How Should We Define Voice Naturalness

Sajad Shirali-Shahreza
Department of Computer Science, University of Toronto
Department of Computer Engineering, Amirkabir University
Email: shirali@aut.ac.ir

***Abstract*—Naturalness is a commonly used criteria in Text-To-Speech (TTS) evaluations. The goal is to measure how close generated voice is to real human voice. This is measured through listening tests by human participants. However, no definition for naturalness is provided to participants. In this paper, we aimed to identify what definition participants used when they rank the naturalness. We conducted a user study similar to TTS evaluations and analyzed their responses. We noticed that users have different and sometimes contradictory definitions about it and a major dimension for them was how close it sounds to a real human. Our results show that we should explicitly define the naturalness for the participants. Furthermore, we should ask separate questions for different dimensions of naturalness such as clarity and having accent.**

Keywords-Text-to-Speech (TTS); Naturalness; Evaluation.

I. INTRODUCTION

Text-to-Speech (TTS) system evaluations usually have two parts: naturalness measurement via Mean-Opinion-Scores (MOS) and intelligibility measurement via transcription error rates of Semantically Uninterpretable Sentence (SUS). TTS systems are usually compared with each other and/or real human voices. As a result, naturalness is now regarded as an ordinal dimension of speech quality in its own right.

Recent advances in TTS systems results in deep-learning-inspired systems, such as Tactron [1] are almost indistinguishable from real-human voice. The way that such claims are presented is through MOS results that are almost 5. Mean opinion scores for TTS naturalness are generally calculated on a scale between 1 (worst) and 5 (best) as a subjective assessment by a human listener as to how *natural* a sample sounds, with no definition of what *natural* means, nor a provision of context within which the sample occurs, out of concern that it may prime the listeners [2]. Samples for this task are generally one sentence long.

What is interesting, however, is that up until about 1995, "*natural speech*" was the preferred technical term for describing human-generated speech. There was no discussion of an abstract *naturalness* that synthesizers could approximate on a scale from 1 to 5. There was a very detailed discussion, on the other hand, about the quality of synthesized speech, and indeed the earliest ITU-T P.85 standard [3] for evaluating speech synthesizers was equipped with three so-called Q-type scales that were designed to measure just that. The first mention of *naturalness* that we can find was actually in the speech coding literature [4], where it was used to describe degradations in subjective quality and speaker recognizability that did not also affect intelligibility.

The earliest Blizzard challenges [5] faithfully measured naturalness, along with another feature called *similarity*, in a context in which every synthesized prompt could be compared to a gold-standard recording of the same prompt by the same voice on which the synthesizer itself had been trained, and so every synthesized sample could be interpreted as an approximation of a human-generated sample. The connection to speech coding was very clear.

Our recent [6] comparison of the naturalness of TTS systems and ordinary human users shows that, by the empirical standards of the present-day TTS research community, TTS systems had reached statistical parity with human speech in its degree of naturalness at some point prior to 2013. This forces us to conclude that either the more recent quest for human-like speech quality by deep learning researchers is simply moot or that the concept of abstract naturalness is not well-founded.

One of the results of our previous study was that users rank accented speech as less natural. That was similar to an old study [7] that reported similarities between degradation due to synthesized speech and degradation due to foreign-accented speech. That was observed through a dimensionality reduction of more ecologically valid performance measures in the context of speech interfaces for pilot's cockpits by the United States Air Force [7].

In earlier Blizzard challenges (as per the recommendations for the ITU-Q scales), it was not uncommon to find considerably longer prompts, with very vertically directed instructions on how to establish one's impression:

"Overall impression: Please try to imagine what your reaction would be if this were an actual telephone message from a mail order house or a request for information from a travel agency."

"Acceptance: Please indicate whether or not you find that the voice you heard would be acceptable for such an automatic answering service by telephone."

However, these are not precisely defined instructions or definitions, as they require introspection on the part of the listener. This is in contrast to the transcription tasks for measuring intelligibility, in which the listener's accuracy is objectively measured.

In this paper, we try to see how users who participate in TTS evaluation implicitly define naturalness for themselves and then use it to perform the evaluation. We conducted a user study that mimics the usual evaluation of TTS systems and at the end, explicitly asked the participants to define the naturalness (Section II). Then, we coded the answers and extract concepts from them using grounded theory [8] (Section III). After that, we analyze our observations from coded data, identify some potential problematic area related to naturalness definition, and propose some potential solutions for them (Section IV).

