INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Sustained Vowel Game: a computer therapy game for children with dysphonia

Vanessa Lopes, João Magalhães, Sofia Cavaco

NOVA LINCS, Departamento de Informática
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2829-516 Caparica, Portugal

vs.lopes@campus.fct.unl.pt, {jmag, scavaco}@fct.unl.pt

Abstract

Problems in vocal quality are common in 4 to 12-year-old children, which may affect their health as well as their social interactions and development process. The sustained vowel exercise is widely used by speech and language pathologists for the child's voice recovery and vocal re-education. Nonetheless, despite being an important voice exercise, it can be a monotonous and tedious activity for children. Here, we propose a computer therapy game that uses the sustained vowel exercise to motivate children on doing this exercise often. In addition, the game gives visual feedback on the child's performance, which helps the child understand how to improve the voice production. The game uses a vowel classification model learned with a support vector machine and Mel frequency cepstral coefficients. A user test with 14 children showed that when using the game, children achieve longer phonation times than without the game. Also, it shows that the visual feedback helps and motivates children on improving their sustained vowel productions.

Index Terms: childhood dysphonia, voice disorders, speech-language therapy, sustained vowel exercise, child-computer interaction

1. Introduction

According to recent studies on vocal analysis, dysphonia (*i.e.* voice disorders) affect approximately 4 to 38% of 4 to 12 year-old children [1, 2, 3]. Since speech is one of the most important ways of communication, an abnormal voice can cause communication difficulties and trigger embarrassment, shame, exclusion and other negative feelings, which may affect the quality of life of these children. In addition, speaking with an improper voice, in terms of loudness or pitch, can damage the vocal folds, which may lead to nodules and cysts, and cause a hoarse voice and aphonia [4].

In order to correct these pathologies, children with voice disorders may have to resort to speech therapy. The sustained vowel exercise (SVE) is commonly used by speech and language pathologists (SLP) to treat different types of dysphonia [5, 6, 7, 8]. The success of this exercise depends on its repetition, which, for children, may turn into a monotonous activity. In order to keep the children motivated, engaged and collaborating with the SLP, it is important to transform monotonous exercises into appealing activities [9, 10]. Computer-based therapy games can be used for this purpose.

A few computer games have been proposed to complement therapy of speech and language disorders (SSD). These include LittleBeeSpeech [11], Falar a Brincar [12] and sPeAK-MAN [13], which focus on articulation problems, Vithea [14], which was designed for the treatment of aphasia, Flappy Voice [15] for apraxia and Carvalho's interactive game for training the Portuguese vowels [16].

The Lee Silverman Voice Treatment companion system is a computer tool that uses the SVE exercise to complement voice treatment sessions and assessment [17, 18]. The tool focuses on the treatment of Parkinson's disease, and other neurological pathologies, including dysarthria or cerebral palsy. It allows practicing the SVE and other continuous speech exercises. Children can use the tool to practice these exercises. Nonetheless, the tool does not offer an engaging interactive interface for them, since it was designed to be a tool, not a game.

Despite the potential of computer therapy games, and while a few games have been proposed to complement therapy of SSD, there is a lack of such tools to assist therapy of voice disorders. As a contribution to improve the motivation of children on doing the sustained vowel exercise, we have developed the *sustained vowel game*, a serious computer game for this exercise. The game uses attractive scenarios and characters, as well as rewards to motivate the children on playing it, and, as a result, practicing the SVE. The movement of the characters gives visual feedback about the child's performance. With the right visual feedback from the exercise scene, children can recognize their failures and improve their sound productions until they accomplish the exercise goal. Moreover, SLPs can parameterize the game according to the needs of each patient and the games uses a dynamic difficulty adjustment model to keep the children engaged on playing it [19].

In order to control the character's movements, we measure voice parameters and use a vowel classification model based on Mel frequency cepstral coefficients (MFCC) to determine if the child's vowel productions are within the expected values. A user test with 14 children showed that the proposed game motivates children on doing the SVE with a good performance and helps them understand how to improve their sustained vowel productions.

