A Game with a Purpose for Automatic Detection of Children's Speech Disabilities using Limited Speech Resources

Reem Elhady, Mohamed Elmahdy, Injy Hamed and **Slim Abdennadher**
German University in Cairo, Cairo, Egypt
Faculty of Media Engineering and Technology
reem.elhady@student.guc.edu.eg,
{mohamed.elmahdy,injy.hamed,slim.abdennadher}@guc.edu.eg

Abstract

Speech therapists and researchers are becoming more concerned with the use of computer-based systems in the therapy of speech disorders. In this paper, we propose a computer-based game with a purpose (GWAP) for speech therapy of Egyptian speaking children suffering from Dyslalia. Our aim is to detect if a certain phoneme is pronounced correctly. An Egyptian Arabic speech corpus has been collected. A baseline acoustic model was trained using the Egyptian corpus. In order to benefit from existing large amounts of Modern Standard Arabic (MSA) resources, MSA acoustic models were adapted with the collected Egyptian corpus. An independent testing set that covers common speech disorders has been collected for Egyptian speakers. Results show that adapted acoustic models give better recognition accuracy which could be relied on in the game and that children show more interest in playing the game than in visiting the therapist. A noticeable progress in children Dyslalia appeared with the proposed system.

1 Introduction

Speech and language disorders can affect a persons ability to talk, understand, read, write and express himself/herself. In this study, we are concerned with Dyslalia speech disorder. Dyslalia is an articulatory disorder in which children, or adults, do not pronounce the sounds clearly, sounds are changed or distorted or they replace one sound for another, e.g. a person may have a lisp use of the /T/ instead of the /s/ sound[1]. It is a result of having the sound pronounced from an incorrect part of the vocal tract. It may be also due to delayed speech, hearing impairment or learning disability. Mental retardation can also cause Dyslalia.

In the process of therapy, speech therapists use a variety of strategies including oral motor therapy, articulation therapy, and language intervention activities. In the step of oral motor therapy, the therapist uses a variety of oral exercises, including facial massage and various tongue, lip, and jaw exercises, to strengthen the muscles of the mouth. The therapists physically show the child how to make certain sounds, such as the /r/ sound, and may demonstrate how to move the tongue to produce specific sounds. During the language intervention activities, the therapist interacts with a child by playing and talking. They may use pictures, books, objects, or ongoing events to stimulate language development. The therapist may also model correct pronunciation and use repetition exercises to build speech and language skills.

Our proposed system plays the role of language intervention activities. Therapists not only have to

[1]Throughout the paper, SAMPA notation is used for phonetic transcriptions (Wells, 2002).

Proceedings of Natural Language Processing and Information Retrieval Workshop, pages 28–34,
Varna, Bulgaria, Sep 7 2017.
https://doi.org/10.26615/978-954-452-038-0_004

provide a variety of materials for different tasks for each therapy session, but they also have to keep a record of the childs performance during the tasks. Moreover, the therapist usually has problems to manage the recording details of these sessions for further analysis and to prepare appropriate materials that address the required treatment process.

Several studies mention the importance of computer-based systems that aim at supporting such therapies (Murray and Parker, 2004; Cagatay et al., 2012; Tobolcea and Danubianu, 2010). All of these studies have focused on Latin languages like English and Romanian.

There were specific criteria chosen in the proposed game design (Koster, 2013) for best effect: attractiveness, curiosity, immediate and accurate feedback, the issue of control, challenge feeling, and automatic system adaptation to user's performance. The experiments were done on the most common problems in the Arabic Egyptian language which are the replacement of the (س) /s/ phoneme to (ش) /S/, (ث) /T/, (د) /d/, or (خ) /x/ ;and the replacement of (ر) /r/ phoneme to (ل) /l/, (ي) /i:/ , or (غ) /G/ (Danubianu et al., 2009; Kotby M, 1979).

