ROBUST SCORING OF VOICE EXERCISES IN COMPUTER-BASED SPEECH THERAPY SYSTEMS

Mariana Diogo[1], *Maxine Eskenazi*[2], *João Magalhães*[1], *Sofia Cavaco*[1]

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

[2]Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

ABSTRACT

Speech therapy is essential to help children with speech sound disorders. While some computer tools for speech therapy have been proposed, most focus on articulation disorders. Another important aspect of speech therapy is voice quality but not much research has been developed on this issue.

As a contribution to fill this gap, we propose a robust scoring model for voice exercises often used in speech therapy sessions, namely the sustained vowel and the increasing/decreasing pitch variation exercises. The models are learned with a support vector machine and double cross-validation, and obtained accuracies from approximately 73.98% to 85.93% while showing a low rate of false negatives. The learned models allow classifying the children's answers on the exercises, thus providing them with real-time feedback on their performance.

Index Terms— speech therapy, robust scoring, support vector machines, cross-validation

1. INTRODUCTION

Speech sound disorders (SSD) are a common problem in children across the world. These disorders can have a negative impact on children's lives as they affect the children's communication skills and may cause negative feelings such as embarrassment, shame, and frustration. In addition, if these problems are not detected and treated early, they may persist or lead to a worsening of the situation depending on the disorder. Speech therapy plays an important role to overcome these difficulties since it allows detecting the disorders as well as to correct and improve the children's speech.

However, since speech therapy needs to resort to the repetition of the same words or sounds, the therapy sessions can become monotonous and tedious. One technique that speech and language therapists (SLT) can adopt to motivate children to collaborate and do the therapy activities, is to use games or gamification elements such as rewards, scores and other encouragement techniques. Software systems can be an effective way to apply these concepts and maintain children motivated and interested in the therapy activities [1].

Several tools for articulation disorders have been developed in recent years. These include exercises that encourage children to produce phonemes, syllables, or words correctly. Some of the developed interactive environments focus on specific speech and/or language disorders, like aphasia [2, 3], apraxia [4], articulation [5, 6], dyslalia [7], cleft palate or lip [8], among others [9, 10]. Furthermore, certain systems, such as VisualSpeech and Parnandi's tool, also allow SLTs to monitor and schedule future sessions [4, 6].

Most systems are focused on issues related to the pronunciation of sounds (phonemes and words), in particular to ensure that all different types of phonemes are said correctly. Nevertheless, this is not the only problem affecting children's speech. Another important issue to consider is voice quality. It is quite common for children to scream or speak in a tone of voice that damages the vocal cords. This may lead to the development of nodules, hoarseness, or even aphonia. In these cases, it is necessary that SLTs help children carry out exercises to improve voice quality and teach children the problems that might arise with a continuing mistreatment of the vocal cords.

Pitch variation and sustained vowel exercises are frequently performed to improve voice quality. These exercises consist of saying a sound with increasing or decreasing pitch, or saying a vowel for a few seconds with a steady voice. These exercises can contribute to voice quality improvement and at the same time provide important information about voice quality, such as: maximum phonation time, jitter, shimmer, and harmonics-to-noise ratio [11–14]. Despite the importance of these exercises, they can be difficult to perform and tiresome. In addition, their incorporation in computer tools for speech and language therapy can also be difficult since it requires sound recognition and real-time analysis of the exercise so that the tool can provide feedback and motivate the children. IBM Speech Viewer is the only tool for voice exercises that we are aware of [15]. It was developed for deaf people and is not longer available. This tool performed vocal onset and voice activity detection, and it could be used for varying intensity as well as pitch exercises. However the SLT should manually adjust the parameters for each patient.

Given the importance of voice exercises, our goal is to include them in VisualSpeech [6]. This is an interactive en-

vironment that gives visual-feedback to children in a motivational way. Live recordings of a webcam are displayed on a computer screen to provide visual feedback. The visual feedback lets the child watch himself and the SLT so that the child can observe and correct the orofacial movements (figure 1). This environment also includes elements for articulation exercises and several motivational elements (like the ice cream in figure 1 that works as a progress bar) that aim to increase the child's motivation on doing the speech therapy exercises. A usability study showed that VisualSpeech can be used to improve the children's performance and motivation on following the SLT's instructions. In spite of being so well received by SLTs, VisualSpeech is limited to articulation exercises. We are currently extending it to include voice exercises.

