3200 individual judgements. This was analyzed using PASW Statistics 18.0 (formerly SPSS Statistics) as described below.

As an initial step of analysis it was investigated, whether the similarity-difference judgements and the MOS collected from NSs and NNSs come from the same distribution. A Kolmogorov-Smirnov Z test³ indicated that the MOS of NSs and NNSs are from one population, as are their judgments in part 1.

To further investigate how NNSs performed in the evaluation, MDS analysis was employed: Identity Euclidean Distance MDS was performed on the ordinal level, untying ties, applying transformations to each point individually. Stress 1 is 0.1879, which is an acceptable fit [3]. NNSs data are more variant and introduce more stress to the MDS representation. A Kolmogorov-Smirnov Z test testing whether there was a difference in stress values introduced by NSs and NNSs was significant at $\alpha < 0.05$. For both groups the distributions of stress values are positively skewed; the range of stresses for NNSs is much larger than in NSs with the lowest NNS stress value being below the lowest NS value, but also with the highest NNS stress value exceeding those of NSs by far. This suggests that NNSs dilute our results by introducing higher amounts of variance in the data than NSs.

To examine at how this affects MDS representations, separate MDS graphs were drawn for NSs and NNSs.

For both representations Stress-1 was below .20, which constitutes an acceptable fit. MDS representations for NSs and

²Generally, in MOS tasks, measures between 1 and 5 are used. [7, 6, 9] However, for this experiment, a larger range was chosen to try to encourage a wider dispersal of ratings, in the hope that this would generate bigger gaps between the systems' ratings and that distinct ranks could be clearly established.

³Kolmogorov-Smirnov Z tests will be our test of choice in most of our analysis of the effect of native language status: it is non-parametric and thus not sensitive to (the lack of) normal distributions and homogeneity of variance. Also, it is preferable to the more common Mann-Whitney test, because of the low sample size of only 16 listeners per group [5, p. 529]

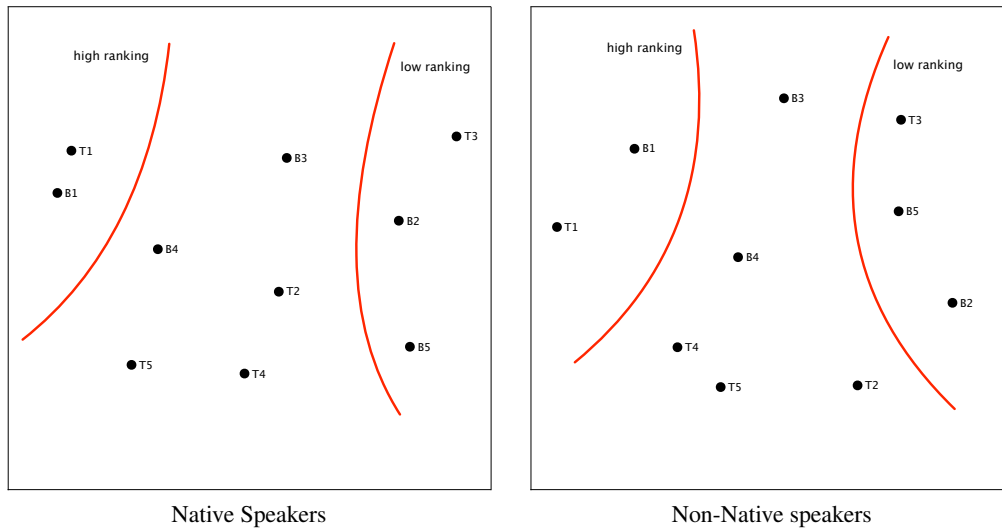


Figure 1: MDS representations of speech quality judgements by native speakers and non-native speakers

NNSs, shown in Figure 1, are similar in distribution of points, and some overall tendencies are visible:

1. The two natural recordings, T1 and B1, are clustered together clearly distinct from the other stimuli. Hence we can deduce that experimental participants perceived a clear distance between those and the synthesized stimuli. This supports [7]'s claim that "even the best examples of speech from TTS systems are unlikely to be mistaken for natural speech". (p. 107)
2. There is a central section of the space, constituted by B4, B3, T5, T2 and T4, in which the largest number of systems are located.
3. There is a group of systems lagging behind, being T3, B2 and B5.
4. There is a potential trend that stimuli become less smooth and more choppy along an diagonal axis from top-left to bottom-right, which may relate to either bad joins or bad durations.

