Development of a voice-based rhythm game for training speech motor skills of children with speech disorders

D Umanski[1], D Kogovšek[2], M Ozbič[2], N O Schiller[1]

[1]Leiden University Centre for Linguistics, Leiden Institute for Brain and Cognition (LIBC), Leiden University, THE NETHERLANDS

[2]Department of Speech Therapy and Special Education, Faculty of Education, University of Ljubljana, SLOVENIA

daniil.umanski@gmail.com

ABSTRACT

In this project, we deal with the development and evaluation of a new tool for conducting speech rhythm exercises. A training methodology is proposed, based on a schedule of exercises, each presenting a sequence of syllables arranged in a specific rhythmic pattern. In order to assist the therapists with conducting speech rhythm exercises with children, we have developed a computer game prototype which implements the training, by providing the exercises, visual feedback and evaluation of performance. The game prototype was further evaluated in a usability study involving children with various speech disorders. We discuss the limitations of the current system and propose improvements for further development.

1. PRINCIPLES OF SPEECH MOTOR PRACTICE

1.1 Speech rhythm skills

Recent studies have argued for the presence of a central timing deficit in children with speech disorders, expressed across modalities and across types of timing measures (Peter and Stoel-Gammon, 2008). Particularly, a problem with timing accuracy in speech is thought to be implicated in childhood apraxia of speech (CAS), and in dysarthria (Liss et al, 2009), conditions jointly known as speech motor disorders (Duffy, 2005). Fluency disorders have also been extensively studied from the speech motor skill perspective (Van Lieshout, 2001).

Since speech motor skills involve an intricate timing and coordination of motor events, we can describe speech rhythm skill as one of the elements comprising it. Concretely, speech rhythm skill can be defined as the ability to reproduce consecutive speech segments with a precise onset and offset timing of each segment.

1.2 Motor Learning

The work-frame of motor skill learning indicates that repeated mass practice, provided with appropriate tasks, goals, schedule and feedback, results in consolidated skills (Maas et al, 2008). As a direct consequence, it can be expected that the refinement of speech rhythm skills by means of repeated practice would lead to improved speech motor skills, and therefore empower people with speech disorders to make progress in their treatment. An insightful design of speech motor training exercises is warranted in order to achieve an optimal learning process, in terms of efficiency, retention, and transfer levels (Namasivayam and Van Lieshout, 2008).

1.3 Speech motor practice

It is now generally agreed that speech motor exercises should involve simplified speech tasks. The use of non-sense syllable combinations is a generally accepted method for minimizing the effects of higher-order linguistic processing levels (Smits-Bandstra et al, 2006). Therefore, our training methodology consists of a schedule of exercises, each presenting a sequence of syllables arranged in a specific rhythmic pattern. This is somewhat similar to a form of jazz-singing known as ‘scat-singing’ (see Figure 1 for an example). Further, the training schedule is designed, in which speech items are selected and the rhythmic complexity of the

exercises is manipulated according to speci fic principles (Maas et al, 2008).

One important principle in the design of exercises is the selection of speech items, as it has been shown that manipulating syllable sets for speech training has an affect on motor learning (Staiger and Ziegler, 2008). In an effort to accommodate this insight into our training methodology, we have constructed, in previous work, an interface for clinicians for selecting speech items based on individual articulatory needs of their patients (Umanski et al, 2010). This application allows to produce lists of syllables which can be directly utilized as speech material in speech rhythm exercises.

1.4 Current rhythm skills training system

A new genre of rhythm video games has recently emerged. In those games, players must press buttons at a precise time corresponding to a sequence dictated by the game. Doing so will cause the game's protagonist to dance or play their instrument correctly, thus achieving a greater score. Although commercial rhythm games often feature novel input devices, no voice-based games have so far been conceived and applied for training speech rhythm skills. Therefore, in this project we aim to construct a practical tool which will provide the appropriate conditions for learning and training of speech rhythm skills.

