Item Response Theory to Evaluate Speech Synthesis: Beyond Synthetic Speech Difficulty

Chaina Oliveira[1], Ricardo Prudêncio[1]

[1]*Universidade Federal de Pernambuco, 1235 Prof. Moraes Rego, Recife, Brazil*

Abstract

Artificial Intelligence (AI) systems have been increasingly developed and improved. In this sense, one of the main challenges is to evaluate and compare them. However, traditional assessment methods do consider some hidden factors that may influence the quality of these systems that can be helpful in their discrimination (e.g., between poor and good techniques). Previously, we developed a work that uses Item Response Theory (IRT) to simultaneously evaluate speech synthesis and recognition. IRT is a paradigm from psychometrics to estimate the cognitive ability of human respondents based on their responses to items with different levels of difficulty. One of the measures we estimated in that previous work was the synthesized speeches' difficulties, in turn, the factors that influence that measure were not deeply explored. So, in this paper, we navigate far on this topic and investigate what explains a synthesized speech difficulty. We found out that some of the factors that may influence are: the sentence, the locale and the service used to generate the speech. Also, we performed a preliminary study to investigate the viability of predicting the synthesized difficulty using machine learning models. So, we trained some regression models using the speech synthesis parameters as features and the difficulty as the label. The best result was achieved using a Random Forest, in which we got 0.31 as normalized R2 score.

Keywords

Item Response Theory, Speech Synthesis Evaluation, Synthesized Speech Difficulty, Speech Quality Measurement

1. Introduction

Progress in speech synthesis and recognition research changed the way we communicate and interact with machines. These techniques can be used as a communication way in diverse applications. It is common to see mobile users who opt for using command voices instead of the device's keyboard to execute some task (e.g., call someone, do a google search, write an e-mail). Those kinds of systems have been developed and improved more and more, but we have not seen many advances in how to evaluate them. In a previous paper, we proposed Item Response Theory (IRT) from psychometrics to evaluate speech synthesis [1] and in other, we assessed speech synthesis and recognition [2].

IRT is commonly used in educational testing to estimate the latent ability of respondents and the difficulty of items. Recently, this methodology of evaluation has been adopted in other contexts, including in the evaluation of AI systems. In supervised learning, IRT was explored by [3], [4] and [5] to evaluate the ability of classifiers based in their answers to a set of instances (what class each instance belongs to). [4] and [5] investigated the importance of analyzing the particular problems in which good techniques fail (e.g., a classifier with good performance does not hit an instance class that a poor one does). For instance, they clarified that it is unfair to evaluate classifiers using just the number of instances they hit, it is also important to analyze the difficulty of instances classified by the models under test. Furthermore, IRT was also adopted to evaluate regression models abilities in [6].

A more recent way of estimating IRT difficulties was proposed by [7]. The authors suggested that we could predict the difficulty of new items using a regression model trained with the problem features, using the difficulty as target. They trained a regression model for a set of domains (i.e., Supervised Learning, Audio Processing, Computer Vision and so on) and the results showed that using this methodology in that context is promising.

Recently, we developed a work that adopted IRT evaluate speech synthesis and speech recognition [2], which its main goal was to estimate the latent ability of Automatic Speech Recognition systems, the quality of speakers and the difficulty of synthesized speeches and sentences. So, firstly, we extracted 100 benchmark sentences from VoxForge [8] and synthesized them using English voices from four services using different variation of pitch and rate. It resulted in a set of synthesized audios that were given as input to four ASR systems to be transcribed. After this, we calculated the accuracy of all transcriptions using the word accuracy rate ($WAcc$). The $WAcc$ become the input to our IRT model (i.e., the responses). To estimate the IRT parameters (e.g., synthesized speech difficulties), we adopted the β^3-IRT model proposed by [3].

