Linnæus University

Sweden

Degree Project

Real-time interactive visualization aiding pronunciation of English as a second language

Author: Dorina Dibra
Supervisor: Nuno Otero
Co-supervisor: Oskar Pettersson
Examiner: Marcelo Milrad
Date: 2013-10-11
Subject: Media Technology
Level: Master
Course code: 5ME10E

Abstract

Computer assisted language learning (CALL) comprises a wide range of information technologies that aim to broaden the context of teaching by getting advantages of IT. For example, a few efforts have been put on including a combination of voice and its visual representation for language learning, and some studies are reporting positive outcomes. However, more research is needed in order to assess the impact caused by specific visualization styles such as: highlighting syllables and/or wave of sound. In order to explore this issue, we focused at measuring the potential impact that two distinct visualization styles and its combination can have on teaching children the pronunciation of English as a second language. We built a prototype which was designed to assist students while learning pronunciation of syllables. This system was employing two different real-time interactive visualization styles. One of these visualization styles utilizes audio capturing and processing, using a recent technology development: Web Audio API.

We evaluated the effect of our prototype with an experiment with children aged from 9 to 11years old. We followed an experimental approach with a control group and three experimental groups. We tested our hypothesis that states that the use of a combined visualization style can have greater impact at learning pronunciation in comparison with traditional learning approach.

Initial descriptive analyses were suggesting promising results for the group that used the combined visualization prototype. However, additional statistical analyses were carried out in order to measure the effect of prototype as accurately as possible within the constraints of our study. Further analyses provided evidence that our combined visualizations prototype has positively affected the learning of pronunciation. Nonetheless, the difference was not big comparing to the system that employed only wave of sound visualization. Ability to perceive visual information differs among individuals. Therefore, further research with different sample division is needed to determine whether is the combination of visualizations that does the effect, or is the wave in itself. Slitting groups based on this characteristic and perform the testing will be considered for the future research.

Eventually, we can be confident to continue exploring further the possibility of integrating our proposed combination of two visualization styles in teaching practices of second language learning, due to positive outcomes that our current research outlined. In addition, from a technological perspective, our work is at the forefront of exploring the use of tools such as Web Audio API for CALL.

Keywords

Second language learning (L2 learning), Technology enhanced language learning (TELL), Computer Aided Language Learning (CALL), Visualization of sound wave, highlighting syllables, pronunciation teaching, visualization of pronunciation, Web Audio API.

Acknowledgements

On my journey as a master student I have had beside me a lot of people, who always will have my appreciation and gratitude.

As I hear people complimenting me about my work, I feel as an obligation and strong desire to thank all those who have stayed close to my academic development. I am deeply grateful to Professor Arianit Kurti, who tirelessly guided me throughout all my master studies. I could not have imagined more valuable support. I own very important debt to my supervisor Dr. Nuno Otero, who diligently supervised my research all the time. His constructive and meticulous feedback encouraged me to work always more starting from refinement of initial idea to the completion of this thesis. His guidance both so constructive and so inexplicit at the same time has inspired me to become creative and capable of conducting a research. I am grateful for the assistance to Oskar Pettersson. His valuable guidance helped me complement my work which has been complimented while ongoing. I am convinced and I appreciate a lot that my supervisor's desire to guide throughout the realization of this thesis has been larger than the obligation to observe my work. Their guidance raised my confidence, in lack of which this thesis would not be possible. Professor Marcelo Milrad is among the tutors whose advice helped me to become more productive and to complement my thesis. Words cannot express the gratitude and appreciation for my professors, compared to what I learned from them.

I would like to highlight among these paragraphs the support of my family. Their advices and faith in me, made me realize that with great desire and work can be achieved all desired success. My brothers' jokes were the best medicine every time I felt tired. Someone who has always cheering me up and has been along with me in this was my fiancé. He has the merits for teaching me to go through long years of education with serenity and patience. I feel blessed to be surrounded by such people on my educational journey who together with my friends' phone calls made me feel loved and supported.

Finally, I want to express my gratitude to the managers of the education institute in Debar, Macedonia and professors of English language: Vjollca Dibra, Natyra Greva, Elona Cami, Fatjona Pocesta, Ardit Fetahu together with their lovely students, who accepted to cooperate and helped us to realize this project.

*To my parents,
for their endless love, support and encouragement.*

*Dedikuar prindërve të mi,
për dashurinë, mbështetjen dhe inkurajimin e pafund.*

Table of contents

1 Introduction	1
1.1 Motivation	2
1.2 Aims	3
1.3 Objectives	3
1.4 Overview	3
2 Literature review	5
2.1 Advantages of Technology Enhanced Language Learning (TELL)	5
2.2 Importance of learning language pronunciation	6
2.3 Approaches used for teaching pronunciation	7
2.4 Evaluation of pronunciation / Pronunciation scoring technologies	8
2.5 Importance of visual representation in learning	9
2.6 Existing solutions in the same problem domain	10
2.7 Feature listing	11
3 Research question and specific hypothesis	13
3.1 Methodology	13
3.2 Methodological framework	15
4 Technical approach and implementation	19
4.1 Motivation for technical platform choices	20
4.2 System architecture	21
4.3 Properties of each component in the system	22
4.3.1 <i>Source node</i>	23
4.3.2 <i>Filtering module</i>	23
4.3.3 <i>Destination node</i>	25
4.3.4 <i>Highlighting style of visualization</i>	26
4.3.5 <i>Scenario</i>	26
5 Experiment and analysis results	29
5.1 Initial exploration of the data:	29
5.2 Testing results taken all groups together	29
5.2.1 <i>Test results split by groups</i>	30
5.3 Testing the differences between the groups	31
5.4 Inquiring the teachers about the sessions run	33
5.4.1 <i>Results</i>	33
6 Discussion	35
7 Conclusion and future work	37
7.1 Limitations	37
7.2 Future work	38
8 References	40
9 Appendixes:	VI

List of figures

Figure 1. a) VocSyl application b) Vocabulary.co.il	2
Figure 2. Research overview diagram.....	4
Figure 3. Picture taken during training sessions.....	17
Figure 4. Application client-server architecture	19
Figure 5. System architecture components	22
Figure 6. Quantization - conversion of analog sound into digital (S mus, 2013)	22
Figure 7. Sound wave shown in time and frequency domains (S mus, 2013)	24
Figure 8. Audio graph of live input scenario	24
Figure 9. Simple audio context that connect source to destination (S mus, 2013)	25
Figure 10. UML - Sequence Diagram: User picks the word for exercising	25
Figure 11. UML - Sequence Diagram: User plays demonstration	26
Figure 12. UML - Sequence Diagram: Highlighting syllables	26
Figure 13. Screenshot of prototype: first interface	27
Figure 14. Screenshot of prototype: second interface. Child can choose which word to exercise.....	27
Figure 15. Screenshot of prototype: third interface. Child is listening to the demonstration.	27
Figure 16. Screenshot taken while child is trying to pronounce syllables of motherhood. Next highlighted part (-er-) is in turn to be pronounced.....	28
Figure 17. Screenshot of prototype: third interface after pronunciation	28
Figure 18. Normality plot for sum of grades after training	30
Figure 19. Plotted means of groups for scores before (left) and after (right) the training	32

List of tables

Table 1. Features of reviewed tools on literature review	11
Table 2. Methodological framework (Scaife et al., 1997).....	16
Table 3. Normality test for grades before training	30
Table 4. Calculated mean for grades. Results grouped based on training approach	31
Table 5. Output from One-Way ANOVA test	31
Table 6. Descriptive Statistics for sumAfter variable	33
Table 7. Es timates - adjusted means of sumAfter.....	33

1 Introduction

The inclusion of technology in the learning activities is becoming widespread with the promise of potential improvements to the learning process. Some Information and Communication Technologies (ICT) innovations are dedicated to the teaching and learning of foreign languages, a specific area that can be termed as Computer-Aided Language Learning (CALL). The use of multimedia in learning foreign languages dates since half century ago where teachers used multimedia on their teaching process such as: gramophone records, film-strip projectors, videocassette recorders and so on. In many cases the evolution of technology has shaped these learning tools into self-learning web solutions, application and even virtual language centers (Wikipedia, 2013). Human language technologies (HLT) is another term that encapsulates a vast amount of products that aim to contribute toward learning a second language. Some of these technologies are especially focused on facilitating learners the learning of proper pronunciation (Gupta and Schulze, 2011).

Several studies summarized by Neri et al., (2008) are outlining the importance of learning pronunciation at early ages. They claim that younger learners are more likely to learn more accurate speech in comparison with others that try to learn a second language during adulthood. Such indications have attracted researchers and developers of CALL to put a considerable effort on creating innovative ways to help children learn pronunciation using speech technology. The ability to combine syllables and pronouncing them is an important stage related to communication development (Velleman, 2002). Plenty of studies and solutions strongly suggest that voice visualization improves the ability of children to pronounce words. For example, Ainsworth (1999) on her research outlines that visual representation helps children to understand on their own the process of learning a subject, give information about the content and help them construct better knowledge around that subject. Additionally, Levis and Pickering (2004), reviewed several visualizing tools for exercising pronunciation and also assessed the impact of their *Computerized Speech Laboratory* (CSL) program. All these investigations proved that visualization tools increased learner's ability to control their voice and pronounce words correctly.

Our present contribution inscribes itself in this field. More specifically we aim at providing a system which, after giving certain word to pronounce to the child, will be able to capture the audio input by him/her and visualize that together with additional visualization style. Thus, it can be understandable by him/her and inform about the performance. In our research we aim to evaluate the performance of children that use our system and compare the achievements with children that learn using a traditional approach. Teachers will be involved to observe the entire process, in order to make sure that we are meeting pedagogical aims while giving to students a system that is designed to teach pronunciation. If experiments' results outline positive impact, we will have in hand a system that will assist children toward learning pronunciation. Eventually, our research queries the impact of specific combination of visualization methods and compares them with traditional approach where no technological solutions are used.

1.1 Motivation

Learning a foreign language such as English has become part of most early stages curricula in most educational systems, especially for European countries (BEC, 2002). In fact, in a globalized world being able to speak more than one language of the more currently spoken ones is becoming a valuable asset.

Another relevant trend concerns the growing importance of after school activities to complement the regular school teaching. In relation to the learning of a foreign language, however, not every child has always the support of well skilled supervisor to help in learning pronunciation as in school environment. This brings the potential need to have a system that will help children to learn how to pronounce English words, somehow supplementing the learning activities carried out in schools. Due to these facts we choose to explore a new approach for dealing with this problem.

A lot of solutions in the market which are designed to support children who learn pronunciation, employ speech technology and different visualization techniques. However, we believe that more can be done regarding the design of simple and attractive multiple visualizations to help the learning of English words pronunciation. More specifically, this research explored how the combination of real-time visualizations presented coherently can help children to learn pronunciation of English language words. Our system is able to take as input the speech voice of the child and meanwhile represent in graphically the wave of the sound and together with that highlight the words. As illustration, we can see some graphical representations (visualization) in the figure below, which are currently existing solutions for children to learn English language.

Figure 1. a) VocSyl application b) Vocabulary.co.il¹

First image is a screenshot from the application called VocSyl (Hailpern et al., 2012) that helps children with speech delays to learn pronunciation whereas the second one comes from web based games provided by Vocabulary.co.il. Similarly in our case visualizations are simple and colorful so that the child will find it easy to perceive what the screen is showing.

Despite plenty of research that have been conducted on this field and is summarized in following section, there are still gaps, in particular regarding the impact of combined visualizations. For example, we do not know much about the possible beneficial effect of combining wave-form and highlighted words representations on children that learn pronunciation of foreign language.

¹ <http://www.vocabulary.co.il/syllables/syllable-lesson/> Accessed: 2013-03-15

1.2 Aims

From the discussion above we fleshed out the following aims:

- Develop a tool that allows the coherent visualization of the audio input and the highlighting syllables of the word in real-time.
- Evaluate whether these two kinds of exercises (visualizing the audio input and the highlighting of syllables) help children learn pronunciation of English language.

1.3 Objectives

Once the aims are established we proceeded with the specification of a set of objectives that we considered necessary to be followed in order to be able to answer our research question:

- Create a prototype consisting of a demonstration, real-time visualization of pronunciation and highlighting syllables that employ theoretical perspectives for learning pronunciation.
- Perform testing of prototype with children in Macedonia, in three sessions of practice.
- Compare changes in pronunciation between children that used traditional approach and the ones that used the prototype and parts of it.
- Perform analysis of result and determine best approach to learn pronunciation between: traditional approach and the proposed solution.
- Evaluate the impact of two visual representations approaches in learning pronunciation.

It is important to mention that linguist related issues are not among the objectives of our research. In this thesis our main focus is to explore from media technology perspective features that serve for pronunciation learning. Previous researches in this field have also elaborated related features, such as scoring technologies, acoustic similarities etc, that help learning pronunciation of a language; however these features are not considered in this research. Wave of sound visualization and highlighting syllables, are the unique combination which impact we aim to investigate.

1.4 Overview

This thesis is divided in 5 parts: literature review, research design, technical approach, analysis of efforts and conclusion and future consideration.

A set of relevant papers is summarized in the chapter: Literature Review. In order to make clearer the outline of studies that are summarized, they are divided into six subsections, each of them treating a unique perspective that we considered for our research. We start by explaining the importance of technology in language learning, importance of pronunciation learning, and then summarizing some approaches that are used for learning it. Next subcategory summarizes studies that include techniques for scoring technologies i.e. how pronunciation is evaluated. After this we proceed to explain the importance of visual representations while learning and then continue to describe some of existing products in the same problem domain, that use technology to teach pronunciation. Having an overview about previous studies done and considering the aims of our research, we summarized into a table all features that existing solutions for teaching pronunciation pose. In the same table we have mentioned features that our prototype contains, thus making even clearer the difference between existing solutions and our proposed solution.

Having proposed a solution, we continue to explain how our research was conducted in order to assert the putative benefits of the concepts and designs considered. Thus, the research design module starts by presenting the research question and hypothesis and continues with the research

methodology. The research methodology describes how experiment was conducted: design of experiment, samples, measurements and what kinds of analysis were performed.