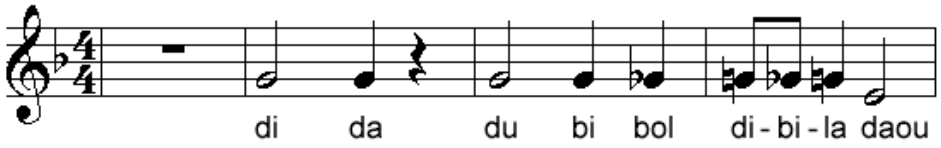

Figure 1: *Louis Armstrong's 'Heebies Jeebies' (1925)*

2. DEVELOPMENT OF A SPEECH TRAINING SYSTEM

2.1 A speech rhythm game

In order to experiment with the methodology of speech rhythm exercises, we have built a first prototype of a speech rhythm game. The theme of this game is a downhill slalom ski competition. The player uses microphone input to control the movement of a slalom skier, so that each syllable onset causes the skier to change direction on his way down the piste. The goal of the player is to produce a sequence of syllables with precise timing, so that each syllable onset corresponds with taking a well-timed turn at a flag.

Each exercise is realized in the game by positioning the flags in such a manner that their relative distances faithfully represent the rhythmic pattern to be practiced. In other words, the spatial distance between flag *i* and flag *i*+1 corresponds to the temporal note duration of element *i* in the exercise. Such spatial correspondence provides the basis for predictive feedforward control, which has been shown to support the motor initiation of rhythmic productions (Zatorre et al, 2007). Each flag is further labeled with the syllable which needs to be produced at that specfic moment in the exercise. For the sake of visual clarity, the exact points of expected syllable onsets are marked with a cross. The game follows the 'call and response' format, in which the correct rhythmic pattern is first demonstrated by means of animating the skier to perform an 'optimal' slalom ride, synchronized with an audio playback1 of the given rhythm. The player is then prompted to perform the exercise.

2.2 Adaptability

In the case when a player produces an inaccurate interval, the game will dynamically adapt the motion of the skier in order to sustain the synchrony between the next expected note duration and the visual representation of that interval. For example, if the onset is produced later than the expected moment, the trajectory of the skier towards the next flag will be animated in such a way that the skier will arrive at the next flag exactly within the next expected interval. This way, the player is not 'punished' for less then precise productions, and has a better chance to synchronize back to the correct rhythm after going 'out of sync'. Furthermore, the clinician can set the appropriate speed for a child, so that both the rhythmic pattern and the motion of the skier are scaled in time. Reducing the risk of frustration, and keeping the player at an optimal challenge level is meant to maintain motivation and flow during gameplay.

Proc. 8th Intl Conf. Disability, Virtual Reality & Associated Technologies
Viña del Mar/Valparaíso, Chile, 31 Aug. – 2 Sept. 2010

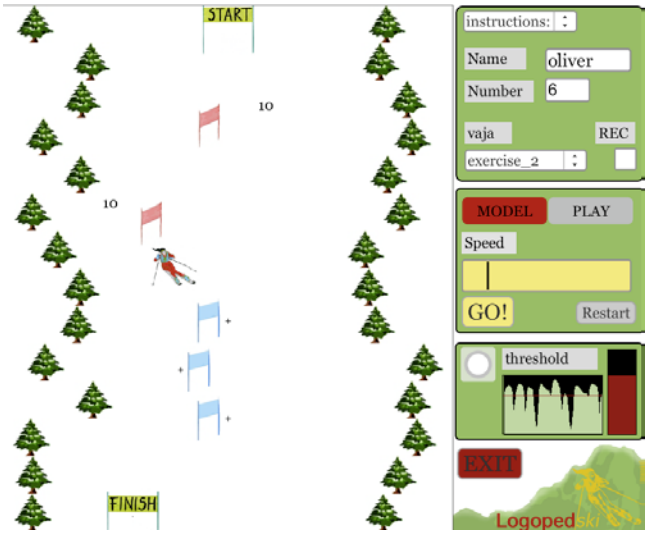

Figure 2: *A screenshot from the rhythm game prototype*

2.3 Evaluation and feedback

In order to evaluate the rhythmic pattern produced by the player in relation to the given exercise, we detect syllable onsets and calculate the inter-onset-intervals (IOI). Each interval is compared with the expected note duration for that specific syllable, and their absolute difference is obtained. The evaluation occurs in real-time, so that immediate visual feedback indicates whether the current syllable onset has been produced with accurate timing. The summing up of individual interval deviations results in an overall score for the exercise and is stored for further evaluation and progress monitoring.

