24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Item Response Theory to Estimate the Latent Ability of Speech Synthesizers

Chaina Santos Oliveira[1] and **Caio C. A. Tenório**[2] and **Ricardo B. C. Prudêncio**[3]

Abstract. Automatic speech recognition (ASR) systems have become popular in different applications (e.g., in virtual assistants). In order to ensure its robustness, ASR systems should be tested with diverse speech test data (i.e., with different sentences and speakers) in order to simulate different real scenarios of use. Relying on human speakers to test ASR systems is time-consuming and expensive. An alternative is to use text-to-speech (TTS) tools to synthesize audios from a set of sentences given as input. The ASR under test then receives the synthesized audios as test data, and the transcription errors are recorded for evaluation. Despite the availability of TTS tools, not all synthesized speeches have the same quality for all speakers and sentences. So before testing the ASR, it is important to evaluate the usefulness of the speakers as well as to determine which sentences are more relevant for ASR evaluation. In this paper, we propose the use of Item Response Theory (IRT) to evaluate speech synthesizers. IRT is a paradigm developed in psychometrics to estimate the cognitive ability of human respondents based on their responses to items with different levels of difficulty. In our context, an item is a sentence to be synthesized and a respondent is a speaker. In turn, each response is the transcription accuracy observed when a given sentence and speaker are adopted for testing the ASR. An IRT model identifies latent patterns of responses to estimate the difficulty of each sentence. In turn, the performance of each speaker is estimated by taking into account its responses to difficult sentences. In order to verify the viability of the proposal, a case study was developed, in which an ASR was used to transcribe voice commands (e.g., "set to vibrate mode") for a mobile virtual assistant. IRT was applied to evaluate 62 speakers (from four TTS tools) and to characterize the difficulty of 12 different sentences. We present interesting insights about the relevance of the synthesized audios for the ASR under test by inspecting the estimated parameters of difficulties and abilities. The knowledge acquired from such analysis can be applied to optimize the task of ASR testing.

1 INTRODUCTION

Automatic speech recognition (ASR) has been a subject of great interest in recent years, changing the interaction between machines and humans by converting speech into text. In fact, ASR systems can improve human-machine interaction in different applications like virtual assistants, smart houses, gaming and voice-based searching, also supporting communication in different native languages, accents and types of voices [18]. In order to ensure the robustness of an ASR system in a real application, its evaluation should consider as many types of user voices and scenarios as possible. ASR testing can be performed by using as test data a set of human speech audios, recorded from people from different places, genders and accents. However, producing such test data from human speakers usually costs a lot of money and time. So, an alternative is to adopt synthesized audios for testing ASR systems, which can be generated by text-to-speech (TTS) synthesis tools.

Current synthesis tools can produce audio records by considering different settings like genders, languages, accents and speakers. This diversity can be very convenient since one can produce test cases to cover different scenarios of the actual use of the ASR system. However, the quality of synthesis can vary a lot, depending for instance on the speech synthesis tool, the adopted speaker and the input sentence. In fact, some input sentences, for instance, can be difficult to be synthesized by any TTS tool, while others may be hard for some synthesizers due to specific reasons. As another example, a synthesizer can produce audio records with good quality for male speakers while presenting a bad quality for female speakers. Hence, it is important to evaluate the audio records produced by the speech synthesizers before actually using them to test the ASR system.

In this paper, we proposed a new framework which adopts Item Response Theory (IRT) for evaluating speech synthesis. IRT is a paradigm to measure latent skills of humans, widely adopted in educational and healthy applications. More recently, in Artificial Intelligence (AI), researchers have been adopting IRT as a new methodology of performance evaluation. A motivation to do so is that the conventional evaluation in AI considers that tasks have the same level of difficulty to be solved. It does not differentiate between hard and easy AI tasks when evaluating a technique. The application of IRT in AI domains can be useful to measure latent skills of AI techniques in the same sense that it has done to humans.

In [12] and [13], for instance, IRT was applied to evaluate classifiers and estimate the difficulty and discrimination of dataset instances. In that work, the respondents are the classifiers and the items are instances in a test dataset. They performed experiments with a set of datasets and diverse classification algorithms and investigated, for instance, how negative discrimination can be an indication of noisy instances. In [3], IRT is extended to learn ensembles of classifiers, in such a way that higher voting weights are assigned to classifiers which correctly predict hard instances. In [2], a model named β^3-IRT was adopted to fit probabilities returned by supervised machine learning models. In [10], IRT was applied to measure the performance of NLP methods. In the analysis, they showed that high accuracy scores do not always mean high ability scores. These previous works demonstrated the feasibility of adopting IRT to AI domains. It

[1] Universidade Federal de Pernambuco, Recife (PE), Brasil, email: cso2@cin.ufpe.br
[2] Motorola Mobility, email: caiocat@motorola.com
[3] Universidade Federal de Pernambuco, Recife (PE), Brasil, email: rbcp@cin.ufpe.br

looks promising as well as in the context of speech synthesis since each sentence may have a different level of difficulty to be synthesized, which must be taken into account for ASR evaluation.