We continue to explain then the technical approach followed for creating prototype. In this chapter we describe technologies used for creating our prototype. We explain the architecture of our prototype, getting into deeper details about their modules and motivation for choosing them. We finish this chapter by describing a scenario how this application was used.

The methodological framework section explains in details all stages of evaluation phase: who was the contributor, what their input was and what we had as output. Here we intentionally try to summarize all details of evaluation stage, in order to make sure that we follow a logical workflow and have a detailed overview for all steps.

Having a prototype to evaluate and a methodology to perform the evaluation, we run a particular study where the results are presented in the next chapters. Here we explain analysis of all gathered data from evaluation stage. By performing different analysis of data we continue to answer our research question in next chapters.

Concluding paragraphs are the ones that better explain the outcomes of our efforts. Discussions and future considerations for our work are finalizing the content of the thesis documentation followed by referring relevant works used in our research and appendixes.

Figure 2 below gives a graphical overview of the thesis documentation structure:

Figure 2. Research overview diagram

2 Literature review

Since English is the most used language all over the world (Wikipedia, 2013) a great technological effort has been put to find different solutions that will enable children to learn this language without the need of professional assistance. Pronunciation of words is a stage that greatly affects the speech development process (Highman et al., 2008). In addition, some researchers (Neri et al., 2008) are strongly suggesting that computer assisted learning of speech development results in desirable outcomes, meaning that it has positive impacts. A lot of existing solutions designed for children aim to contribute toward their development on this stage, by creating attractive applications and other solution that are supposed to help them learn pronunciation.

Several studies done in the 1980s on computer-aided language learning technologies used the stimulus-response model: behavior of querying and answering, to create new approaches for learning by focusing on grammatical properties or discrete-point learning (Conrad, 1996; Johnson, 1992; Pusack & Otto, 1997). A common feature of these applications was that they were focused on creating new approaches to transmit learning content to young learners. Jonson (1992) reviews most of these tools and concluded that their focus was to learn grammatical forms and plenty of them were not useful. The rapid development on technology, especially in multimedia field, attracted the interest of many developers and researchers to create new ways of learning by broadening the horizon of learning content that can be transmitted with the use of contemporary technologies. Examples of usage of multimedia approaches include: videocassettes, video players, computers with CD-ROM and following to the most current technology such as web based and/or native applications for mobile devices. These innovations in technology leave a considerable space to find ways for integrating it with language learning curriculums which now we can refer to as CALL (computer aided language learning), TELL (technology enhanced language learning) and other relevant terminologies. The following subcategories are giving insights about the advantages that come along with the usage of technology at language learning, the importance of learning proper pronunciation of language and approaches used to learn it. Including technological perspectives, later on, we review tools for scoring technologies, the importance of visual presentation and some existing solutions in the market, that are used for pronunciation learning.

2.1 Advantages of Technology Enhanced Language Learning (TELL)

Development of web and mobile technologies allowed the creation of plenty of possibilities that enhance learning approaches. Many researchers are praising them due to the improvement they bring in education. Learners are exposed to learning content not only on education institutions but also outside them. As Ogata and Yano (2004) pointed out, continuous technological development enables learning activities to be embedded in daily life. He noted that vocabulary teaching experiments outside classrooms resulted with faster acquirement from children. Starting from these facts he put efforts on combining mobile technologies and devices in order to assess the achievement of language learning outside learning environments (Ogata et al., 2006). According to him, being flexible during learning process helps children acquire knowledge faster than in school environments. The application used on this experiment also aimed to improve communication skills of children and results were favorable. Both children and teachers who supervised them, found this solution to be useful in practicing verbal skills in a real social

context. Teachers argue that children's confidence for speaking has risen after using the mobile solution given to them.

Adair-hauck et al., (1999) assessed the integration of technology into the activities of learning a second language. They chose to replace one class per week the traditional approach of learning by including multimedia technologies. They aimed to evaluate skills such as: listening, speaking, reading, and writing skills of college level students in French course. Multimedia technologies that were used for treatment group included: computerized reading tool, glossaries (English-French, French-English), grammar notes and so on. Second, the tool enabled the teacher to create forms for querying students about the learning content example multiple choices, completing sentences etc. Furthermore, video cassettes were used by students to exercise speaking and listening. In the design of their study, both the experimental and control group had the same instructor and learning content. Findings of this research aimed to guide language departments for the inclusion of technology in the learning process. The results of the research outlined, were very encouraging according to the researchers. Quantitative metrics showed that writing and reading skills were much improved in comparison with listening and speaking, while qualitative questionnaires outlined that: students were very satisfied with tools they used and, as a consequence, they became more fluent speakers of second language. The use of multimedia technologies promoted some encouraging enhancements in language learning, therefore leading to a promising field of research.

Tanner and Landon (2009) elaborated on the impact of computer-aided pronunciation (CAP) platform, which provided with self-directed computer-assisted practice. Their tool included cued pronunciation reading (CPR). Studies conducted tested the effect of this tool on treatment group by evaluating patterns such as: pausing, word stress, and sentence-final intonation. Tasks were self-directed letting students to have an overview of above mentioned patterns and exercise them. In addition to this gave the possibility to evaluate the learners' ability of comprehension. Results of their quasi-experimental research outlined that students who used computer-assisted platform had significant improvements on their pronouncing skills. Basically, by using their system the students were able to listen to prerecorded words in order to identify the pausing or syllables, record themselves while reading the words, and compare the results. This approach indicated improvements in pronunciation and increase in ability to stress syllables.

2.2 Importance of learning language pronunciation

On her book Rogerson-Revell (2011) covers fundamental aspects that an English learner should consider in order to learn pronunciation. She points out that the majority of English speakers have difficulties distinguishing between words such as 'sheep' and 'ship'. Therefore, she notes that what is captured while hearing is speech sounds and based on these auditory notions we have reaction about the language. The importance of pronunciation can be portrayed with several facts and examples. She argues that most of breakdowns in communication between non-native English speakers occur due to the distorted pronunciation. The lingual variety that exists between second language (L2) speakers is due to the differences in pronunciation, and both misunderstandings and intelligibility are caused because of that, she continues. Reviewing the research of Jenkins (2002), she concludes that the greatest barrier of communication is pronunciation. Rogerson-Revell explains that there are factors such as phonological and organization of talk which differ between learners. She illustrates by comparing Scandinavian and French learners, claiming that Scandinavian people have to put less effort in order to speak

fluent English. This due to the fact that their native language pronunciation is very related to English in comparison with French learner, which have similarities more in grammatical aspects. She argues that although we have plenty of variety in ways of pronouncing, there are concepts and regulations of pronunciation according to linguistic, which if followed can impact the communication language of every individual. Some of them are summarized in the next subcategory.

Neri et al., (2008) argues that learning pronunciation at early ages requires less, if any, effort in comparison with in adulthood. Therefore, a lot of researchers have focused on exploring ways that can help and positively affect learning pronunciation at early ages. She notes, as the use of L2 has become a fundamental requirement due to the multilingualism in many countries, pronunciation educators and researches must invent ways for providing training methods for pronunciation. In addition, Levis and Pickering (2004) claim that key factors that hold the meaning of language are modulations of voice. Movements of pitch, voice quality and way of expressiveness are very important for speaker's attitude and are very tied to the intonation. All studies reviewed on this section, emphasize the importance of learning pronunciation as a stage that greatly affects the fluency of speaking a second language. Good pronunciation can increase ability of spelling and reading skills of children (Lai et al., 2007). Boves et al., (2001) points out that children with high sensitivity to differences in pronunciation are more likely to learn English language faster than children without such skill.

2.3 Approaches used for teaching pronunciation

Since pronunciation is rated as a valuable ability that strongly influences the proficiency of speaking a language, plenty of researchers and educators have attempted to find ways for teaching it apart from traditional approaches used at schools. There are plenty of studies that invented innovative ways for learning pronunciation and since technology is a trend today, most of these innovations are created relying on technological approaches. However, before attempting to find best ways to teach pronunciation, it is important to know phonetic rules that have to be followed in order to learn the correct pronunciation.

Rogerson-Revell (2011) on her book also covers some notions that are important in pronunciation teaching. She argues that drilling, dictation, noticing and phonetics training are among the best approaches that help L2 learners to exercise pronunciation. She explains that the learner should refer to aspects such as: stress-timing, rhythm, vowel reduction and intonation while learning. Pronunciation problems are more related to the sound. Exercises designed to teach pronunciation aim to exercise the sound, right after having proficient knowledge about letters. She continues to argue that manipulation with such parameters can help in controlling the speech production. Articulatory settings are very important for achieving intelligibility therefore the quality of voice at early stages is among the suggested ways to achieve proficiency in L2 pronunciation she concludes.

Lai et al., (2007), points out that on traditional way of learning where teacher asks the learner to practice pronunciation, results in inaccurate pronunciation. He argues that this approach of teaching avoids the recognition capability of right and wrong pronunciation. If learner is aware of pronunciation of the word phonemes, he can easily pronounce the word by disassembling the word tones. Motivated from this fact they proposed a Multimedia Learning system that uses HMMs (Hidden Markov Models) to analyze the pronounced word and provide feedback about

the pronunciation, intonation, rhythm and volume. This system performs phonemes' analysis based on which the feedback is given. The algorithm followed is: taking the input from speech signal and each feature of phoneme is analyzed in cluster based and phoneme based model and phoneme with highest probability is displayed. So each word goes through the speech recognition, phoneme recognition and verification stage. The error analysis module improves the pronunciation of learner by determining the right pitch and volume. In a quasi-experiment this tool was given to 56 students aged 8-9 years old and results showed that low scoring students had improvement on their pronouncing abilities whereas others more in spelling and reading skills.

Su et al., (2006) argues that most of existing works that deal with pronunciation learning are not considering personalized problems for learners. Most of the tools are giving quantitative feedback of performances, denying the opportunity to improve their language skills. They point out that the secret of the proficiency for having a good learning tool of pronunciation is having a good evaluation approach. Thus they propose a fuzzy based evaluation model that aims evaluate pronunciation skills based on hard to be distinguished phonemes (HDP) model. They propose to provide learners with a robust evaluation model which gives specific feedback and thus increasing the ability to learn the correct pronunciation. Basically, this model checks for the correspondence between the pronounced phonemes with the standard speech corpus recorded by native speakers of English language. Phonemes occurred in speech corpora are compared with phonemes recognized by the system and total number of recognition is displayed. Correct and error rate are calculated for each phoneme. This calculation is applied after the user records sound for given sentences. After this HDP recognition results are displayed. Results of experiment with students with different speaking pronunciation showed that this model was giving stable and reliable evaluation results, however remains unmeasured the impact of this model on pronunciation abilities.

2.4 Evaluation of pronunciation / Pronunciation scoring technologies

Almost all of the pronunciation computer based tools that we have summarized include approaches for evaluating the pronunciation. We could notice that most of them are focused on assessing the performance based on phoneme recognition, pitch and volume. In addition to these studies there are also researches that invented a different way for assessing pronunciation.

Dong et al., (2004) describe a new approach for assessing the pronunciation quality that can be divided into two categories: text-depended and text-independent approaches. They also point out that automatic speech recognition systems are not designed to teach pronunciation therefore other approaches for evaluating this are required. The methods proposed by Dong et al., are not using scoring technologies that follow HMM algorithms or other probability scoring systems that rely on machine evaluation of pronunciation. Instead they propose the text-depended algorithm which evaluates the pronunciation based on acoustic similarities and the text-independent method which compares energy, pitch and frequency variations. Their database is consisted from four groups with same content which differ in pronunciation quality. Based in comparison between pronounced word and the one recorded on database, the system is able to grade the performance. Dong et al., claim that measuring the distance of acoustic parameters and comparison of pitch frequency between input and demonstrative speech, are among the most used methods to determine the quality of pronunciation. Results of their experiment outlined that the text-independent approach where the pitch frequency was considered, was more reliable method when

comparing with text-dependent. They argue that text-dependent approach has to consider features such as gender and age in order to make the correct assessment of input.

Hinks R. (2003) reviewed on her research several technologies that are used for teaching language. She elaborates among the pronunciation teaching approaches and the features that are considered for feedback and evaluation of pronunciation. She notes as the inclusion of new approaches for learning pronunciation are growing; teachers and researchers are looking for an empirical evaluation of pronunciation. According to her reviews, most of teachers exercise the production and perception of the sound when teaching pronunciation of second language (L2). The visual representation of audio signals therefore, is a beneficial way to follow for teaching both perception and production. For this reason, researches that emerge the visualization of pitch property result with positive outcomes of evaluation. However, the progress of speech technology (or automatic speech recognition – ASR) made software and tools to include the speech recognition feature for language learning. According to Hinks, ASR again makes comparison of signals with additional mathematical ways to process the input. ASR makes a comparison between *added signal* and *hundreds of variations* spoken by native speakers where the phoneme with highest probability is chosen. This scoring is then used to tell the learner how far is from the target. She adds that for these kinds of evaluation approaches, gender and age does not matter.

2.5 Importance of visual representation in learning

The effect of multiple visualizations of learning content has been proved as effective in many studies. As Ainsworth (1999) elaborated on her research, MERs (multiple external representations) can assist learning from three perspectives: to complement, to constraint and construct. In first perspective the representations are supposed to give information about understanding the process. Visual representation helps learner to construct required understanding about a certain subject. The second perspective deals with representations that aim to constraint from misinterpretations. Familiar interpretations assists the learner to understand better the new given learning content, thus helping him/her to avoid misinterpretation or misunderstanding about that subject. Third perspective, helps constructing a deeper knowledge of the situation through familiar representation by accompanying new content to learn.

