Evaluating Voice Quality and Speech Synthesis Using Crowdsourcing

Jeanne Parson[1], Daniela Braga[1], Michael Tjalve[1,2], and Jieun Oh[3]

[1] Microsoft, USA
[2] University of Washington, Seattle, USA
{jeannepa,dbraga,mitjalve}@microsoft.com
[3] CCRMA, Stanford University, Stanford, USA
jieun5@ccrma.stanford.edu

Abstract. One of the key aspects of creating high quality synthetic speech is the validation process. Establishing validation processes that are reliable and scalable is challenging. Today, the maturity of the crowdsourcing infrastructure along with better techniques for validating the data gathered through crowdsourcing have made it possible to perform reliable speech synthesis validation at a larger scale. In this paper, we present a study of voice quality evaluation using the crowdsourcing platform. We investigate voice gender preference across eight locales for three typical TTS scenarios. We also examine to which degree speaker adaptation can carry over certain voice qualities, such as mood, of the target speaker to the adapted TTS. Based on an existing full TTS font, adaptation is carried out on a smaller amount of speech data from a target speaker. Finally, we show how crowdsourcing contributes to objective assessment when dealing with voice preference in voice talent selection.

Keywords: voice quality evaluation, speech synthesis, Text-to-Speech (TTS), crowdsourcing (CS), voice preference, gender preference.

1 Introduction

Online crowdsourcing (CS) marketplaces provide an environment for fast turn-around and cost-effective distributed outsourcing, at a statistically meaningful scale, leveraging human intelligence, judgment, and intuition. Such services are typically used by businesses to clean data, categorize items, moderate content and improve relevancy in search engines. However, over the past several years, the Speech Science community has also adopted them as a novel platform for conducting research that offers a more scalable means of: a) measuring speech intelligibility [1] and naturalness [2], b) of collecting data [3] and c) of processing that same data, either through annotation for natural language tasks [4] or audio transcriptions for automatic speech recognition [5]. For a complete review on the use of CS for speech-related tasks and anticipated challenges for the future of CS for speech processing, see [6-7]. In all the described experiments, we used Microsofts Universal Human Relevance System (UHRS) as the crowdsourcing platform. UHRS is a marketplace that connects a large worker pool with human intelligence tasks. Tasks can be distributed to workers within a specific country. The UHRS workers are

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 233–240, 2013.

provided by several vendors across world-wide markets, providing many thousands of unique workers. Through these studies, we demonstrate the potential of CS to obtain subjective evaluation of real and synthesized speech. This paper details how CS can be used to evaluate human voice quality and synthesized speech in the context of text-to-speech (TTS), using subjective ratings from listeners and users. We survey voice gender preference (Section 2), examine the effects of voice adaptation technology on the perception of synthesized speech (Section 3), and evaluate voice preference to select the ”best” voice from recordings of several voice talents (Section 4).

2 Surveying Voice Gender Preference

We had two goals with this experiment. One was to understand end users' voice gender preference when using Text-to-Speech (TTS) systems on mobile phones across three different scenarios: 1) instructions and confirmations, 2) read-out of text (SMS and/or tweets), and 3) driving directions, and covering eight locales. The second goal was to ascertain if we could extract reliable data for speech validation using CS. We are aware that there is considerable discussion around the effect of gender vis-a-vis users' preferences in all types of voice response systems [8], however it was not a goal in this experiment to investigate beyond the scope of our scenario.

We first framed the experience with a statement intended to ensure, as much as possible, that the crowd judges understood the meaning of TTS and its general usage on a mobile phone: *”You probably know that many smartphones have the ability to talk back to you. For example, if you ask to ”Call Anna”, the phone might talk back to you and say ”Call Anna, at home or on the mobile phone?” Or the phone might have GPS and could read driving directions to you. For this survey, we will use ”TTS” to refer to the phone's ability to talk back to you.”*

We then presented the judge with a short series of questions: *1) Do you use a mobile phone that has TTS? 1.a. If ”yes”, do you use the TTS feature?; 1.b. If your previous answer is ”sometimes”, when do you use the TTS feature?; 2) If the TTS is confirming actions for you, do you prefer a male or female voice?; 3) If the TTS is reading a text message or tweet to you, do you prefer a male or female voice? 4) If the TTS is reading driving directions, do you prefer a male or female voice?*

We also asked the judges' to briefly describe their impression of TTS voices, and we asked the judges' gender. The text-input responses for question 1b and the follow-up question asking for a brief description of judges' impression of TTS served as a spam filter, and also provided opportunities to validate if judges' who did not have a mobile phone with TTS understood what TTS is. As with all CS, we also knew that it was not possible to control the balance between male and female respondents and that results would need to be normalized for gender.