In this paper, we present a deep analysis of the predicted synthesized speeches' difficulties estimated in [2]

EBeM'22: Workshop on AI Evaluation Beyond Metrics, July 25, 2022, Vienna, Austria

cso2@cin.ufpe.br (C. Oliveira); rbcp@cin.ufpe.br (R. Prudêncio)
0000-0002-4891-1802 (C. Oliveira); 0000-0001-7084-1233 (R. Prudêncio)

in order to understand if they can be explained by the sentences or the synthesis parameters used to generate the speeches. So, we deeply analysed the data produced by these previous work and found that the synthesized speech difficulty can be affected by the sentence and some speech synthesis parameters (e.g., speaker, locale, pitch, rate and gender). We also aimed to know if we could use any regression model to predict the IRT difficulty in this context. So, we trained MLP, Linear Regression and Random Forest models using the synthesis parameters as features and the difficulty as the label. The Random Forest outperformed the others, getting 0.50 as normalized MAE and 0.31 as normalized R2.

The proposal of this paper fits with the AI Evaluation Beyond Metrics workshop's goal once both aim to investigate and give visibility to new robust approaches to AI systems assessment. As the workshop's goal, we desire to explore new assessment methods to try to cover some limitations of the traditional ones. The approach of evaluation used in this work (i.e., IRT) has been already adopted to evaluate other kinds of AI systems such as classifiers, NLP systems, and so on. Here, we explore the analysis of using IRT in a new context - to evaluate speech synthesis and recognition.

2. Item Response Theory

IRT is a methodology from psychometrics that aims to estimate the latent abilities of respondents in tests [9]. It models the responses to testing items based on their difficulties and the skills of the respondents who answered them. This section presents a classical IRT model (i.e., the binary) and a more recent model (i.e., β^3-IRT). This last one was the one we adopted in this work.

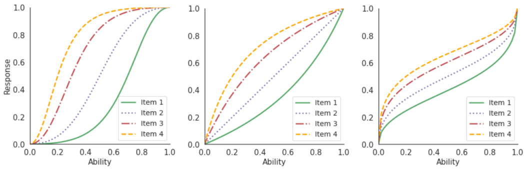
2.1. Binary Model

The binary model, also known as dichotomous, is usually used when a response to an item is positive or negative. In this category, we have the the 3-parameter (3PL) IRT model and the 2-parameter (2PL) IRT model. In 3PL, the probability of a correct response is defined by a logistic function of the respondent ability and the item's difficulties, discrimination and guessing. This model returns the Item Characteristic Curve (ICC), which is modeled according to the function below:

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

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).
- θ_j is the ability of respondent j.



(a) $a_i = 2$ (b) $a_i = 1$ (c) $a_i = 0.5$

Figure 1: Item characteristic curves of β^3 model with different values of difficulty and discrimination. Items 1, 2, 3 and 4 have difficulty of 0.2, 0.3, 0.5 and 0.7, respectively [3].

It is important to emphasize that when using IRT, different from traditional evaluation methods, the respondent's ability is not necessarily estimated only by the number of questions he answers correctly. It depends on the number of difficult items he hits. Similarly, the difficulty of an item is measured by the number of respondents who answer it correctly. In other words, to estimate these parameters, we consider the sets of items and respondents under analysis.

2.2. β^3-IRT Model

The binary IRT model is applied when the response can be correct or incorrect. In turn we have this more recent model that deals with continuous responses, the β^3-IRT [3]. The authors of β^3-IRT applied it in two contexts. The first one was to estimate the responses given by students to items, a typical application of IRT. The second application was in supervised machine learning, in which classifiers and instances were respondents and items, respectively. In turn, the responses were the probability of the classifiers assigning the correct class to an instance. The expectation of the correct responses can be calculated by:

$$E[r_{ij}|\theta_j, \delta_i, a_i] = \frac{1}{1 + (\frac{\delta_i}{1-\delta_i})^{a_i}(\frac{\theta_j}{1-\theta_j})^{-a_i}} \tag{2}$$

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.

