Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation

Rasmus Dall[1], *Junichi Yamagishi*[1 2], *Simon King*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]National Institute of Informatics, Tokyo, Japan

r.dall@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk, simon.king@ed.ac.uk

Abstract

In this paper we present evidence that speech produced spontaneously in a conversation is considered more natural than read prompts. We also explore the relationship between participants' expectations of the speech style under evaluation and their actual ratings. In successive listening tests subjects rated the naturalness of either spontaneously produced, read aloud or written sentences, with instructions toward either conversational, reading or general naturalness. It was found that, when presented with spontaneous or read aloud speech, participants consistently rated spontaneous speech more natural - even when asked to rate naturalness in the reading case. Presented with only text, participants generally preferred transcriptions of spontaneous utterances, except when asked to evaluate naturalness in terms of reading aloud. This has implications for the application of MOS-scale naturalness ratings in Speech Synthesis, and potentially on the type of data suitable for use both in general TTS, dialogue systems and specifically in Conversational TTS, in which the goal is to reproduce speech as it is produced in a spontaneous conversational setting.

Index Terms: speech synthesis, evaluation, naturalness, MOS, spontaneous speech, read speech, TTS

1. Introduction

In speech synthesis research there are two generally used methods for evaluation, namely intelligibility and naturalness. Intelligibility is a metric which has robust measures such as semantically unpredictable sentences (SUS) [1] and synthesis systems perform well compared to natural sentences [2, 3]. Naturalness on the other hand is a less defined concept, although it is generally always used e.g. in the Blizzard challenges [2, 4, 5]. It is also used to evaluate prosody and is the focus of this paper.

Naturalness is normally evaluated as a Mean Opinion Score (MOS) where participants rate the quality of the synthetic speech on a 5-point scale ranging from 1-Very Unnatural to 5-Very Natural. The scale itself has not been much investigated, however the Blizzard 2008 [2] evaluation gave support to the scale being treated, by listeners, as an interval rather than ordinal scale by comparing it to scores obtained using an unnumbered slider. While systems tend to perform well on intelligibility they are generally lacking behind natural speech in terms of naturalness. One assumption made in several conversational speech synthesis studies is, that spontaneous conversational speech is more natural than read speech [6–8]. Thus, it is assumed, synthesis based on conversational speech will similarly increase the system's naturalness. However, it has not been shown that people actually find conversational speech more natural than read speech, and earlier studies using spontaneous recordings have not managed to increase the perceived naturalness of synthetic speech [6, 9]. People can distinguish the two modes of speech with high accuracy despite lexical equivalence [10], so it is likely that people will be able to pick up upon and judge according to this distinction when asked. This study attempts to test this by obtaining naturalness ratings of natural speech from the same speakers, of speech produced spontaneously in a conversation and when reading aloud. We hypothesise, as has been done before, that conversational speech is considered more natural.

It is also likely that 'naturalness' as a concept is underspecified. That is, we do not have an exact definition of what naturalness is. In fact differing studies give participants differing instructions. The Blizzard 2013 evaluation [11] instructs participants to give a score which "should reflect your opinion of how natural or unnatural the sentence sounded. You should not judge the grammar or content of the sentence, just how it sounds." In contrast [12] explains the meaning of naturalness as if it is "likely that a person would have said it this way?" (p.470). The two stand in contrast to each other, the one asking to disregard grammar and content, and the other to judge the 'way' it was said - including content and grammar. If listeners do find it to be underspecified then people's perceptions should be be influenced by their expectations of what naturalness means in any given context. We therefore attempt to influence the prior expectations of listeners by slight variations in instructions to bias them toward either conversational or read speech, and compare this to the general case with no further instructions.

Note that there are genuine worries about the ecological validity of MOS-scale naturalness tests of isolated sentences presented in very controlled noise environments. It is not the purpose of this paper to attempt to rectify these, but rather to explore current means and enable further detail in their application. Section 2 describes our first listening test, in Section 3 we attempt to separate audio and text and Section 4 discuss the overall implications, before concluding in Section 5.