2 Speech Engine

In order to have a game that improves child Dyslalia, there should be first a speech engine that can accurately and automatically detect the problem in the child pronunciation, then the game shall take its role. So, we have worked separately on speech engine, and after we had successfully ensured recognition accuracy of the engine, we have merged it to an attractive game to test the effect of the game therapeutically. A major problem in Arabic speech recognition is the existence of quite many different Arabic dialects. Every country has its own dialect and sometimes there exist different dialects within the same country. There are many speech data resources for MSA, but unfortunately, the available resources for dialectal Arabic are very limited. That is why there are only limited researches done in the area of dialectal Arabic speech recognition. In this paper, we are proposing a cross-lingual acoustic modeling approach for dialectal Arabic, where we can benefit from existing MSA speech resources, in order to improve dialectal Egyptian Arabic recognition rate.

2.1 Speech corpora

2.1.1 MSA corpus

The existing MSA speech resource we used is the Nemlar news broadcast speech corpus. It was chosen in training MSA acoustic models (Yaseen et al., 2006). The corpus consists of 33 hours of MSA news broadcast speech. The broadcasts were recorded from different radio stations. All files were recorded in linear PCM format, 16 kHz, and 16 bit. The total number of speakers is 259 and the lexicon size is 62K distinct words with a phoneme set of 34 phonemes. This corpus was mainly selected because the transcriptions are fully vowelized and manually reviewed, and hence there were accurate phonetic transcriptions. The Nemlar corpus excluded speech segments with music or noise in the background. Cross-talks and segments with truncated words were excluded as well.

2.1.2 Egyptian Arabic Corpus

In order to adapt MSA and make it Egyptian dialect dependent and hence improve the recognition rate, we have used a previously collected Egyptian corpus (Elmahdy et al., 2010); a database of most frequently used words and utterances. The database includes utterances from different speech domains like greetings, time and dates, words spelling, restaurants, train reservation, Egyptian proverbs, etc. The diversity of speech domains ensures good coverage of acoustic features. A lexicon of 700 words was used with accurate phonetic transcription using the dictionaries (Hinds and Badawi, 2009) and (Stevens and Maurice, 2005). The total number of speakers is 22 native Egyptian speakers with tri-phones coverage of 15K distinct tri-phones. Every speaker was prompted to read 50 utterances chosen randomly from the database. All recordings were performed in linear PCM, 16 kHz, and 16 bits. The Egyptian corpus was used as a training set to train the Egyptian baseline acoustic model and in adapting existing MSA acoustic model.

2.1.3 Experimental Testing Corpus

Since the main objective is to test if the implemented engine detects the defected phonemes or not, subjects for this experiment were adults and children. This was intended to make sure that the engine works generic on all people suffering from Dyslalia not only children. We had ten subjects, six adults; three males and three females, and four

children; three girls and one boy, aged from 7 to 10 years old.

Subjects were males and females who responded to a general request for participation in an Automatic Speech Recognition experiment. All subjects were normal people without any speech problem diagnosed. There were no any age or gender restrictions. The reason why the chosen subjects were not suffering from dyslalia, is the difficulty of exactly identifying the wrong phoneme pronunciation since the patient may have a problem in more than one phoneme in the word. Our strategy was to test specific phoneme each time, so that would have given us wrong indication.

The tool being used for speech collection is Audacity; a free audio editor and recorder. The recording was in a closed room free from any noise. Subjects used a Microsoft LifeChat LX-3000 microphone while recording for better and clear audio files. Files are recorded in PCM, mono channel, and sampling frequency of 16 kHz. Each subject was asked to record specific words that consist of the (س) /s/ and (ر) /r/ phonemes, where the phoneme is placed in the beginning, middle and ending of the word. The subjects were asked to record each of these words several times, but with a replacement of the correct phoneme to its Dyslalia replacements. These were the data that have been used for testing recognition accuracy. Each of the collected recorded words was associated with the actual phonetic transcription and the expected correct phonetic transcription. This allows us to identify whether the word is pronounced correctly, or the Dyslalia replacement in case of wrong pronunciation. An example from the collected testing set is shown in Table 1 where the two words سماعه and بسكويت are pronounced correctly with the /s/ phoneme (/sma?:h/ and /bskawi:t/ respectively), and all common Dyslalia replacements with phonemes /S/, /T/, /d/, and /x/.