The resulting rotation and scaling of resulting MDS spaces are somewhat arbitrary, and as presented the resulting NNSs space is rotated through 180 degrees to have an orientation similar to that of the NSs space.

In general NSs and NNSs agree on the make up of these groups, which suggests that there is a rough agreement on what makes a natural sounding stimulus. However, there are a few disagreements on the fine-tuning of the ranks inside these groups.

From these representations, ranks can be computed by sorting the distances of each data point to the natural stimulus T1. (T1 was chosen instead of B1, as T1 got higher MOS by NSs as well as NNSs). These ranks are shown in the leftmost columns of Table 2 which confirms what we have already seen in the graphical representation: The rough ordering of systems is the same for both listener groups. However, what is probably most striking is the rank difference of stimulus T4: while NSs rank it in the bottom of the middle field, NNSs put it straight to the top. This rank difference is also observable in MOS (rightmost columns of Table 2), although not to the same extent. MOS

Rank	MDS		MOS	
	NS	NNS	NS	NNS
1	T1	T1	T1	T1
2	B1	B1	B1	B1
3	B4	T4	B4	T5
4	B3	B4	T5	T4
5	T5	T5	T4	B4
6	T2	B3	T2	B3
7	T4	T2	B3	T2
8	B2	B5	B5	B5
9	T3	T3	B2	B2
10	B5	B2	T3	T3

Table 2: Table comparing rank order of stimuli computed from MDS distances and MOS judgements

scores themselves are shown in Figure 2 where it can be seen that there is a general agreement between NSs and NNSs, however there is a slight tendency for NSs to utilise the more extreme ends of the scale than NNSs. Stimulus T4 itself sounds smooth and continuous and not buzzy, but does sound a little strange in terms of prosody and voice quality.

To get a feeling for *why* these differences might arise, the dimensions of the NSs' MDS graph were interpreted. Analysis is performed visually and auditorily. We attempt to organize the stimuli into clusters according to their auditory features. Obviously, as already discovered above, the two natural recordings build one cluster. The other clusters, however, are not divided in the same way as our high-middle-low representation above: B4, T5 and B3 all are characterized by good prosody, as the intonation is vivid and not flat. Thus they can be clustered together. However, B3 has some problems with joins, so that while the overall intonation is good, there are little "jumps in pitch" in between. Problems of joins can also be identified in B2, B5 and T3, so these are clustered together. T4, T2, B2, and B5 build a cluster of stimuli with bad intonation. The "'sentence melody'" of B5 resembles that of a NNS who speaks English rather well, while transferring their own language's "sentence melody" into English. B2 has rather flat intonation and the final

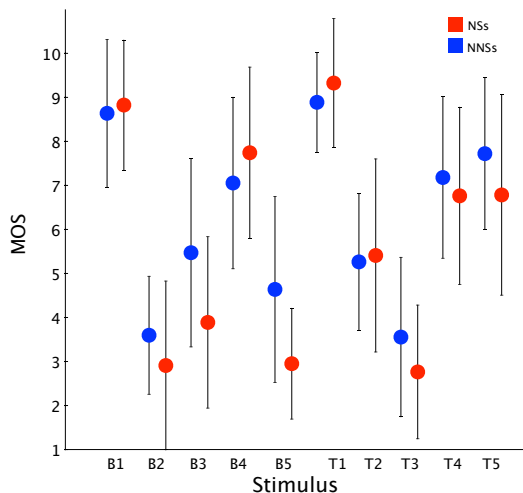


Figure 2: MOS of speech quality judgements by native speakers and non-native speakers. Means are plotted with 1 std. dev. error bars.

segment in *pigsty* sounds somewhat clipped. T5 is unique in that it has a sort of echoing quality. Based on these clusters we can identify two axes along which stimuli differ. One axis goes diagonally from the left top to the right bottom, along which the goodness of prosody changes, being the most natural in the left and the least natural in the right. Perpendicular to it is an axis that describes the goodness of joins, decreasing the further it gets to the top right.