An IRT model works as follows: initially, a set of items with different difficulty levels (e.g., questions, exams) are given to the respondents. Based on the observed responses, an Item Characteristic Curve (ICC) is fit to each item, predicting the probability of response to that item given the respondents' ability. An ICC is usually a monotonic parameterized function and its parameters characterize the item's difficulty and discrimination [6]. Abilities, in turn, are estimated in such a way that the higher is an individual's ability, the higher is the probability of a good response to a more difficult item.

In our context, the items are the sentences to be synthesized. In turn, the respondents are the speakers available in different speech synthesis services. Each response records the observed recognition quality, measured by comparing the original sentence and the one transcribed by the ASR system. IRT will estimate the difficulty of sentences and also identify which ones are more useful to discriminate between good and bad speech synthesizers. Ability in our context is a performance measure for evaluating speech synthesizers by taking the difficulty of the sentences into account.

As a case study, we evaluate the speech synthesizers adopted for testing an ASR system in a real mobile application. In the experiments, we consider 12 English sentences and 62 speakers adopted for testing the ASR system. For each input sentence, the pool of speakers was used to produce the corresponding audios. The produced audios were then given as input to the ASR system under test. The word accuracy rate ($WAcc$) measured for each sentence and speaker was recorded as a response. In our experiments, we adopted the β^3-IRT model proposed by [2] to estimate the abilities and item parameters. This model is adequate to treat responses in a continuous scale. The proposed framework provided interesting insights regarding, for instance, which sentences are more relevant to choose the best synthesizers as well as which speakers are more robust for ASR testing. To the best of our knowledge, the investigation of IRT methods to evaluate speech synthesis is original in the literature.

This paper is organised as follows. Section 2 provides a background on speech synthesis evaluation. Section 3, in turn, introduced IRT and presents the model adopted in our experiments. Section 4 details our proposal as well as the developed case study. Finally, Section 5 concludes the paper.

2 SPEECH SYNTHESIS EVALUATION

A speech recognition system can understand what someone has said and translate it into a machine-readable format, usually from speech to text [17, 11]. This type of communication between the user and the machine prevents other communication methods from being used, such as the keyboard, buttons, screen, or even gesture communication [4]. There are different techniques adopted in literature for speech recognition, including audio enhancing procedures. Also, the variety of applications and environments of use makes it mandatory to perform a robust evaluation of ASR systems.

In this paper, we focused on the use of speech synthesis for evaluating ASR systems, since this alternative is considerably inexpensive and the speeches can be produced in a more controlled way (for instance, with no noise). Some companies offer useful synthesis services like the Amazon Polly [1], the Google Text to Speech API [7], the IBM Watson Text to Speech [9] and the Microsoft Azure Text to Speech [14]. Each service generates speeches by adopting different *speakers*, each one associated to a different voice type, language, accent and genre. In this work, we focused on the English language. Table 1 shows the number of locales (English accents) and speakers that each service offers. Google Text to Speech provides the largest quantity of speakers, while IBM Watson provides the lowest amount.

Table 1. Synthesizer Services

Service	Number of Locales	Number of Female Speakers	Number of Male Speakers
Amazon Polly	5	10	6
Google TTS	3	13	13
IBM Watson TTS	2	3	1
Microsoft TTS	6	11	5

Due to the variety of speakers available in the different synthesis services, it is relevant to evaluate them before ASR testing. According to [16], the quality of speech synthesis can be measured by analyzing two main aspects: the intelligibility and the naturalness of the output. The first aspect refers to how understandable the speech is. The second aspect, in turn, considers how similar the synthesized speech is to the human voice. In this work, we focused on evaluating the intelligibility of speech synthesis.

The intelligibility of synthesized speeches can be evaluated by calculating the transcription error, which can be done by deploying humans or ASR systems. The measure commonly used to evaluate the intelligibility of speeches is the Word Error Rate (WER) [8, 15]. This metric is also used to evaluate ASR systems. It measures the transcription error rate of a given speech, comparing the original sentence with the transcribed one (the sentence understood by the recognizer), according to the Eq. 1:

$$WER = \frac{S + D + I}{N} \tag{1}$$

in which:

- S = number of substitutions;
- D = number of deletions;
- I = number of insertions;
- N = number of words words in the original sentence.