The total crowd size for 6 of the 8 locales was large enough to provide solid data. For en-CA and it-IT, with less than 29 judges, the text comments validated that the responses from these locales were still highly valuable (the comments from Italian judges were especially robust and insightful). As expected with CS, the gender distribution of

Table 1. Demographics distribution per locale

Locale	Total Crowd	Female	Male
en-US	100	69	31
en-CA	23	14	9
en-GB	54	28	26
de-DE	38	21	17
fr-FR	30	12	18
es-ES	31	6	25
it-IT	18	5	13
es-MX	29	9	20

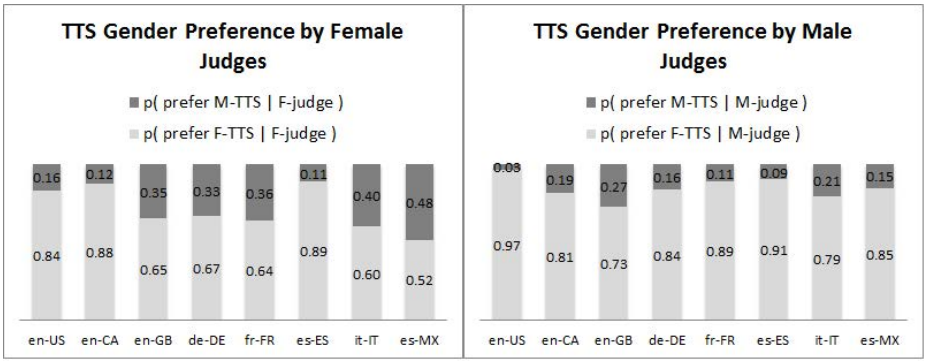

Fig. 1. Gender preference by female (left) and male (right) judges. Light gray denotes preferring female TTS voice, and dark gray denotes preferring male TTS voice.

the judges was uneven (Table 1). When answers to questions 2-4 were aggregated, we observed that both female and male judges in all 8 locales tended to prefer a female voice to a male voice, although the extent to which female voice was preferred was stronger by male judges than by female judges, for all locales except en-CA (Fig. 1).

Judges' responses to questions 2 and 3 varied by less than 2 except for de-DE and es-MX, which varied by 3 and 4 respectively. Because the responses for these two scenarios (confirming actions and reading a text message) were so aligned, we combined the results for both questions. For all locales, there was a strong preference for a female TTS voice in both of these scenarios (Table 2). For driving directions, there was more variance in the gender preference gap for some locales (Table 4), and especially for judges from Mexico. Since Mexico also had the largest gap for questions 2 and 3, it is worth looking at Mexico separately (Table 3).

For the Mexican market, there were only 9 female judges. Only 3 of the female judges preferred a female TTS voice for the driving directions scenario, but 7 preferred female for confirming actions, and 4 preferred female for reading text. These numbers reflect the overall pattern, regardless of the gender of the judge, indicating no preference based on judge's gender.

Table 2. Gender preference for confirming actions and reading text messages

Locale	Total Crowd	% prefer Female
en-US	100	89.0%
en-CA	23	91.1%
en-GB	54	70.3%
de-DE	38	73.7%
fr-FR	30	80.0%
es-ES	31	96.8%
it-IT	18	77.8%
es-MX	29	82.3%

Table 3. Mexico gender preference by scenario

(TTS gender preferred)	Confirm Actions	Read Text	Driving Directions
Female	26	22	17
Male	3	7	12