Researchers that used the IBM SpeechViewer software tool highlighted the importance of visualization when learning pronunciation. Stenson et al., (1998) examined the impact that the visual representation of speech had in teaching assistants for exercising their pronunciation. They believed that the feedback given by technology could help them to learn pronunciation by comparing their performance with the one of native speakers. The SpeechViewer software was designed to give feedback about different types of vocalization so that it could detect and suggest specific parts of pronunciation. Basically this software was able to process and give feedback about pitch curve, amplitude, wave of sound and so on. It was designed to help young learners without known disabilities in speaking and hearing. However, they used this tool to assess its effectiveness for older target groups such as: teaching assistants (adults). They had difficulties with timing, loudness and other voicing problems while pronouncing English words. Results of the experiment where international teaching assistants (ITAs) where supposed to use the tool for 10-20 minutes in 8 sessions, outlined that SpeechViewer did not have significant effect on pronunciation skills. Researchers argue that the reason for this could be the short time that international teaching assistants had disposable for exercising with the software. Despite this, the

motivation of international teaching assistants to use this software was high enough and according to Stenson et al., this was also the reason for having changes in pronunciation quality although not in high parameters.

Levis and Pickering (2004) also outlined on their research that visualization techniques are widely used for teaching languages especially intonation. They review several programs such as SAP, PRAAT, VisiPitch and SpeechViewer that have been widely used for teaching intonation. Visualization techniques are dating since 1960s and the positive effect that they have, makes language technology advance more and more. Summarized researches from 1980s, have showed results upon which they conclude that L2 learners who receive audio-visual feedback demonstrated improvements in producing and perception speaking patterns. In addition they were more sensitive in pitch movements. In order to continue exploring the impact of visual representation importance on L2 learning, they used *Computerized Speech Laboratory* (CSL) program. They recorded the sentences and then used the program to analyze phrasing and pitch. Users were allowed to exercise and practice discourse-level of speech, whereas the tool was analyzing all pitch movement. Data gathered from this experiment were analyzed and outlined that this way of visualization increased learners' ability to understand how to behave with intonation and pitch movements in real communication environments.

2.6 Existing solutions in the same problem domain

In addition to the tools already mentioned there are also some solutions currently available on the market that intend to solve similar problems.

Hailpern et al., (2012) also considers speech production and vocalization as an important development stage of communication for children with and without impairments. The visualization of information has therefore emerged in many applications that aim to learn language proficiency. Inspired from successful outcomes of researchers that examine the impact of visualization, Harlipen continued to explore on this field aiming to examine the visual impact on children with impairments. As result they designed VocSyl software, a real-time audio visualization system. This syllable based system, gives a visual representation of timing, pitch and volume of the sound as a word is being pronounced. They conducted several rounds of testing, by assessing different perspectives of speech. On first two round they were more focused on assessing the interaction between children and software. They were focused on choosing the desired visualization method for following exercises. Since VocSyl offers several ways to modify the display they argue that all children were enthusiastic for choosing their way of learning, and as different options were chosen, they were trying to exercise they pronunciation to match the score as much as possible with sample given. Phonetic accuracy increased and children tent to achieve highest correctness, without receiving any other speech therapy.

As Mak et al., (2003) explicate on their paper speech recognition technology is mainly used in CALL for two purposes: for learning accurate pronunciation of foreign language and evaluation the pronunciation quality. Persuaded from these facts, they also developed multimedia tool called PLASER. Their interface contained a pronunciation video-clip with picture aiming to teach students the movements that should be done in order to pronounce correctly. In addition, pre-recorded samples were available for him/her to listen before trying on their own. The feedback was given in visualized representation where first one was calculating the overall score of phonemes spelled and second one which showed in colors the accuracy of each phoneme spelled.

Research results showed that this system improved the pronunciation to 77% of students that participated the testing and 99% of teachers approved that. Such positive results are motivation for future studies which aim to contribute on similar field.

2.7 Feature listing

A lot of researches are dedicated to evaluate the impact of innovations that aim to help language learners. The above mentioned ones are field studies that complement foundation of our work. We have seen a lot of features that different tools employed in order to enable pronunciation improvements by measuring different metrics. In Table 1 below we can see a summarized view of all features that we mentioned until now and how different tools combined them.

	T1	T2	T3	T4-txt dep.	T4 - txt indep.	T5	T6	T7	S
Listen to pre-recorded sample	✓	✓	✓	✓		✓	✓		✓
Repeat without evaluation	✓								
Exercise writing abilities	✓								
Record/Listen own pronunciation		✓							
Visual feedback presentation			✓			✓	✓		✓
Visual representation of sample			✓				✓		✓
Acoustic similarities				✓					
Pitch		✓			✓	✓	✓		
Frequency					✓				✓
Energy					✓				
Wave of sound						✓			✓
Loudness control							✓		
Timing mechanisms			✓	✓					✓
HMM/ASR			✓	✓	✓				
Real-time feedback							✓		✓
Video demonstration								✓	
Comparison				✓	✓				✓

Table 1. Features of reviewed tools on literature review

Explanation on attributes:

T1 = Tool reviewed by Adair-Hauck et al., (1999)

T2 = Tool reviewed by Tanner and Landon (2009)

T3 = Tool reviewed by Lai et al., (2007)

T4 txt dep. = Tool reviewed by Dong et al., (2004) – text depended approach

T4 txt indep. = Tool reviewed by Dong et al., (2004) – text independed approach

T5 = Tool reviewed by Stenson et al., (1998)

T6 = Tool reviewed by Hailpern et al., (2012)

T7 = Tool reviewed by Mak et al., (2003)

S = our proposed solution

This table can give us a clear overview of the features that different tools/software had. Foundation of our research is motivated from the fact that the impact of these features combined differently is still unmeasured. For example, the last column shows the features that are included in our application which are also uniquely combined in comparison with previous works. Therefore, we are focused to continue exploring this unique combination of these features for enabling children a new approach for learning pronunciation. Some of the tools put in table, are excluded in our work. For instance we will not measure the impact of features that exercise writing abilities, record own performance, repeats without giving evaluation of previous performance, finding acoustic similarities etc. More precisely, in our research we will evaluate only the combination of visual representation of the sample which gives a graphical overview of the frequency and wave of sound, as well as timing mechanisms by highlight with colors each syllable. On the technical approach section we describe in details how these features are implemented.

3 Research question and specific hypothesis

Based on the description of the problem that we described and the solution that we aim to investigate, our work can be summarized with the following general research question:

Does the combination of real time a) highlighting syllables visualization and b) wave form of speech visualization, increases the pronunciation skills of children learning English language?

Hypothesis 1: The use of a system that presents both (a) the visualization of the wave form of the spoken word and (b) the highlighting of syllables, will increase the children's ability to pronounce a given set of English words when comparing with traditional approach and systems that only employ one of the forms proposed.

3.1 Methodology

In order to be able to answer our main hypothesis we followed an approach that starts with a literature review. We continued to further identify the problem space and the planned a solution how to address it. These stages included literature reading about usage of technology in language learning, especially their impact on pronunciation of foreign language learning such as English. We were focused on finding studies that elaborate on the impact of visual representation while learning. We tried to identify specific problems and issues that have not been elaborated before. After having identified the problem we continued to further plan an experiment upon which we could gather data to solve such issues. Technical aspects were also elaborated in order to find a reliable framework that will help us to create the prototype with which we had to conduct the experiment. The solution found will help to elucidate the potential benefits of combining the two forms of visualization stated. Gathered relevant data and outlined results, will help to answer the hypothesis and suggest further issues to investigate within the confines of the research question.

Design of experiment:

In our present work we followed an **experimental approach** where we tried to compare the results from groups that went through distinct conditions (some used our solution and parts of it and a control group that did not use any system for learning pronunciation). A **between subject design** was followed where conditions were assigned to different groups which means that a single group could not be exposed to two different conditions. In order to be more accurate and ensure that we do not have differences between groups we randomly divided the whole sample into subgroups. Pre and post testing on each session of learning words also helped us to determine the difference in more accurate way. Statistical analysis of the outcomes helped to outline the difference in their pronouncing and based on them answer our research question.

Sample:

Our sample included students (children) from primary school that are at the early stages of learning English language. Their age varies from 9-11 (other demographical characteristics are not being considered in this work). It is important to mention that in our sample we excluded children with known disabilities such as speech delays, autism and so on. The size of the sample was 7 students per group (28 students in total), where participants were randomly assigned to treatments. Children were given 6 words to learn on three sessions, which in total corresponds to 18 words learnt. This particular choice was in accordance to teachers, who considered that if more than 5- 6 words per session were given, children would get bored. The experiment took

place in Macedonia due to easy access to participants and careers willing to collaborate for conducting this research.

The four experimental groups were:

Control group:

- Traditional Group – was the group that learned the words in traditional way without intervention of any system that employs technological approaches.

Experimental groups:

- Highlight Group – was second sample who learned using only the highlighting visualization style of the prototype.
- Wave Group – was third sample who was using only that visualization style which was drawing the wave of sound based on the sound produced.
- Complete Group – was fourth sample that had combined in their prototype the visualization of two above mentioned styles.

Measurements:

The four groups were tested before the session of learning words and after that. With the help of teachers we evaluated the changes in the pronouncing skills of all participants. The teachers involved in this task graded the student's pronunciation for each word, always considering timing, loudness and intelligibility that the child produced while pronouncing.

It should be mentioned now that prior to the main study with children, teachers were asked to assess the accuracy of the system. After completing the functional requirements and operability testing of the solution with teachers, regular training with participants started. An important fact to mention is that teacher who helped us to assess the differentiation in pronouncing skills of children was not aware about the platform used to teach the pronunciation (blind evaluator). As we mentioned, during three sessions of learning words, we had four groups that learned the same content in four different approaches.

The whole evaluation stage lasted 5days: the first day was set for pilot testing, the following three days for training sessions and finally on the last day teachers were be asked to fill questionnaires or participate in interviews in order to collect further data. Regarding the activities of the fifth day, we tried to elicit their overall opinions and recommendations about the solutions. Teacher used a five-point rating scale to assess the correctness of pronunciation. Grades used varied from 1 to 5 (where 1-not sufficient, 2-sufficient, 3-good, 4-very good, 5-excellent). The reason for using this grading approach was because teachers in Macedonia are already familiar with this grading scale.

Analysis:

After gathering the data following the approach referred to above, we focused on measuring the impact that our system had in learning pronunciation. Statistical analyses were conducted in order to compare the children's performance. Some initial descriptive statistics helped to form a first impression of potential trends whereas inferential statistics allowed the construction of a more robust conclusion. These analyses and results are further elaborated in Experimental Analysis chapter.

3.2 Methodological framework

Scaife et al., (1997) created a methodological framework that helps gain some understanding of each develop/design phases involved in the creation of educational software/technologies. He claims that using this approach possible generated values are known in prior, regardless if activities occur in sequential or parallel way. Similar approach we decided to use in our research, in order to better identify on each stag the contributors and their input. Our framework is summarized in the table below whereas phases are briefly described after that.

Phases of development and evaluation	Contributor	Input	Method
Phase 1 - Define the problem	Researcher	Read existing publications related to same problem domain identify	Reading, collecting, analyzing and classifying
	Teachers	Describe the learning process identifying main goals	Teacher interviews
Phase 2 – Translation of specifications	Researcher	Analyzed existing solutions and multimedia tools they use. Define the scope of interactivity	Reading, collecting, analyzing and classifying
	Software/graphic designer	Define specifications of system and functional requirements	Sketching, interactivity and scenario design
Phase 3 – Design low tech materials	Researcher	Investigate on how to design application for children	Reading, collecting, analyzing and classifying
	Software/graphic designer	Built main functional specification	Make low-tech materials
	Software/graphic designer and HCI analyst	Test feasibility of the system	Asses the functionality of the system
Phase 4 – Develop and test the prototype	Software graphic/ designer	Flesh out prototype specifications and validate design and functional aims based on output from prior phases	Prototype hi-tech solution using multimedia tools and programming environment
	Teachers	Test validity of pedagogical aims	Try out the prototype suggest immediate improvements if needed
Phase 5 – Implement	Software graphic/ designer	Test the functionality of the system and improve the solution based on output from prior stage	Prototype hi-tech solution using multimedia tools and programming environment
	Teachers	Provide input to serve as example for children	Using recording multimedia tools to record demonstration
Phase 6 – Test and evaluate the prototype	Teachers	Test validity of pedagogical aims	Try out the prototype
	Teachers and Children	Pilot testing of prototype	Ensure that current design and materials we capture necessary data.
	Children	Interact (use) with prototype to learn the given content	Learning tasks divided into 3 sessions
	Teachers	Assess the changes that occur in pronunciation and intonation of children	Grades given by teachers to assess the pronunciation quality.
Phase 7 – Evaluate the performance of children	Teachers	Verify whether prototypes brought improvement over existing method. Suggest future improvements	Discussion and questionnaires. Use data from previous grading scale.

Table 2. Methodological framework (Scaife et al., 1997)

In the following we will briefly explain each phase.

Phase 1 – Define the problem: outcomes of this stage are summarized in literature review section. We tackled a certain problem and we proposed a solution, always considering prior

contributions that have been done on this domain. Except this we organized a meeting with teacher in order to get more information about learning approach and goals in English course. The likelihood that teachers from Macedonia (Debar) will accept collaboration was greater, therefore since the beginning we discussed with them regarding the way English course is held.

Phase 2 – Translation of specifications: Having in consideration the problem that we aimed to solve and also referring to prior solutions, we started classifying and categorizing potential tools that could be used to build our prototype. This phase involved studying different technological approaches. The choice fell on the possibility to create a web based application for helping children to learn pronouncing and we moved on by starting to define functional requirements that our solution should be able to handle. At the end of this stage we had sketches and mockups of prototype and UML diagrams defining basic capabilities of prototype.

Phase 3 – Design low-tech materials: Having in mind main the functionalities that our solution needed to have, we started thinking about how to make it attractive for children. This stage included studying about designing an application for the specific target audience. After this, we proceeded building the main functionalities by converting sketches into a functional prototype. In collaboration with an HCI analyst we reviewed the prototype assessing its functionalities and feasibility. Thus, as it can be understood as output of this stage we had a functional prototype with the basic capabilities built. In addition, we kept note of all comments done by HCI analyst.

Phase 4 – Developing and testing the prototype: Having the feedback from the prior stage, we continued towards finishing the development of the prototype. Once finished we went to Macedonia in order to try out the prototype together with teachers and have their feedback. As outcome of this stage we had fully functional prototype.