In order to have the voice exercises automatically classified, we use classification models that decide if the exercises are correct. Here we propose robust classification models to score three types of voice exercises: the sustained vowel, the gradual increase of pitch, and gradual decrease of pitch. The three classification models are learned using support vector machines (SVM) and acoustic features that have shown to be meaningful to the three voice exercises (section 3). We trained the SVMs with recorded voices of children doing the exercises with European Portuguese (EP) phonemes (section 2).

The three models have high accuracies (from $73,98\%$ for the decreasing pitch exercise to 85.93% for the sustained vowel exercise) and are very robust, showing a low rate of false negatives. In order to create such robust models, the way the SVMs are trained is important. We used a one-child-out experiment for assessing the models' results, and trained the models with a 5-fold cross-validation on the data of the remaining children. The main novelty of this work is the use of robust classification models learned with double cross-validation on voice exercises for speech therapy.

2. DATA

The data used in this study consists of audio recordings from children doing the sustained vowel and gradual pitch change exercises. The recordings were done in an elementary school and with the presence and help of an SLT. A total of 21 children aged 8 and 9 years old (12 boys and 9 girls) participated in the recordings. Those ages were chosen with the advice of SLTs: at these ages, children's voices are already sufficiently stable due to the speech organs' development and children are already able to say all EP phonemes (for instance, at the age of four and sometimes five, many children still do not pronounce some phonemes correctly, such as the rhotic consonant, and they still misplace some phonemes in the words) [16]. Children with and without SSD participated in the recordings. Some of the children could have non-diagnosed SSD. Most children were EP native speakers. There was one bilingual and one whose first language was not EP.

Fig. 1. VisualSpeech interactive environment.

The recordings were done in a small room, the school library, with a digital recorder and an external microphone. While the library was a quiet room, many of the recordings were done during recess time and thus these include background noise of children playing in the playground or chatting and running in the corridors. The signals were recorded with a 48000 Hz sampling rate and mono format.

Before starting the recordings, the SLT explained and exemplified each exercise at a time. The recording protocol was discussed with SLTs and we used the same phonemes that SLTs would use for these exercises in speech therapy sessions. Even though about six EP consonant phonemes can be used for these exercises, the SLTs advised us to use vowel phonemes because they can last longer. We used only three phonemes for the sustained vowel and two phonemes for the gradual pitch change exercises so that the children would not get tired and would keep collaborating with us. During the recordings, it was necessary to encourage and motivate the children not to give up.

For the sustained vowel exercise, we asked the children to take a deep breath and then to sustain a vowel for as long as possible without being in struggle. The vowels selected were the EP sounds for the vowels *a*, *i* and *u*, which correspond to the phonemes /a/, /i/ and /u/. To illustrate and to ensure that the child correctly understood the requested exercise, we exemplified it using the vowel phoneme /ɛ/ (for the EP *e*).

When conducting this exercise we observed that three types of cases can occur: the children loose their voice, the children gradually decrease the intensity of the voice; or the children perform the exercise in the requested manner. The first two cases described are classified as incorrect.

With regard to the gradual pitch change recordings, we asked the children to gradually increase or decrease the pitch while saying the vowel phonemes /a/ and /u/. These gradual changes (increase and decrease) are considered as two different exercises. To help the child understand how to do the exercise correctly, we exemplified and explained it. For example, for the decreasing pitch exercise we told the children to imagine that they had a character toy that was falling from a high place and that they should make the falling sound. These exercises were considered incorrect when children varied the pitch in the wrong way or when they varied the intensity in-

Exercise	Correct	Incorrect	Total
Sustained Vowel	127	39	166
Increasing Pitch	69	15	84
Decreasing Pitch	43	42	85

Table 1. Total number of audio recordings for each exercise.

stead of the pitch.