2.4 Creating speech rhythm exercises

In order to offer clinicians a modular way of creating individual exercises for their patients, we have built an interface which allows the compilation of custom rhythmic sequences. The interface, illustrated in Figure 3, includes a graphic representation of note durations associated with each syllable in the exercise. As the clinician selects a sequence of note durations for a certain exercise, a graphical representation of the exercise is updated. When the exercise is compiled, it can be saved to a file, and the process repeated as long as necessary. The created exercises are automatically added to the exercise-menu of the game.

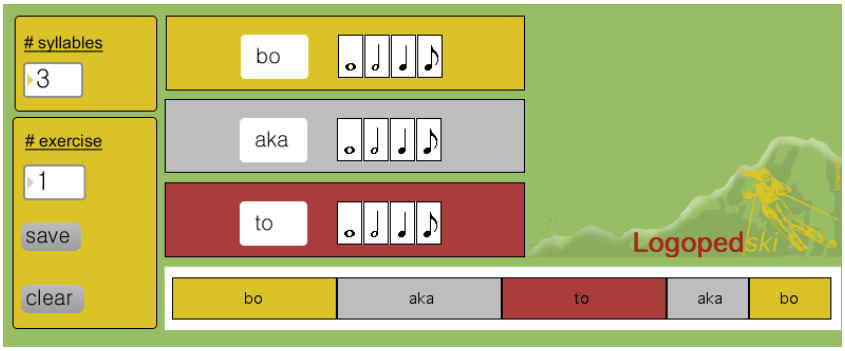

Figure 3: *A screenshot from the rhythm exercise- making interface*

2.5 Implementation

The main technical challenge in our development is the task of syllable onset detection in real-time. Because the phonetic realization of co-articulated syllables tends to blur the acoustic markers of syllable boundaries, the task of detecting each syllable becomes far from trivial. For example, semivowels are produced without complete closure of the vocal tract or frication, and have similar spectral characteristics with vowels. Therefore, a syllable such as /la/ would be difficult to distinguish from its preceding syllable. A similar problem arises with nasal consonants, such as /n/ and /m/. In general, the speech signal does not allow for a

straightforward detection of boundaries when speech is continuous.

Most research in the area of automatic onset detection focuses on real-time onset detection of musical events per se. On the other hand, the studies which investigated syllable nuclei or syllable boundary detection in speech rely on large analysis windows (around 800 ms), and are therefore not appropriate for real-time implementation (Kochanski and Orphanidou, 2008).

2.5.1 Onset detection function. In the technical literature, there are two main approaches to the onset detection task. The most straightforward detection function is based on the energy of the signal, which is associated with perceived loudness. A logarithmic energy function is considered more psycho-acoustically relevant (Klapuri, 1999). In our exercises, energy rising above a specific threshold corresponds to syllable nuclei, and therefore allows us to track the rhythmical pattern of consecutive syllables. While this function works well for separated syllables, it runs into trouble with some of the syllables produced in coarticulation, as discussed above.

Another approach is a detection function based on spectral difference, which is defined as the sum of the spectral bin magnitude differences between adjacent audio frames (Kumar et al, 2007). This function is more sensitive to variations on the phoneme level, but tends to produce more 'false onsets'. The application of this function demands a smoothing of the difference signal, and the amount of smoothing strongly affects the detection robustness. In general, both detection functions require parameters settings in order to achieve reasonable performance, and no universal settings, in terms of smoothing factors or thresholds, can be readily set up. We have experimented with both detection functions, which exhibited similar performance once parameters are manually set. We have then concluded that the energy threshold function is more appropriate to be used in the current application, since setting for energy threshold is more transparent (in terms of human-computer-interface and the user group in mind) than tweaking smoothing factors in the spectral difference function.