Table 4. Preference for female gender for confirming actions/reading text [Q2-3] vs. driving directions [Q4]

	en-US	en-CA	en-GB	de-DE	fr-FR	es-ES	it-IT	es-MX
Q2-3	89.0%	91.1%	70.3%	73.7%	80.0%	96.8%	77.8%	82.3%
Q4	85.0%	73.9%	68.5%	78.9%	76.7%	80.6%	72.2%	58.6%

3 Perception of Speaker Mood in Adapted TTS Fonts

3.1 Overview and Methodology

Voice Adaptation is a technique in text-to-speech (TTS) that generates a new voice (target voice) based on the training of a source voice [9]. The adaptation technology takes an existing TTS font and ”adapts” it to the voice of a new target speaker based on a smaller quantity of speech data of the new speaker. Thus, we designed a study to better understand the extent to which voice qualities of the human target speaker impact that of the adapted font; our underlying motivation was to estimate the potential of improving certain problematic voice qualities such as friendliness or mood of existing TTS fonts through adaptation. We employed two types of listeners: non-expert crowd workers through UHRS, and language experts who are validated native speakers in the language. This study investigated the impact of adaptation on the perceived voice qualities of a newly generated TTS font in the es-ES locale. Four voices (2 human and their respective synthetic voices) were used: *Helena-human* was used to generate *Helena-TTS*, and *Laura-human* was the target speaker for generating the adapted font, *Laura-TTS*.

We conducted analysis in terms of the following four **voice qualities**: Listener perception on speaker's naturalness of prosody, Listener perception of speakers mood, voice color (timbre) preference by listener, and General preference by listener.

We generated 12 unique tasks using different scripts. In choosing our scripts, we ensured that the script was available in both of the human voice recordings, and that the semantic content of the script was neutral. We deployed the tasks using 1) crowdsourced non-expert workers and 2) recruited language experts (LEs). For CS, we created two versions of each of the 12 unique HITs (tasks), reversing the order of speech stimuli. We allowed for up to 12 worker responses per HIT-order for a total of 288 HITs available in the marketplace. Similarly, for deployment to LEs, we created two versions of 12 unique surveys (tasks), reversing the order of speech stimuli. Twenty medium-skilled LEs each completed the 12 surveys, resulting in 10 LE responses per order-pattern. The LE surveys were implemented as a web application which matched the format of the CS surveys.

We designed a task consisting of three sets. In Set 1, listeners compared between the two human recordings, *Helena-human* and *Laura-human*. In Set 2, listeners compared between two synthesized speech, *Helena-TTS* and *Laura-TTS*. Set 3, in which listeners compared between *Helena-human* and its derived font *Helena-TTS*, was designed as a catch trial to check on the quality of workers responses, with the assumption that a non-spam response would show a clear preference for the human recording over the synthesized font. For each set, two speech stimuli under comparison were presented, followed by 11 questions (translated into Spanish) addressing the perception of our four voice qualities: prosody, mood, timbre, and general preference.

We hypothesized that adaptation would result in a font (*Laura-TTS*) whose voice qualities match that of the target speaker (*Laura-human*) rather than the parent speaker (*Helena-human*). That is, for a given voice quality, we hypothesized that (1) if a listener prefers *Laura-human* over *Helena-human*, then s/he would prefer *Laura-TTS* over *Helena-TTS*; and inversely, (2) if a listener prefers *Helena-human* over *Laura-human*, then s/he would prefer *Helena-TTS* over *Laura-TTS*.

