Rajs Arkadiusz, Banaszak-Piechowska Agnieszka, Drzycimski Paweł. Speech synthesizer based on the project MBROLA. Journal of Education, Health and Sport. 2015;5(12):160-164. ISSN 2391-8306. DOI http://dx.doi.org/10.5281/zenodo.35266
http://ojs.ukw.edu.pl/index.php/johs/article/view/2015%3B5%2812%29%3A160-164
http://pbn.nauka.gov.pl/works/678876
Formerly Journal of Health Sciences. ISSN 1429-9623 / 2300-665X. Archives 2011–2014
http://journal.rsw.edu.pl/index.php/JHS/issue/archive

Deklaracja.
Specyfika i zawartość merytoryczna czasopisma nie ulega zmianie.
Zgodnie z informacją MNiSW z dnia 2 czerwca 2014 r., że w roku 2014 nie będzie przeprowadzana ocena czasopism naukowych; czasopismo o zmienionym tytule otrzymuje tyle samo punktów co na wykazie czasopism naukowych z dnia 31 grudnia 2014 r.
The journal has had 5 points in Ministry of Science and Higher Education of Poland parametric evaluation. Part B item 1089. (31.12.2014).

The authors declare that there is no conflict of interests regarding the publication of this paper.
Received: 10.11.2015. Revised 25.11.2015. Accepted: 10.12.2015.

SPEECH SYNTHESIZER BASED ON THE PROJECT MBROLA

PhD. Arkadiusz Rajs
University of Science and Technology in Bydgoszcz
Faculty of Telecommunications and Electrical Engineering,
Al. Prof. S. Kaliskiego 7, 85-796 Bydgoszcz, arajs@utp.edu.pl
PhD. Agnieszka Banaszak-Piechowska
Institute of Physics, Faculty of Mathematics, Physics and Technique Science, Kazimierz Wielki University,
85-052 Bydgoszcz, Powstańców Wlkp. 2, agnb@ukw.edu.pl
Eng. Paweł Drzycimski
student, University of Science and Technology in Bydgoszcz
Faculty of Telecommunications and Electrical Engineering,
Al. Prof. S. Kaliskiego 7, 85-796 Bydgoszcz, pdrzycimski@gmail.com

Summary: Speech synthesizers find wider application in IT systems to ensure easy Communications with the Man and machine. In this article focused on speech synthesizer of polish language realization based on international project MBROLA.

Keywords: speech synthesizers, IT systems, man, machine, polish language, MBROLA.

Introduction

The pace of technical changes in nowadays world is stunning. Currently buys much more computers than cars. Multimedia technologies reach to the offices, schools and also houses. Connection of text, sound, static image and video creates a entirely new possibilities in the transmission of various kinds of information and technology management. Today user don't have to even strain his arms and even eyes to check his mails or check communicator. For full voice communication with the computer we need two things – possibility of understanding by computer human speech, and also saying sentences by computer, it mean speech synthesis. It should be noted that when the first of these is the relatively young, that the second one – speech synthesis – is the field over which the study began in XVIII century. Thanks to computers, technological solutions are coming closer to precise mapping human speech. Speech synthesizers are widely used in today's world, although not everyone of us must be aware of this issue.

The main purpose of this work is to create speech synthesizer of polish language based on MBROLA project and assessment in terms of generated signal quality.

Speech synthesis methods

Methods for signal processing and analysis of speech must be based on knowledge of its structure, the structure and substantially depends of its generation method. Until recently production of speech signal was the domain of natural systems, that is human articulation organs. Currently, in addition to natural signal sources, we need to consider technical speech signal sources (Fig. 1.). Sometimes artificial systems that generate a signal are pretending natural process of speech articulation, occurring in the human voice track (synthesized speech). However, nowadays much more popular and commercial succeeded are systems based on reconstruct samples chosen from natural speech signal and recorded in memory of generating system [1].