2.5.2 Manual parameters settings. With these considerations in mind, we have included a manual setting of the noise floor and the onset threshold levels in the interface of the game (see Figure 4). The task of the clinician is to set the appropriate levels in accordance to the acoustic situation of the training - the noise floor level according to the static noise in the room where the exercises take place, and the onset threshold according to the voice of the specific child. An indicator is added in the interface (top left corner in Figure 4), which lights up when the signal energy threshold is passed, so that the clinician can visually verify the detection function on produced syllables. Integrating these manual settings into the game reflects a somewhat problematic tradeoff inherent in all systems, which rely on microphone input in an unpredictable environment. On the one hand, for optimal usability, the system should require as least manual settings as possible, and try to adjust its internal settings automatically. On the other hand, if internal settings are not optimally set for a specific situation, the system robustness and reliability would be compromised.

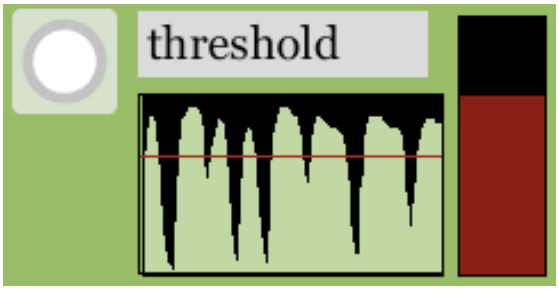

Figure 4: *Manual threshold setting*

2.5.3 Potential system improvements. One possibility in advancing the detection function is to consider a multimodal setup, in which visual input from the player is integrated with the audio input to produce better predictions of onsets. Furthermore, a learning mechanism can be employed, which learns to map acoustic features of individual players to onset detection events. In this scenario, the clinician might provide external cues to the system in the learning phase, in which appropriate mappings are stored.

3. USABILITY STUDY WITH THE RHYTHM GAME PROTOTYPE

3.1 Study goals

We have conducted an initial study to access the usability of the current prototype of the speech rhythm game. The study was performed in the School for the Deaf and Hard of Hearing in Ljubljana, Slovenia (http://www.zgnl.si). The clinicians of the school provide speech therapy not only to hard of hearing children, but also work with ambulatory patients with a wide variety of speech and language disorders, including stuttering, speech motor disorders, phonology and articulation disorders. The goal of the study was to perform an initial evaluation of the suitability of a speech rhythm game as a valuable tool for speech therapist's work. In addition, we wished to collect the impressions of clinicians as to which target groups of children the proposed game could eventually target, in terms of diagnosed speech disorders and age.

3.2 Procedure

During the study period of six weeks, clinicians used the game to practice speech rhythm skills with patients for whom such training made sense from the therapy program point of view. The guideline for the clinicians was to perform the training with children having various disorders and various ages in order to get a better overall impression on the usability aspects of the system. At the end of the evaluation period, clinicians were presented with a questionnaire which included 10 items aiming at evaluating various aspects of the training experience. The questionnaire was adapted from Öster (2006) for evaluating the usability of the 'Box of Tricks' software. For each item, the clinicians chose the most appropriate answer on a 4-point scale, and added their own comments on that question, where appropriate. In total, 6 clinicians, having worked with 26 children have completed the study period. Each child completed 2-4 practice session with the game, on separate occasions, each lasting for 10-15 minutes. During each practice session, the clinicians were free to choose the rhythmical exercises which best fit to the level of the child, as well as the syllables, which are to be produced, in accordance with the child's abilities.

3.3 Subjects

The group of children included 7 girls and 19 boys, aged 4-10, with an average age of 6.6 years. Among these children, 9 have articulation problems, 7 stutter, 7 have an expressive language disorder, 2 children are hard of hearing, one is diagnosed with Asperger's syndrome, and one has a cleft palate (among these subjects, a few children are diagnosed with multiple disorders).