The total number of recordings made was 335. We tried that every child performed each exercise at least twice correctly with each vowel (thus the total number of recordings per child varied). Table 1 shows the number of correct and incorrect recordings for each exercise. For all the exercises, the expected vowel is not important, if we asked the child to perform the exercise with /a/ and she used an /ɛ/ instead, we did not consider this incorrect. In addition, recordings with sounds such as moving a chair, children's speech, and silence were also used to ensure a more robust model. These samples did not contain the voices from the exercises, they were just added to have normal sounds that can happen in an SLT office and that do not correspond to exercises. These audio files are classified as incorrect and comprise a total of 19 files.

Lastly, all recordings were manually labelled as incorrect or correct. The labelling was mostly done by us, but with some help of the SLT who was present at the recording session.

3. ROBUST SCORING OF VOICE EXERCISES

In order to develop the proposed scoring models, we trained SVMs with the voice of children doing the sustained vowel and the pitch variation exercises. Below we discuss in more detail the required steps, namely the feature extraction and the creation of the classification models.

3.1. Feature Extraction and Analysis

Since one of our goals is to develop models that can be used in real-time during speech therapy sessions, we did not use a brute force approach. Instead, we tried to find a small set of features with which it is possible to discriminate the exercises correctly. In other words, instead of starting with a very large set of features and perform multivariate data analysis to reduce the size of that set, we looked for specific features that provide relevant information for each exercise.

In order to choose the best features to create the classification models, we started using a set of features commonly used in analysis of automated recognition of speech and then iteratively improved this set. The initial feature set contained 16 low level features, delta coefficients, and 12 statistical measures of these low level features [17]). We used openSMILE for feature extraction.

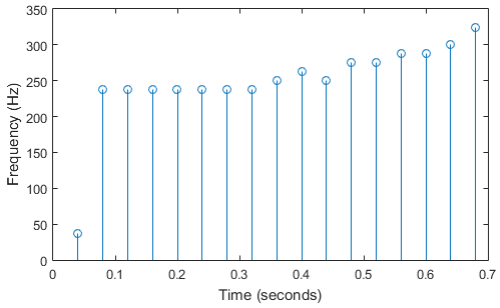

Fig. 2. Values of F0 over time of a gradual increase of pitch exercise of the phoneme vowel /a/ from an eight year old boy.

With this initial set the classification results were not satisfactory, with accuracies around 50%. In order to improve the results, we performed several experiments varying the structure of the feature set, and changed the set taking into account the results obtained, the type of exercises we were dealing with and the properties that were being evaluated. In this manner, we were able to select a small set of meaningful features, *i.e.*, that carry relevant information to the voice exercises we are addressing and that lead to improved results.

Since MFCCs are an appropriate feature to delineate the vowel part of the recordings, our final feature set for all three exercises used 13 MFCCs and statistical measures of these 13 time-varying vectors. These include the minimum and maximum values of the MFCCs in each frame, the arithmetic mean, slope and offset of the linear approximation, quadratic error, standard deviation, skewness and kurtosis.

In addition to these features, the gradual pitch change exercises also used the fundamental frequency (F0) and statistical measures of the F0. (The same statistical measures as described above for the MFCCs were used with F0.) F0 was chosen since it is the main acoustical cue to pitch perception.

In the gradual pitch change exercises it is important to detect the variation of the pitch over time. As an illustration, figure 2 shows the F0 over time from a recording of the gradual increase of pitch exercise. Each point in the graph shows the F0 value for a 0.03 seconds segment of the recording. As it can be observed in the figure, as desired the pitch of the signal in this exercise increases over time. In order to have a measure of the signal's time variation, in addition to the features mentioned above, we also used the delta regression of both F0 and MFCCs for the gradual pitch change exercises.