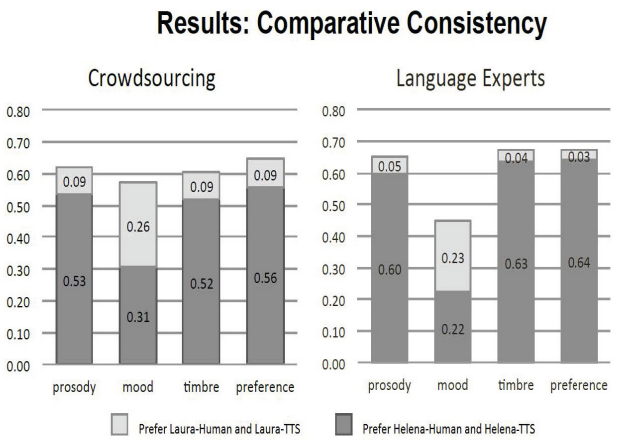

Fig. 2. Results for crowd-sourced workers and language experts

Fig. 2 summarizes the percentage of responses that follow our hypothesis, categorized by four voice qualities. The upper bar, in light gray, represents responses preferring *Laura-human* and *Laura-TTS*; the lower bar, in dark gray, represents responses preferring *Helena-human* and *Helena-TTS*. The sum of two bars represents the percentage of responses that met our hypothesis, which was shown to be greater than 50 percent for all of the voice qualities, with the exception of "mood" ratings by LEs. More interestingly, speaker's mood was shown to be a voice quality most positively impacted by adaptation in our experimental context; judges perceived *Laura-human* and *Laura-TTS* to be in a better "mood", as shown by the height of the light-gray bar for "mood". Though the trained LEs responses are slightly more pronounced, the graphed results clearly show that results from the crowd are consistent with LE results. This validates that crowdsourcing can be a reliable resource for perception of voice qualities.

4 Evaluating Voice Preference

4.1 From Top 10 to Top 3

The goal of this study was to contribute to a decision to narrow a pool of 8-10 voice talent candidates (voices) down to 2 or 3 finalists. A listening survey was deployed to 100 judges via a customizable tool for audio listening and judgment collection from the crowd. The survey had two main sections: 1) the judge listened to sets of 3 voices and chose a favorite, and 2) the judge chose his/her favorite and least favorite voice overall, and provided comments as to why s/he made that choice. To ensure no bias based on listening order in the first section, comparison sets were pre-mapped on a matrix such that each voice was: 1) heard three times by each judge 2) compared in different orders (e.g., first, middle and last) and 3) compared an equal number of times against each other voice. For all listening samples, we used a recording of the human voice speaking the same sentence. It was not expected that the results of the first section would exactly match the results of the second section because a judge may have had to choose between 3 voices in a set, none of which were their favorite. Judge's comments offered insight into their favorite and least favorite voice.

Table 5. Results of the crowds

Voice	1	2	3	4	5	6	7	8	9	10
Total votes	57	51	109	78	74	114	79	84	93	111
Favorite	6	6	11	9	11	15	3	4	9	11
Least Favorite	13	32	3	6	9	5	5	3	7	5

In total votes from the first section, Voices 3, 6 and 10 were the most preferred, with a clean margin between the 3rd choice and 4th choice (Voice 9). When voting for a single favorite (second section of the survey), users gave Voice 6 the most votes, however Voices 3, 5 and 10 tied for 2nd place. In terms of the least favorite, Voice 2, there was consistency between the least favorite score (e.g. 32) and the lowest score in total votes (e.g. 51). In the favorite category, Voice 2 did receive more votes (e.g. 7) than Voices 7 and 8. However, a score of less than 7 in the Favorite category marks the mid-to-low

range. This survey was used as one of three streams of input for selecting 3 top finalists for recording a new TTS font. It was coupled effectively with objective analysis from TTS developers who ranked each voice using algorithmic measurements in a separate study, and expert opinion from audio designers experienced in TTS voice production.

4.2 Voice Talent Selection - From Top 3 to the Best

This study had two types of speech assets available for review: 1) utterances from a 1500 sentence recorded corpus of each candidate, and 2) a prototype unit selection based TTS font (font). A pair of surveys was created: Survey #1 compared the human voices, while Survey #2 compared samples generated from the fonts. The surveys were identical with the exception of whether the source was human or TTS. Each CS survey contained two listening sets, and each set consisted of one sentence with matched content spoken by each (human) voice or font (presented in varying order). Judges were asked to choose which voice they preferred for each set, choose an overall favorite and least favorite, and also to give reasons for their preference. 100 judges were polled for each survey. A few surveys came back with incomplete results. For spam detection, we checked two things: a) we reviewed all comments and b) we made sure no judges chose the same voice as both favorite and least favorite. Although unusual, we felt confident that there were no spam responses (judges' comments support this conclusion). A total of 95/100 and 93/100 complete responses were collected for each survey. We examined responses for sets 1 and 2 in relation to the choice for favorite overall and found strong correlation: the mean score for Set 1 + Set 2 is ≤ 5 of the overall favorite score. We found this correlation to be consistent enough that we used only the scores for favorite and least favorite for conclusions.