3.4 Results and discussion

In this section, we report and discuss the experiences and impressions expressed by the clinicians in relation to each of the items in the usability questionnaire (see Figure 4 on the next page for charted results):

1) Concerning previous computer knowledge required for using the game, the replies suggest that among the clinicians who participated, this issue presented no obstacle. Additionally, an interesting insight comes from the following fact: Initially, eight therapists engaged in the usability study. However, two have stopped shortly after, as a result of being uncomfortable with using a computer-based system. These two therapists are the eldest in the team, and their lack of comfort with using the computer is an important factor for consideration. Development of software for speech clinicians as the target group might need to face the choice between applying specific usability principles for the 'technology-intolerant' sub-group, such as older generations of clinicians, or targeting the younger generations of clinicians per se.

2) The question whether children understood the visual feedback provided by the game remains inconclusive. Clinicians report that while children understood that they have to 'say sounds' in order to make the skier turn, as well as the changing of the colors of the flags, some did not understand the point of the numbers presented on the screen. It might be the case that for the younger children, presenting the score in a numerical form is too complex, and a more simplified visualization scheme is needed.

3) Concerning the motivation of children to practice with the game, the clinicians report, rather unanimously, that children were motivated to practice only during the first few attempts, and their enthusiasm faded after a few trials. This is not surprising if considering the importance of variation in game design and the rather static nature of the current game prototype, which features one screen only. For children who regularly play modern computer games, the current setup is not able to satisfy the variation standards they are used to. In future work, we will need to provide a game environment which optimally preserves the flow of game-play, by including variable graphical themes, while controlling for adaptable levels of challenge.