2. The sustained vowel game

The proposed game integrates the SVE. In order to detect and treat dysphonia symptoms, SLPs use the SVE to correct improper pitch and loudness levels [5, 6, 7, 8]. This exercise consists of saying a vowel for as long as possible while maintaining the voice intensity level stable.

In order to play the game and control the main character, children have to perform the SVE with a good performance. The game uses multiple scenarios and characters combined with a set of drawings to reward the child's success (figure 1). Figure 1a shows one of the nine available scenarios with one of the eleven available characters. In addition, the game includes visual cues that give children feedback about their vocal pro-

http://dx.doi.org/10.21437/Interspeech.2019-3017

Figure 1: *The sustained vowel game. (a) One of its scenarios and characters. (b) The rewards: a collection of drawings.*

ductions and helps them understand how to correct themselves. The main character moves towards the target while the child is producing a specific vowel, chosen by the SLP, with a stable intensity level. If the voice production's intensity level starts decreasing or increasing, the character starts falling or going up, respectively. This visual feedback is important to allow the child to understand why he/she is failing the exercise, and to understand what to do in order to perform the exercise correctly.

Apart from recognizing if the child is producing the expected vowel, the game controls other parameters of interest to the SVE: intensity level and stability, and maximum phonation time. Children with different pathologies may have to practice the SVE at different intensity levels. The SLP can choose the appropriate intensity level. The goal of the SVE is to practice the vowel production with a stable voice at the chosen intensity level. Small variations in intensity are allowed. Thus, the game analyses the vowel production intensity over time, to determine if it is within the allowed intensity interval.

Maximum phonation time (MPT) is the maximum time (in seconds) a person can sustain a vocal sound (for example a vowel), after taking a deep breath and producing the sound with a comfortable level of intensity. The game allows the SLP to choose the initial expected MPT, which we call MPT_e. The distance between the main character and the target gives a visual cue for MPT_e. The character will reach the target when the child performs the vowel production for as long as MPT_e and within the allowed intensity interval.

3. Vowels automatic recognition

The game uses an EP vowel model to classify the child's speech productions and recognize if the child is producing the vowel chosen by the SLP. The model in the game's current version is trained to recognize the five European Portuguese (EP) vowels: *a, e, i, o, u*, which correspond to the phonemes /a/, /ɛ/, /i/, /ɔ/, /u/ [20].

3.1. Automatic vowel classification

We compared the performance of different features and classification algorithms, to decide which techniques to use to develop the game's vowel model. The different approaches were trained and validated with a combination of two EP vowel data sets with children's data: (1) Ferreira's vowel data set, which includes the phonemes /a/, /ɛ/, /i/, /ɔ/, /u/ [21], and (2) the BioVisualSpeech's vowel data set, which includes the sustained vowels /a/, /i/, and /u/ [22].

The samples are 100 ms long with a 48 kHz sampling rate. By combining these two data sets, we have data from 48, 4 to 10-year-old children.

Our feature vectors consist of filter bank and MFCCs based feature vectors (which were computed with a 20 ms window size and 10 ms overlap). We experimented with six different types of feature vectors, which we call FV1 to FV6 and which are composed of the following features (extracted with the Librosa python library [23]):

FV1 - MFCCs (12 dimensional vectors, 12D vectors for short);

FV2 - MFCCs, MFCCs delta and double-delta (36D vectors);

FV3 - Mean and standard deviation for each coefficient of the MFCCs, MFCCs delta and double-delta (72D vectors);

FV4 - Mean and standard deviation for each coefficient of the MFCCs (24D vectors);

FV5 - Filter banks (40D vectors);

FV6 - Mean and standard deviation of the filter banks (80D vectors).

Each sound sample is represented by several feature vectors FV1, FV2 and FV5, but only one feature vector FV3, FV4 and FV6.