When evaluating the performance of ASR system, sometimes word accuracy rate ($WAcc$) is used instead of the WER, for convenience. $WAcc$ (Eq. 2) was the measure adopted in this work.

$$WAcc = 1 - WER \tag{2}$$

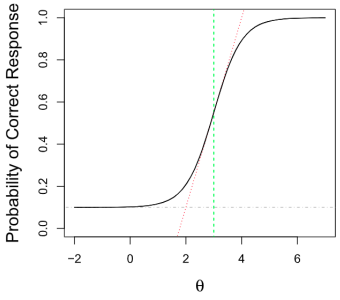
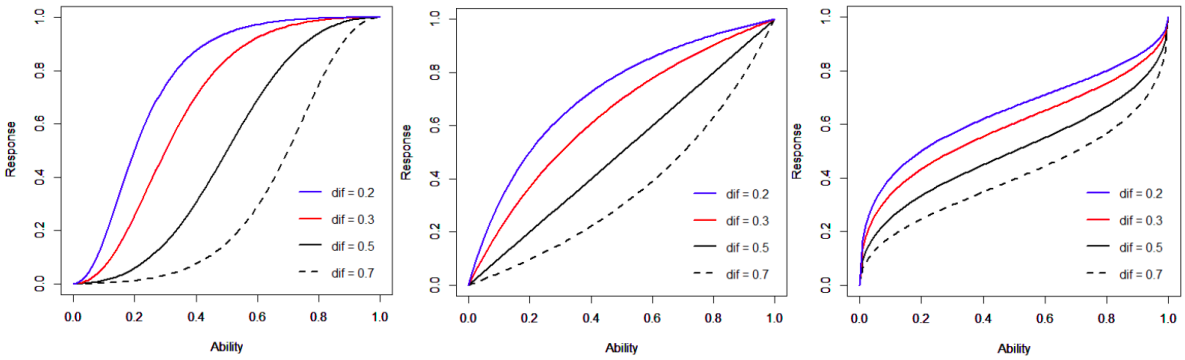


Figure 1. Example of item characteristic curve of a 3PL model [12]



(a) $\mathbf{a}_i = 2$. (b) $\mathbf{a}_i = 1$. (c) $\mathbf{a}_i = 0.5$.

Figure 2. Example of item characteristic curve of β^3 model with different values of difficulty and discrimination [12]

In [8], intelligibility was measured by comparing the use of either native or non-native speakers to transcribe the synthetic speeches. In [15], the authors compared the use of ASR systems against humans to evaluate intelligibility. They concluded that ASR systems could replace humans in this task, which in fact, saves resources. Hence, in this paper, we opted to use an ASR system to perform the transcription task and then evaluate the speech synthesis.

Measuring the quality of speech synthesis only relying on average WER or $WAcc$ has its limitations. In fact, like many other experimental procedures, there are random factors that may deviate the *observed* quality measure (in our case, the $WAcc$) from the *true* quality. For instance, the ASR performance and the difficulty of the sentences are factors that can influence the observed recognition errors. Fortunately, similar issues have been treated in psychometrics by the use of latent models, in which the responses are modeled as random variables conditioned by a latent trait. In our case, the observed recognition accuracy is modeled as random variables conditioned by the quality level of synthesis. To do that, we applied Item Response Theory models, introduced in the next section.

3 ITEM RESPONSE THEORY

Item Response Theory (IRT) is a methodology in psychometrics for measurement of latent skills [6, 5]. It is commonly used in educational testing, combining items and respondents' latent characteristics to predict observed responses. This paradigm assumes that the probability of a correct answer is associated with latent characterizations of individuals (e.g., ability) and items (e.g., discrimination and difficulty). In this section, we introduce the standard IRT model widely adopted in literature for dealing with binary responses (correct or incorrect responses). Followed, we present a more recent IRT model, adequate for dealing with continuous responses.

3.1 Dichotomous IRT Model

There is a large number of IRT models, which can vary depending on the kind of response. The most common IRT model is dichotomous, where the response to an item is either correct or incorrect, i.e., an observed response is defined with value 1 if the respondent correctly answered an item, and defined as 0 otherwise. In the 3-parameter logistic (3PL) IRT model, the probability of a correct response is a logistic function of the latent individual's ability and some item's parameters (difficulty, discrimination and guessing). The probability function for an item across the ability scale is called the Item Characteristic Curve (ICC) [6, 5].