Phase 5 – Implementation: Since no improvements were needed we skipped the first step of this phase, heading to the recording of the demonstration material, which was done by a teacher. Materials that were recorded were included in prototype and served as an example to children on how a word should be pronounced. So at this stage we added demonstration material to our prototype - the main outcome of this stage.

Phase 6 – Testing and evaluating the prototype: After including the recorded material, teachers reviewed again the whole prototype to assure that we are conforming to the pedagogical aims. Pilot testing was conducted in order to ensure that during the real testing procedure we would be able to gather all needed data for conducting the analysis. After this task, the regular sessions started with the children using the prototype to learn the given content, while a teacher was grading their pronunciation before and after each training session. Figure 3 on the right side is a photograph taken during the training session.

Phase 7 – Evaluating the performance of children: After the training sessions ended, the teachers were given questionnaires to fill. We also conducted some informal interviews with the

Figure 3. Picture taken during training sessions

teachers, taking notes regarding comments that they made related to the prototype and the overall experiment. As we can see as outcome of this stage in addition to questionnaires, we had also notes extracted from informal discussions. A more detailed description regarding the method that we used to gather data related to our prototype will be explained in the next section.

4 Technical approach and implementation

Recalling Table 1 “Features of reviewed tools on literature review” in Chapter 2; the last column lists features of our proposed prototype along with other solutions that are reviewed on background research section. Foundation of properties of our prototype derived from the combination of features that we proposed to include in our tool. In Table 1, we affirmed that our system will be comprised of features such as: listening to demonstration, visual representation of demonstration and performance, frequency and sound wave, timing mechanism, real-time feedback and comparison of performance and demonstration. This chapter describes the various aspects of the implementation of these features in our prototype. Consequently, from technical perspective this means a tool that implement following features:

- visual representation of both: recorded and produced sound wave,
- real-time feedback – by drawing a sound wave at the moment of speaking,
- timing mechanisms – by highlighting each syllable at the time when it should be pronounced,
- ability to listen to a prerecorded pronunciation demonstration,
- comparison between performance and demonstration – which can be seen by looking at drawn sound wave of recorded and produced sound wave.

These features were built using technologies such as HTML5 with its Web Audio API for sound processing and playback, as well as jQuery for interchangeable actions in our application. In context of web programming, modules of this application runs only on client side. Web Audio API enables audio processing and synthesis that run on client side, whereas servers were used only as storage for application files and demonstration playback sounds. AJAX technologies were used to asynchronously retrieve needed data and use them on client side application. Components of our client-server architecture are illustrated in Figure 4.

Figure 4. Application client-server architecture²

² <http://www.ct2013.co.uk/exhibitors/richardferris/project-image.jpg> Accessed: 2013-04-09

The following section describes the motivation of technological platform choices, system architecture, and a detailed description of each component of the system. The functionality of the system is then illustrated in a scenario.

4.1 Motivation for technical platform choices

This section motivates the choice to make the application web based, the reasons for using HTML5 and the Web Audio API in conjunction with other technical frameworks needed for delivering and evaluating the prototype. The literature review shows that applications that enable graphical representation of live audio input, playback of samples and highlighting words are commonplace.

In order to implement functionalities such as: visual representation of produced sound and playback of recorded demonstration, we needed a tool which should be capable to:

- a) record input from a peripheral device such as microphone,
- b) process in real time incoming audio source,
- c) efficient real-time frequency analysis,
- d) ability to produce frequency data that will be later used for visualization and
- e) stream audio from an existing source (playback).

As Rogers Ch. (2012) states in official documentation of Web Audio API, it has potentials like:

- Sound playback from `<audio>` tag,
- audio stream and synthesis in JavaScript,
- Efficient real-time frequency and time-domain analysis / music visualizer support.

These are most fundamental and important features that a tool has to pose in order to serve us for: playing a pre recorded sound and visualize wave of sound which can be done only after processing incoming sound. In addition, he claims that this is a high-level JavaScript API, which we see as positive feature since we aimed at creating a solution in web environment.

In addition, among the features of our prototype that we mentioned above was drawing the wave of sound at moment of speaking. In order achieve drawing using scripting language (as JavaScript) we used the HTML5 Canvas element. The reason for using this tool was the easiness for generating dynamic graphics in a certain region defined in an HTML page by using JavaScript³. Also the ability to put a background picture in canvas' area would later be helpful for making comparison between sound wave of demonstration and student's live input (illustrated in Figure 16 and 17 in Scenario subcategory).

Timing mechanisms we decided to bring them by highlighting with colors each syllable in sequence. The moment when the color of syllable changed, was actually the time when it was supposed to be pronounced. Likewise, we aimed to help child pronounce each syllable at the right time. For developing this highlighting module we used client side libraries such as jQuery, since this feature neither required communication with the server nor linking to other devices or functionalities such as microphone, audio processing etc. As a result, we found jQuery as the appropriate library for creating low-level interactions in our prototype.

As mentioned above listening to pre-recorded sample was enabled using HTML5's `<audio>` tag, due to the easiness in making it compatible with other modules of the system.

³ http://en.wikipedia.org/wiki/Canvas_element Accessed: 2013-04-10

Except the advantages that we gained from developing perspective by using HTML5 and Web audio API we also aimed at using a new technology which has not been used previously for achieving goals as we have defined for our research. We already saw that there are many solutions that exercise pronunciation but none of them used HTML5 or Web Audio API. So together with the objective to create a tool for exercising pronunciation through graphical representation of the sound, we wanted to reveal the power of new technologies.

Web Audio API, was not the only tool that offers sound processing. A similar tool such as Sphinx-4⁴, which is also known as state of the art at speech technology, is entirely written in Java. However, we do not see this as favorable options since creating a web-based application that uses Java libraries require the use of Servlets. In addition, speech recognition was a feature that we do not include in our application; therefore sophisticated tools that make such analysis of the sound are useless. For capturing the pitch and wave form of the sound there should not be applied many calculations and measurements of input. Web Audio API has an interface for all sound analysis that we needed, and they can be simply called by using JavaScript methods.

Admittedly, deciding to use Web Audio API lead toward choosing web environment as framework that would enable us to develop required functionalities. Web based application can take advantage for easy accessibility and mobility, because the prerequisites to access them are having internet connection and a web browser. Also, Web Audio API as the tool that fulfilled our requirements was supported mostly on web browsers, whereas it was partially supported at few mobile browsers. In addition, screen size is another issue with mobile devices. Since for the Complete group we had to display syllables as they were highlighting and at the same time wave of sound, we thought mobile screen sizes are not sufficient to render all of them. Due to these reasons we decided to create our application upon web based technologies. Reciprocally, with this platform we are required to use tools that can be supported by web framework.

In the following we are going to present a schema of our system architecture and continue to analyze each module.

4.2 System architecture

Visualization of live input and highlighting words required the combination of different components. The Web Audio API is created based on Audio Routes concept⁵, which means different nodes, can be added and connected to each other with edges. Based on functionality, our application can be logically divided into three nodes. The nodes and their relations are depicted in Figure 5.

⁴ <http://cmusphinx.sourceforge.net/sphinx4/> Accessed: 2013-04-10

⁵ <http://creativejs.com/resources/web-audio-api-getting-started/> Accessed: 2013-04-12

Figure 5. System architecture components

As we can see from Figure 5 our architecture has three nodes. The following section details the functionality of each of these components.

4.3 Properties of each component in the system

As enlighten above, Web Audio API is a powerful audio API tool that enables manipulation with sounds. In order to explain the audio functionality (how we captured sound, processed and visualized audio data), we must first understand how audio is converted from analog to digital data which is differently known as quantization. Figure 6 depicts the conversion process from analog to digital audio in a computer.

Figure 6. Quantization - conversion of analog sound into digital (Smus, 2013)

The figure shows that in each bit of data or *samplerate*, there are a lot of numerical values that represent wave of a sound for one second. Basically, after quantization computer treats sound as a long an array of numerical variables (integers), which later can be processed (Smus, 2013).

According to the documentation of Web Audio API, for performance reasons they use block processing. This means that number of sample-frames is fixed for each audio node which is defined to be 128 sample-frames. So in an unsigned 8bit array we have $8 \times 128 = 1024$ places with normalized values from 0 - 255⁶.

4.3.1 Source node

Web Audio API is built around *AudioNodes*, where we can determine the audio source and output and also the modules in between that will be used for audio processing or synthesis. The *SourceNode* has two sub-modules: buffering source and live input.

Buffering source is an array where audio file resides after it is being loaded from the server. When the sound from memory is triggered, buffer array populates with audio file data and starts decoding. For playback (of recorded demonstration) the format of sound could be anything that is also recognized from <audio> tag of HTML5. In this case the *SourceNode* is directly connected to speakers which are in *DestinationNode*, since no filtering of sound is necessary to be done.

Live input on the other hand can be generated from for example a microphone. In order to use the microphone to start audio capture (which to client is known as metaphor of pronunciation), he must allow his browser to access his microphone. For initiating processing of audio, the API needs to know the source of coming audio. We access the user's microphone using `getUserMedia()` API⁷, which access user's local devices such as: camera, microphone, web cams etc. There are two methods as callback functions *OnStreaming()* and *OnStreamingError()*, where the first one gives a callback about successful streaming and allows further processing while audio is being captured and the second which notifies about streaming errors in case they occur. Once we got successful audio capture we precede it to Filtering module. This is an *AudioNode* that can be connected between source and destination nodes. The sound processing that is done in this module will be described in the following.

4.3.2 Filtering module

After audio input has been captured from microphone, it is transmitted to another module such as filtering module, where processing occurs. The *AnalyzerNode* interface of Web Audio API is the filtering module, which enables to do real-time processing of audio characteristics based on FFT (Fast Fourier Transform) analysis. This is a digital signal processing technique for transforming (converting) a signal (in time or space) to a sum of frequencies and vice-versa⁸. Transformed (converted) data are normalized between 0 and 255 and stored into an array from where they can be interpreted differently. In our case we have used 8-bit unsigned array to store sample of data.

A sound wave can be inspected over time-domain or frequency-domain (Smus, 2013). Figure 7 illustrates these two approaches.

⁶ <https://dvcs.w3.org/hg/audio/raw-file/tip/webaudio/specification.html#AudioNode-section> Accessed: 2013-05-23

⁷ <http://dv.w3.org/2011/webRTC/editor/getusermedia.html> Accessed: 2013-05-24

⁸ https://en.wikipedia.org/wiki/Fast_Fourier_transform Accessed: 2013-05-24

Figure 7. Sound wave shown in time and frequency domains (Smus, 2013)

In order to have a sound wave that was representing the entire audio input, we had to visualize a wave of entire input from left to right ordering sound values in same sequence as produced. Therefore, we used time domain analysis using *getBytesTimeDomainData* method. In the figure above we can see how time domain values which belong to a range between -128 and 128, are represented before being normalized between 0 and 255. Despite this, Frequency domain graph changes the entire wave quickly, giving a snapshot of values for each capture in a certain time slot. Such graph draws all available values at a second and changes rapidly when new values are processed, therefore was not appropriate to be used for our prototype.

In order to visualize data of continuous speech we query the analyzer periodically, so that sound properties are returned on a certain interval. This can be done using *setTimeout()*, *setInterval()* or *requestAnimationFrame*. In our case we used *requestAnimationFrame* for performance⁹ reasons. As Smus (2013) also notes, since *requestAnimationFrame* is experimental, we use a prefixed version. This means that if it not supported by the browser, we have to make sure that we handle this error through callback function that provides the same functionality using *setTimeout* as alternative.

As mentioned earlier, we have used 8-bit array to store analyzed data. Considering the fact that this array is populated quite often with new data, we have added additional calculation before representing data graphically. Each time the array is populated with time-domain data, we call the function *peak()*. On every single *Uint8Array* we apply the method *peak*. So, if a *Uint8Array* has 1024 samples in it, we find the sample with the greatest magnitude and draw that. The reason for this is that the number of samples in the array is too numerous for practical visualization. In a sample-rate from 44.1 KHz we would have approximately 44000 values between 0 and 255 in a second.

Furthermore, the rendered input was not connected to destination node. If it was, the user would listen to the microphone. For this case the audio context is depicted in Figure 8.

Figure 8. Audio graph of live input scenario

As Figure 8 shows, the source node that is the live input is connected to a rendering module whereas the destination module is not.

⁹ <http://creativejs.com/resources/requestanimationframe/> Accessed: 2013-05-24

4.3.3 Destination node

In order to enable user to hear some sound from speakers we have to connect any generated audio to a destination node. For visualizing our audio input we did not need to connect the input to the destination, since the user does not need to listen himself/herself while pronouncing a word. However, destination nodes were useful for enabling user to hear the recorded demonstration. As illustration we can take the word “misspell”, which was given to students to learn. When the application was launched, before starting to exercise the pronunciation of “misspell” the child had the opportunity to hear the demonstration. This was a pre-recorded sound stored on our server. For playback cases, the *audioContext* was taking the sound from source instead of microphone device. Since processing of sound was not required for playback we simply connected the sound (source) to the destination node (which where speakers) as illustrated in the figure below:

Figure 9. Simple audio context that connect source to destination (Smus, 2013)

If we needed to process the sound before the user hears it, this graph would have additional nodes between source and destination as illustrated in Figure 8.

Once the page is initiated and certain word is selected for exercising, using XMLHttpRequest with a response type *arrayBuffer*, the array is loaded with the corresponding audio content from the resident source as shown in a part of sequence diagram below. Please find in appendix complete sequence diagram.

Figure 10. UML - Sequence Diagram: User picks the word for exercising

A variable was assigned to contain audio file data. The source could be played then using a button that handles paying. The following UML describes the sequence of activities that are triggered when “Play” button is pressed:

Figure 11. UML - Sequence Diagram: User plays demonstration

4.3.4 Highlighting style of visualization

On literature review chapter, we summarized the importance of timing while learning pronunciation as explained by professionals. The second method of visualization which is highlights the words, aims to help children stick to the timing while pronouncing. This will help to spell the syllables correctly in terms of time slots. For example if a three-syllable word needs to be pronounced highlighting of syllables one after the other for a certain period of time, will inform the child that he should proceed spelling next syllable. To achieve this technically we used jQuery, by assigning an index to each syllable and timeslots when they should change the color. In the following sequence diagram we show sequence of actions that occur as a child starts to exercise using live input and highlighting the syllables:

Figure 12. UML - Sequence Diagram: Highlighting syllables

The following sub category illustrates how our prototype worked during a training session.