Table 6. Set preferences compared to overall favorite for both voice and font

Voice/TTS	Voice 1	Voice 2	Voice 3	TTS 1	TTS 2	TTS 3
Set 1	43%	46%	11%	22%	43%	35%
Set 2	42%	45%	13%	28%	58%	14%
Overall Fav.	46%	47%	7%	26%	49%	25%

Table 7. Crowd results: favorite and least favorite

	Voice 1	Voice 2	Voice 3
Favorite / human	46%	47%	7%
Favorite / TTS	26%	50%	24%
Least favorite / human	30%	28%	42%
Least favorite / TTS	33%	17%	50%

The first trend to note is that of least favorite: there is a clear least favorite in this study - Voice 3. Next, looking across the scores for favorite and least favorite TTS, it is evident that the font from Voice 1 was not as well-liked as the human Voice 1. In terms of picking a "best" voice, when both human and TTS versions of the voice are considered, the crowd chose Voice 2. Like the previous study, this survey was used as just one of three inputs for selecting a new TTS voice talent. Its input was combined with analysis from TTS developers and the insights of audio designers to make the final decision.

5 Conclusions

In this paper, we have presented results from three studies on evaluating human voice talents and synthesized speech. The focus of each study was quite different, but all three are unified by the use of CS. The first study shows the power of CS to query not one but many countries about voice gender preference. The second one shows how CS helped to uncover one voice quality (”mood”) which is more strongly carried over by TTS adaptation technology. The third study demonstrates 2 methodologies to effectively query general voice preference, depending on the size of the sample set. Considering the variety of focus across the 3 studies and the meaningful and relevant data uncovered by each, it is a reasonable conclusion that CS holds a wealth of potential feedback for developers of TTS voices and other applications of voice output.

Executing and reporting on a variety of CS experiment types helps us understand the strengths and weaknesses of CS apropos to research on human and synthesized speech. In this way, we can build reliable and repeatable experiment templates for use in crowdsourcing and tap the power of the crowd to hone and improve voice talent selection, steer gender or other country-dependent application decisions, and identify which areas of TTS technology innovation have the greatest impact with end-users.

References

1. Wolters, M., Isaac, K., Renalds, S.: Evaluating Speech Synthesis intelligibility using Amazon’s Mechanical Turk. In: Proc. 7th Speech Synthesis Workshop, SSW7 (2010)
2. King, S., Karaiskos, V.: The Blizzard Challenge 2012. In: Proc. Blizzard Challenge Workshop 2012, Portland, OR, USA (2012)
3. Lane, I., Waibel, A., Eck, M., Rottman, K.: Tools for Collecting Speech Corpora via Mechanical-Turk. In: Proc. of Creating Speech and Language Data with Amazon’s Mechanical Turk, pp. 185–187 (2010)
4. Snow, R., O’Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics (2008)
5. Marge, M., Banerjee, S., Rudnicky, A.: Using the Amazon Mechanical Turk for transcription of spoken language. In: Proc. IEEE-ICASSP (2010)
6. Parent, G., Eskenazi, M.: Speaking to the Crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges. In: Proc. of INTERSPEECH 2011, pp. 3037–3040 (2011)
7. Cooke, M., Barker, J., Lecumberri, M.: Crowdsourcing in Speech Perception. In: Eskenazi, M., Levow, G., Meng, H., Parent, G., Suendermann, D. (eds.) Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, pp. 137–172. Wiley, West Sussex (2013)
8. Lee, J., Nass, C., Brave, S.: Can computer-generated speech have gender?: an experimental test of gender stereotype. In: Proc. CHI EA 2000, CHI 2000 Extended Abstracts on Human Factors in Computing Systems, pp. 289–290. ACM, New York (2000)
9. Masuko, T., Tokuda, K., Kobayashi, T., Imai, S.: Voice characteristics conversion for HMM-based speech synthesis system. In: Proc. of ICASSP, pp. 1611–1614 (1997)