Formally let θ_j the (unknown) ability of a respondent *j*. The 3PL ICC for a given item *i* can be modelled as the logistic function:

$$P_{ij}(r_{ij} = 1|\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - \delta_i)}} \tag{3}$$

in which:

- r_{ij} is the response of respondent *j* to item *i*;
- δ_i is the item difficulty (the location parameter of the ICC);
- a_i is the item discrimination (the slope of the ICC);
- c_i is the guessing parameter (the asymptotic minimum of the ICC).

An example of ICC is shown in Figure 1, with discrimination ($a_i = 2$), difficulty ($\delta_i = 3$) and guessing ($c_i = 0.1$). In the 2-PL IRT model, the guessing parameter c_1 is assumed as zero for simplicity. In practice, abilities and item parameters can be estimated via maximum likelihood in such a way that the ICCs fit the observed responses for a pool of respondents and items.

The ability of a respondent is not directly associated with the number of questions it responds correctly. In other words, it is not always true that the respondents who get more right answers have higher ability. The performance of a respondent is associated with the right responses it gives to difficult items. On the other hand, difficult items are the ones that are better answered by the most skilled respondents. In turn, discrimination means how the probability of a correct response changes if the ability increases. For higher values of discrimination, a small change in the ability can result in a big change in the probability of a correct answer.

Dichotomous models are adequate to be adopted when the responses are binary (i.e., a response of a multiple-choice question which is either correct or incorrect). In the current work, the responses are not binary, as they are transcription accuracy rates, measured in the interval $[0, 1]$. So, we adopted the β^3-IRT model [2], which is appropriate to deal with responses measured in a bounded continuous scale.

3.2 β^3 IRT Model

The β^3-IRT model was proposed by [2], originally to deal with continuous responses in any bounded interval. In that work, it was applied to a psychometric task of students from an online platform.

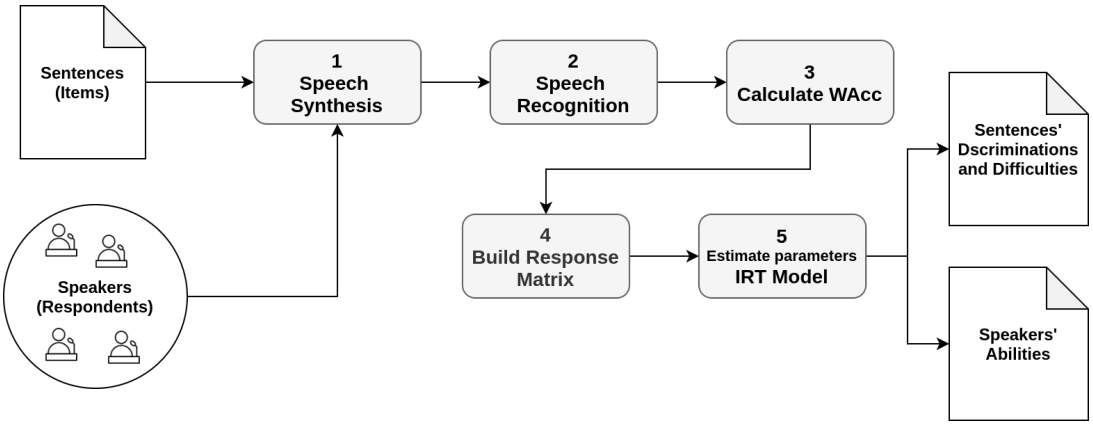


Figure 3. Overview of the Speech Synthesis Evaluation with IRT

The data was composed of answers given by students to questions. In this case, it is a traditional application of IRT. In the same work, this model was also applied to evaluate the ability of machine learning classifiers, where a respondent is a classifier and an item is an instance. In turn, a response is the probability of a classifier hits the instance class.

The β^3-IRT model can generate abilities and difficulties in the $[0, 1]$ range, which is more easily interpreted than standard logistic IRT models in which abilities and difficulties have infinite support (Figure 1). In the β^3-IRT, continuous responses are modelled by the Eq. 4:

$$P[r_{ij}|\theta_j, \delta_i, a_i] = \frac{1}{1 + (\frac{\delta_i}{1-\delta_i})^{a_i}(\frac{\theta_j}{1-\theta_j})^{-a_i}} \tag{4}$$

in which:

- $r_{ij} \in [0, 1]$ is the response of respondent j to item i;
- δ_i is the difficulty of the item i;
- θ_j is the ability of the respondent j;
- a_i is the discrimination of the item i.

Figure 2 shows examples of β^3-ICCs with different values of difficulty and discrimination. The ability values adopted were 2, 1 and 0.5, respectively. When $a_i > 1$ the curve has a sigmoid shape similar to traditional logistic IRT models. When $a_i = 1$, it assumes a parabolic curve. In turn, when the discrimination is a number between 0 and 1, the curve has anti-sigmoidal behavior. This flexibility is convenient for a better fit of continuous responses.