3.2. Model Estimation Methodology

A classification model for scoring each voice exercise was learned independently. To learn these three models we used SVM with a Gaussian radial basis function kernel. (We used the LibSVM library.) In order to obtain robust models we performed a double cross-validation on the voice recordings: We run a *one-child-out* experiment to assess the models' results, and used a 5-fold cross-validation to train the models and find

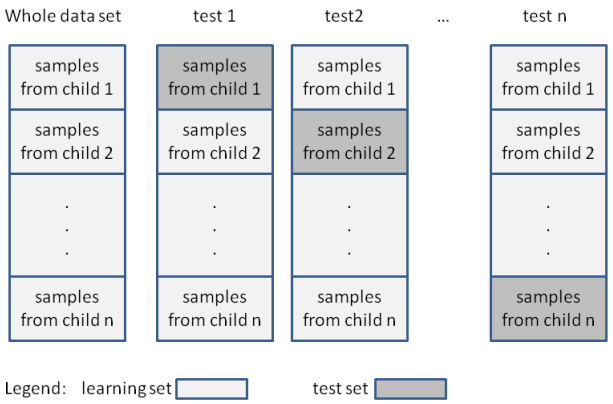

Fig. 3. One-child-out scheme. Each box respresents the set of all samples from one child.

the best parameters C and γ. We explain this double cross-validation in more detail below. We start by discussing the one-child-out experiment and then the 5-fold cross-validation.

The one-child-out experiment consists of n tests, where n is the number of children participating in the recordings (figure 3). In each test, there is a test data set that consists of the data samples from only one child. The data samples of the remaining children, which we call the learning set, were used in the 5-fold cross-validation, explained below. In other words, test i used a test set with all the data samples from child i. The learning data set for test i contained the samples from all the other children. Naturally, we performed a separate one-child-out experiment for each of the exercises. Table 1 shows the audio samples used in each of the three experiments.

As mentioned above, in order to learn robust models, our training algorithm performed a 5-fold cross-validation with the learning data for each test from the one-child-out experiment. In other words, to learn the model for test i, the algorithm divided the learning set of that test into 5 buckets and the cross-validation was done with these 5 buckets. As a result, we obtained 5 classification models in each of the n tests from the one-child-out experiment. The classification model with the highest accuracy in the validation fold was the one selected for each test.

4. RESULTS

The average results for the voice exercises are presented in table 2. The table shows the exercise (column 1), the accuracy of the classification model, i.e., the accuracy of the best model obtained in the 5-fold cross-validation (column 2), followed by the accuracy of that model when tested against the test set, that is, the samples from child i (column 3). The fourth column shows the average rate of false negatives.

The accuracy averages for the sustained vowel exercise are the highest while the accuracy averages for the gradual decrease of pitch exercise are the lowest. The main reason that led to a lower accuracy for the gradually decreasing pitch exercise is that children have greater difficulty in carrying out this exercise and this was reflected in the recorded samples. Nevertheless, the results are very positive and the margin of error in regards to false negatives is very small.

	Model Accuracy	Test Accuracy	Average Rate of False Negatives
Sustained Vowel	85.9%	84.7%	4.2%
Increasing Pitch	83.7%	82.2%	4.2%
Decreasing Pitch	74.0%	64.7%	20.5%

Table 2. Average accuracies and false negatives.

Taking into account that the main goal of this study is the creation of a classification model for voice exercises to be used in real-time and embedded in a program for speech therapy with children, it is of utmost importance that the classification has low false negatives, i.e., well done exercises that the system considers to be wrong. As shown in table 2, in average the rate of false negatives is very low (from 0.04 to 0.20). For example, for the sustained vowel exercise, most tests had zero false negatives, some had one false negative, and only one test had two false negatives. The average rate, avg_{fnr}, was obtained in the following manner: we took the rate of false negatives for each test in the one-child-out experiment, and then we took the average over those results

$$avg_{fnr} = \frac{\sum_{i=1}^{n} f_i/c_i}{n},$$

where n is the number of tests in the one-child-out experiment, f_i is the number of false negatives in test i and c_i is the number of samples in the test set of test i (that is, the number of samples from child i). For instance, for the increasing pitch exercise there was 1 false negative for a child who made one recording and 2 false negatives for a child who made 6 recordings. 20 children participated in this exercise. Thus, $avg_{fnr} = (1/2 + 2/6)/20 = 4.2\%$.

5. CONCLUSION

Voice exercises are very frequent in speech therapy sessions because of their importance to analyze the voice quality. As a contribution to fill the gap in automatic classification of voice exercises for speech therapy, here we proposed robust scoring models for the sustained vowel, the gradual increase of pitch and gradual decrease of pitch exercises.