In order to decide on the number of MFCCs, we compared the accuracy scores using 5 to 16 MFCCs with a support vector machine (SVM) and feature vectors FV1 to FV4. We used a random split of the data at this stage. We observed that the score increase stagnates at 12 MFCCs with 96% accuracy for FV1 and FV2 (figure 2). Thus we are using 12 MFCCs to compute the feature vectors FV1 to FV4.

In order to decide on a classification algorithm, we compared the performance of the following algorithms: SVM with a radial basis function (RBF) kernel (parameters C and γ were carefully chosen with grid-search cross-validation), random forest (RF), for which we tested different combinations for the number of estimators and the maximum depth, and quadratic discriminant analysis (QDA).

The data was split into training and test sets following a leave-one-child-out approach. More specifically, we run n tests, where n is the number of children in our data set. For each test i, the test data set contains all data from child i, while the training set uses the remaining data. With this technique, we ensure that the learned models are not biased with samples from the same child in both training and test sets.

3.2. Evaluation

In order to decide on which vowel model to use in our game, we compared the results from the models learned with the techniques described in section 3.1. Let us first look into type of feature vectors. Figure 3 shows the average accuracy scores obtained with a SVM for the different types of feature vectors. The accuracy scores obtained with features FV1, FV2 and FV5 show that the models with these features are highly effective on classifying EP vowels. The lower accuracy scores obtained with FV3, FV4 and FV6 (which are the feature vectors with the mean and standard deviation) show that loosing the time definition has a negative effect on the models' performances. Sound samples

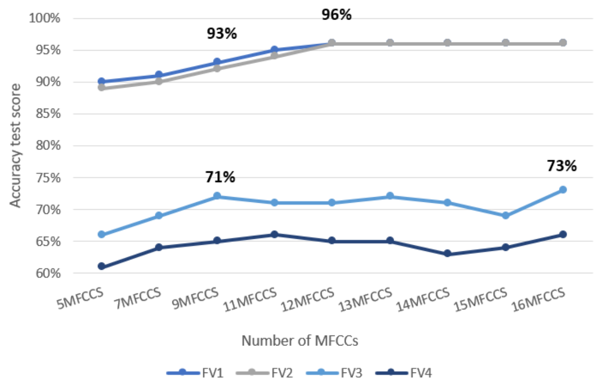

Figure 2: *Accuracy scores of FV1 to FV4 (MFCC based feature vectors) with a SVM.*

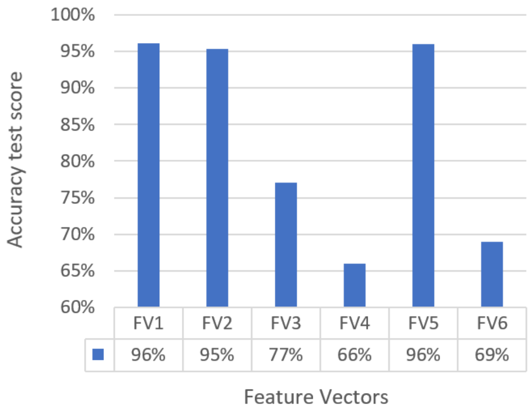

Figure 3: *Average accuracy scores for FV1 to FV6 with a SVM.*

are represented by only one feature vector for FV3, FV4 and FV6, while they are represented by several feature vectors for FV1, FV2 and FV5. Since many sound samples include background noise, having many feature vectors representing each sound sample may lead to more robust classification models as the no-noise segments of the data are better represented this way. The results obtained with RF and QDA have lower accuracy scores but FV1 and FV5 still lead to the best accuracy scores. Given that the models learned with FV1 (with MFCCs) and FV5 (with filter banks) have similar accuracy scores, we could chose any of these two feature vector types to train the game's final model. We chose to proceed with FV1 because these vectors have a smaller dimension.