4 IRT to SPEECH SYNTHESIS EVALUATION

In this paper, we adopt IRT in the speech synthesis context, in which the items are sentences and the respondents are synthetic speakers. We aim to estimate how hard is a given sentence to be synthesized with good quality, allowing it to be transcribed with a low error rate. Besides that, we can also measure the ability level of the respondents. By incorporating IRT, we can learn more about the characteristics of the sentences, considering that not all of them have the same level of difficulty to be synthesized.

4.1 Methodology

As said before, the items in this work are sentences adopted to test an ASR system in a real environment of testing. The sentences adopted in the experiments are presented in Table 2. These 12 sentences have been adopted in a real industrial scenario to test the ASR system of Motorola mobile devices. We adopted these sentences in our experiments to reproduce the same testing settings used in this industrial scenario. The respondents are the English speakers that can be chosen when synthesizing audios. A total number of 62 speakers were adopted, considering the four synthesis services presented in Table 1. Thus in our work, we estimate the ability of the synthesis services for each speaker to produce audio test files that can be well recognized by the ASR system.

Table 2. Sentences used in the experiment

Id	Sentence (item)
1	Do i have any friends nearby?
2	Do i need an umbrella?
3	Do not disturb!
4	Find me a popular restaurant near me
5	How many calories are in 3 eggs 2 slices of bacon and 1 slice of pizza?
6	Set to vibrate mode!
7	Tell john im on my way
8	Turn projector on
9	Watch stranger things
10	What is stock price of Apple, Google and Microsoft?
11	What's my day look like?
12	What's playing?

In order to create the response matrix (the input for the IRT algorithm), we followed steps 1 to 4 showed in Figure 3:

- Initially, given a sentence i and a speaker j, we generated the corresponding audio (step 1);
- Following, the synthesized audio is given as input to the ASR system (step 2);
- After that, the $WAcc$ measure is computed (step 3) and stored as the response r_{ij} (step 4). Then the response indicates the success of the speech recognition process for item i and speaker j (the higher is the response, the better is the speaker for that item).

We perform the steps above for all 12 sentences synthesized and 62 speakers from all service. As a result, a $12X62$ matrix of responses is built, which will be given as input to the β^3-IRT model

(step 5 in Figure 3). The responses in the input matrix are values between 0 and 1, where 0 means the worst response and 1, in turn, means the best response (i.e., no recognition errors were observed). For each sentence, the β^3-IRT model builds a characteristic curve which returns the expected response for each value of speakers' ability, which also varies in a [0, 1] range. The estimated difficulty and discrimination parameters, as well as the ability values, will be discussed in the next sections.

4.2 Analysis of the Items' Parameters

After applying the β^3-IRT model, 12 ICCs were produced, one for each sentence. Table 3 presents the difficulty and discrimination values as well as the average $WAcc$ obtained across the speakers for each sentence. Figure 4 presents the ICCs with positive discrimination parameter, while Figure 5 presents the ICCs with negative discriminations. The X-axis indicates the speakers' ability, while the Y-axis represents the probability of a response. In the figure, each star represents a speaker's response, colored by the service the speaker belongs to (see Table 4). The ICC's parameters will be discussed in the next subsections.

Table 3. Items Parameters

Item	Difficulty	Discrimination	Avg. $WAcc$
1	0.98	-0.85	0.98
2	0.32	1.04	0.99
3	0.56	1.55	0.79
4	0.35	1.38	0.85
5	0.34	1.06	0.92
6	0.36	1.29	0.88
7	0.56	1.31	0.79
8	0.98	-0.58	0.88
9	0.47	1.19	0.78
10	0.48	1.13	0.84
11	0.36	1.24	0.93
12	0.36	1.24	0.85

Table 4. Colors that represents each service in the figures

Color	Service
★	Amazon Polly
★	Google Text to Speech
★	IBM Watson Text to Speech
★	Microsoft Azure Text to Speech

4.2.1 Discrimination

In our experiments, from 12 sentences, 10 had ICCs with positive slopes (Figure 4), meaning that the speaker's abilities are positively related to the response probability. Sentence 3 presented the highest discrimination value among all sentences, which means that this sentence had more power to discriminate between good and bad speakers at a certain ability level. In fact, we can observe in its ICC in Figure 4 that most of the speakers with ability lower than the 0.55 presented a very low response value (very close to zero). In turn, the speakers with higher abilities had a very high response value (close or