2. Naturalness Ratings of Spontaneous and Read Speech

A simple way of testing if there is a preference for conversational over read utterances is to mimic the standard naturalness test setup. In such a procedure the common instruction is for the participant to listen to one sentence at a time, rating how natural they find the sentence. That is people are only told to rate what sounds 'natural' with no further qualification. If naturalness is an underspecified concept it should be possible to influences people's ratings by slightly changing the given instructions, and as we are concerned with the difference between conversational and read speech we attempt to influence people's perceptions in these directions. Instead of closely matching the content of these sentences by rating the same sentences either spoken in a conversation or read aloud (see Section 3), it was decided to ini-

10.21437/SpeechProsody.2014-191

Read	Spontaneous
Challenge and errors both go well.	It's kinda ridiculous, but it was funny at the time.
Author of the Danger Trail Philip Steel etc.	When I was younger I... loved uhm Ang Lee.
How funny is your funniest joke?	Absolutely, I'm sure there are evil kings with rotten voices.
Officials have no evidence yet that the plane could have been sabotaged.	And at the point where it goes into the park, the tunnel goes underneath at that point.

Table 1: Example sentences.

tially use sentences representative of the respective styles to see if a difference was to be found in a fairly unconstrained setting.

2.1. Data

Studio recordings of conversational and read-aloud data from two differing speakers, one male and one female, was used as the stimuli. For each speaker 30 conversational and 30 read sentences were selected. For the read sentences the female data included mainly read news text and the male data was the first 30 sentences of the Arctic prompts [13]. The conversational utterances were chosen from recordings of the speakers having an unscripted conversation with an experimenter. The sentences were chosen so as to be complete sentences with no initial or final disfluency, although disfluencies were allowed in the sentences. Where the read-prompts had a distinct third-person perspective most conversational sentences in the database were first person. To reduce this mismatch, conversational sentences were chosen to generally be about something rather than the speaker him/herself. Sentences in both conditions were also matched for length with the shortest being about 2s long and the longest about 6s. Table 1 provides a few example utterances and audio samples are available.[1]

2.2. Method

32 paid native speakers of English were recruited, mainly students at the University of Edinburgh. 11 participants rated general naturalness (GenNat), 10 conversational naturalness (ConvNat) and 11 participants reading naturalness (ReadNat). Participants were instructed to rate the sentences in the standard TTS paradigm and they were instructed to “Listen to each sentence and rate it according to how natural you find the sentence from a scale of 1 - Very Unnatural to 5 - Very Natural” in the GenNat case, in the ConvNat the sentence ”if you were having a conversation” was added between ”sentence” and ”from”; in the ReadNat case ”if somebody was reading aloud” was added in the same place. This difference in instruction was the only difference between conditions. Each participant rated all 120 sentences once, in a randomised order of presentation for each participant. Each participant also rated an additional 5 sentences as a trial run to get accustomed to the methodology. After the trial run participants were encouraged to ask clarifying questions before proceeding to the main part of the test. The test was performed in a soundproof room with the participants wearing good quality headphones. The test took about 15 minutes to complete. There were three groups of participants (GenNat, ConvNat and ReadNat) and two types of audio (conversational or read).

	GenNat		ConvNat		ReadNat	
	Read	Spont	Read	Spont	Read	Spont
N	660	660	600	600	660	660
Mean	2.98	4.23	2.62	4.04	3.67	3.74
SD	1.192	1.131	1.291	1.189	1.182	1.466
p	p<0.0001		p <0.0001		p=0.352	

Table 2: Condition descriptives. The shown significances are between spontaneous and read sentences for each condition.