Correct	سماعه	بسكويت
س /s/	/sma?:h/	/bskawi:t/
ش /S/	/Sma?:h/	/bSkawi:t/
ث /T/	/Tma?:h/	/bTkawi:t/
د /d/	/dma?:h/	/bdkawi:t/
خ /x/	/xma?:h/	/bxkawi:t/

Table 1: Sample from the subjects' experiment for the Phoneme (س) with all possible Dyslalia replacements

2.2 Adaptation and Results

The whole amount of the MSA corpus was used to train the MSA acoustic model with a typical number of tied-states and Gaussians of 3,000 and 8 respectively. The MSA acoustic model has been adapted in order to make it dialect-dependent and hence improve the recognition rate. The MSA acoustic model was adapted using the Egyptian training set along with the normalized transcriptions. CMU Sphinx has been used in this work (Elmahdy et al., 2012). Below are the three adaptation techniques that were evaluated. In all the adaptation techniques, we compared word recognition accuracy, phoneme recognition accuracy, and their normalization recognition accuracy. This comparison is calculated using the process of force-alignment which takes an existing transcript and finds out which, among the many pronunciations for the words, or each phoneme in the word occurring in the transcript, are the correct pronunciations. For the phoneme, the output is written into a file with .phsegdir file name extension in sphinx3_align and it contains each phone start and end positions in terms of frames on time scale (100 frames per second) along with the log likelihood acoustic spectral match score. For the whole word, the output is written into a file with .wdsegdir file name extension in sphinx3_align and it contains also the word start and end positions in terms of frames on time scale along with large negative acoustic spectral match score. For the normalization, the match score of the targeted phoneme is divided by the match score of the whole word. We have used Confusion Matrix (CM) method for Word, Phoneme and their normalization as well in testing the results. The idea in the three scenarios is that we get the final match score; e.g. the Phoneme CM, for each word tested, we get the score of the phoneme we want to test with its correct audio reference and start to compare it with other known defects of this phoneme. The lowest score in a row is the best match, and hence the replaced phoneme is detected. Table 2 shows a sample of phoneme CM confidence scores where the word is detected as correctly pronounced if the largest log likelihood value is on the diagonal.

2.2.1 Maximum A-Posteriori (MAP) Adaptation

As shown in Tables 3 and 4, the MAP adapted model resulted in 86.2% recognition accuracy in case of (س) /s/ phoneme-based CM acoustic modeling, and 83.1% recognition accuracy in case of (ر) /r/ phoneme-based CM acoustic modeling, which are actually worse than the baseline with 3% and 4% absolute. This result was almost predictable since MAP adaptation requires large data set for adaptation which was not the case in this experiment.

2.2.2 Maximum Likelihood Linear Regression (MLLR) Adaptation

As shown in Tables 3 and 4, MLLR adaptation was found to give better results when adapting all acoustic model parameters: Gaussian means, variances, mixture weights, and transition weights. In the case of phoneme-based CM acoustic modeling, the adapted MSA model performed recognition accuracy of 92.3% in case of (س) /s/, and 91.2% in case of (ر) /r/ which are actually better than the baseline with 3% and 4% absolute.

2.2.3 Combined MAP & MLLR Adaptation

The combination of MAP and MLLR resulted in the best recognition accuracy As shown in Tables 3 and 4. In the case of phoneme-based CM acoustic modeling, the adapted MSA model performed recognition accuracy of 93.4% in case of (س) /s/ as shown in Table 3, and 95.6% in case of (ر) /r/ as shown in Table 4 which are actually 4% and 8% absolute increase compared to the baseline.