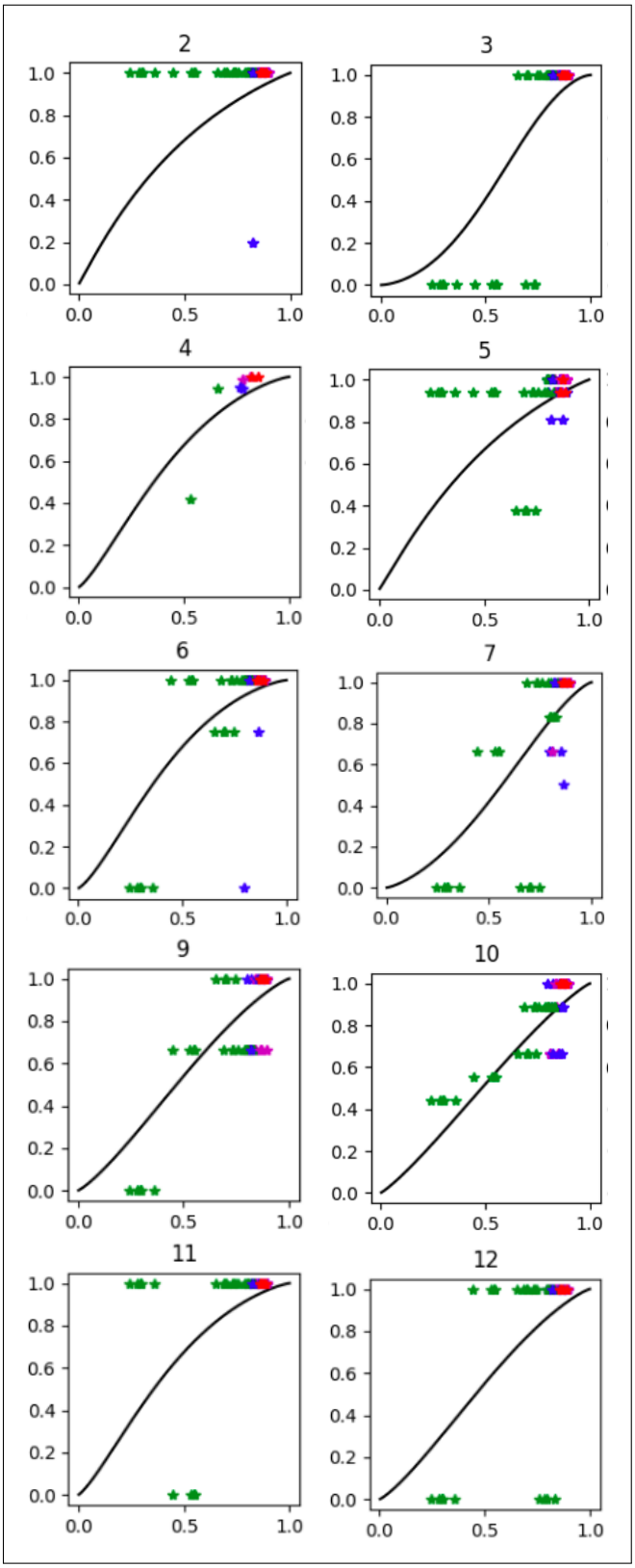
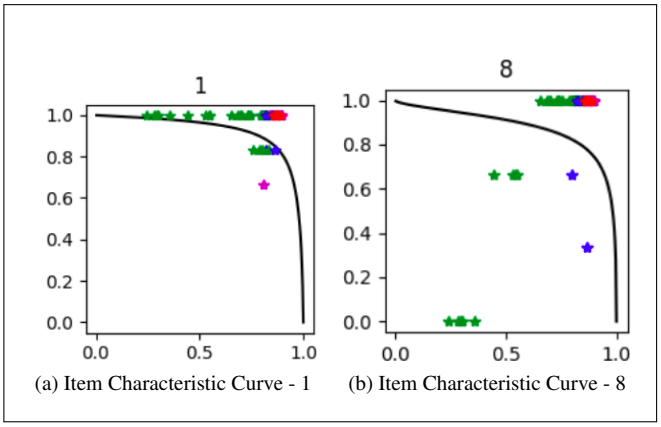


Figure 4. Examples of ICCs with positive discrimination

equal to 1). There were uncertain responses when ability approaches 0.55. The estimated response probability is 0.5 in this case.

When applying IRT to evaluate speech synthesis, the discrimination parameter can be used to measure the capability of a sentence to differentiate between groups of speakers with different abilities. For instance, sentence 9 is useful to discriminate speakers below the ability equals difficulty 0.47. Relying on discriminating items in different ability levels can be useful to perform iterative testing, which is actually widely performed in educational applications. This procedure will be adapted in our context in future work.