The average accuracy scores obtained with the models learned with FV1 and the three classification algorithms were: 0.964 for SVM, 0.942 for RF, and 0.932 for QDA. All three models achieve high accuracy scores. The best results are from the model learned with a SVM ($C = 1.0$ and $\gamma = 0.1$).

Some data samples have low quality, due to silence and low volume segments or background noise. The segments with these problems result in misinterpretations by the classification models, which affect the final score. We could remove these outliers and produce a 2% higher accuracy score. However, these cases allow the model to capture small variations in the data that are relevant for future predictions, since the game will certainly have future speech productions with silence, background noise and lower volume.

To conclude, the proposed game uses the best model obtained in these experiments. Due to its high accuracy score, this is the model learned by a SVM with $C = 1.0$ and $\gamma = 0.1$, and feature vectors FV1 with 12 MFCCs.

4. Validation

The goal of the proposed work is to help SLPs on motivating children on doing the SVE and to help children on achieving a good performance in this exercise. To validate if our work can achieve these goals, we run a user test with 14 children.

4.1. Protocol

The objective of the user test is not only to determine if the proposed game is appealing, intuitive and feasible for the children, but also, and more importantly, to determine if the game helps children on doing the SVE with a good performance. In other words, we want to assess if the game helps the children on doing the exercise with a stable intensity and a good phonation time.

We run this study at a kindergarten, with the help of the school's SLP. There were 14 children (6 boys and 8 girls) from 4 to 5 years old participating in this user test. Three 5-year-old participants had diagnosed dysphonia, as determined by the school'a SLP after performing the auditory-perceptual voice quality evaluation [24, 25]. There was a child with rough voice (child 14), a child with asthenia, *i.e.* week voice (child 3), and a child with abnormal and unsteady loudness (child 9).

We performed individual user test sessions in a quiet room, with the presence of the SLP. The equipament used during the test was a laptop running the game (HP Pavilion x360 intel core i5-6200U dual core) and a Blue Yeti USB microphone. All children performed the SVE exercise three times: once without the game and twice with the game. The protocol consisted of the following steps, where MPT_i is the MPT measured for trial i:

1. We explain the game to the child (a journey through different scenes where they can be rewarded with gifts if they conclude the tasks).
2. **Trial 1**: The SLP asks the child to perform the SVE for as long as possible. For this trial we do not use the game. The sound production is recorded with Audacity 2.3.0.
3. The SLP opens the game and inserts the basic information about the child (like age, gender, etc.).
4. **Trial** i: We repeat the process from step 4a to step 4d twice, in order to register two trials using the game (trial 2, and trial 3).
   (a) The SLP inserts the parameters for this trial. The expected MPT, MPT_e, is chosen as follows: If this is trial 2, then $MPT_e = MPT_1 + \Delta t$. If this is trial 3 and in the previous trial the child performed the exercise without breaks, then $MPT_e = MPT_2 + \Delta t$, otherwise MPT_e remains the same. (The maximum MPT_e is 10 s.) If the child has no dysphonia then $\Delta t = 4$s, otherwise $\Delta t = 2$ s.
   (b) The child chooses the scene and character.
   (c) When the child enters the scene, the SLP tells her/him to produce the chosen vowel, until the character reaches the target.
   (d) At the end of the trial the child chooses a reward.

We use vowel /a/ for the first and second trials because the SLP considered that this to be the easiest vowel for the SVE. The child can choose another vowel for the last game trial. We start with the vowel /a/,