3 "Kalemni Aktar"

"Kalemni Aktar" ("Talk to me more") is our proposed web GWAP that is used mainly to help Dyslalia children improve their Dyslalia. It also works as a tool for therapists to monitor their patients progress.

3.1 Game Design

"Kalemni Aktar" is a web game application; it is either a one player game or a two player game for the sake of competition. There are three main interfaces; Player, Physician, Admin. The game is divided into three rounds. In the player interface, once the player is logged in and the game starts, he/she chooses a theme to continue the game with

Figure 1: *Kalemni Aktar* ("Talk to me more") GWAP graphical user interface

by selecting between the following: Zoo, Kitchen, or Around the world. In the 1st round of the game, the player is directed to all the Arabic Phonemes to pick a phoneme out of the Arabic 28 phonemes. A round of random three words with the selected phoneme on the beginning, middle, and end appear to the player in bubbles. The player must click on the 3 bubbles but in any order he/she wants, to explode with an image, text, and a stored audio of the target word. For each round, there is a timer of 180 seconds set. This is done to increase the challenge for the players. Within this interval of time, it is possible to have 5 trials for each word. For each trial, the system detects if the phoneme in the word is pronounced correctly or not. If it is correct, the system automatically moves to the next word in the round, else the system recognizes the said word as a feedback to the player to help him in the other 4 trials. In Fig. 1 the player should have said سمكه /smkh/ (word appearing in green color), but the player said ثمكه /Tmkh/ (word appearing in red color) instead.

Since one of the incentives in GWAPs that engage and motivate players is the score (Von, 2006). The score for each round is calculated as follows:

- If the word is pronounced correctly from the first trial, the player gets 10 Points.
- If the word is pronounced correctly from the second trial, the player gets only 8 points.
- If the word is pronounced correctly from the third trial, the player gets only 6 points.

Table 2: Sample for phoneme confusion matrix confidence score.

Audio-Text	سماعه /sma?:h/	شماعه /Sma?:h/	ثماعه /Tma?:h/	دماعه /dma?:h/	خماعه /xma?:h/
سماعه /sma?:h/	-2269102	-2514996	-2327558	-2453191	-2482886
شماعه /Sma?:h/	-2427326	-2319749	-4686261	-2399874	-2325771
ثماعه /Tma?:h/	-2262885	-2322751	-2222441	-2263326	-2266508
دماعه /dma?:h/	-2310795	-2370847	-2233933	-2231769	-2278456
خماعه /xma?:h/	-2247120	-2301651	-2480661	-2224442	-2110647

Table 3: Recognition accuracy (%) for the different adaptation techniques for (س) /s/ phoneme CM results of the subjects' collected data.

Technique	Phone	Word	Norm.
Baseline	89.2	86.1	90.4
MAP	86.2	82.1	88.1
MLLR	92.3	86.1	92.3
MAP + MLLR	93.4	89.6	92.4

Table 4: Recognition accuracy (%) for the different adaptation Techniques for (ر) /r/ Phoneme CM results of the subjects' collected data.

Technique	Phone	Word	Norm.
Baseline	87.1	86.3	86.2
MAP	83.1	82.2	83.2
MLLR	91.2	89.1	89.4
MAP + MLLR	95.6	93.1	94.2

- If the word is pronounced correctly from the fourth trial, the player gets only 4 points.
- If the word is pronounced correctly from the fifth trial, the player gets only 2 points.
- Finally, if after the 5 trials the player still pronounces the word incorrect, he/she does not achieve any point.

After each round, a feedback appears to the player with the achieved score. This is repeated till all phonemes are pronounced. At the end of this level, a summarized report appears to the player with the feedback of the detected phoneme problems of all phoneme rounds. The 2nd round is the automatic system adaptation to the performance in the 1st level. It consists of some exercises with random words on those detected phoneme issues. The 3rd round is a tongue twister, a proficiency level stressing on certain phoneme which evaluates several words at the same time. In the physician Interface, speech therapists have access to their patients profiles to monitor the progress of the cases they have; so that they can assist in the sessions. For the Admin Interface, it is a basic interface for managing game content. By going through the game, all proven important game design factors from attractiveness, curiosity, immediate and accurate feedback, the issue of control, challenge feeling, and automatic system adaptation to user performance were covered (Umanski et al., 2008).