II. Data Collection

As we mentioned earlier, it is common in TTS evaluation to measure the naturalness of generated speech. They ask the user to express how natural an example prompt is. For example, in the Blizzard challenge, they ask the user to:

*"Now choose a score for how **natural** or **unnatural** the sentence **sounded**. The scale is from **1 [Completely Unnatural]** to **5 [Completely Natural]**."*

The assumption is that the user already knows the definition of "natural". In our work, we designed a user study that closely mimics the usual TTS evaluations studies, such as the Blizzard challenge. Considering that we have both human and TTS-generated voices in our study, out ethics board did not allow us to tell the participants that they are evaluating TTS generated voices (which is common in TTS evaluations). Instead, they allowed us to use the phrase "evaluate computer-generated speech."

A. User Study Structure

Our user study had 5 main parts:

1) Consent

The first part is the welcome page that provides an overview of the study. The user is provided communication options, such as email address and phone number that they can use to obtain more information about the study before deciding whether they want to participate or not. If they decide to participate, they should indicate their acceptance of the rules which is used as the consent form.

2) Demographic questionnaire

After expressing the desire to participate in the research study, they will fill out a questionnaire that collects general information about the user. It includes items, such as their age range, whether they are native English speakers, how they would rate their English reading/listening/speaking/writing ability, etc. The information that collected is similar to what is collected in TTS evaluations, such as Blizzard challenge.

3) Individual prompt naturalness

We ask the user to perform two types of naturalness assessment. In the first type, they should assess the naturalness of a single prompt. This is how usually TTS evaluations, such as Blizzard ask the user to assess a TTS system. Considering that TTS evaluation tasks also ask the user to transcribe prompts (which is used to measure the intelligibility of the generated voices), we created a combined question for each prompt that first asks the user to transcribe the text, followed by a question that asks them to assess the naturalness. We tried to use question and prompt that closely resemble those that are used in previous Blizzard challenge and other TTS evaluations. Here is the instruction that we show to them for the naturalness assessment:

*"Now rate how **natural** or **unnatural** the sentence **sounded**"*

There were 5 options to select from:

1. Completely Unnatural
2. Mostly Unnatural
3. In Between Natural and Unnatural
4. Mostly Natural
5. Completely Natural

4) Pairwise Naturalness Comparison

We also added an extra section that asks the user to perform pairwise comparison of naturalness between prompts that are generated by different systems. TTS evaluations usually do not include this because they would need larger number of participants and longer study sessions inorder to have enough data to perform data analysis. However, it provides better evaluation between prompts because even if we assume that everyone have the same definition for naturalness, their expectations are not aligned. For example, one user may be more sensitive to small deficiencies and rank a prompt with a lower score, while another user gave them the same score.

For this part, the user could only listen to each prompt once, and they should also listen to prompt A first and after that prompt finished, they can listen to prompt B. After listening to both prompts, they should compare their naturalness. Here is the instruction that we gave them:

"Please listen to the following two voices and compare their naturalness. You should ignore the meanings of the sentences and instead concentrate on how natural or unnatural each one sounded. You can listen to each utterance by clicking on the play button beneath it. Note that you can only listen the utterance once, and you should listen to voice A at first."

We used almost identical wording to refer to the naturalness in this part. They should select one of the five options as the answer:

1. Voice A is significantly more natural than voice B
2. Voice A is slightly more natural than voice B
3. Their naturalnesses are similar
4. Voice B is slightly more natural than voice A
5. Voice B is significantly more natural than voice A

5) Naturalness Definition

After completing the naturalness assessment of different prompts from different speakers, we ask our main question as a single post-study questionnaire:

"Please define the naturalness definition that you used to rank the naturalness of voices in this user study:"

Our goal was to let the user complete the evaluation of the prompts as they would in other TTS evaluations and then ask them to define the naturalness that they used earlier.

B. Speaker Prompts

We included 25 different speakers in our study: 5 professional speakers (one from the original training data of Blizzard 2013 and 4 other professional speakers), 5 native Indian speakers, 5 native (but not professional) North American speakers, and 5 TTS systems from Blizzard 2013 (systems B, C, D, H, K). We selected our samples from the Blizzard 2013 challenge because it was the last year that they used English as their main task.