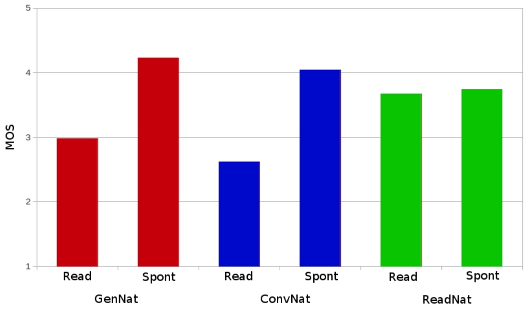

Figure 1: Overall ratings by category.

2.3. Results

As noted in Section 1 we have evidence that the 5-point MOS scale is used as an interval scale and not in an ordinal fashion, therefore we can meaningfully compare the means instead of the medians of the ratings [14]. No null responses were recorded and all ratings were used in the analysis. A significant difference was found between the read (M=2.98, SD=1.192) and conversation (M=4.23, SD=1.131) sentences in the GenNat group (t(1318)=19.644, p <0.001), this was also the case for ConvNat (Read: M=2.62, SD=1.291; Conv: M=4.04, SD=1.189; t(1198)=19.848, p <0.001) but not the ReadNat condition (Read: M=3.67, SD=1.182; Conv: M=3.74, SD=1.466; t(1318)=0.93, p=0.352). In other words, when asked to rate what they found natural with no further instruction, or instructions toward conversation, participants preferred the spontaneous utterances, however there was no preference when rating naturalness for reading aloud. See Table 2. Across instruction conditions one-way ANOVA's were run for each speech type. An effect for both read (F(2,1917)=122.285, p <0.001) and spontaneous utterances (F(2, 1917)=25.509, p <0.001) were found. Bonferroni correction showed all differences to be significant at the p <0.001 level for the read speech and for the spontaneous speech all differences were significant at the p <0.001 level except GenNat and ConvNat which was significant at p <0.05. Thus different instructions gave different ratings. It is possible that the findings are speaker specific or gender specific. Repeating the tests by speaker we find that the effects are slightly smaller for the male speaker and larger for the female, however both speakers exhibit the same tendencies with the same significant differences suggesting that, at least in this small sample, neither speaker or gender affects the results.

3. Separating Acoustics and Text

While we see a difference in a fairly unconstrained setting, it is clear that the content of the read and conversational sentences was quite different despite ensuring that each spontaneous utterance was “complete”. It is therefore possible that the prefer-

[1] http://rasmus.dall.dk/SP2014Samples.zip

	GenNat		ConvNat		ReadNat	
	Read	Spont	Read	Spont	Read	Spont
N	248	246	249	249	247	246
Mean	2.79	4.29	2.99	4.09	3.36	4.07
SD	1.292	0.915	1.292	0.938	1.114	1.145
p	*p*<0.0001		*p* <0.0001		*p* <0.0001	

Table 3: Descriptives for the audio data. The significances are between spontaneous and read sentences for each condition.

	GenNat		ConvNat		ReadNat	
	Read	Spont	Read	Spont	Read	Spont
N	399	399	399	400	395	397
Mean	3.36	3.72	2.73	3.81	3.76	3.07
SD	1.385	1.286	1.280	1.213	1.139	1.404
p	*p*<0.001		*p* <0.0001		*p* <0.0001	

Table 4: Descriptives for the textual data. The significances are between spontaneous and read sentences for each condition.

ences found are not due to differences in articulation or speech mode - but rather due to differences in content. The opposite, however, is also possible, that is, the content has nothing to say and only the acoustic differences matter. In order to tear this apart further we need to isolate the two possibilities. This is possible in the following way, firstly in order to test whether it is purely the content of the utterance which affect people's perception, we can elicit ratings from people based on text only. That is by comparing normal written text - e.g. from newspapers or novels - with transcriptions of conversational speech we can avoid the acoustic component entirely and focus purely on the content. Secondly we can isolate the acoustic component by recording a speaker in a conversational setting and then, at a later time, ask the same speaker to re-read transcriptions of their own earlier utterances. The content of the utterances will be the same however the mode of speech will differ. In this way we can tear apart the effects of content and mode.