4.3.5 Scenario

To illustrate how we combined the above explained components of architecture in order to have the desired functionalities, we are going to describe a scenario where a child uses the application to exercise the pronunciation of the word “motherhood” on first session of training.

First, the url where application is hosted, is typed and the interface as in Figure 13 appears. When child is ready to start, he/she presses the button “Start” next interface as in Figure 14 appears:

Figure 13. Screenshot of prototype: first interface

Figure 14. Screenshot of prototype: second interface. Child can choose which word to exercise.

As we can see, the Start button redirects to the first session. Redirecting link was changed in the file in server for each day, corresponding to the training session to be done. In this screen, user picks the word that has to exercise (we assume clicks “motherhood” for this example) and is being redirected to the next interface:

Figure 15. Screenshot of prototype: third interface. Child is listening to the demonstration.

Here we can notice that the word motherhood is split into syllables “moth-er-hood” and when play button is pressed, highlighting of word starts together with the playback of demonstration. These two are coordinated so that neither highlighting nor pronunciation of syllables, occur without being accorded with each other. The green line also appearing inside pink canvas shows how the sound should be “shaped” when pronouncing “motherhood”.

After the child hears the demonstration, he/she can continue pressing the microphone button and start speaking. As we explained highlight and wave of sound are accorded with each other. In the following picture we can see how the first syllable “moth” has been highlighted and also pronounced. The next syllable just changed the color and the wave is showing that pronunciation is about to be captured from microphone. This is illustrated in Figure 16. The red line which is drawn while speaking is visualization of sound, i.e. the sound recognized by microphone. Visualization after the first try is depicted in the Figure 17.

Figure 16. Screenshot taken while child is trying to pronounce syllables of motherhood. Next highlighted part (-er-) is in turn to be pronounced

Figure 17. Screenshot of prototype: third interface after pronunciation

Based on the visualization and also on what they heard, teachers assessed whether the attempt was successful or not (i.e. if the produced sound is similar to the demonstration). If refresh arrows at the right are clicked, the page is refreshed giving user the opportunity to start trying over and over.

The interface is deliberately made simple and illustrated with colors in order to easily determine what the user should do. The goal is to make our prototype easy to use and kind to be seen by children. Half-transparent letters in the background, aspire to give the feeling of learning content related to words. Buttons and Canvas are positioned in the center, in order to be found easily. Three buttons are illustrated with images, in order to avoid giving information to read. As it can be understood from their image, they serve to trigger three main actions which required by students: listening to demonstration, start trying and repeating (start-over). The Canvas also is placed in the center, aiming to facilitate focusing on it. The entire interface is designed to be easily understood, so that children can adapt to it without efforts.

5 Experiment and analysis results

Different statistical analyses have been performed in order to analyze the data collected and produce results that can give us confidence answering the research questions posed. The analysis started with an exploration of the data based on descriptive statistics and focusing on measures of central tendency and variation. We also performed standard statistical tests for the analysis of the data distribution. This procedure was performed for both scoring results: prior to training and after training.

We chose statistical models as One-Way ANOVA and ANCOVA due to several reasons. Between-subjects design of our experiment, the ability to show the difference between their means and variance, and post-hoc analysis are among the reasons that lead us toward choosing these two models for understanding trend of our data. Considering the fact that pre-test and post-test are a fair comparison between groups, by using test like one-way ANOVA we can estimate the effect of experiment and make inferences for each of these groups.

We have a pre-testing and a post-testing design and a treatment that is being applied between these two stages. We start exploring dataset by testing normality of data that we had from scores before and after treatment. Knowing the distribution of our data, then we could choose between regression models to use for generalization of results. We wanted to know, whether the groups were similar before training. After treatment we again explored the data about all the groups together and we to see if there is a difference between these two testing stages. Difference between means showed that no matter what the treatment was changes occurred. Further explorations were done in order to reveal whether the change occurred because of a chance or it was the impact of experiment. Basically, we aimed to find the statistical significance between trained groups. We start exploring dataset by testing normality of data.

5.1 Initial exploration of the data:

In order to find an instant picture about the spread of our data we used descriptive statistics. We also used normality test to see how data are distributed among groups. We began analysis by finding how data are distributed considering the central position. Central tendency of data was measured using mean which is the average score of the population given.

5.2 Testing results taken all groups together

First, we started to compute descriptive analysis by calculating the mean of all grades before and after training. The mean of sum of grades before training which was $\mu=43.25$ is a determiner that reveals that some changes had occur when we compare it with the mean for sum of grades after training which was $\mu= 80.53$. Standard deviations of (before training) $\sigma =0.81$ and (after training) $\sigma = 0.67$ were showing us that data are spread around the average value, which means that we do not have significant differences between participants. To prove this we continued to explore on our data by performing normality tests by using Kolmogorov-Smirnov and Shapiro-Wilk test. These tests compare the scores of groups to a set of scores that are normally distributed. Results of this test would tell us if our data between groups are deviating and whether we can expect observed values to fall within the range of normal distributed data. We performed this for total (or sum of) grades given before and after training and also for each word separately. In table below we have results of normality test for the sum of grades before testing:

Tests of Normality						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
sum Before	.135	28	.200	.955	28	.270

Table 3. Normality test for grades before training

As we can see from the table above, results had no statistical significance ($p=0.27$ in both tests) between groups, prior to training, which means that groups did not differ before training. These numbers tell us that data are distributed in such way that there is no significant difference between them, meaning we have normal distribution. According to Field (2009), if this significance value was less than $p < .05$ in that case it would mean that groups were differentiating a lot from each other, and as a result we would not be able to measure the effect of experiment appropriately. Similarly we conducted normality test for sum of grades after training and we had similar results. The figure below is a graphical representation of normality test:

Figure 18. Normality plot for sum of grades after training

With a significance level of $p=0.24$ and also from the graph we can see that after the experiment all groups as whole did not differ a lot, which is not surprising fact since they were all exposed to a certain experiment that aimed to teach them (therefore all of them scored better). Taken all the groups together we saw that these variables are normal.

5.2.1 Test results split by groups

SPSS environment allows computing different statistics by splitting the results according to a certain variable. We used this option to split our data based on groups while calculating their mean for all grades before and the after training. This means that as a result we got a mean for each group before training and then a mean for each group for the sum scores after training. Results are shown in table below:

	Traditional G	Highlight G	Wave G	Complete G
sumBefore	44.71	46.14	40.14	42.00
sumAfter	82.00	75.71	79.85	84.57

Table 4. Calculated mean for grades. Results grouped based on training approach

As we can see from the table above, averages of grades are approximately the same for groups before and after training. The difference between means of sumBefore and sumAfter variables, are demonstrating that there have been some changes in pronunciation, since students scored higher grades, however inferential statistics will help us to reveal the impact of experiment that lies in this difference.

5.3 Testing the differences between the groups

As we explained, regression models were used to explore the experiment effect and extract inferences about them. ANOVA is a useful statistical model that examines the variability amongst the means and compares it against the variability within each mean in terms of individuality of groups (Field, 2009). We used ANOVA tests for grades before and after. We expected not to have huge differences considering all group together. This was also determined by results of partial eta squared explained that 46% of variance is accounted by the treatment approach, which is a big percentage. This means that there was a difference between our groups somewhere. Table 5 below is one of the outputs from ANOVA testing:

ANOVA

sumAfter

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	294.964	3	98.321	6.820	.002
Within Groups	346.000	24	14.417		
Total	640.964	27			

Table 5. Output from One-Way ANOVA test

A one-way ANOVA revealed a significant effect of the treatment approach on error rate $F(3, 24) = 6.8$ and $p = .002$.

The weakness of this model is that although it can reveal if there is a difference between groups, it cannot tell where this difference lies. For this reason together with ANOVA we computed Post-Hoc tests like Bonferroni, which was comparing each group with the other (Field, 2009). We performed ANOVA and Post-Hoc test for grades before and then after training. Results of ANOVA were outlining that the variance between groups was significant (both before and after training), whereas the post-hoc test determined which means differed. Results (which you can find in the table in Appendix) determined that before training Highlight group was very different from three other groups by having statistical significance with third (Wave group) at the post hoc of before training grades, with a significance level of $p = .038$. If we go back at Table 4, we can see that Highlight group was starting from a higher score whereas the Wave group was starting from a lower score in comparison with Highlighting. Traditional, Wave and Complete group, although different their significance level did not exceed .05. The same tests were applied using after training grades. Results outlined the Highlighting group again as the most different one in comparison with others, whereas Traditional, Wave and Complete groups had less statistical significance between each other. Recalling the mean of scores from Table 4, we can reveal that Highlighting group although it started from an upper level of knowledge they scored worse than

other groups. Graphically the difference in means before and after training would look as in following:

Figure 19. Plotted means of groups for scores before (left) and after (right) the training

We can see from the graph that Highlight group started from a higher score in comparison with other groups, and scored lower after the treatment. Outlined results made obvious that highlighting the syllables was definitively the worst approach to teach pronunciation.

If we continue to check the difference between remained groups, we can see that Wave and Complete groups were very similar and both had less statistical significance with Traditional group. The fact that Wave and Complete are very similar, it might happen that the wave of sound is doing the effect. However, if we recall the mean of scores at Table 4, we can see that Complete group scored much better than Wave group with a mean of 84.57 compared to 79.85. These results make us confident that highlighting approach on itself was not helpful; however in combination with the sound wave it had better effect.

In order to increase statistical power of our results we also run ANCOVA (Analysis of covariance) testing. ANCOVA analysis is an ANOVA analysis but including covariates. According to Field, by including covariates we want to measure if there are effects other than from the experiment that influence final result. For example in our case we added before grades as a covariate variable, because we expected that student's score after treatment might be also affected by his knowledge prior to the treatment. This means that when we calculate the effect of experiment based on sumAfter variable, we assume that all groups are the same, therefore the result are determining the real effect of experiment. Part of this process was checking the violation of assumption¹⁰, which results we are not going to describe separately, instead summarize that none of the assumptions was violated. Outcomes of our ANCOVA testing (which are also put on Appendix) outlined that taken all groups together there was no significance ($p = 0.55$). However, comparing groups against each other showed that there was a significance between our groups at a level $p = .00$ when we control for sumBefore. Post-hoc analysis again outlined Highlighting group as less effective and also with high significance level in comparison with other. Also at Table 6 and 7 below, we represent the adjusted means which we will compare with means from initial descriptive statistics.

¹⁰ <https://statistics.laerd.com/spss-tutorials/ancova-using-spss-statistics.php> Accessed: 2013-05-26

Group	Mean
Traditional	82.00
Highlight	75.71
Wave	79.85
Complete	84.57
Total	80.53

Table 6. Descriptive Statistics for sumAfter variable

Group	Mean
Traditional	81.28
Highlight	74.30
Wave	81.37
Complete	85.81

Table 7. Estimates - adjusted means of sumAfter

If we compare these unbiased means of each group, we can see that adjusted means at Table 7 are higher for the Wave and Complete groups. This means that the effect of the experiment for these two groups has increased. Estimates computed by ANCOVA are just imaginary means, predicting what the results would be if all groups would start from same level. If we compute the difference for each pair of mean we can see that mean of Wave group increased more ($81.37 - 79.85 = 1.52$) in comparison with others. In discussion section these presented results are summarized in order to answer our research question.

5.4 Inquiring the teachers about the sessions run

A simple questionnaire was used to inquire teachers about their views of the sessions run and to what extent the tool developed was perceived usable and useful. Questionnaires have been used for a long time to evaluate the usability of user interfaces (Root & Draper, 1983). The particular questionnaire used was based on Perlman's (2011) examples of usability rating questionnaires. Questionnaires were created using Google Spreadsheets and were distributed to teachers through their Facebook accounts after all training sessions ended. We gave them two days to respond and meanwhile we met teachers informally and discussed about the whole evaluation stage. First ten questions were querying about and user interface satisfaction, usefulness and ease of use of the application. Since only three teachers were monitoring testing sessions using our prototype, we had three responded questionnaires. Due to this low number of answers we did not use any sophisticated tool to analyze them. Results are elaborated in the next subsection. Please check Appendix to have a complete view of the questionnaire. Beside the information that we extracted from questionnaires, informal conversations with children helped us to know their opinion about the prototype. In general, all children were satisfied with the tool they given in terms of usage and understanding. Their past experience with computer games helped them to easy adapt with interface and its use. Eventually, learning school content with tools that are simple and entertaining was attractive for them.

5.4.1 Results

Based on answers we saw that students were very excited to use the application and were satisfied with all its functionalities. According to their observation, teachers claimed that student's attempt to learn a word was mostly from 3 to 7 times. This, according to them was common whereas the teacher that used Highlighting approach of visualization pointed out that her student's attempt was mostly from 1 to 3 times. Teachers evaluated usability features with maximal grade, which means that application was easy to be understood, characters could be easily read and elements were arranged properly.

Last few questions helped us to reveal why the group that used Highlighting approach did not put a lot of effort on learning pronunciation. Teacher that was observing this grouped stated that: "There was no way to teach loudness, intelligibility and also determine the wave sound using this

part of prototype, which I think are essential when learning pronunciation. Students achieved to perfectly stick to the timing as the syllables were highlighted and for that not a lot of effort was needed. ” Teacher that observed Wave and Complete group agreed that timing, loudness, intelligibility and wave of sound were enough parameters to consider while learning pronunciation.

The teacher that was observing Complete group answered that not always students were capable to watch at both visualizations at the same time. This information will later become very meaningful during our statistical analysis. They were also asked to express their agreement with the statement: “As teacher I need to put extra effort for teaching pronunciation after using this application”. Teacher that used the Wave visualization was neutral; teacher observing the Complete group disagreed, whereas teacher training the Highlight group strongly agreed.