This is the way few times indicated earlier but not covered by the relevant notion. Concatenation method is currently the most common method of speech synthesis, taking into account also synthesizers of synthesized speech. Model of this one, developed since the 70', has gained much popularity thanks to possibility of generating very natural, and well-sounding and understandable speech in quite simple way. First synthesizers was generating low quality speech, because it did not sound natural and was not understandable. Progress in technology field enabled to reach better effects. Concatenation speech synthesis generates speech by bonding together acoustic elements formed from natural speech (phone, diphone, triphone). Main advantage this kind of speech synthesis is the small size of database, due the small capacity of acoustic units. The smaller size of database, the speech synthesis will be faster and hardware requirements will be smaller.

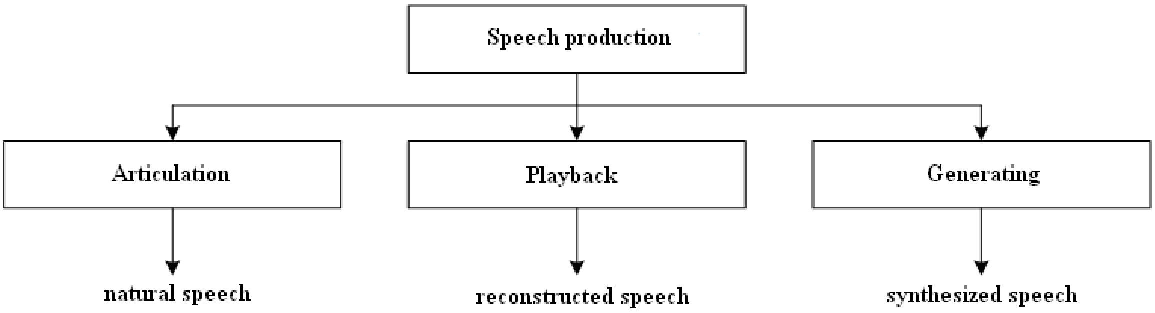

Fig 1. Methods of speech synthesis

Concatenation method of speech synthesis has also few disadvantages, belong to them:

- Which units chose to record
- Units concatenation recorder in different contexts
- Recorded segments compression problem
- Prosody modification, problem of intonation and duration

Today concatenation speech synthesizers generates very high speech quality. Therefore it become the object of interests such application as telephone services, computer education or speaking toys. Synthesizer made in practical part of this work also belongs to this group of synthesizers, and database of acoustic units is compose of diphones only.

MBROLA Project

MBROLA is an international project, which was started by Thierry Dutoit in 1995 in Belgium at the University of Mons. Project is realization of speech synthesis from phonetic text to speech. Main goal of the project is propagating speech synthesis and arouse interest of this branch especially in academic society. Objective of the MBROLA is create speech synthesis system covering as much languages of the world as it possible. Generally, MBROLA was regarded as very important step forward in production multi-language speech synthesis, since 1996 to 2002 were adding over 10 more diphone databases where half of this was in new languages, there is no other synthesizer that can work with more number of languages at year 2011. Looking at the present can be seen, that possibilities of cooperate with MBROLA by free software speech synthesizers, for example eSpeak engine or Festival system, is recognized as very important advantage.