For each speaker, we selected to different sentences to mimic the different sentences that are used in TTS evaluations to eliminate the effect of text on the performance. The sentences were selected from the set of the sentences that were used for Blizzard 2013 challenge.

C. *Participants*

We wanted our user study to be similar to the Blizzard challenge. Therefore, it was designed as a web-based study that could be completed over the internet. We recruited participants from Amazon Mechanical Turk [9]. 175 participants completed the study. Amazon Mechanical Turk was used in various Blizzard challenges and also by different researchers for evaluation of TTS systems, which is why we used the same approach to participant recruitment.

III. CODING

We used the grounded theory and emergent coding approach that is presented in chapter 11 of [8] for coding.

A. *Codes*

We started with one pass of analyzing all definitions and extracting codewords from them. Each time we see a new keyword in an answer, we add it to the code list and consider it for the remaining answers. The output of this phase is 146 codewords, while each answer has in average 4.85 keywords.

Most of these codewords only appear in a few answers. For example, more than 85% of them appeared in less than 10 answers, two/third of them appeared in less than 5 answers, and more than one third of them only appeared in a single answer.

Then, we started to combine codewords that were closely related to each other or were used in the same context for the same meaning. For example, we grouped codewords *Tell* and *Express* together because both describes the same action. Another example is grouping of codewords *Human*, *People*, *Person*, *Mind*, *Everyone*, and *Someone* that were used to refer to a human user speaking. At the end of this pass, we reduced the number of codewords to 39 codes. Each answer has an average of 4.61 codes.

The reason that the average number of codes is reduced is because some answers were using related codewords that were combined during this process. For example, a user mentioned “reading from a paper” in their answer. In the first pass, we added both “reading” and “paper” as codewords. At the end, only 3 answers had codewords “reading” and no other answer had keyword “paper”. Furthermore, both codewords were referring to the concept of a written text that is being read. So, we combined these two codewords (along with the codewords *Scripted* that was mentioned in another answer). This resulted in reduction of the codes that are assigned to this answer from 5 to 4.

B. *Concepts*

After finalizing the set of codes, we grouped similar and related codes into *concepts* [8]. We performed multiple iterations of grouping to finally come up with five concepts. The concepts and related code are presented in Table 1, along with the number of answers that have that code.

TABLE I. CONCEPTS AND CODES

Concept	Code	Answer Count
Speech Properties	Accent	20
	Clarity	30
	Emotion	4
	Flow	11
	Noise	5
	Pause	11
	Pitch	3
	Pronunciation	11
	Tone	14
	Smoothness	8
	Speed	3
	Understand	24
Classes	Computer	48
	Everyday	9
	Generated	24
	Human	64
	Mechanical	10
	Normal	22
	Reading	7
	Real	18
Adjective/ Adverbs	Adjective	18
	Adverb	13
Defining Process	I	53
	Feel	4
	How	15
	Comparison	18
	Like	45
	Mean	14
	Quality	5
	Rank	11
	Should	7
	Whether	14
Receiving Information	Hear	23
	Speak	26
	Speech	21
	Sounded	83
	Tell	4
	Understand	24
	Voice	66
	Word	21

1) *Speech Properties*

The first concept consists of codes that describe the properties of speech. Half of all answers (88) had at least one of these codes. This shows that for at least half of the users, the naturalness relates to speech properties.

In this concept, the main codes were Clarity (34%), Understand (27%), Accent (23%) and Tone (16%). This shows that users usually focus on the clarity of the voice and whether they can understand it. However, these two properties (especially the understand one) are more closely aligned with the intelligibility of the voice that is usually measured in TTS evaluations with transcription error.

Another important code here is *Accent.* 11% all answers have a word that express this code. Users used other words, such as *Native*, *American*, *Indian*, and *Foreign* to refer to this concept.

There was also a clear disagreement between users about whether they should or should not consider the accent as part of naturalness. While most users think that having an accent does not reduce the naturalness (such as saying “*Even if it* *were foreign, it could still sound natural*" or "*... not including accents which I did not use as a basis.*"), while another user says (such as a user that equate naturalness with speaking with American accent and says "*If the person had an American accent, I thought it was more natural than an Indian accent.*") However, this is in contrast to what our other study [6] shows: people rank speakers with Indian accent as less natural than speakers with North American accent.