Two sentences had negative discrimination: the sentences 1 and 8, with ICCs presented in Figure 5. Negative discriminations in IRT usually indicate unexpected behaviors. In our context, some synthesizers with good ability do not synthesize these sentences well. This is an unstable behavior that should be investigated with attention before using such sentences for ASR testing. If a good synthesizer is chosen for ASR testing (which is reasonable or course), adopting sentences with negative discrimination may result in underestimated ASR performance.



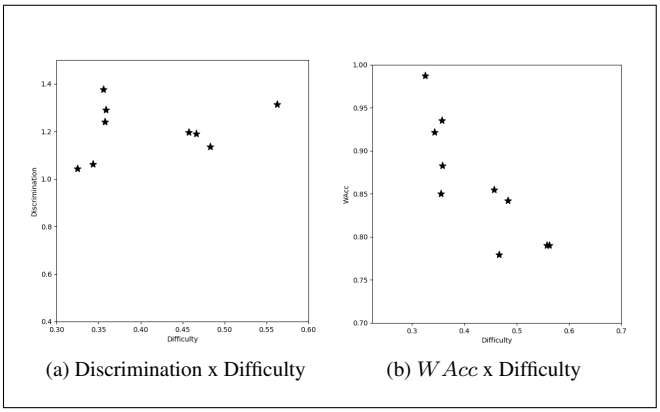
(a) Item Characteristic Curve - 1 (b) Item Characteristic Curve - 8

Figure 5. Examples of ICCs with negative discrimination

4.2.2 *Difficulty*

The difficulty parameter in our context indicates how difficult a sentence is to be synthesized in quite good quality, comparing with the other sentences in the experiment. This parameter can vary from 0 to 1. Considering only the sentences with positive discrimination, the most difficult item is the sentence 7 with difficulty 0.562 with a relatively high discrimination value (see Table 3). As it can be seen in Figure 4, there is actually a high variance in response, as the ability increases. Sentence 2, in turn, was the easiest item with difficulty 0.324. In fact, apart from a single speaker, almost all speakers had a response close to 1 for that sentence.

Figure 6 plots the relation between difficulty, discrimination and average $WAcc$ across speakers for each sentence. In this plot, we discard the sentences with negative discrimination (1 and 8), as they can distort the interpretation of difficulty. By considering the items with positive discrimination, we observe a strong correlation (-0.84) between difficulty and $WAcc$. This is expected since better responses are expected to be observed for less difficult items. Sentence 2, for instance, is the easiest one considering both the difficulty parameter



(a) Discrimination x Difficulty (b) $WAcc$ x Difficulty

Figure 6. Relation between difficulty with discrimination and $WAcc$

and $WAcc$. This correlation shows that the latent difficulties estimated by the IRT model are consistent with the observed responses.

By comparing difficulty and discrimination, in turn, we also observe a strong correlation (0.51), but not so high as compared to $WAcc$. In fact, difficult items tend to be more discriminating, especially in the higher levels of ability. However, difficulty and discrimination bring different information.

4.3 Analysis of the Abilities

Figure 7 plots ability against the average $WAcc$ obtained for each speaker across sentences. In other words, this plot compares the speakers' performance considering both their latent trait and the observed behavior. Both measures are consistent, as the correlation between average $WAcc$ and ability is strong (0.9). As expected, the higher is the ability of a speech synthesizer system, the better are its observed responses.

Another observation is that speakers with the most inferior abilities belong to the Google service (green stars), taking into account the other synthesizer tools and also the adopted sentences. Most of these speakers have a female voice. From the 20 speakers with the worst performance, 13 have a female voice (and all of them are from Google). In contrast, the best-performing speakers from that service have a male voice.

We highlight that IRT evaluates synthesized speeches according to the ones that are in the pool (under evaluation). It has some advantages when considering the context of using synthetic voices in ASR software testing. Depending on the stability of the software, the tester engineering can choose a specific group of speakers or sentences to be used in the test campaign, considering their respective estimated parameters (e.g., ability, discrimination and difficulty).