MBROLA “can not” read text, we need for that frontend application or program wrapping MBROLA in larger engine. Program ‘Mbroli’ illustrate very well, how can we work with “clear” MBROLA features. We can use it just after install ‘MBROLA Tool’. Each letter of word in phonetic notation (SAMPA transcription) we need put in new line, after letter we have to write how long, in millisecond, each phone has to persist. Then, in brackets, we put two other parameters – first (percentage unit) determine place, where phone reach the basic tone frequency (F0), the second one determine the basic tone [6][12]. Now it’s obvious, that using MBROLA synthesizer in that way is a quite problem even for a person, who is familiar with speech synthesis subject. MBROLA provide “only” units concatenation from the base, which we indicate by phonetic notation, and possibility of sets few sound parameters. The word ‘only’ was marked by quotation marks cause we can consider MBROLA features in two ways. If we are more interests of theoretical field of speech synthesis we say that, MBROLA provide the most important part of synthesizer – search through the base and concatenation voice samples. But if we look at from practical perspective, we can easily say that we have to do some additional work to make synthesizer intuitive usage. In this article we ventured statement that both arguments are right. Good quality of synthesis largely depends on searching and concatenation algorithm, but on the other hand nobody won’t use it if we have to care about such details as frequency of base tone or phone duration. It’s not enough to write user interface if we want to build application based on MBROLA, we have to make translator from orthographic text to phonetic notation, take care about correct settings loading etc. Therefore using MBROLA is fair compromise between build synthesizer from scratch, and ready to use engines, which do not require knowledge about speech synthesis.

Synthesizer implementation and evaluate the quality of the generated signal

Polish speech synthesizer was implemented using Cpp programming language and a set of portable libraries of Qt Framework (Graphic user interface). Polish diphone database was downloaded from official MBROLA website and is named “pl1”.

In synthesizer, engine space has been strongly demarcated from interface space (Fig. 2.). Such architecture provide more comfort with work with code, for example when we are trying to find error in text parsing method we don’t have to analyze code opening or creating the window. Using such fragmentation we don’t have to worry that we cause an error in engine part of code when we’ll be during interface code modification. Finally, a developer who will inherit the project, will have less work with program understanding and will have easier task to build better interface using his favorite technology, which don’t have to be Qt.

The hole project was divided into three logical layers, according with the Model-View-Controller assumptions. “Model” layer is responsible for the proper program logic. “View” layer contains procedures responsible all that, what user see. “Controller” layer manages the work of other layers, other words that is space of the code, form which layers “Model” and “View” getting informations how to behave.

When we builds speech synthesizer we have to consider completely different criteria for speech difficulties that in human pronunciation case. Machine don’t have problem with saying sentences, which ones causes troubles for human, for example synthesizer won’t have any problem with say sentence “W czasie suszy szosa sucha” or “Jola lojalna, jola nielojalna.”, assuming, that text input is put correctly. If we check quality of synthesizers output signal quality and implementation class of it, we need to look at the tool like on machine and predict, where the problems can occur.

In the describing here speech synthesizer included such derogations between speech text and orthographic text:

- Numerals – reading single numbers
- Phonetic rules of polish language – the most common relationship between sounds
- Abbreviations – speech rules and expand some of them
- Custom chars as input – skipping chars like "@#$"