3.2 Subjects and Experiment

Subjects for this game were twenty children; ten boys, and ten girls aged from seven years old to ten years old. They are patients in Ain Shams Specialized Hospital. The subjects were chosen suffering from same level of Dyslalia specifically from the phonemes (س) /s/ and (ر) /r/. The twenty children were not able to pronounce the targeted phoneme correctly. Subjects were divided into two groups; each group consists of five boys, and five girls. The 1st group was having basic normal sessions with the speech therapist. The 2nd group was introduced to the implemented game in the session. The sessions were held for thirty to sixty minutes twice a week. The duration of the experiment was two months. The experiment had two main targets. The 1st target was to detect the accuracy of the game to see whether it could be really relied on for detecting the defected phonemes with its wrong replacement or not. The 2nd target was to compare the progress of the child speech when playing the game than when having normal sessions with the speech therapist. Throughout the two months experiments, the subjects of the two groups were monitored and interviewed about their feedback, level of interest and motivation.

3.3 Experimental Results

All subjects were introduced to a new speech therapist, who hasn't been involved earlier in the experiment, to test the children's level of dyslalia in the targeted phonemes after the two months therapy. This was to make sure that results will be double blinded. After gathering all test results for the two groups, results showed that both groups were been able to develop the (س) /s/ and (ر) /r/ phonemes better compared to the beginning of the experiment; however the 2nd group determine faster progress in speech than the 1st group does.

For the phoneme (س) /s/, nine out of the ten children of the second group were able to pronounce the (س) /s/ correctly, however one child still pronounces (ث) /T/ instead. In comparison to the first group where only seven out of the ten children were able to pronounce it correctly as shown in Fig. 2. For the phoneme (ر) /r/ , nine out of the ten children of the second group were able to pronounce the (ر) /r/ correctly, however one child pronounced (ي) /?:/ instead. In comparison to the first group where only six out of the ten were able to pronounce it correctly as shown in Fig. 3.

The children in the second group reported very good feedback, they were interested in using the application and switching between different themes, and phonemes. They found the interfaces friendly with suitable colors, and the characters of the game attracted their attention. Some mentioned that they were very curious to explore the different levels, and to achieve the highest score possible. While the children in the 1st group reported quite negative feedback compared to the others, some mentioned that they got bored in the therapy, some wanted to return home, and some got unmotivated throughout the session.

The use of such computer-based methods during various phases of the speech therapy determines a new psychological and pedagogical situation by creating a special interesting learning environment, and by facilitating a new, superior method for correcting and developing speech.

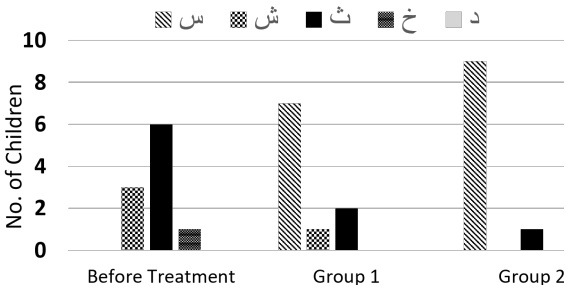

Figure 2: The progress after two months for phoneme (س) /s/

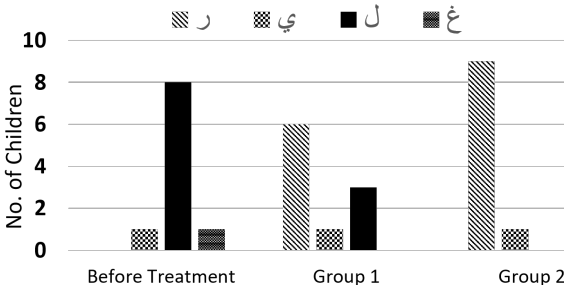

Figure 3: The progress after two months for phoneme (ر) /r/

4 Conclusion and Future Work

We have presented a speech recognition based system for Dyslalia children called Kalemni Aktar. The aim is to provide assistance to Dyslalia children to improve their speech. Results showed that Kalemni Aktar reached its goal of providing a suitable and useful environment for the Dyslalia child to develop his/her speech. It has been shown that MAP + MLLR combined adaptation technique has best recognition results with accuracy reached 93.4% for (س) /s/ phoneme and 95.6% for (ر) /r/ phoneme.