Our models are learned by the SVM algorithm with a small set of acoustic features that have shown to carry relevant information. In order to obtain robust models, our learning algorithm performs a double cross-validation: it runs a 5-fold cross-validation of the learning data for each test within a one-child-out experiment.

The learned models are suitable for use in speech therapy applications since they have high accuracy and a low rate of false negatives, which ensures that there is no lack of motivation in patients due to this factor. The sustained vowel

model is quite effective, delivering a precision of approximately 85.93% and a low number of false negatives (most tests of the one-child-out experiment had 0 false negatives). The gradual increasing pitch exercise also shows a high accuracy for the overall classification model: 83.66%. The accuracy of the model for the gradual decreasing pitch exercise is slightly lower: 73.98%. This can be due to the difficulty that children have in performing this exercise, which is reflected in the recorded samples.

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/04516/2013).

We thank Sofia Moniz and the children from St. James School for the recordings. Thanks also to Isabel Guimarães and Margarida Grilo for help on elaborating the recording protocol and Rita Pires for help with the recordings.

6. REFERENCES

[1] T. G. Murray and V. Parker, "Integration of computer-based technology into speech-language therapy," *Educational Technology*, vol. 31, pp. 53–59, 2004.

[2] J. Galliers, S. Wilson, S. Muscroft, J. Marshall, A. Roper, N. Cocks, and T. Pring, "Accessibility of 3D game environments for people with aphasia: an exploratory study," in *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2011, pp. 139–146.

[3] A. Abad, A. Pompili, A. Costa, and I. Trancoso, "Automatic word naming recognition for treatment and assessment of aphasia.," in *INTERSPEECH*, 2012, pp. 1055–1058.

[4] A. Parnandi, V. Karappa, Y. Son, M. Shahin, J. McKechnie, K. Ballard, B. Ahmed, and R. Gutierrez-Osuna, "Architecture of an automated therapy tool for childhood apraxia of speech," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2013, p. 5.

[5] O. Engwall, O. Bälter, A. Öster, and H. Kjellström, "Designing the user interface of the computer-based speech training system artur based on early user tests," *Behaviour & Information Technology*, vol. 25, no. 4, pp. 353–365, 2006.

[6] A. Grossinho, S. Cavaco, and J. Magalhães, "An interactive toolset for speech therapy," in *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*. ACM, 2014, p. 36.

[7] E. Q. Rivas and E. S. Molina, "A proposal for a virtual world that supports theraphy of dyslalia," in *Proceedings of the 6th Euro American Conference on Telematics and Information Systems*. ACM, 2012, pp. 371–374.

[8] Z. Rubin and S. Kurniawan, "Speech adventure: using speech recognition for cleft speech therapy," in *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2013, p. 35.

[9] 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 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*. ACM, 2013, p. 28.

[10] M. Danubianu, S. Pentiuc, O. A. Schipor, M. Nestor, and I. Ungureanu, "Distributed intelligent system for personalized therapy of speech disorders," in *Computing in the Global Information Technology, 2008. ICCGI'08. The Third International Multi-Conference on*. IEEE, 2008, pp. 166–170.

[11] R. Speyer, H. Bogaardt, V. Passos, N. Roodenburg, A. Zumach, M. Heijnen, L. Baijens, S. Fleskens, and J. Brunings, "Maximum phonation time: variability and reliability," *Journal of Voice*, vol. 24, no. 3, pp. 281–284, 2010.

[12] 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, pp. 675–e1–4, 2012.

[13] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition.," in *INTERSPEECH*, 2007, pp. 778–781.

[14] P. J. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2866–2881, 1999.

[15] F. Destombes, B. A. G. Elsendoorn, and F. Coninx, "The development and application of the IBM Speech Viewer," 1993.

[16] C. Amorim, "A aquisição das consoantes líquidas em português europeu: contributos para a caracterização da faixa etária 4;0 - 4;11 anos," *Revista de Estudos Linguísticos da Universidade do Porto*, vol. 9, pp. 59–82, 2014.

[17] B. Schuller, S. Steidl, and Batliner, "The interspeech 2009 emotion challenge," in *INTERSPEECH*, 2009, pp. 312–315.