3.1. Data

One acoustic and one textual dataset was obtained. The acoustic data consisted of studio recordings of 50 sentences initially produced in a longer conversation by a female speaker with one of the experimenters. From this conversation 50 complete (as above) sentences were identified and transcribed. The speaker was then, a few days after the first recording, asked to re-read the sentences by having them given as prompts.[2] The textual data consisted of 120 sentences. Half were taken from transcriptions of spontaneous data and the other from written sources. The transcribed data was obtained 50/50 from two generally available corpora of spontaneous data (AMI [15] and Switchboard [16]). The written data contained 30 sentences from the Arctic [13] scripts and the last 30 sentences were from News data taken from prompts used in the Edinburgh Voicebank Project [17]. For both types, novels and news, names and quotes were avoided as none were included in the spontaneous and their length matched to the spontaneous in terms of numbers of words. The choice of using various sources for both written and spontaneous data, and the inclusion of disfluencies, was to enable analysis of the possibility of internal variation depending on the style of the textual data but this analysis is not presented here due to space constraints, however we note that it does not significantly affect the presented results. An example sentence of each type can be found in Table 5.

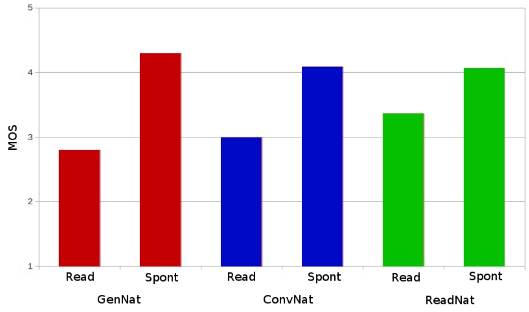

Figure 2: Naturalness ratings for the audio.

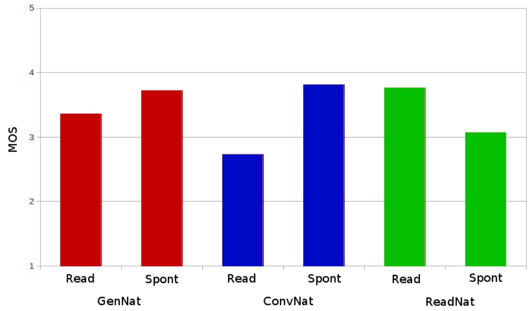

Figure 3: Naturalness ratings for the text.

3.2. Method

30 paid native speakers of English, mainly students at the University of Edinburgh, were recruited to take part. The general method was similar to the first experiment except as noted below. As before, each participant was assigned one of three groups - general naturalness (GenNat), conversational naturalness (ConvNat) or reading naturalness (ReadNat). The test had two sections. Section 1 consisted of the 50 audio samples and 4 test samples, two spontaneous and two read. Section 2 contained the 120 textual samples and 6 test samples, one of each text type. Except for test samples all presentation was randomised for each participant. In section 1 participants were asked to rate for naturalness according to their group as in experiment 1. In section 2 participants were asked to imagine that the sentence was either ”spoken aloud” (GenNat), ”said in a conversation” (ConvNat) or ”read aloud” (ReadNat), and then judge how natural the sentence would be. In total the test took about 15 minutes to complete.

3.3. Audio Results

15 responses (1%) null responses were excluded. For the GenNat (t(492)=14.864, *p* <0.0001) and ConvNat (t(496)=10.837, *p* <0.0001) groups we see a repetition of the previous results with spontaneous speech being significantly preferred over read prompts (Table 3). Contrary to earlier we now have a significant difference for the ReadNat group (t(491)=6.888, *p* <0.0001) - that is *spontaneous* speech is significantly preferred over read speech (see Figure 2). Again one-way ANOVA's were run for each speech type across groups. Here we find that no difference exists for read speech (F(2, 746)=2.693, *p*=0.068) - ratings

[2] Samples are available at http://rasmus.dall.dk/SP2014Samples.zip

Source	Example
AMI	Yeah, but you can appreciate the way they look.
SB	I do try and regulate how much exercise I get a week.
Arctic	Unconsciously, our yells and exclamations yielded to this rhythm.
News	The current deployment is designed as a deterrent.