In addition, all teachers claimed that this approach would be useful to complement the traditional approach but not replace it. Together with statistical analysis, these results will be summarized in the Discussion chapter.

6 Discussion

In order to measure the impact of treatment in our experimental groups we collected and analyzed our data for both stages: before and after the experiment. Initial descriptive statistics gave us an overview about normality of our data, their mean, standard deviation and so on. Using ANCOVA analysis we aimed to discover how would student's scores change in relation to training approach.

Initial descriptive statistics outlined that we had normal distribution of the data regarding their means. Said otherwise, means of each group were correctly representing average of our data and we did not have extreme values that would require further processing of data. The difference in scores' average before the experiment and after that was an indicator proving changes that occurred. Descriptive results also outlined that one of the groups (i.e. Highlight group) started from higher level of scoring and scored worse after the treatment in comparison with others. In addition, the significance level calculated after using statistical models as one-way ANOVA and ANCOVA was proving that there was a significant difference between this and other groups before treatment. Accompanying to these fact we have the results that we had from our surveys where teacher was stating that this approach was quite effective for teaching student to stick to the timing. However, "...there was no way to teach intelligibility and loudness..." – she added. Due to these facts we can conclude that highlighting visualization style was not effective. Even though students' knowledge prior to treatment was better, the portion of prototype that they used did not help them to achieve even the equality with other groups.

After excluding the highlighting group from our focus, we examined the difference between remaining groups: Traditional, Wave and Complete. Means of these groups were closer in comparison with Highlight group. The fact that traditional approach was successful for learning pronunciation did not surprised us, since it is widely used in schools around the world therefore its effect should not be underestimated. However, in order to be able to prove our hypothesis either right or wrong, we had to focus on which approach between wave visualization style and combined visualization style would have better effect in learning pronunciation.

Descriptive statistics showed us that these groups were starting from lower level of scores that is 40.14 for Wave group and 42.00 for Complete group. Inferential statistics together with post-hoc tests were showing that there was no significance between these two groups after the treatment $p > 0.5$ ($p = 0.17$). Also prior to the treatment, these two groups were even less significant $p = 1.00$. This means that they were much more similar prior to the experiment rather than after that. Therefore, we can consider the effect of combined visualization as more successful in comparison with wave of sound visualization. The power of these outcomes was also augmented by ANCOVA results which are controlled by sumBefore variable. The significance level between Wave and Complete group shifted from 0.174 to 0.306 and also recalling Table 7 imaginary means were increased for both groups. The shift of significance level means that if groups started from the same level their scoring would be even more similar. In addition to these we have survey and interview results where teachers claimed that these approaches were useful to teaching pronunciation since they were exercising its features such as: timing, loudness and wave of sound. We will also recall teacher's comment about students that used Complete group, which was stating that "...some of the students were not able to look at both visualization styles at the same time." However, it might be the case that those students that were capable of looking at both styles scored much better than those who did not.

Considering our analyses and recalling our main hypothesis which was stating that:

“The use of a system that presents both (a) the visualization of the wave form of the spoken word and (b) the highlighting of syllables, will increase the children's ability to pronounce a given set of English words when comparing with traditional approach and systems that only employ one of the forms proposed”,

We can conclude that combination of visualization styles was quite more effective when comparing highlighting approach visualization style. In comparison with remained approach which is using the wave of sound visualization, there was not a significant difference. We do not have enough data to prove that the wave of sound was in fact affecting the whole improvement of pronunciation; however this remains as a gate for further investigation. Our results are determining that complete and wave approaches are relatively close to each other and considering the fact that the Wave group started from lower score makes us tied to the opinion that this approach was quite effective as well. Despite this, when we consider adjusted results of ANCOVA presuming what would the results be if groups were starting from same level, we see that Complete group in such circumstances will again perform better even though without a big difference. Meanwhile recalling teacher's comment about students' inability to look at both visualizations, we suggest to continue exploring on issues such as: measuring child's ability to perceive visualization or other skills that in our experiment could be the case. As a consequence, the answer of our research question will be another research question that will elaborate on these above mentioned cases.

Eventually, more investigation is required to complement our current conclusions. Currently we are optimistic about a strategy which we think will be the appropriate to determine what we doubt so far. These strategies are explained in future work chapter.

7 Conclusion and future work

Plenty of researches have indicated the importance of technology when learning foreign languages. We broaden studies in this field by examining the effect of web based application at second language learning. In this thesis we reported a research which in conclusion revealed the impact of combined real-time visualization styles of syllables that aim teaching pronunciation of English as foreign language. We developed a tool which key features were: real-time visualization of live audio input and highlighting syllables of a word.

Following experimental methodology we indicated that combination of visualization styles that: highlight syllables and draw the wave of sound coherently, has positive impact in learning timing, loudness, intelligibility and wave of sound. All these features, which theoretically are important when learning pronunciation, were incorporated in functionalities of our prototype.

Four groups of young learners in Macedonia, aged between 9 to 11 years old, were given our solution to experiment with it and statistical analysis were applied to gathered data. Results although not highly consistent, are outlining that combination of visualization styles, indicates higher scores in comparison with groups that used parts of the system and control group (using traditional approach). Surprisingly highlighting the syllables indicated that was less effective, despite the fact that a lot of existing tools for teaching pronunciation are employing that approach.

Above all, although the main hypothesis was claiming that the complete style will help pronouncing, results are not giving strong evidence that combination of visualizations is the case, but again we have to consider that there is a wave which possibly did the effect. Even though results are outlining combination of visualizations as effective, further research is needed in order to identify type of students that benefit from combined visualization styles. Nonetheless, it is important to mention that all approaches were quite effective except highlighting the syllables. As such, wave of sound on itself and also combined with highlighting style can be integrated at learning practices due to their positive effect. Outcomes of this thesis, were further validated after being peer reviewed in the form of the research paper included in Appendix E.

In order to have stronger evidence that will clearly distinguish the impact of combined approach with wave of sound, further study should be done. Aspects such as student's integration with visual representation, ability to leak these two sources of information can be further investigated. Student's capabilities to perceive information and learn, which we did not consider in our research, can be causer of good or bad performance and thus shape the results.

7.1 Limitations

Although we reached research's aims there are some unavoidable limitations that we could not correct even though the research was planned carefully. Apart from technical issues that we have in our prototype we were browbeaten by our sample characteristics. While performing statistical analysis we had to change the format of our data. A lot of tests were not significant therefore we could not distinguish the effect of experiment among the groups. According to Field (2009), having not a significance level it does not mean that you have to accept the null hypothesis which says the experiment did not have effect, instead it might be the case that we do not have enough data to test it. By improving the format of data we were able to perform all planned analysis for our data. Therefore, we are considering for the future to have bigger samples in order to generalize in broader community.

In addition, when performing the experiment we randomly chose participants without considering their individual capabilities to learn. We did not consider the fact that some students

are able to learn faster and some of them slower; therefore we did not apply any division during the treatment. There are possibilities that such division can clearly identify which system is more beneficial for a particular group. Perhaps, due to this limitation we cannot clearly measure the difference between sound wave and combined visualization based on this study alone. However, bigger sample and division based on individual characteristics are considered for future work that can be conducted on basis of this research.

Technical issues were another limitation in our research. As we have explained in our technical approach chapter, we manually set the milliseconds when highlighting of syllables will happen in order to accord it with sound of demonstration. Although we used advanced audio software to detect the time when a certain syllable is pronounced again we cannot claim that it was perfectly matching. Nevertheless, this issue was not highly prioritized since the difference in milliseconds was almost incapable to be noticed. In addition, there were no tools to adjust microphone volume, therefore when sound was too loud audio visualization was clipped inside the canvas. On one hand, volume balancing tool are highly affecting visualization of sound, on the other hand, loudness is an important factor when learning pronunciation. Therefore further study should be done in the future in order to elaborate on possibilities for optimizing volume of audio input and its impact on pronunciation learning. Improving these features will be considered for the future along with other research limitations.

As listed in Table 1 (Chapter 2), there are plenty of features that can be included when building tools that aid learning pronunciation of English language. Their impact was measured either separately or combined with others features. Our research delineates positive impact when highlighting syllables and wave of sound visualization are combined. However, there are still possible combinations that can be further examined. For instance, wave of sound visualization and acoustic similarities can be a potential combination, which effect can be measured. All these properties, upon which a prototype can be built, are a potential gate for further research.

7.2 Future work

The space for further investigation comes from the fact that combination of visualization styles in comparison with visualization of sound wave ensue approximate scores. We still have not proven that the approach taken in this work scales to all types of students. Bigger sample size is needed in order to generalize the results for broader community. Therefore, for future consideration is important to have bigger samples and also divide them based on their individual learning capabilities. If we had more students we could split the sample in to -low and -high i.e. students that are not good learning English (that might have different needs) and students that are good at English learning, we could now how this reflects to the system. Instead of four groups we would have 8 groups: high and low group for traditional, high and low group for highlight group and so on. Then we run another cycle of analyses comparing all of these groups again to see if there was a particular group that benefit from a certain system. Thereafter, we could say for example students that had difficulties at learning English benefit from the wave and the students that were good at learning English benefit from the complete functionalities. Although, we do not have enough students for this case we still think that students will react differently depending on their abilities. However, in order to prove that we will need a bigger sample in split into low and high group as explained above. We still need to continue investigate if individual differences do play a role at finding out what kind of system is more useful.

In addition, the features of the prototype may be expanded to match the notions discussed above. For example, based on their individual skills, students might need different timing and visual options in order to learn the content properly. Such capabilities can be built upon current existing

prototype and further evaluations can be done. Moreover, a progress tool could be another important feature of our prototype. We mentioned on literature review, that several tools also include scoring technologies. Similarly, we can add features that would inform student about their scores and progress over time. Undoubtedly, implementing this tool would include studies about existing scoring technologies and their level of impeccability. Eventually, derived issues from our statistical analysis stage and limitations can be addressed by conducting another research on basis of this study.

By creating a tool on basis of current prototype we can be implement it in real world cases; however, it still can be extended including the above-mentioned suggestion. As discussed in the introduction, positive outcomes of our teaching approach can make our prototype a valid option for teaching pronunciation practices. Since statistical analysis outlined that combination of visualizations and sound wave helped children as much as traditional approach of teaching did, we can convincingly say that our application can complement regular teaching practices. Depending which approach they find more useful, an additional feature to our tool can be the ability to choose visualization styles. Thus, our application can offer more flexibility by corresponding to individual capabilities. Therefore, we see the future of our application as a tool accessible from web environment, with expanded corpus (learning material) that will assist young learners. Moreover, further studies could reveal additional features from which individuals can benefit and further improvements can be made.

8 References

- Adair-hauck, B., Willingham-mclain, L., & Youngs, B. E. (1999). Evaluating the Integration of Technology and Second Language Learning. *CALICO Journal*, 17(2), 269–306.
- Ainsworth Sh. (1999). The function of multiple representation, *Computers & Education*, 33(2-3), 131–152., Elsevier Science Ltd., Oxford, UK. Retrieved from: [http://dx.doi.org/10.1016/S0360-1315\(99\)00029-9](http://dx.doi.org/10.1016/S0360-1315(99)00029-9) , accessed: 06.02.2013
- BEC - Barcelona European Council. (2002). Presidency Conclusions, Barcelona European Council 15 and 16 March 2002, Retrieved from: http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/ec/71025.pdf, accessed: 07.02.2013
- Boves, L., Neri, A., Cucchiari, C., Strik, H. (2001). The pedagogy-technology interface in Computer Assisted Pronunciation Training, 1–14.
- Conrad, B. K. (1996). CALL: non-English L2 instruction. *Annual Review of Applied Linguistics*, 16, 158–181.
- Dong, B., Zhao, Q., Zhang, J., & Yan, Y. (2004). Automatic assessment of pronunciation quality. SympoTIC '04. *Joint 1st Workshop on Mobile Future & Symposium on Trends In Communications (IEEE Cat. No.04EX877)*, 137–140. doi:10.1109/CHINSL.2004.1409605
- Field, A. P. (2009). *Discovering statistics using SPSS (and sex and drugs and rock "n" roll) (Third.)*. London EC1Y 1SP: SAGE Publications Ltd.
- Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll (4th ed.)* London: Sage. Retrieved from: <http://www.statisticshell.com/docs/onewayanova.pdf>. Accessed: 11.07.2013
- Gupta P., Schulze M. (2011). Human Language Technologies (HLT). Module 3.5 in Davies G. (ed.) *Information and Communications Technology for Language Teachers (ICT4LT)*. Slough, Thames Valley University. [Online]: http://www.ict4lt.org/en/en_mod3-5.htm
- Hailpern J., Harris A., Botz R. L., Birman B., Karahalios K. (2012). Designing visualizations to facilitate multisyllabic speech with children with autism and speech delays. *In Proceedings of the Designing Interactive Systems Conference (DIS '12)*. ACM, NY, USA.
- Highman C., Hennessey N., Sherwood M., and Leitˆao S. (2008). Retrospective parent report of early vocal behaviours in children with suspected childhood apraxia of speech (scas), *Child Language Teaching and Therapy* 24, 3, 285.
- Jenkins, J. (2002). A Sociolinguistically Based, Empirically Researched Pronunciation Syllabus for English as an International Language. *Applied Linguistics*, 23(1), 83–103. doi:10.1093/applin/23.1.83
- Johnson, D. M. (1992). *Approaches in second language learning*. White Plains, NY: Longman Publishing Company.
- Lai, Y.-S., Tsai, H.-H., Yu, P.-T., & Hsu, J.-M. (2007). A Multimedia Learning System Using HMMs to Improve Phonemic Awareness for English Pronunciation. *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, 1064–1069. doi:10.1109/MUE.2007.30
- Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32(4), 505–524. doi:10.1016/j.system.2004.09.009