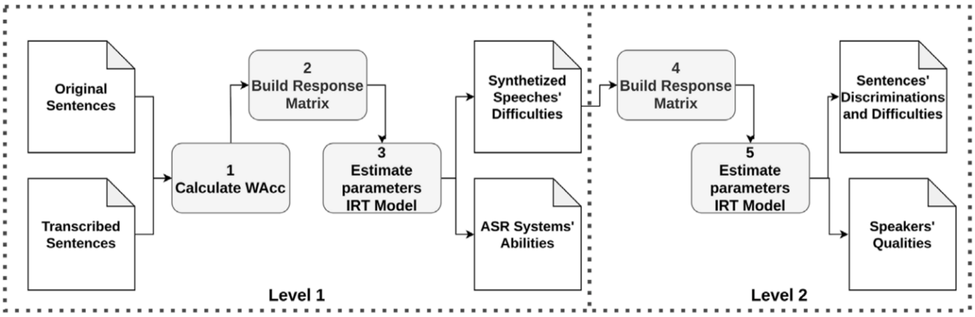


Figure 2: Speech Synthesis and Recognition Evaluation Using IRT [2].

Some ICCs that can be modeled by the Eq. 1 is shown in Figure 1. Each plot shows the curve with different values of difficulty and discrimination. When $a_i = 2$, the curve assumes a sigmoidal shape. If the discrimination is 1, the curve is parabolic, but if that parameter is between 0 and 1, the ICC assumes an anti-sigmoidal behavior.

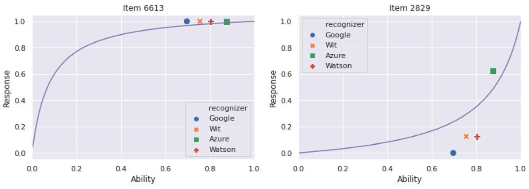


Figure 3: Examples of ICCs of two Synthesized Speeches [2]. Each mark represents the Response ($WAcc$) from each recognizer with a certain ability to that synthesized speech. The recognizers ability were estimated in our previous work.

3. IRT to Evaluate Speech Synthesis

In a previous work ([2]), we developed a two-level IRT model to evaluate speech synthesis and recognition. This model is illustrated on Figure 2. In the first level, an item is a synthesized speech produced from a given sentence and a speaker. In turn, the respondent is an ASR system. Each response is the transcription accuracy observed when a synthesized speech is adopted as an input the ASR system (i.e., $WAcc$). An IRT model identifies latent patterns of responses to estimate the difficulty of each synthesized speech and the ability of each ASR system. In the second first level, the synthesized speech's difficulty is decomposed into two latent factors: the sentence's difficulty and the speaker's quality. In this current work, we focus on the first level. Our main goal is to find characteristics that may influence the estimated synthesized speech's difficulty. So, in this paper, we focus on analyzing and using data generated and estimated on Level 1 presented in [2].

Figure 3 shows two ICCs of synthesized speeches with low and high difficulty, respectively. In the first one (i.e., 6613), all Automatic Speech Recognition (ASRs) systems got a high response value to that item. However, almost all (3 of 4) ASRs got a low response value for the most difficult item (i.e., 2829).

A variety of sentences, speakers and automatic speech recognizers were used by [2] as presented below:

- Sentences: The sentences were extracted from VoxForge [8], an open speech dataset. A total of 100 English sentences of different sizes were adopted. Figure 4 shows the distribution of those sentences size (number of characters) with median of 51.5. The shortest sentence has 12 characters, and the biggest has 134.
- Speakers: The speakers are from four different services: Amazon Polly [10], Google Text to Speech API [11], IBM Watson Text to Speech [12] and Microsoft Azure Text to Speech [13]. Each service has speakers with different English accents, genders, pitches and rates.
- Automatic Speaker Recognizers: The recognizers adopted in this work were: Google Speech to Text [14], Microsoft Azure Speech to Text [15], IBM Watson Speech to Text [16] and Wit [17]. They were responsible for receiving a synthesized speech and generating a transcription (the sentence the recognizer understood) of the referred audio.