5 CONCLUSION

As ASR systems have become largely usable, it is important to ensure its robustness by exploring diverse scenarios and voice variations during the testing process. Using synthesized speeches instead of human voices in this task looks promising. However, before giving a synthetic speech as input to test an ASR system, it is important to ensure its quality. So, in this work, we proposed a new approach in order to evaluate its performance. Traditional methods of speech

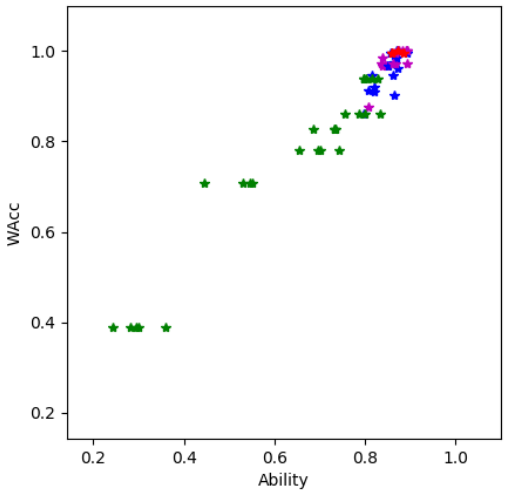


Figure 7. $WAcc$ X Ability

synthesis evaluation do not consider some factors that may influence the synthesis quality as sentence difficulty. In this work, we use a method of evaluation that takes this into account. We have introduced Item Response Theory from psychometrics as an alternative method to evaluate speech synthesis performance.

In the analysis, we showed that IRT is promising when applied in the speech synthesis evaluation context. It can identify sentences with different levels of difficulty and discrimination power between good and poor synthetic speakers. This information can be considered during an ASR system testing process optimizing this task.

Synthesized speeches from speakers with bad abilities can be eliminated from the ASR system test campaign, for example. And also, depending on the software version, we can use sentences with different levels of difficulty. If the software is in the first stages of development, it can be tested with sentences with low difficulty. On the other hand, if the ASR system is stable, difficult sentences can be used to test it.

Future work can extend the analysis adopting a large number of sentences to be investigated, with different sizes and a variety of words. More than one ASR system can be included in the experiments in order to make $WAcc$ rate more reliable and robust. A similar methodology (with IRT) can also be applied to evaluate ASR systems instead of synthesizers. Other possible work is comparing the performance of synthesized speech with human speech, analyzing the difference between the abilities of synthetic and real speakers.

ACKNOWLEDGEMENTS

This work was supported by CAPES, CNPq and FACEPE (Brazilian funding agencies) and Motorola Mobility. We thank the Research and Voice Test Automation teams (in special to Isadora Soares, Joao Cordeiro and Marlom Oliveira) from CIn/Motorola Project for all support and feedback given.

REFERENCES

[1] Amazon Web Services AWS, 'Amazon polly', (2019). Access in: 25/09/2019.
[2] Yu Chen, Telmo Silva Filho, Ricardo B. C. Prudêncio, Tom Diethe, and Peter Flach, 'β^3-irt: A new item response model and its applications', in *Proceedings of Machine Learning Research*, volume 89, pp. 1013–1021, (2019).
[3] Ziheng Chen and Hongshik Ahn, 'Item response theory based ensemble in machine learning', *arXiv:1911.04616*, (2019).
[4] Hubert Crepy, Jeffrey A. Kusnitz, and Burn Lewis, 'Testing speech recognition systems using test data generated by text-to-speech conversion', number US6622121B1, (2003).
[5] Rafael Jaime De Ayala, *The theory and practice of item response theory*, Guilford Publications, 2013.
[6] Susan E. Embretson. and Steven P. Reise, *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Inc., 2000.
[7] Google, 'Cloud text-to-speech', (2019). Access in: 25/09/2019.
[8] CT Justine Hui, Sahil Jain, and Catherine I Watson, 'Effects of sentence structure and word complexity on intelligibility in machine-to-human communications', *Computer Speech & Language*, **58**, 203–215, (2019).
[9] IBM, 'Watson text to speech', (2019). Access in: 25/09/2019.
[10] John P Lalor, Hao Wu, and Hong Yu, 'Building an evaluation scale using item response theory', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 648. NIH Public Access, (2016).
[11] Navin Kumar Manaswi, in *Deep Learning with Applications Using Python*. Apress, (2018).
[12] Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Making sense of item response theory in machine learning', in *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1140–1148. IOS Press, (2016).
[13] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Item response theory in ai: Analysing machine learning classifiers at the instance level', *Artificial Intelligence*, **271**, 18–42, (2019).
[14] Microsoft, 'Azure text to speech', (2019). Access in: 25/09/2019.
[15] Eli Pincus, Kallirroi Georgila, and David Traum, 'Which synthetic voice should i choose for an evocative task?', in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 105–113, (2015).
[16] Youcef Tabet and Mohamed Boughazi, 'Speech synthesis techniques. a survey', *7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA)*, 67–70, (2011).
[17] C. Vimala and Radha, 'A review on speech recognition challenges and approaches', in *World of Computer Science and Information Technology Journal (WCSIT)*, pp. 1–7, (2012).
[18] Dong Yu and Li Deng, *Automatic Speech Recognition*, Springer, 2016.