- Mak B., Siu M., Ng M., Tam Y. C., Chan Y. C., Chan K., Leung K. Y., Ho S., Chong F. H., Wong J., and Lo J. (2003). PLASER: pronunciation learning via automatic speech recognition, In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2* (HLT-NAACL-EDUC '03), Vol. 2, Association for Computational Linguistics, Stroudsburg, PA, USA, 23-29.
- Neri A., Mich O., Gerosa M., Giuliani D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5), 393-408. Routledge.
- Ogata, H., and Yano, Y. (2004). Knowledge awareness map for computer - supported ubiquitous Language – learning, *Proc. of WMTE'04*, pp.19-26, IEEE Computer Society Press.
- Ogata, H., Paredes J., R. G., Saito, N. a., Yano, Y., Oishi, Y., & Ueda, T. (2006). Supporting Mobile Language Learning outside Classrooms. *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, (3), 928–932. doi:10.1109/ICALT.2006.1652595
- Perlman, G. (2011). User Interface Usability Evaluation with Web-Based Questionnaires. Retrieved from: <http://hcibib.org/perlman/question.html>. Accessed: 15.04.2013
- Pusack, J. P., & Otto, S. (1997). Taking control of multimedia. In M. D. Bush & R. M. Terry (Eds.), *Technology - enhanced language learning* (pp.1-46). Lincolnwood, IL: National Textbook Company.
- Rogerson-Revell, P. (2011). *English phonology and pronunciation teaching*. London: Continuum. ISBN: 978-0-8264-2403-7.
- Root, R. W., Draper, S. (1983). Questionnaires as a Software Evaluation Tool Interface Design 4 - Analyses of User Inputs. *Proceedings of ACM CHI'83 Conference on Human Factors in Computing Systems 1983-12-12 p.83-87*. Retrieved from: <http://hcibib.org/bs.cgi?searchtype=question&query=C.CHI.83.83>. Accessed: 15.04.2013.
- Smus, B. (2013). *Web Audio API*. M. B. Simon St. Laurent, Ed. First Edit. Sebastopol. CA., O'Reilly Media, Inc.
- Scaife, M., Rogers, Y., Aldrich, F. and Davies, M. (1997). Designing for or Designing With? Informant Design for Interactive Learning Environments. *CHI'97: Proceedings of Human Factors in Computing Systems*. ACM: New York, 343-350.
- Stenson, N., Downing, B., Smith, J., & Smith, K. (1992). The Effectiveness of Computer-Assisted Pronunciation Training. *CALICO*, 9(4), 5–19.
- Su, P., Chen, Q., & Wang, X. (2006). A FUZZY PRONUNCIATION EVALUATION MODEL FOR ENGLISH LEARNING. *Machine Learning and Cybernetics, 2006 International Conference on* (pp. 13–16).
- Tanner M. W., Landon M. M.. (2009). The Effects Of Computer-Assisted Pronunciation Readings On Esl Learners' Use Of Pausing, Stress, Intonation, And Overall Comprehensibility, *Language Learning & Technology*. Brigham Young University, Provo, UT, USA
- Velleman S. (2002). Phonotactic therapy, In *Seminars in Speech and Language*, vol. 23, 35–46.
- Wikipedia - Computer aided language learning (2013). Retrieved from: http://en.wikipedia.org/wiki/Computer-assisted_language_learning, accessed: 08.02.2013
- Wikipedia – English Language. (2013). Retrieved from: http://en.wikipedia.org/wiki/English_language#cite_note-4, accessed: 08.02.2013

9 Appendixes:

Appendix A. UML Sequence Diagram:

Appendix B. Table of post-hoc results from one-way ANOVA

Multiple Comparisons

sumAfter

Bonferroni

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Traditional	Highlight	6.28571 [*]	2.02954	.030	.4506	12.1208
	Wave	2.14286	2.02954	1.000	-3.6923	7.9780
	Complete	-2.57143	2.02954	1.000	-8.4066	3.2637
Highlight	Traditional	-6.28571 [*]	2.02954	.030	-12.1208	-.4506
	Wave	-4.14286	2.02954	.314	-9.9780	1.6923
	Complete	-8.85714 [*]	2.02954	.001	-14.6923	-3.0220
Wave	Traditional	-2.14286	2.02954	1.000	-7.9780	3.6923
	Highlight	4.14286	2.02954	.314	-1.6923	9.9780
	Complete	-4.71429	2.02954	.174	-10.5494	1.1208
Complete	Traditional	2.57143	2.02954	1.000	-3.2637	8.4066
	Highlight	8.85714 [*]	2.02954	.001	3.0220	14.6923
	Wave	4.71429	2.02954	.174	-1.1208	10.5494

*. The mean difference is significant at the 0.05 level.

Appendix C.

Post-hoc results for ANCOVA

Pairwise Comparisons

Dependent Variable: sumAfter

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Traditional	Highlight	6.982 [*]	1.836	.006	1.682	12.282
	Wave	-.087	2.005	1.000	-5.872	5.699
	Complete	-3.895	1.885	.302	-9.337	1.546
Highlight	Traditional	-6.982 [*]	1.836	.006	-12.282	-1.682
	Wave	-7.069 [*]	2.130	.018	-13.216	-.922
	Complete	-10.878 [*]	1.972	.000	-16.570	-5.185
Wave	Traditional	.087	2.005	1.000	-5.699	5.872
	Highlight	7.069 [*]	2.130	.018	.922	13.216
	Complete	-3.809	1.849	.306	-9.146	1.529
Complete	Traditional	3.895	1.885	.302	-1.546	9.337
	Highlight	10.878 [*]	1.972	.000	5.185	16.570
	Wave	3.809	1.849	.306	-1.529	9.146

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Appendix D. Questionnaire:

5/7/13

Thesis

Edit this form

Thesis

Hello:

You are invited to participate in our survey to help us know more about the web-based application that aims to teach L2 pronunciation. In this survey, 4 people will be asked to complete a survey that asks questions about the overall evaluation stage of our application. It will take approximately 10 minutes to complete the questionnaire.

Your participation in this study is completely voluntary. There are no foreseeable risks associated with this project. However, if you feel uncomfortable answering any questions, you can withdraw from the survey at any point. It is very important for us to learn your opinions.

Your survey responses will be strictly confidential and data from this research will be reported only in the aggregate. Your information will be coded and will remain confidential. If you have questions at any time about the survey or the procedures, you may contact Dorina Dibra at dorina.dibra@gmail.com.

Thank you very much for your time and support.

1. What was student's level of knowledge before the experiment?

- ☐ Poor
☐ Average
☐ Good

2. Did students understand the goal of application?

- ☐ Yes
☐ No

3. How was the interaction of children with application?

1 2 3 4 5

Poor ☐ ☐ ☐ ☐ ☐ Excellent

4. Reading characters on screen was:

1 2 3 4 5

Hard ☐ ☐ ☐ ☐ ☐ Easy

5. Sequence of screens was:

1 2 3 4 5

Very Confusing ☐ ☐ ☐ ☐ ☐ Very Clear

6. Position of elements on screens was:

1 2 3 4 5

<https://docs.google.com/forms/d/1xVqq7yhEJB-jelVdlIJrXr19cqCHtG40FkQDYDc/viewform>

1/3

Inconsistent ☐ ☐ ☐ ☐ ☐ Consistent

7. Learning how to use the system was:

1 2 3 4 5

Difficult ☐ ☐ ☐ ☐ ☐ Easy

8. How often did errors appear on application:

1 2 3 4 5

Never ☐ ☐ ☐ ☐ ☐ Always

9. Did students in your group felt excited about using this application?

☐ Yes

☐ No

10. What was the level of satisfaction after using the application?

1 2 3 4 5

Not satisfied ☐ ☐ ☐ ☐ ☐ Very satisfied

11. Did children asked to have additional tools in the application? If yes, please specify!

12. Average student's attempts to learn a new word was:

☐ 1 to 3 times

☐ 3 to 7 times

☐ more than 7 times

13. The idea of this prototype was to teach pronunciation considering: timing, loudness, intelligibility and wave of sound. Do you think there are other metrics that should be considered when learning pronunciation? If yes please specify.

14. Students that were using prototype with all functionalities were able to look at both visualization styles all the time:

1 2 3 4 5

Strongly disagree ☐ ☐ ☐ ☐ ☐ Strongly agree

15. This approach of learning pronunciation can substitute or complement the traditional approach?

- ☐ Substitute
☐ Complement

16. As teacher I need to put extra effort for teaching pronunciation after using this application:

1 2 3 4 5

Strongly disagree ☐ ☐ ☐ ☐ ☐ Strongly agree

17. What functionality you would add to this prototype?

Never submit passwords through Google Forms.

Powered by
Google Drive

This content is neither created nor endorsed by Google.

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Appendix E. Published research paper

Dibra, D., Otero, N., Pettersson, O. (2014). Real-Time Interactive Visualization Aiding Pronunciation of English as a Second Language. *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on.* vol., no., pp.436,440. doi: 10.1109/ICALT.2014.131

Real-time interactive visualization aiding pronunciation of English as a second language

Dorina Dibra, Nuno Otero, Oskar Pettersson

*School of Computer Science, Physics and Mathematics (DFM) Linnaeus University
dorina.dibra@gmail.com, (nuno.otero, oskar.pettersson)@lnu.se*

Abstract - Some studies have been reporting encouraging results concerning the possibilities of combining voice and its visual representation for language learning. Following this line of investigation this paper explores the potential impact of two distinct visualization styles for the learning of English pronunciation of syllables for non-native speakers: a) highlighting syllables and b) the visualization of the produced sound wave. In order to evaluate the benefits of the two different styles three distinct digital tool prototypes were created in order to test four study conditions. The conditions under evaluation were a) teaching syllables without the support of any digital tool; b) teaching using a prototype that highlighted the syllables under study; c) using a prototype that displayed the sound wave of the syllable to be learnt and d) a prototype that combined the functionality of b) and c). Results suggest that the combined approach seems to be as effective as the traditional classroom approach of teaching the syllables. However, more research is needed in order to consolidate the findings, being able to explore in more detail how is the learning process occurring and to what extent the tools developed can be integrated into classroom practice.

Keywords- *computer assisted language learning; second language learning; visualization styles of syllables; pronunciation teaching; Web Audio API.*

I. INTRODUCTION

The use of multimedia for the teaching/learning of a second language is not a recent trend: gramophone records, film-strip projectors, videocassette recorders and so on have been utilized with, probably, different degrees of effectiveness. Recently, however, the evolution of digital and web technologies has shaped these learning tools into self-learning web solutions, application and even virtual language centers. Furthermore, some of these technologies are especially focused on fostering the learning of proper pronunciation [1].

Several studies are outlining the importance of learning pronunciation at an early age [2]. Such indications have attracted researchers and developers of Computer-Aided Language Learning (CALL) to put a considerable effort on creating innovative ways to help students learn pronunciation using speech technology.

Despite plenty of research that have been conducted on this field and is summarized in following section, there are still gaps, in particular regarding the impact of combined visualizations. For example, we do not know much about the possible beneficial effect of combining wave-form and

highlighted words representations on children that learn pronunciation of foreign languages.

The present contribution inscribes itself in this field. More specifically, the system under investigation intends to support the learning of pronunciation by allowing the student to check her performance pronouncing a specific word using certain combination of visualization styles. In other words, the feedback being provided by the visualization styles will scaffold the learning of the correct pronunciation in an intuitive way. In this research we will also compare the performance of students using the systems developed (considering that we will test more than one prototype as explained further ahead) with the achievements of students that learn using a traditional approach. Furthermore, in order to make sure that we were meeting reasonable pedagogical aims while giving to students a system that is designed to teach pronunciation, teachers of English as second language were involved in the entire process of tool design and teaching activity. In a nutshell, our research queries the impact of specific combination of visualization methods and compares them with traditional approach where no technological solutions are used.

II. RELATED WORK

Much effort has been put to find different technological solutions that enable students to learn second languages without the constant need of professional assistance. Pronunciation of words is a stage that greatly affects the speech development process [3].

A. Learning language pronunciation

Rogerson-Revell [4] argues that drilling, dictation, noticing and phonetics training are among the best approaches that help second language (L2) learners to exercise pronunciation. She explains that the learner should refer to aspects such as: stress-timing, rhythm, vowel reduction and intonation while learning. Pronunciation problems are more related to the sound. Articulatory settings are very important for achieving intelligibility therefore the improving voice skills at early stages is among the suggested ways to achieve proficiency in L2 pronunciation she concludes. Furthermore, learning pronunciation at early ages requires less, if any, effort comparing with in adulthood [2].

Another key factor that enhances communication intelligibility is the modulation of voice [5]. Movements of pitch, voice quality and way of expressiveness are very important for speaker's attitude and are very tied to the

intonation. In all, the studies reviewed on this section emphasize the importance of learning pronunciation as a stage that greatly affects the fluency of speaking a second language.

A. Strategies for pronunciation learning

Lai et al. [6] point out that traditional ways of teaching where teacher asks the learner to practice pronunciation, results in inaccurate outcomes. He argues that this approach of teaching avoids the recognition capability of right and wrong pronunciation. If the learner is aware of the word phonemes pronunciation, he can easily pronounce the word by disassembling the word tones. They proposed a Multimedia Learning system that uses HMMs (Hidden Markov Models) to analyze the pronounced word and provide feedback about the pronunciation, intonation, rhythm and volume. Their system performs phonemes' analysis based on which the feedback is given. In a study run the tool was given to 56 students aged 8-9 years old and results showed that low scoring students improved their pronouncing abilities, whereas others scored better in spelling and reading skills.

B. Visual representations for the learning of pronunciation

Some studies suggest that providing multiple visualizations of the content to be learned is beneficial for learners. A thoughtful use of Multiple External Representations (MERs) can assist learning by being able to explore the potential representational capabilities of different types of external representations. Ainsworth's taxonomy addresses the important issue of the distinct types of relationships that different external representations might establish between themselves in a learning system and how these might be used to foster learning. She considers that a crucial part of the evaluation of a CBLE that uses multiple-external representations is to understand clearly which functions are present and how they fit with the learning goals. Designers need to understand how to combine distinct external representations *to complement, to constraint and construct* the topic under scrutiny [7].