In [2], a total of 15,000 synthesized speeches were produced. Each one was generated from a single sentence and a speaker setting. The IRT model estimated the difficulty of each speech, with distribution presented in Figure 5. The difficulty lies between 0 and 0.9. The majority part of the speeches has difficulty between 0.2 and 0.6. Also, we do not see a representative peak. It means

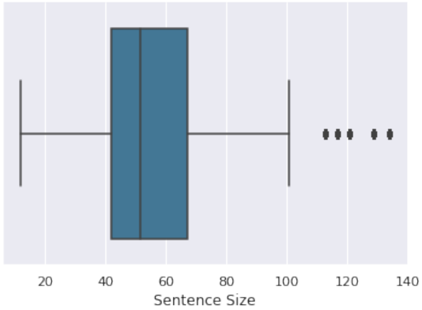


Figure 4: Sentence size distribution in terms of number of characters.

that there is not a specific difficulty value shared by a big part of synthesized speeches.

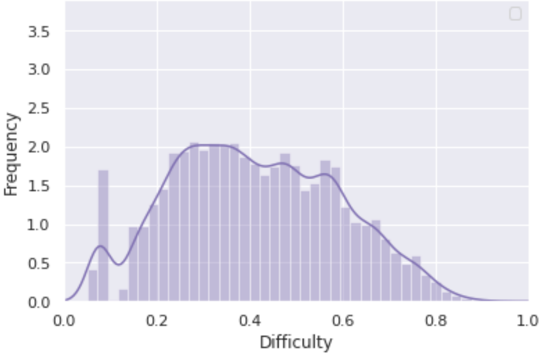


Figure 5: Distribution of the synthesized speeches' difficulties with pitch and rate variation [2]. The difficulty can vary from 0 to 1.

4. Experiments and Results

The IRT model provided in [2] estimated the difficulty value of each speech, but the aspects that impacted the difficulty across speeches were not deeply investigated. In this paper, we deeply explored the synthesized speeches' difficulty inferred, aiming to observing its relation to speech synthesis parameters and sentence features. For instance, may the length of a sentence influence the difficulty? Are bigger sentences easier or more difficult to synthesize than short ones? Is gender somehow related to difficulty? So, in Section 4.1, we explore the relationship between specific synthesis parameters and the difficulty. We show the difficulty distribution among the groups of each feature and also performed statistical tests to see the significant differences between them. For instance, we present the difficulty distribution of each gender and performed the statistical test among the difficulty values of male and female speeches. We also aimed to know if we could predict the synthesized speech difficulty. Thus, in Section 4.2, we present insights and results of a preliminary predictive model we developed to predict the difficulty, using the synthesis parameters as predictor attributes and the difficulty as the target attribute.

4.1. How Synthesis Parameters Influence the Difficulty of a Synthetic Speech?

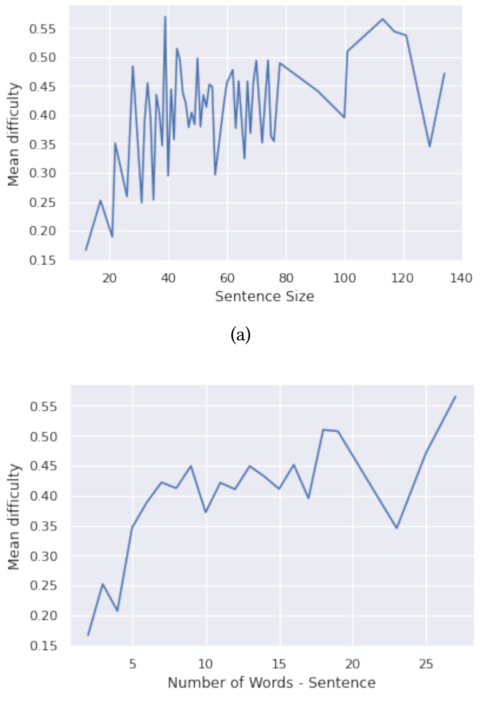
Initially, we aimed to understand if the size of the sentences has any relation to the difficulty. We noticed that the bigger the sentence's size or the number of words, it tends to be more difficult, as seen in Figure 6. We inspected some cases and found out that, depending on the parameters used to synthesize the speeches, they are not fully transcribed by some recognizers. It directly affects the $WAcc$, the response used as input to the IRT model. Table 1 shows examples of transcriptions of two of the longest sentences of our dataset. See that just a part of them was transcribed. It is impacting on the mean difficulty of those sentences.