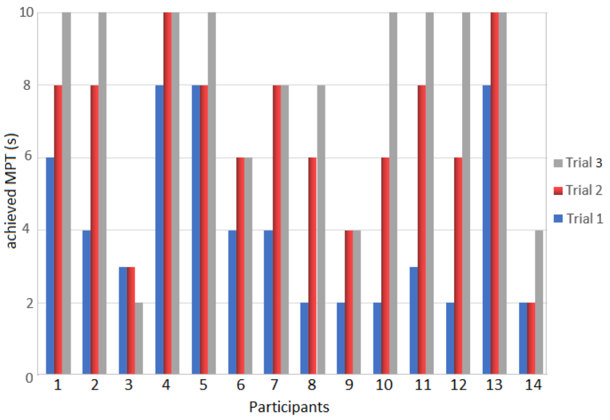

Figure 4: *The achieved MPT for each child during the experiment. Participants 3, 9 and 14 have diagnosed dysphonia.*

4.2. Results

Figure 4 shows the achieved MPT for each child during the user test. The left bar for each child (in blue) shows MPT_1, the achieved MPT without using the game. Tavares *et al.* determined that the MPT normative value for 4 to 7-year-old children without voice pathologies is 6.12 ± 1.89 s [26]. As can be observed in figure 4, MPT_1 ranges from 2 to 8 s and there are four children (participants 1, 4, 5 and 13) who achieve phonation times above 6 s. All the other children, including children without dysphonia, achieve phonation times below 6 s.

The middle bar for each child (in red) shows MPT_2, the achieved MPT for the second trial, which was the first time the participants played the game. As it can be observed in figure 4 most children improve their achieved MPT in trial 2. There are only 3 children (participants 3, 5 and 14) for whom $MPT_1 = MPT_2$, that is, they maintain the same achieved phonation time. From these 3 participants, two have dysphonia and one (participant 5) has no dysphonia, but his/her achieved phonation time (8 s) is already close to the normative value upper bound ($6.12 + 1.89$ s).

The right (gray) bar for each child shows MPT_3, the achieved MPT for the last trial, which was the second time the participants played the game. There are 8 children (participants 1, 2, 5, 8, 10, 11, 12 and 14) who achieve longer MPT values in this trial ($MPT_3 > MPT_2$ and $MPT_3 > MPT_1$). Five children maintain the same MPT value as in the previous trial ($MPT_3 = MPT_2$).

There were 8 participants achieving an MPT of 10 s, which was the maximum value we were testing and which is above the normative value. Two participants achieved the 10 s in trial 2, and the remaining 6 achieve this MPT in trial 3. While during trial 1, most children achieved phonation times below the 6.12 s normative value, children without diagnosed dysphonia did not have problems on overcoming this value while using the game.

Overall, apart from child 3 who has asthenia, all the children improved their achieved MPT progressively during the user test. In some cases, the improvement was drastic, which shows that the game gave these children the motivation on doing the SVE: from trial 1 to trial 3, participants 10 and 12 achieved an improvement of 8 s, participant 11 achieved an improvement of 7 s and participants 2 and 8 achieved an improvement of 6 s.

While it is relevant to see that children in general achieve better phonation times while using the game, it is important to check if this pattern is also observed in children with dysphonia. In fact, participants 9 and 14, who have dysphonia, managed to improve their achieved phonation times while using the game. Both these children started with a MPT of 2 s and ended with a MPT of 4 s. Although child 9 was dealing with an unsteady loudness while doing the SVE and child 14 had a rough voice, they were able to achieve the 4 s either at trial 2 or 3.

Child 3 is the exception, as this child did not improve the MPT. This child started with a MPT of 3 s for trial 1. After trial 1, the SLP set $MPT_e = 4$ s. During trial 2, the child did not increase the achieved MPT value and could not make the character reach the target. We asked her/him if the trial was hard, and she/he agreed. Thus, the SLP decided to decrease the MPT_e to 2 s. The child was able to conclude the task successfully during trial 3 for this MPT_e.

Apart from concluding that the game can help children achieve better phonation times, it is also important to check if children react well to the game and if the game motivates them to do the SVE. During the user test, we observed that when we presented the game to the children, all children (including children with dysphonia) showed interest on finishing the exercise to bring the character to the target and avoid having it falling. They showed far more motivation on doing the SVE exercise while they were using the game than during the first trial (without the game). At the end of the user test, the children's feedback was positive and unanimous. Moreover they wanted to keep on playing to conquer rewards and try different scenarios.