2) Classes

The second concept was the classes that users use to define the naturalness. They usually consider two classes of speech such that one is natural and the other is not natural, and then use terms to describe them. Sometimes, they only refer to one of the classes and say that it is natural if it belongs to or not belongs to that group. Two-third of all answers use at least one code to describe these classes.

In general, these two groups are *Humans* (55%) and *Computers* (42%). In addition to using those nouns, they also used adjectives for speech to express this: *Generated* (21%) and *Mechanical* (9%) for computers and *Normal* (19%), *Real* (16%), *Everyday* (8%), and *Reading* (6%) for humans.

3) Receiving Information

The third concept consist of words that describe how they receive the information from the prompt. The majority of answers (85%) have at least one such code. Top codes in this concept are *Sounded* (56%), *Voice* (45%), *Speak* (18%), *Understand* (16%), and *Hear* (16%).

They can be grouped into two sub-concepts: those that related to how the speech is generated by the speaker (Speak, Speech, Tell, Voice, and Word) and how it is received by the listener (Hear, Sounded, and Understand).

4) Defining Process

The fourth concept is the group of words that describe how they define the naturalness. Two-third of answers have at least one such code. The most common one is *I* (46%) (that is a combined code for words such as I, my, me, we, etc.) which shows that users express how they would define naturalness. Only a few users' answers (6%) have code *Should* that represents what is the global definition of naturalness.

Other common codes in this concept are *Like* (39%), *Compare* (15%), and *Whether* (12%). They are used along class concept codes to express that users consider naturalness to be measurable by finding the class (human or computer) that it belongs to. In other words, users consider natural to be equivalent to human generated and unnatural with computer-generated. For example, one user says "*Naturalness to me means something that comes out of the person.*".

5) Adjective/Adverbs

The fifth concept was adjective and adverbs that they use to better express their idea. For example, they may say "*understand easily*". 17% of answers have such an adjective/adverb.

IV. Analysis

In the previous section , we provided the concepts and main codes that could be used to describe how users define naturalness. Now we provide a summary of what themes we observed from these codes and how they can help us to better define the naturalness.

A. Is it necessary to define it?

The first observation is that even for some users, the question of defining naturalness seems to be unnecessary, because they the consider the naturalness to be a primitive fact. Here is one example of a user's answer:

"What? if it sounded more natural I voted it to sound more natural. what is this question asking?"

Or another user says:

"I'm not sure what you are asking, we were asked to rank which voices sounded more natural and less like computer-generated voices."

While this may signal that maybe everyone already knows what naturalness means, our results show that is not completely correct.

B. Is there a universal definition?

Our second observation was the use of codes from the defining process concept, such as *I* and *Mean* that shows that the users express how they define the naturalness and not how it is actually defined. This shows some sort of doubt and ambiguity about the universal definition of it. We could remove this ambiguity by providing a clear and concise definition of the naturalness.

Another result of this lack of universal definition is the difference between people in assigning numerical (or level-based) scores to prompts. Some users may be more precise and reduce score even for small errors and easily mark a prompt as somehow natural in that case, while other users are more lenient and still mark a prompt as completely natural even if it has small errors, such as s single mispronounced work. This can be referred to as the normalization problem.

One way of preventing this error is to ask the users to compare the naturalness of two prompts together (rather than asking them to individual rank them). This way, we will not have the need to normalize their rankings.

C. Do people agree on specific factors?

Third, we noticed that users may have contradictory opinions about how different properties of speech affects naturalness. One such aspect is the speaker's accent. Referring to the accent of the speaker was a common theme, although they used different terms to do it (such as accent, native, Indian, American, Foreign, etc.). And while most of the people who used such a term were trying to say that this factor is NOT affecting naturalness, we also has examples that users explicitly say that they consider foreign speakers who have an accent to be less natural.

Furthermore, our other research [6] revealed that even if most of the users think that they are ignoring accent, in practice they would rank speech samples with accent as less natural.

One possible solution is to explicitly pick a subset of criteria and ask the user to rank the samples base on that. For example, we can ask the user to rate the clarity of the voice, whether it had accent, how fluent it is (while also clearly define fluency to prevent mixing it with foreign speakers who have accent and are not considered fluent), how is the speakers pronunciation (which can include pronunciation errors that is present in a TTS system due to OOV words, or in real human speakers if the user do not know the correct pronunciation).