Two of the speech parameters we explored were pitch and rate. We generated speeches with three different pitch values (e.g., low, medium and high). Figure 7 shows the distribution of the synthesized speech difficulty for each pitch group. Each box represents 50% of the difficulty values of the respective group. In turn, the lower and upper whiskers represents the difficulties outside the box. It also indicates the variability of the data outside the lower and upper quantiles (i.e., the lower and upper box lines). The line that divides each box into two parts is the median. It means that a half the difficulty values are greater than or equal to that value, and half are less. For instance, Figure 7 shows that speeches with low or medium pitch tend to be easier than the ones with high pitch, once the difficulty of 50% of the synthesized speeches with high, medium and low pitch is 0.42, 0.38 and 0.37 (the median of each group), respectively. It is also possible to see that speeches with high pitch are the ones that tend to be more difficult whilst the ones with a low pitch are the easiest. Regarding the rate (Figure 8), we noticed that speeches with a fast rate tend to be more difficult. In turn, the ones with medium pitch are the easiest.

As we used four services to synthesize the speeches, we aimed to investigate if speeches from a specific synthesizer are more difficult than the ones generated by the others, and we confirmed that as shown in Figure 9. The speeches from Azure are the most difficult. In turn, the ones from Watson are the easiest. In the middle, we have Google and Polly, with this last one tending to generate easier synthesized speeches than the service from Google.

Table 1
Examples of sentences and their transcriptions

Original Sentence	Examples of Transcriptions
It was really nice talking to you this week. I hope I could provide you with information sufficient for making the right decision.	It was really nice information sufficient.
	It was really nice talking to you this week.
Hope you all are doing fine. I was on jury duty three days last week, really interesting, but totally screwed up my schedule at work.	I'm doing fine.
	Doing fine last week really interesting my schedule at work.



(a)

(b)

Figure 6: (a) Mean difficulty of the synthesized speech per sentence size (number of characters). (b) Mean difficulty of the synthesized speech per number of words of the sentences.

The relation between gender and locale (i.e., type of English) with difficulty was also analyzed. Figures 11 and 10 show the synthesized speech difficulty distribution by gender and by locale, respectively. Following, Figure 12 shows the mean difficulty of each gender by locale. We see that female voices are more difficult than male ones. Regarding the English type, synthesized speeches with English from the United States are the easiest ones. In turn, speeches from Australian English are the most difficult, followed by British English and Indian English, respectively. Furthermore, we can see that female voices

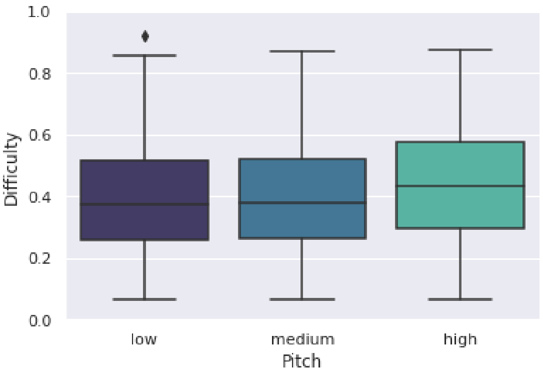


Figure 7: Distribution of synthesized speech difficulty for each pitch category.