This leaves T3 an interesting case: it is located at the edge of the NSs' stimulus space, as one of the worst stimulus samples. It is fairly natural sounding in general, but it has one grave duration error, which is rendered more salient by its position: the voice-onset time of the *t* in the word *controllers* is so long, it almost seems like a break half-way into the word. A break like this would not occur in naturally spoken English, though, since even when breaks are made within words, they tend to be made between syllables, but never in a syllable's onset, as it is the case in T3. Thus it can be assumed that this kind of error is one that influences NSs' perception very strongly, and affects them more than generally bad joins.

When now mapping these axes onto the NNSs' MDS representation, it is striking that while they agree on the whole, some of the components are assessed to a different degree: staying with stimulus T3, it is higher up on the prosody axis, suggesting that NNS do not find the error as offensive as NSs. T4, which has the biggest rank difference between NSs and NNSs is much higher up on the NNSs' intonation axis than on the NSs'. As opposed to the NSs', the NNSs prefer the unconventional intonation of B5 to the flat intonation of B2.

4. Discussion

The statistical analysis have confirmed our intuition that NNSs are a fairly inhomogenous group and that it thus does not make a lot of sense to generally test non-native speakers when purely evaluating the quality of speech generated by a TTS systems, as their judgements are likely to make the data more noisy. As a consequence, a larger number of participants would obviously be needed to get clear results, which in turn would increase the cost of testing. However, it is conversely also true that native speakers would not necessarily make a suitable subject

pool when evaluating how non-native speakers perceive synthetic speech.

We have also shown that in the evaluation of speech quality, there are no problems analogous to those of the evaluation of speech comprehensibility. The systems NSs approve of do also find approval of NNSs. The data even suggest that NNSs are more lenient and accepting than the NSs where utterances with prosodical errors are produced. It can be hypothesized that NSs prefer flat intonation to faulty one, whereas NNSs are not as susceptible to the latter. Thus it can be assumed that positive results from speech quality testing by NSs can be transferred to NNSs without further testing. In other words: the data suggest that what NSs find natural, NNSs will deem natural too.

5. Acknowledgements

The authors would like to thank The Blizzard Challenge organisers and participants for providing the data used for this work.

6. References

- [1] E. Axmear, J. Reichle, M. Akamsaputra, K. Kohnert, K. Drager, and K. Sellnow. Synthesized speech intelligibility in sentences: a comparison of monlingual english-speaking and bilingual children. *Language, Speech, and Hearing Services in School*, 36:244–250, 2005.
- [2] C. L. Bennett. Large scale evaluation of corpus-based synthesizers: results and lessons from the blizzard challenge 2005. In *Proc. Interspeech 2005*, Lisbon, Portugal, September 2005.
- [3] I. Borg and P.J.F. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Verlag, 2005.
- [4] APM Coxon, J.E. Jackson, PM Davies, HV Smith, L. Sachs, and J. Schmee. *User's guide to multidimensional scaling*. Heineman Education books, 1982.
- [5] Andy Field. *Discovering statistics using SPSS*. SAGE Publications Ltd, London UK, 2005.
- [6] J.L. Hall. Application of multidimensional scaling to subjective evaluation of coded speech. *Journ. Acoust. Soc. of America.*, 110:2167, 2001.
- [7] JN Holmes. *Speech synthesis and recognition*. CRC, 2001.
- [8] ITU-T Recommendation P.85. A method for subjective performance assessment of the quality of speech output devices. International Telecommunications Union publication, 1994.
- [9] D. Jurafsky and J.H. Martin. *Speech and language processing*. Prentice Hall, 2008.
- [10] C. Mayo, R.A.J. Clark, and S. King. Multidimensional scaling of listener responses to synthetic speech. In *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.
- [11] Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.
- [12] Sanders J. van Wijngaarden, Herman J. M. Steeneken, and Tammo Houtgast. Quantifying the intelligibility of speech in noise for non-natives. *Journ. Acoust. Soc. of America.*, 111(4):1906–1916, 2002.
- [13] M. Viswanathan and M. Viswanathan. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech and Language*, 19:55–83, 2005.