D. *Sounds like a human*

Our fourth observation was how people equate naturalness with being generated by a human. In other terms, people consider something to be natural if it is said by a human. Although users referred to aspects such as understandability (which is usually referred to as intelligibility by researchers) and clarity, their main criteria seems to be their belief of whether it is generated by a human or not. I.e., is it a real human speaking or is it created by a computer.

This is an important point to consider. If people equate naturalness with human-generated voice, then any computer-generated voice will not be completely natural. And it becomes unclear less useful to ask them to mark its naturalness if they do not consider it to be completely natural.

In other terms, if we tell the user that the speech sample is NOT said by a real human (e.g., by saying these are the output of different TTS systems), we already biased them to believe that it is not natural. Therefore, it would not be surprising that our results will show that our system is not completely natural (i.e., get a score of 5) and instead we would see a significant different between the naturalness score of our TTS systems and real human samples.

One potential wat to resolve this problem is to ask users to rate how close our system sounds like a real human (instead of asking them how natural it is). In that case, even if they know that it is not a real human and therefore not completely natural, they may still say that it sounds almost like a real human.

V. Conclusion and Future Work

In this paper, we visited the problem of what we mean by naturalness when are evaluating the TTS systems. We conducted a study that asked to users to perform similar TTS evaluations and at the end ask what naturalness definition they used. We coded the answers and groups them into concepts.

Our analysis revealed that even some users believe that it is not necessary to try to define the naturalness, yet most of them provide answers that show the definition that they provide is how they would define it (rather than a universal definition that they quote). Furthermore, their short definitions were contradicting each other about how should we consider properties such as accent.

Overall, our results show that not providing a precise definition for the naturalness caused confusion for the users and results in them using varying and conflicting definitions for the naturalness.

We provided some potential approaches to resolve these issues, which are primarily focused on how we can better define the naturalness for the participants. We plan to conduct a follow-up study to assess how these solutions may resolve the problems. For example, we can provide precise definition of different criteria such as clarity, understandability, fluency, and accent to the users and ask them to rank the samples base on them and then measure the correlation between them. This can be done by replacing a single naturalness question with multiple questions for each aspect of that. For example, one question just asks for the ranking of the prompts based on their clarity, while another one asks for the presence of the accent in the sample.

The main takeaway of our paper for TTS researchers is the need to clearly define the naturalness for the participants. Furthermore, considering different aspects of naturalness for human users, it is better to evaluate aspects such as clarity and accent individually rather than combine all of them as a single measurement of naturalness.

References

[1] J. Shen, et al.,“Natural TTS synthesis by conditioning wavenet on mel-sectrogram predictions,” Proc. ICASSP 2018, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[2] M. Fraser and S. King, “The blizzard challenge 2007,” in Proc. 3rd Blizzard Challenge, 6th ISCA Workshop on Speech Synthesis, 2007, paper 003, pp. 1-12.

[3] ITU-T, “Telephone transmission quality subjective opinion tests: A method for subjective performance assessment of the quality of speech voice output devices,” ITU-T Recommendation P.85, 1994.

[4] W.B. Kleijn and K.K. Paliwal, “Principles of speech coding,” in Speech Coding and Synthesis, Elsevier Science, 1995.

[5] Speech Synthesis Special Interest Group, Blizzard Challenge, https://www.synsig.org/index.php/Blizzard_Challenge, retrieved: April 2023.

[6] S. Shirali-Shahreza and G. Penn, “MOS Naturalness and the Quest for Human-Like Speech,” 7th IEEE Workshop on Spoken Language Technology (SLT 2018), 2018, pp. 346-352, doi: 10.1109/SLT.2018.8639599.

[7] C.A. Sampson and T. Navarro, “Intelligibility of computer generated speech as a function of multiple factors,” in Proc. National Aerospace and Electronics Conference, 1984, pp. 932–940.

[8] J. Lazar, J.H. Feng, and H. Hochheiser, “Research Methods in Human-Computer Interaction,” 2nd Edition, Morgan Kaufmann, 2017

[9] Amazon, Amazon Mechanical Turk (MTurk), https://www.mturk.com, retrieved: April 2023.