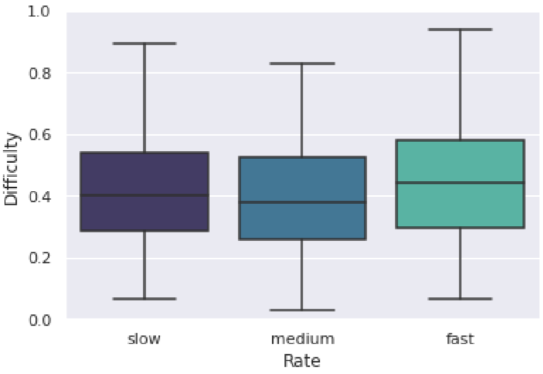


Figure 8: Distribution of synthesized speech difficulty for each rate category.

are more difficult than male voices in all locales (except for Australian English that there is not male voices in our database to compare).

We performed ANOVA statistical test among the groups of each feature shown in this Section's plots (Figures 7 - 11) o see the significant differences between them. The p-value obtained from the analysis in all cases was significant ($p < 0.01$). So we conclude that there are significant differences among them.

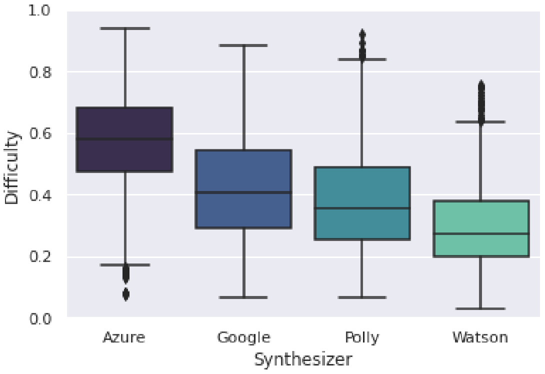


Figure 9: Distribution of synthesized speech difficulty for each synthesizer.

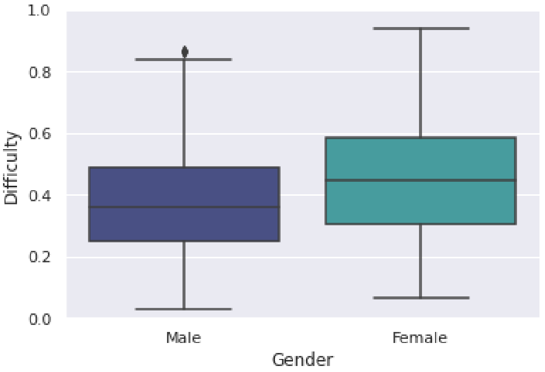


Figure 11: Distribution of synthesized speech difficulty for each gender.

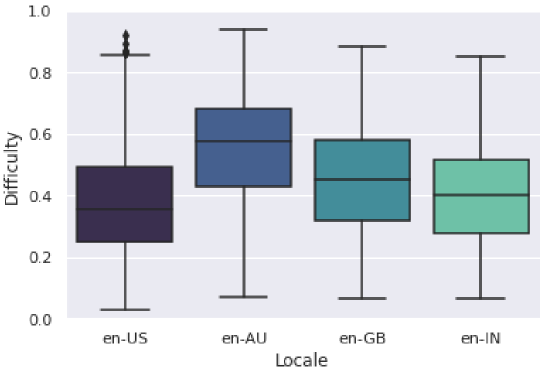


Figure 10: Distribution of synthesized speech difficulty for each locale.

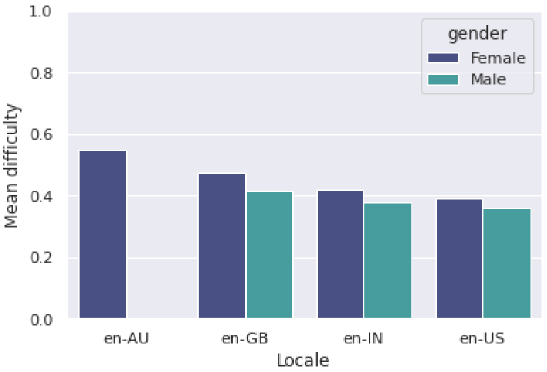


Figure 12: Mean difficulty of each gender by locale.