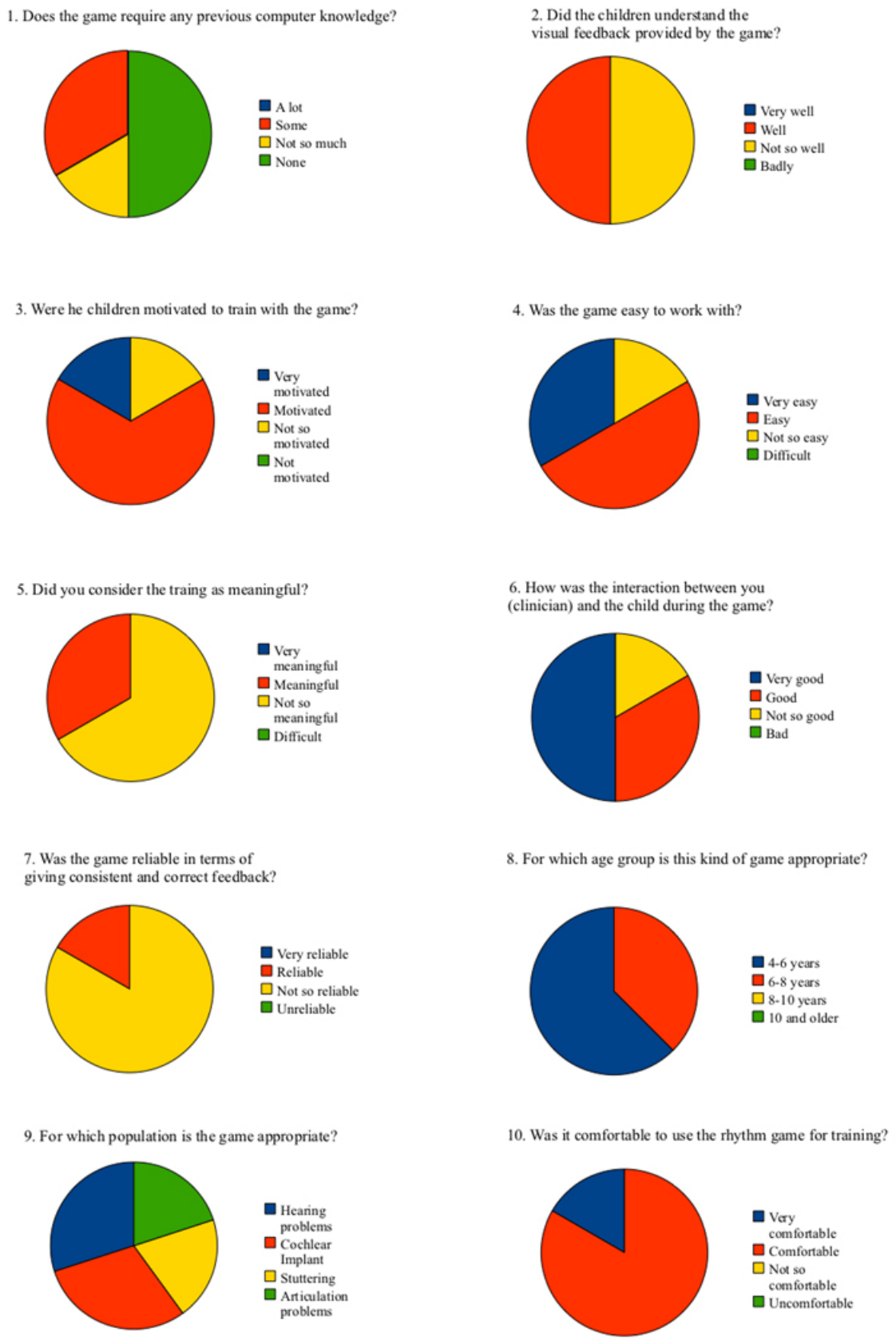

Figure 5. *Results of the usability questionnaire.*

Proc. 8th Intl Conf. Disability, Virtual Reality & Associated Technologies
Viña del Mar/Valparaíso, Chile, 31 Aug. – 2 Sept. 2010

4) As for the game being easy to use for training, although clinicians tend to regard it as being generally so, they do report difficulties with using the manual controls for noise floor level and for the syllable onset detection. This difficulty is understandable, and reflects the inherent tradeoff between the amount of manual settings and system robustness in microphone based applications (as discussed in the 'Implementation' section). This problem demands a special attention when designing applications for speech therapists, as any mechanism less than intuitive could compromise the usability of the system.

5) For the question of the speech rhythm training being meaningful, no clear conclusion was derived. As a critic, one clinician mentioned that other speech skills than rhythm can better be trained with the game, meaning that rhythm per se is not the main element which the children focused on. Another clinician reported that there were too many different rhythms with little difference between them. The lack of graphical variation has also been reported as diminishing the meaningfulness of the training, probably due to its influence on the motivation to practice.

6) The interaction between the child and the clinician during practice have been positively regarded. Clinicians mentioned that children were willing to cooperate with them in the process of playing the game, and that children were ready to listen to instructions and guidelines while playing. Only one clinician reported some children being more occupied with the game than listening to her.

7) The reliability of the game in terms of providing consistent feedback has been observed as not enough reliable. Clinicians report that consistent feedback depends on the manual noise and threshold settings, as well as on the speech items being produced. These observations confirm our discussion in section 'Implementation', where the robustness issues of syllable onset detection are discussed. It is clear therefore, that future versions of the speech rhythm game must include a more robust mechanism of detecting beats in the speech signal.

8) As for the age group most suitable for using the speech rhythm game, clinicians seem to agree that 4-6 years old can best appreciate this form of training. Children who are 6-8 years old are also found to be involved with the game, but to a lesser degree.

9) Concerning the target populations for whom a speech rhythm training is beneficial, clinicians mentioned all the disorders found in the group of children involved in the study. Mostly, articulation problems and hearing problems were pointed out, although stuttering and children with cochlear implants are also believed to gain from this training.

4. CONCLUSIONS

In this paper, we have presented a new method for conducting speech rhythm exercises. The theoretical background and the clinical motivation for developing a computer-based intervention for speech rhythm training were outlined and discussed. We have then described a game prototype which aims to implement the proposed methodology. The speech rhythm game provides a training platform in which each exercise is visually presented, and the rhythmical pattern needs to be matched by the player through producing syllable sequences in accurate timing. The mechanisms of the game prototype have been detailed, which allow for adaptability, evaluation of performance, visual feedback and score presentation. Furthermore, a special interface which allows clinicians to create speech rhythm exercises has been described. The system implementation revolves around the non trivial task of detecting syllable onsets in real-time. The difficulties and approaches to this task have been suggested.