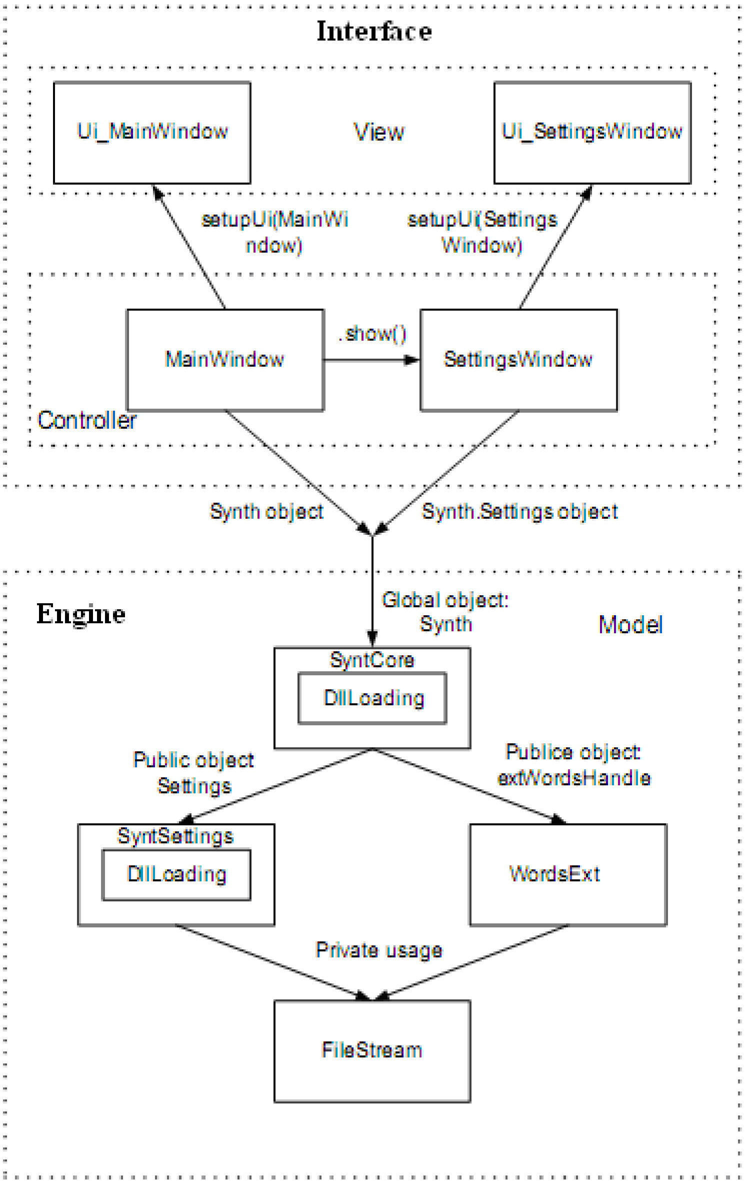

Fig 2. Flowchart of the speech synthesizer

Synthesizer in its current form has no mechanism implemented responsible for prosody simulation. However MBROLA core provide such possibility, and modular structure of the engine makes rescale some elements easier. If we wanted to add prosody module to the engine, we should start from took back of the user possibility of set parameter named "Length of the pitch". While accent is simulating (one of the part of prosody) there is necessary to changing this parameter dynamically, so it mustn't be "rigidly" set to each phoneme. We should also consider changing volume parameter before accented syllable. Volume rate shouldn't be significant, half of the unit is max. Volume rate also should be based on the volume value set by the user. Other thing, is solve problem of the words accent exceptions, we should study proper literature to get know which words belongs to them and how many is them.

Next step in prosody simulation is base tone frequency rate dynamically changing in various sentence types. In declarative sentences that frequency falling in the finals fragments of the sentences, mostly on the last or the last two words. In the interrogative sentence we can observe opposite situation - frequency rises on the last words. In the exclamative sentence tone frequency falls rapidly on the first words and have quite constant value in the rest parts of sentence. It is hard task to clearly calculation in which point of sentence frequency has to start descending or rising. Literature is not clear in that subject, so it would need to make lot number of tests to make prosody natural at least for usual cases.

Conclusions

Based on engine built on core of MBROLA was formed synthesizer, which provide decent sound quality and some flexibility in realize user requirements (possibility to set own voices, expanding database of abbreviations and phonetic exceptions). Quality of generating signal allows to full understanding output voice without any problem. Engine does not create difficulties in adding prosody module. Add this statement is first task in further development of synthesizer.

References

1. Tadeusiewicz R., Sygnał mowy. WKiŁ, Warsaw, 1988.
2. Szklanny K., Preparing the polish diphone database for MBROLA system.
3. Wierzchowska B., Opis fonetyczny języka polskiego. Warsaw 1967.
4. PJWSTK, types of synthesizers, http://www.syntezamowy.pjwstk.edu.pl/synteza.html
5. Dutoit T., MRBOLA intonation, http://tcts.fpms.ac.be/synthesis/mbrola/mbruse.html
6. Basztura Cz., Komputerowe systemy diagnostyki akustycznej. Warsaw, 1996.