4.2. Predicting the Difficulty of a Synthetic Speech

This section presents the experiments we performed to evaluate the predictability of the synthesized speech difficulty. As we have the sentences and speaker parameters used to generate the speeches (i.e., pitch, rate, speaker, locale), we investigated if difficulty can be predicted using these parameters as predictor attributes (Table 2). Thus, we trained some regression models by assuming difficulty as the target attribute.

The regression models we trained were: MLP, Linear Regression and Random Forest from scikit-learn[1], a machine learning python library. We encoded the categorical features (e.g., sentence, speaker, and so on) using the label encoding method, also from scikit-learn. We also run each model using four different combinations of features (Table 3):

- Combination 1: all features (Table 2).
- Combination 2: all features except the sentence.
- Combination 2: all features except the speaker.
- Combination 2: all features except the sentence and the speaker.

The Random Forest trained with all features outperformed all models. It had normalized MAE and R2 of 0.50 and 0.31, respectively (Table 3). Figure 13 shows the feature's importances. It represents the score of the features we used to train the Random Forest model with all features (i.e., combination 1). The feature that has more effect is the sentence, followed by the size of the sentence (i.e., len_sentence), service, speaker, number of words, pitch, rate, locale and gender, since higher values mean that a feature has more effect on the prediction process. For instance, the feature service is more useful for predicting the synthesized speech difficulty than the rate. In fact, in Section 4.1 we could see that the tendency some services have to generate more difficult speeches is more

[1] https://scikit-learn.org/stable/

Table 2
Features used to train the model

	Feature	Meaning
1	gender	Whether the voice used to synthesize is female or male.
2	len_sentence	The length (number of characters) of the sentence used to synthesize that speech.
3	locale	Whether the English accent is American, Australian and so on.
4	number_of_words	The number of words of the sentence used to synthesize.
5	pitch	The pitch used in the synthesis.
6	rate	The rate used in the synthesis.
7	sentence	The sentence used in the synthesis.
8	service	The service used (e.g., Google, Watson, etc).
9	speaker	The voice used (e.g., Ana, Michael, etc).

Table 3
Results for the estimation of difficulties using each one of the three regression models: MLP, Linear Regression and Random Forest

	MLP		**Linear Regression**		**Random Forest**	
Features	**MAE**	**R2**	**MAE**	**R2**	**MAE**	**R2**
All	0.87	0.82	0.87	0.79	**0.50**	**0.31**
All except 7	0.87	0.79	0.87	0.79	**0.64**	**0.54**
All except 9	0.83	0.75	0.87	0.80	**0.53**	**0.38**
All except 7 and 9	0.84	0.76	0.87	0.79	**0.62**	**0.54**

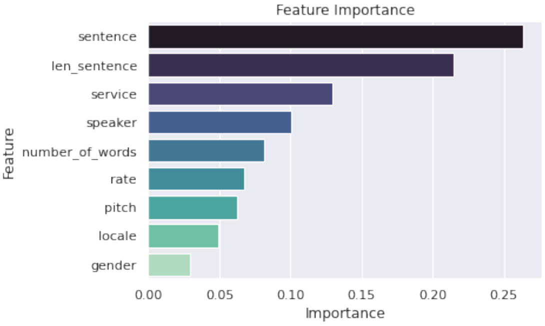


Figure 13: Feature Importance

explicit than some rates do. In other words, the difficulty distribution between the services is more different than the difficulty distribution among the rates.

It was a preliminary study to analyze the viability of using three different types of models to predict the synthesized speech difficulty. The experiments showed that by having a sentence and the synthesis parameters of a new speech we want to synthesize, we can predict its difficulty without having to run an IRT model again. We can use the dataset we already constructed to train a model that would be able to perform that prediction. In the near future, we aim to delve deeper into this and do more experiments and further analysis. We can explore adding more features related to phonemes, of instance. Also we can test our models with a newly .