In order to evaluate the usefulness of the proposed approach and the current game prototype, a usability study was conducted. During this study, clinicians have been using the speech rhythm game with children diagnosed with a variety of speech disorders. In general, clinicians report to have been comfortable with using the speech rhythm game for training with children. They have voiced a number of significant critics on the current prototype of the game. The main limitations at this point seem to be the lack of robustness on detecting syllable onsets, and the lack of graphical variation during game-play. While graphical variation is a matter of extension and elaboration, the problem of robust onset detection will require extensive development, as the state of the art techniques in this field do not yet offer a satisfying solution.

In spite of outlining a number of critics on the usability of the current game prototype, clinicians have expressed a positive attitude towards this line of development. They have concluded that although the game prototype requires improvement, the initiative is very welcome, and further prototypes will be anticipated. The outcomes of the usability study have equipped us with a deeper understanding of the challenges and

difficulties involved in developing this method of training, but also of its potentials. We will be concerned with advancing this development further on.

Acknowledgements: This research is supported with the 'Mosaic' grant from The Netherlands Organisation for Scientific Research (NWO). The authors are grateful for the anonymous reviewers for their constructive feedback.

5. REFERENCES

J R Duffy (2005), Motor speech disorders: Substrates, Differential Diagnosis, and Management, *(2nd Ed.) 507-524. St. Louis, MO: Elsevier Mosby.*

A Klapuri (1999), Sound onset detection by applying psychoacoustics knowledge, *in Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Arizona, pp. 3089-3092.*

G Kochanski, C Orphanidou (2008), What marks the beat of speech? *J. Acoustical Society of America 123, 2780–2791.*

J M Liss, L White, S L Mattys, K Lansford, A J Lotto, S M Spitzer, J Caviness (2009), Quantifying speech rhythm abnormalities in the dysarthrias, *Journal of Speech, Language, and Hearing Research Vol. 52 1334–1352.*

E Maas, D A Robin, S N Austermann Hula, S E Freedman, G Wulf, K J Ballard, R A Schmidt (2008), Principles of Motor Learning in Treatment of Motor Speech Disorders, *American Journal of Speech-Language Pathology, 17, 277-298.*

A K Namasivayam, P Van Lieshout (2008), Investigating speech motor practice and learning in people who stutter, *Journal of Fluency Disorders 33 (2008) 3251.*

A M Oster (1996), Clinical applications of computer-based speech training for children with hearing impairment, *Proc. ICSPL 1996, 157-160. Philadelphia, USA.*

B Peter, C Stoel-Gammon (2008), Central timing deficits in subtypes of primary speech disorders, *Clinical Linguistics & Phonetics, 22(3): 171–198.*

P Pradeep Kumar, P Rao, S Dutta Roy (2007), Note Onset Detection in Natural Humming. *iccima, vol. 4, pp.176-180, International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007).*

S Smits-Bandstra, L F DeNil, J Saint-Cyr (2006), Speech and non-speech sequence skill learning in adults who stutter, *Journal of Fluency Disorders, 31,116136.*

A Staiger, W Ziegler (2008), Syllable frequency and syllable structure in the spontaneous speech production of patients with apraxia of speech, *Aphasiology, Volume 22, Number 11, November 2008 , pp. 1201-1215(15).*

D Umanski, F Sangati, N O Schiller (2010), How spoken language corpora can refine current speech motor training methodologies, *Proc. ACL2010 Intl. Conf..*

P Van Lieshout (2001), Recent developments in studies of speech motor control in stuttering, *In B. Maassen, W. Hulstijn, R. D. Kent, H. F. M. Peters, P. H. H. M. Van Lieshout (Eds.), Speech motor control in normal and disordered speech (pp. 286290). Nijmegen, The Netherlands:Vantilt.*

R J Zatorre, J L Chen, V B Penhune (2007), When the brain plays music: auditory-motor interactions in music perception and production, *Nature Reviews Neuroscience 8, 547-558.*