We also observed that the visual feedback given by the character's movement is efficient. The children tried to avoid having the character falling and they understood that they should not interrupted the vowel production, as the character stopped.

5. Conclusion

We propose a game for voice therapy that uses the sustained vowel exercise. The game aims to motivate children on practicing the SVE often and on trying to achieve a good performance. In addition, it provides visual feedback that helps children understand how to improve their sustained vowel productions. In order to recognize if the child is producing the vowel chosen by the SLP, the game uses a high accuracy classification model learned with a SVM and MFCCs.

As verified in a user test, although most children started with MPT values below the 6.12 s normative value, all children without dysphonia were able to achieve or surpass this value while playing the game. Children with dysphonia struggled more in achieving the expected MPT than other children but two children with dysphonia achieved a final MPT close to the normative value lower bound ($6.12 - 1.89$ s). The user test allowed us to verify that this interactive game motivates children the will to give their best while performing the SVE. Apart form a child with asthenia, all children participating in the user test improved their MPT while playing the game. Nonetheless, this child was interested in continuing playing, which shows that even when the SVE is difficult for the children, the game gives them the motivation to continuing practicing.

6. Acknowledgements

This work was supported by the Portuguese Foundation for Science and Technology under projects BioVisualSpeech (CMUP-ERI/TIC/0033/2014) and NOVA-LINCS (PEest/UID/CEC/04516/2019). We thank Dr. Alfredo Mota kindergarten (Castelo Branco, Portugal) and all the children who participated in the user test.

7. References

[1] M. C. Duff, A. Proctor, and E. Yairi, "Prevalence of voice disorders in African American and European American preschoolers," *Journal of Voice*, vol. 18, no. 3, pp. 348–353, 2004.

[2] R. Oliveira, L. Teixeira, A. C. Gama, and A. de Medeiros, "Análise perceptivo-auditiva, acústica e autopercepção vocal em crianças," *Jornal da Sociedade Brasileira de Fonoaudiologia*, vol. 23, no. 2, pp. 158–163, 2011.

[3] E. L. M. Tavares, A. Brasolotto, M. F. Santana, C. A. Padovan, and R. H. G. Martins, "Epidemiological study of dysphonia in 4-12 year-old children," *Brazilian Journal of Otorhinolaryngology*, vol. 77, no. 6, pp. 736–746, 2011.

[4] J. Herranz, J. Gavilán Bouzas, C. Vázquez Barro, and L. Martín Méndez, *Mucosal Disease of the Glottis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 483–486.

[5] American Speech-Language-Hearing Association (ASHA). (2018) Voice Disorders: Treatment. https://www.asha.org/ (retrieved in Jun/2018).

[6] A. El Sharkawi, L. Ramig, J. A. Logemann, B. R. Pauloski, A. W. Rademaker, C. H. Smith, A. Pawlas, S. Baum, and C. Werner, "Swallowing and voice effects of Lee Silverman Voice Treatment (LSVT®): A pilot study," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 72, no. 1, pp. 31–36, 2002.

[7] L. Mota, C. Santos, J. Vasconcelos, B. Mota, and H. Mota, "Applying the technique of sustained maximum phonation time in a female patient with adductor spasmodic dysphonia: a case report," *Revista da Sociedade Brasileira de Fonoaudiologia*, vol. 17, pp. 351 – 356, 2012.

[8] B. J. D. Swart, S. C. Willemse, B. Maassen, and M. W. Horstink, "Improvement of voicing in patients with parkinsons disease by speech therapy," *Neurology*, vol. 60, no. 3, p. 498500, Nov 2003.

[9] C. Bowen, "Childhood apraxia of speech," in *Children's Speech Sound Disorders*, 2nd ed. John Wiley and Sons, L.td., 2015, pp. 343–350.