Chun [8] claims that multimedia tools that employ automatic speech recognition are quite useful for learning pronunciation of L2, however they do not meet pedagogical aims. According to her, more efforts are needed on improving the mechanism that server for providing feedback in order these tools to be as effective as human listeners.

Another example is the work by Levis and Pickering [5] that explored the impact of visual representation on L2 learning. They used Computerized Speech Laboratory (CSL) program. Sentences were recorded and then used with the program to analyze phrasing and pitch. Users were allowed to do different exercises and practice discourse-level speeches while the tool was analyzing all pitch movement. Data gathered from this study was analyzed and outlined that this way of visualization increased learners' ability to understand how to perform intonation and pitch movements in real communication environments.

C. Existing solutions for teaching pronunciation

Not surprisingly, some researchers also consider speech production and vocalization as an important development stage of the acquisition of communication skills for students with impairments]. The visualization of information has therefore emerged in many applications that aim to learn language proficiency. Hailpern et al. [9] designed the VocSyl software: a real-time audio visualization system. This syllable-based system, gives a visual representation of timing, pitch and volume of the sound as a word is being pronounced. The evaluation conducted comprised several rounds of testing, by assessing different perspectives of speech and the authors suggest that the results obtained are promising.

Speech recognition technology within this field usually has two distinct goals: the support for learning accurate pronunciation of foreign language and the evaluation of the pronunciation quality [10]. For example, a multimedia tool called PLASER contained a pronunciation video-clip with animated pictures aiming at teaching students the movements that should be done in order to pronounce correctly. In addition, pre-recorded samples were available as demonstration. The feedback conveyed in the visual representation comprised calculating the overall score of phonemes spelled and the accuracy of each phoneme spelled. Research results showed that this system improved the pronunciation to 77% of students that participated in testing and that 99% of teachers approved that.

II. RESEARCH QUESTION

Based on the description of the problem considered in the previous sections the research work can be summarized with the following research question:

- *Does the combination of real time a) highlighting syllables visualization and b) wave form of speech visualization, increase the pronunciation skills of students learning English language?*

It is important to mention that linguist related issues are not among the objectives of our research. Our main focus is to explore from media technology perspective features that serve for pronunciation learning. Wave of sound visualization and highlighting syllables, are the unique combination which impact we aim to investigate.

III. IMPLEMENTATION

Foundation of properties of our prototype derived from the combination of features that we propose to include in our tool, such as: listening to demonstration, visual representation of sound wave of pre-recorded demonstration, visualization of produced sound wave, timing mechanism, real-time feedback and comparison of performance and demonstration. These features were implemented as a web app in HTML5 and JavaScript using the Web Audio API for sound processing and playback. Visualization of live input and highlighting words required the combination of different components. Based on functionality, our application can be logically divided into three nodes. The nodes and their relations are depicted in Fig. 1

Figure 1. System architecture

The Web Audio API is created based on the Audio Routes concept. This in practice means that different nodes can be added and connected to each other through edges [11]. The nodes used in the prototype are elaborated upon in the following sections.

1) Source Node

SourceNode determines the source of audio context. In our architecture it has two sub-modules: buffering source and live input. Buffering source is an array where audio file resides after it is being loaded from the server. When the sound from memory is triggered, buffer array populates with audio file data and starts decoding. In this case the *SourceNode* is directly connected to speakers which are in *DestinationNode*, elaborated later on this chapter.

Live input on the other hand can be generated from for example a microphone. In order to use the microphone to start audio capture, the user must grant his browser the right to access the computer's microphone. If this is successfully granted, the captured audio is routed to the Filtering Node. This is an *AudioNode* that can be connected between source and destination nodes.

2) Filtering Node

After audio input has been captured from microphone, it is routed to the filtering module, where processing occurs. The Analyser Node interface of Web Audio API is the filtering module, which enables to do real-time processing of audio characteristics based on FFT (Fast Fourier Transform) analysis. Rendered (analyzed) data is temporarily stored in an array, which is used as input for the wave form visualizer. In order to have a sound wave that is representing the entire audio input, we had to visualize a wave of entire input from left to right ordering sound values in same sequence as they are produced. Therefore, we used time domain analysis through the *getByteTimeDomainData* method.

3) Destination Node

In order to enable user to hear any sound generated by the application, we have to connect the generated audio to a destination node. This node is used for enabling the learner to hear the prerecorded pronunciation used as demonstration.

The highlighting functionality is implemented using an index for the syllables and through that assigning them timeslots for when the change of color should occur.

II. THE EXPERIMENT

Our application was conceived to assist young learners while learning pronunciation of syllables of English words. Fig. 2 is a capture of the application's interface, where apart from section of buttons, we have combined (1) highlighting module where word is divided into syllables and highlighted at the right time when it should be pronounced and (2) canvas where sound wave of the demonstration and produced sound is drawn. In this interface visualization styles are combined.

Figure 2. Screenshot taken during practice to pronounce syllables of motherhood. Next highlighted part (-er-) is in turn to be pronounced

During the experiment, these functionalities were also separated into independent modules as illustrated in Fig. (3) and (4) below. The effectiveness of these visualization approaches was evaluated both (1) separately as illustrated in Fig 3. and Fig 4. below and (2) combined as depicted in Fig. 2. By exercising pronunciation using one of these systems, we expect students to learn correct pronunciation of the given learning content.

Figure 3. Interface containing only highlighting module

Figure 4. Interface containing only sound wave module

A. The study conditions

For this study a between subject design was followed, comparing the results obtained in a pronunciation test from groups that went through distinct conditions. The four conditions were:

1) *Traditional Group* – was the group that learned the words in traditional way without intervention of any digital system.

1) *Highlight Group* - the participants included in this group learned using only the highlighting visualization style of the prototype.

2) *Wave Group* – in this group the participants only had access to the visualization style which was drawing the wave of sound based on the sound produced.

3) *Complete Group* – in this group the participants interacted with the prototype that enabled the visualization of two above mentioned styles: highlighting and showing the wave sound form.

B. The participants and procedure

Our sample included students from primary school aged between from 9-11 years old. There were no students in the sample group with known disabilities that could reflect the result (for example, speech delays, dyslexia, etc). The size of the sample was 7 students per group (28 students in total), where participants were randomly assigned to one of the conditions considered. Every student was given 6 words to learn during three sessions (18 words in total). Teachers did pre- and post- training evaluation for each word for each student with grades from 1 to 5 (1 – very poor, 5 – excellent).

II. ANALYSIS

A. Analyzing the results of the learners in the two pronunciation tests

The results being reported in this sub-section concern the two measures used: the pronunciation tests before (sumBefore) and after (sumAfter) the interaction with the systems under investigation (except with the control condition where the teaching relied on the traditional classroom delivery). Following the research question proposed the idea was to investigate if the use of the distinct systems could show statistically significant differences on the students' achievements after usage. Furthermore, by using a pre-test and post-test design we can flesh out the progression from one stage to the other and also control for potential a priori differences in the composition of the groups formed.

Initial analysis of the overall distribution of the scores indicated no violations of normality. TABLE I, shows the unadjusted means of the scores obtained by the students in the two tests taken. Considering this framing an ANCOVA model, using the pre-test pronunciation scores as the covariate, was run.

TABLE I. NON-ADJUSTED MEANS AND SD FOR THE SCORES ON THE TWO PRONUNCIATION TESTS BY CONDITION.

	Traditional Group	Highlight Group	Wave Group	Complete Group
sumBefore	44.71	46.14	40.14	42.00
sumAfter	82.00	75.71	79.85	84.57

TABLE I shows the differences between the scores' averages before and after the interactions in the different conditions and indicate that changes occurred. The descriptive statistics

also show that one of the groups (i.e. Highlight group) started from higher level of scoring in the pronunciation pre-test but scored worse in the post-test.

TABLE II. SUMMARY OF SOME STATISTICS OF THE ANCOVA MODEL

	df	F	Sig.	Partial Eta
Corrected Model	4	8.12	.00	.59
Intercept	1	54.75	.00	.70
sumBefore	1	6.94*	.02*	.23
Group	3	10.60*	.00*	.58

TABLE II shows that the covariate, sumBefore, and the independent variable Group are statistically significant. This suggests that indeed the different conditions tested produced distinct results in relation to the students' ability to pronounce the words taught.

TABLE III. ADJUSTED MEANS OF SUMAFTER (COVARIATE=SUMBEFORE = 43.25)

Group	Mean	SD
Traditional	81.27	1.31
Highlight	74.30	1.39
Wave	81.37	1.40
Complete	85.18	1.31

TABLE III shows the adjusted means of the model run. In other words, taking into consideration the potential effect of the covariate sumBefore.

TABLE IV. STATISTICALLY SIGNIFICANT PAIRWISE COMPARISONS (ADJUSTMENT FOR MULTIPLE COMPARISONS: BONFERRONI)

Groups	Mean diff.	SD	Sig.
Trad. vs High.	6.98	1.84	.01
High. vs Wave	-7.07	2.13	.02
High. vs Comp.	-10.88	1.97	.00

TABLE IV shows the statistically significant differences from the pairwise comparisons between the conditions. All other comparisons did not produce statistically significant differences. Hence, it seems clear that the less effective teaching method involved the use of the Highlight prototype. This suggests that only showing the highlighting of the syllables is not a way forward.

The answer to our research question is not straightforward since although the complete system seems to produce better results the differences found to the other conditions was not statistically significant.

B. Analysis of the survey with teachers

Teachers claimed that student's attempt to learn a word was mostly from 3 to 7 times, except for the group that used only Highlighting approach of visualization whose attempt was mostly from 1 to 3 times. In our survey, the teacher observing Highlighting group stated that this approach was quite effective for teaching the student to stick to the timing. However, "...there was no way to teach intelligibility and loudness..." – she added. This could be the reason why they

did fewer attempts for exercising a word. Teacher's comment about students that observed Complete group, stated that "...some of the students were not able to look at both visualization styles at the same time." However, it might be the case that those students that were capable of looking at both styles scored much better than those who did not. Nonetheless, our results are not highly consistent to prove that the wave of sound was in fact affecting the whole improvement of pronunciation. Beside the information that we extracted from questionnaires, informal conversations with the participants revealed that they were satisfied with the tool given in terms of usage and understanding.

I. DISCUSSION AND CONCLUSION

Previous work in the field has indicated the potential benefits of technology when learning foreign languages. We broaden studies in this field by examining the effect of a MER focused web-based application for practicing pronunciation in second language learning. This application utilizes real-time visualization of live audio input and the highlight of syllables. Using this implementation, we investigated the impact of combined real-time visualization styles of syllables as a method for teaching pronunciation.

Previous work has indicated that combination of visualization styles such as highlight syllables and visualizing the sound wave has a positive impact in learning timing, loudness, intelligibility and wave of sound. These features were implemented in our prototype. During the experiment, groups of students used the complete system and parts of it. The complete system employed combination of visualization on single interface, whereas modules of the system employed each visualization style on separated interfaces. All groups were given the same content to learn, in three sessions of practice. Results, although not highly consistent, are showing that a combination of visualization styles resulted in higher scores in comparison with groups that used only parts of the system and the control group. Furthermore, only highlighting the syllables was shown to be less effective, despite the fact that a lot of existing tools for teaching pronunciation are employing that approach.

Even though results are showing that a combination of visualizations are effective, further research is needed in order to identify which type of students benefit from the combination of the visualization styles. Nonetheless, the outcome of the analysis has hinted towards the other methods being a promising component for second language learning. However in order to deduce just how effective it is, a larger sample is needed.

In addition, in order to have stronger evidence that will clearly distinguish the impact of combined approach with wave of sound, further studies should be done. Aspects such as student's ability to "read" and integrate the information from the visual representations needs to be further investigated. We believe that the potential differences in student's capabilities to perceive and understand the information conveyed by the distinct forms of representations can have a discernible influence on the results obtained.

REFERENCES

- [1] Gupta P., Schulze M. (2011). "Human Language Technologies (HLT). Module 3.5 in Davies G. (ed.)", Information and Communications Technology for Language Teachers (ICT4LT). Slough, Thames Valley University.
- [2] Neri A., Mich O., Gerosa M., Giuliani D. (2008). "The effectiveness of computer assisted pronunciation training for foreign language learning by children", *Computer Assisted Language Learning*, 21(5), 393-408. Routledge.
- [3] Highman C., Hennessey N., Sherwood M., and Leitˆao S. (2008). "Retrospective parent report of early vocal behaviours in children with suspected childhood apraxia of speech (scas)"
- [4] Rogerson-Revell, P. (2011). "English phonology and pronunciation teaching", London: Continuum. ISBN: 978-0-8264-2403-7.
- [5] Levis J., & Pickering, L. (2004). "Teaching intonation in discourse using speech visualization technology", *System*, 32(4), 505-524. doi:10.1016/j.system.2004.09.009
- [6] Lai, Y.-S., Tsai, H.-H., Yu, P.-T., & Hsu, J.-M. (2007). "A Multimedia Learning System Using HMMs to Improve Phonemic Awareness for English Pronunciation," *International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, 1064-1069.
- [7] Ainsworth Sh. (1999). "The function of multiple representation, *Computers & Education*", 33(2-3), 131-152., Elsevier Science Ltd
- [8] Chun, D. M. (2007). Come ride the wave: But where is it taking us?. *CALICO Journal*, 24(2), 239-252.
- [9] Hailpem J., Harris A., Botz R. L., Birman B., Karahalios K. (2012). "Designing visualizations to facilitate multisyllabic speech with children with autism and speech delays", In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. ACM, NY, USA.
- [10] Mak B., Siu M., Ng M., Tam Y. C., Chan Y. C., Chan K., Leung K. Y., Ho S., Chong F. H., Wong J., and Lo J. (2003). "PLASER: pronunciation learning via automatic speech recognition", In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2 (HLT-NAACL-EDUC '03)*, Vol. 2, Association for Computational Linguistics, Stroudsburg, PA, USA, 23-29.
- [11] Smus, B. (2013). *Web Audio API*. M. B. Simon St. Laurent, Ed. First Edit. Sebastopol. CA., O'Reilly Media, Inc.