Table 5: Example textual sentences. SB = Switchboard.

of reading naturalness did not change with instructions. However for the spontaneous speech a significant difference was found (F(2, 746)=12.197, p <0.0001) and Bonferroni correction showed the read group to be significantly (at p <0.01) different to the general and conversational group, no difference existed between those (p=0.154). In other words, instructions toward rating for reading naturalness changed peoples perception toward a higher preference for read speech.

3.4. Text Results

11 responses (0.5%) null responses were excluded. In both the GenNat (t(797)=3.877, p <0.001) and ConvNat (t(796)=12.207, p <0.0001) groups the transcribed text was significantly preferred. However, the ReadNat group significantly preferred the *written* text (t(790)=7.694, p <0.0001) (see Table 4). When imagining text spoken aloud or said in a conversation people find transcriptions of spontaneous speech over textual sources more natural - but when imagining it read aloud people found written text more natural. One-way ANOVA's support the conclusion that instructions affect peoples perceptions. For the transcriptions (F(2, 1196)=41.058, p <0.0001) Bonferroni correction showed the GenNat and ConvNat groups to differ significantly from the ReadNat group (both at p <0.0001) however not in between themselves (p=1). That is, only when rating for reading naturalness are peoples ratings affected by instructions for transcribed speech, and then towards being less natural (see Figure 3). In the written case there was also a significant effect (F(2, 1196)=58.978, p <0.0001) and with Bonferroni correction all differences were significant (p <0.001). So, when rating written text the instructions consistently affected peoples perceptions, people found written text the least natural when rating for ConvNat, more for GenNat and most natural for ReadNat (see Figure 3).

4. General Discussion

The perception of naturalness changes in the context in which it is rated, by simply adding "if you were having a conversation" or "if somebody was reading aloud" the ratings change. When no instructions were given as to what kind of naturalness to rate, participants find spontaneously produced utterances to be more natural - in line with the assumptions of earlier research. In experiment 1 the ReadNat group showed no preference for either mode of speech, when explicitly asked to rate according to naturalness when reading aloud, participants found spontaneously produced utterances *equally* natural. However, when tearing apart audio and text we see a general acoustic preference for spontaneous speech and a preference dependent on instructions for textual stimuli. Thus spontaneously produced utterances are *always* more natural acoustically than read speech - suggesting conversational speech to be the, generally speaking, most natural of the two modes of speech. If this is true it has consequences for how we should be doing speech synthesis. Assuming improved naturalness is the main current challenge in speech synthesis (in particular HMM-based) then it suggests that we should be utilising the preference for conversational speech by basing our models on such speech. This is particularly true if we wish to synthesise conversational speech, but even if we wish to make the most broadly applicable speech synthesis system we should not assume that read speech is a neutral middle ground, that may in fact be conversational speech. This is also supported by the contextual preference for transcribed speech over actual written sources.

From the second experiment, we can see that combining the general preference for spontaneous speech in the audio and the textual results, in which we see a preference for the written sources only for the ReadNat group, yields us the same picture as given in the first experiment. That is, we have successfully managed to tear apart the difference between the acoustics and the meaning content of the sentence by removing the variables in their respective tests. It is important to note that, for the textual case, we have focused on the spoken word, not the written, by instructing participants to rate it according to how natural it would be in various spoken scenarios and not how natural it would be focusing on it as text. In light of the clear effect of instructions on peoples ratings (more below) we would expect instructions geared toward *written* naturalness to yield a differing result. Both the first and the second tests support the hypothesis that naturalness as a metric can be easily influenced by experimental instructions, and that the influence is dependent on the type of data under consideration. This is likely due to the concept of naturalness in general being under-specified, and so by conditioning the experimental setting we can influence our participants toward various interpretations. Knowing this encourages both caution and enables more detail when evaluating synthetic speech. Caution because we must be diligent with the instructions we give participants so as not to bias them in an unwanted direction. More detail as we can condition the metric toward specific aspects of naturalness.