5. Conclusion and Future Work

In this paper, we investigated what explains the synthesized speech difficulty. We deeply analyzed the data regarding an experiment we performed in [2] focusing on that topic: finding out if the difficulty of a synthesized speech can be explained by the sentence or any other parameter used in the synthesis process (e.g., pitch, rate, speaker).

The results of our descriptive analysis showed that bigger sentences tend to be more difficult. Also, some services or languages generate easier speeches than others. Female voices are more difficult than male ones. We also trained regression models in order to see if we can predict the synthesized speech difficulty. Our preliminary experiment showed that it may be useful to use this approach in this context. So, in the feature, we aim to better investigate this topic, training more robust models and adding more features to see if we have more insights about that and even better results.

Acknowledgments

This work was supported by CAPES, CNPq and FACEPE

(Brazilian funding agencies) and Motorola Mobility.

References

[1] C. S. Oliveira, C. C. Tenório, R. B. Prudêncio, Item response theory to estimate the latent ability of speech synthesizers., in: ECAI, 2020, pp. 1874–1880.

[2] C. S. Oliveira, J. V. Moraes, T. Silva Filho, R. B. Prudêncio, A two-level item response theory model to evaluate speech synthesis and recognition, Speech Communication 137 (2022) 19–34.

[3] Y. Chen, T. S. Filho, R. B. C. Prudêncio, T. Diethe, P. Flach, β^3-irt: A new item response model and its applications, in: Proceedings of Machine Learning Research, volume 89, 2019, pp. 1013–1021.

[4] F. Martínez-Plumed, R. B. C. Prudêncio, A. Martínez-Usó, J. Hernández-Orallo, Making sense of item response theory in machine learning, in: Proceedings of the Twenty-second European Conference on Artificial Intelligence, IOS Press, 2016, pp. 1140–1148.

[5] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, J. Hernández-Orallo, Item response theory in ai: Analysing machine learning classifiers at the instance level, Artificial Intelligence 271 (2019) 18–42.

[6] J. V. Moraes, J. T. Reinaldo, M. Ferreira-Junior, T. Silva Filho, R. B. Prudêncio, Evaluating regression algorithms at the instance level using item response theory, Knowledge-Based Systems (2022) 108076.

[7] F. Martınez-Plumed, D. Castellano-Falcón, C. Monserrat, J. Hernández-Orallo, When ai difficulty is easy: The explanatory power of predicting irt difficulty, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[8] VoxForge, Voxforge, 2022. URL: http://www.voxforge.org/, access in: 08/05/2022.

[9] R. J. De Ayala, The theory and practice of item response theory, Guilford Publications, 2013.

[10] A. W. S. AWS, Amazon polly: Turn text into lifelike speech using deep learning, 2022. URL: https://aws.amazon.com/polly/, access in: 08/05/2022.

[11] G. Cloud, Cloud text-to-speech: Text-to-speech conversion powered by machine learning, 2022. URL: https://cloud.google.com/text-to-speech/, access in: 08/05/2022.

[12] I. Watson, Text to speech, 2022. URL: https://text-to-speech-demo.ng.bluemix.net/, access in: 08/05/2022.

[13] M. Azure, Text to speech: Convert text to lifelike speech for more natural interfaces, 2022. URL: https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/, access in: 08/05/2022.

[14] G. Cloud, Speech-to-text: Speech-to-text conversion powered by machine learning, 2022. URL: https://cloud.google.com/speech-to-text, access in: 08/05/2022.

[15] M. Azure, Speech to text: Convert spoken audio to text for more natural interactions, 2022. URL: https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/, access in: 08/05/2022.

[16] I. Watson, Speech to text, 2022. URL: https://speech-to-text-demo.ng.bluemix.net/, access in: 08/05/2022.

[17] W. AI, Natural language for developers, 2022. URL: https://wit.ai/, access in: 08/05/2022.