"Kalemni Aktar" turned out to be a real help in the therapeutic activity, by providing various exercises that children can do both in the sessions and at home. Computer games are a powerful tool for motivating children to practice speech motor skills. The existence of computer based methods cannot replace the therapists role, but it only helps them in the therapy and helps the children having more exercises at home to develop their speech. The experiments showed high attention and concentration levels of children who practiced, as well as improvement in performance in terms of the game scores. The visual environments used in the prototype game proved to be easy for children to relate to, however more variety is needed to

sustain curiosity.

For future work, we recommend upgrading the application by including all other Dyslalia types. The game may also include detection of other speech problems as stuttering to be fully integrated software for all speech disorders. It is also important to evaluate the application by involving more Dyslalia children in the testing phase, this will help to adapt the application according to their assessed needs. This Game can be of great importance to some nonprofit organizations in the sake of improving the society.

References

Mehmet Cagatay, Pinar Ege, Gul Tokdemir, and Nergiz Ercil Cagiltay. 2012. A serious game for speech disorder children therapy. In *7th International IEEE Symposium on Health Informatics and Bioinformatics (HIBIT)*. pages 18–23.

Mirela Danubianu, Stefan gheorghe Pentiuc, Ovidiu Andrei, Schipor Marian, Nestor Ioan, Ungureanu Doina, and Maria Schipor. 2009. Terapers- intelligent solution for personalized therapy of speech disorders .

Mohamed Elmahdy, Rainer Gruhn, and Wolfgang Minker. 2012. *Novel techniques for dialectal Arabic speech recognition*. Springer Science & Business Media.

Mohamed Elmahdy, Rainer Gruhn, Wolfgang Minker, and Slim Abdennadher. 2010. Cross-lingual acoustic modeling for dialectal Arabic speech recognition. In *INTERSPEECH*. pages 873–876.

Martin Hinds and El-Said Badawi. 2009. *A Dictionary of Egyptian Arabic*. Librairie du Liban.

Raph Koster. 2013. *Theory of fun for game design*. O'Reilly Media, Inc.

N Barakah M Kotby M. 1979. Patterns of dyslalia in Egypt .

Tiffany G Murray and V Parker. 2004. Integration of computer-based technology into speech-language therapy. *Educational Technology* 31:53–59.

Stevens and Salib Maurice. 2005. *A Pocket Dictionary of the Spoken Arabic of Cairo*. The American University in Cairo Press, Second printing.

Iolanda Tobolcea and Mirela Danubianu. 2010. Computer-based programs in speech therapy of dyslalia and dyslexia-dysgraphia. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 1(2):52–63.

Daniil Umanski, Walter A Kosters, Fons J Verbeek, and Niels O Schiller. 2008. Integrating computer games in speech therapy for children who stutter. In *First Workshop Child, Computer and Interaction (WOCCI)*. pages 17–21.

Ahn Von. 2006. Games with a purpose 39(6):92–94.

JC Wells. 2002. SAMPA for Arabic. www.phon.ucl.ac.uk/home/sampa/arabic.htm.

Mustafa Yaseen, M Attia, Bente Maegaard, Khalid Choukri, N Paulsson, S Haamid, Steven Krauwer, C Bendahman, Hanne Fersøe, M Rashwan, et al. 2006. Building annotated written and spoken arabic LRs in NEMLAR project. In *Proceedings of LREC*. pages 533–538.