5. Conclusions and Further Work

We have shown that MOS-scale ratings can beneficially be employed to distinguish the conversationality of speech, in fact spontaneous conversational speech is found more natural by listeners than read prompts. We can affect peoples perception of naturalness by simple conditions in the instructions, enabling greater control over the testing scenario while also cautioning its use. Further work includes rigidly defining what is natural in the general case, but also attempting to utilise the apparent advantages of conversational speech. Our results suggests that read prompts may not be the neutral general speech as previously assumed and that this role is more likely attributable to spontaneous conversational speech. The gathering and use of such speech present many challenges which must be met before it is generally applicable, however we intend to attempt the use of such data by gathering an appropriate spontaneous corpora, but also by utilising existing data not recorded specifically for speech synthesis.

6. Acknowledgements

Thanks to Cereproc Ltd. for providing the female data for the first listening test. This work was funded by the JST CREST uDialogue project.

7. References

[1] C. Benoit, M. Grice, and V. Hazan, “The SUS test : A method for the assessment of text-to-speech synthesis intelligibillity using Semantically Unpredictable Sentences,” *Speech Communication*, vol. 18, pp. 381–392, 1996.

[2] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, “The Blizzard Challenge 2008,” in *Blizzard Challenge Workshop*, 2008.

[3] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, “Statistical analysis of the Blizzard Challenge 2007 listening test results,” in *Blizzard Challenge Workshop*, Bonn, Germany, 2007, pp. 1–6.

[4] S. King and V. Karaiskos, “The Blizzard Challenge 2009,” in *Blizzard Challenge Workshop*, 2009.

[5] M. Fraser and S. King, “The Blizzard Challenge 2007,” in *Blizzard Challenge Workshop*, 2007, pp. 1–12.

[6] J. Adell, A. Bonafonte, and D. Escudero-mancebo, “Modelling Filled Pauses Prosody to Synthesise Disfluent Speech,” in *Speech Prosody*, Chicago, USA, 2010.

[7] S. Andersson, J. Yamagishi, and R. A. Clark, “Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis,” *Speech Communication*, vol. 54, no. 2, pp. 175–188, Feb. 2012.

[8] N. Campbell, “Towards Conversational Speech Synthesis; Lessons Learned from the Expressive Speech Processing Project,” in *SSW6*, Bonn, Germany, 2007, pp. 22–27.

[9] T. Koriyama, T. Nose, and T. Kobayashi, “Conversational Spontaneous Speech Synthesis Using Average Voice Model,” in *Interspeech*, no. September, Makuhari, Japan, 2010, pp. 853–856.

[10] E. Blaauw, “Phonetic Characteristics of Spontaneous and Read-Aloud Speech,” in *ESCA Workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, no. September, Barcelona, Spain, 1991, pp. 1–5.

[11] S. King and V. Karaiskos, “The Blizzard Challenge 2013,” in *Blizzard Challenge Workshop*, Barcelona, Spain, 2013.

[12] J. Adell, D. Escudero, and A. Bonafonte, “Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence,” *Speech Communication*, vol. 54, no. 3, pp. 459–476, Mar. 2012.

[13] J. Kominek and A. W. Black, “CMU ARCTIC databases for speech synthesis,” Carnegie Mellon University, Tech. Rep., 2003.

[14] H. M. Marcus-roberts and F. S. Roberts, “Meaningless Statistics,” *Journal of Educational Statistics*, vol. 12, no. 4, pp. 383–394, 1987.

[15] J. Carletta, “Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus*,” *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[16] J. J. Goodfrey, E. C. Holliman, and J. McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development,” in *ICASSP*, San Francisco, CA, USA, 1992, pp. 517–520.

[17] J. Yamagishi, C. Veaux, S. King, and S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.