[10] E. Eriksson, O. Balter, O. Engwall, A.-M. Oster, and H. Kjellstrom, "Design recommendations for a computer-based speech training system based on end-user interviews," in *Proceedings of the International Conference on Speech and Computer (SPECOM)*, 2005.

[11] H. Hanks, C. Hanks, and M. Suman. (2018) Little Bee Speech. Retrieved from http://littlebeespeech.com/ (retrieved in Dec/2018).

[12] J. Tavares, J. Lopes, M. Cunha, and R. Saldanha. (2017) Falar a Brincar. https://falarabrincar.wordpress.com (retrieved in Sep/2017.

[13] C. T. Tan, A. Johnston, K. Ballard, S. Ferguson, and D. Perera-Schulz, "sPeAK-MAN: towards popular gameplay for speech therapy," in *Proceedings of the Australasian Conference on Interactive Entertainment: Matters of Life and Death*, 2013, p. 28.

[14] A. Pompili, P. Fialho, and A. Abad, "VITHEA: On-line word naming therapy in portuguese for aphasic patients exploiting automatic speech recognition," in *Proceedings of IberSPEECH*, 2012.

[15] T. Lan, S. Aryal, B. Ahmed, K. Ballard, R. Gutierrez-Osuna, R. Gutierez-osuna, and A. Texas, "Flappy Voice: An interactive game for childhood apraxia of speech therapy," in *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*, vol. 2, no. 1, 2014, pp. 429–430.

[16] M. Carvalho, "Interactive game for the training of portuguese vowels," Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2008.

[17] C. M. Fox and C. A. Boliek, "Intensive Voice Treatment (LSVT LOUD) for Children With Spastic Cerebral Palsy and Dysarthria," *Journal of Speech Language and Hearing Research*, vol. 55, no. 3, p. 930, 2012.

[18] A. E. Halpern, L. O. Ramig, C. E. C. Matos, J. A. Petska-Cable, J. L. Spielman, J. M. Pogoda, P. M. Gilley, S. Sapir, J. K. Bennett, and D. H. McFarland, "Innovative technology for the assisted delivery of intensive voice treatment (LSVT(R)LOUD) for Parkinson disease." *American journal of speech-language pathology*, vol. 21, no. 4, pp. 354–367, 2012.

[19] V. Lopes, J. Magalhes, and S. Cavaco, "A dynamic difficulty adjustment model for dysphonia therapy games," in *Proceedings of the International Conference on Human Computer Interaction Theory and Applications (HUCAPP)*, 2019, pp. 137–144.

[20] M. Cruz-Ferreira, "Portuguese (european)," in *Handbook of the International Phonetic Association, A guide to the use of the international phonetic alphabet*. Cambridge, University Press, 1999.

[21] A. J. S. Ferreira, "Static features in real-time recognition of isolated vowels at high pitch," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2389–2404, 2007.

[22] M. Diogo, M. Eskenazi, J. Magalhaes, and S. Cavaco, "Robust scoring of voice exercises in computer-based speech therapy systems," in *European Signal Processing Conference (EUSIPCO)*, 2016, pp. 393–397.

[23] B. McFee, C. Raffel, D. Liang, D. C. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.

[24] A. Behrman, "Common practices of voice therapists in the evaluation of patients," *Journal of Voice*, no. 19, pp. 454–469, 2005.

[25] A. Webb, P. Carding, I. Deary, K. Mackenzie, N. Steen, and J. Wilson, "The reliability of three perceptual evaluation scales for dysphonia," *European Archives of Oto-Rhino-Laryngology*, no. 261, pp. 429–434, 2004.

[26] E. L. M. Tavares, A. G. Brasolotto, S. A. Rodrigues, A. B. B. Pessin, and R. H. G. Martins, "Maximum phonation time and s/z ratio in a large child cohort," *Journal of Voice*, vol. 26, no. 5, 2012.