human speech,
artificial speech generator, blind person.

Piotr PORWIK, Marek SZCZEPANKIEWICZ*

THE VOICE SYNTHETISER OF POLISH TEXT FOR BLIND PERSONS

In this paper we present new method of computer text analyser and computer Polish speech (words) generator. In the described computer program the grammatical characteristics of Polish speech and accents in some words have been taken into consideration. All users' actions are commented by artificial, computer voice. The group of blind students of University of Silesia have examined and tested the presented final program for over one year. Described software tool has in a lot of cases better parameters than others, commercial products.

1. INTRODUCTION

There is a fundamental difference between the system we are about to discuss here and any other talking device. In the sense that we are interested in the automatic production of new sentences. Systems that simply concatenate isolated words of parts of sentences are only applicable when a limited vocabulary is required and when the sentences have very restricted structure, as is the case for the announcement of arrival at railway station for instance. Obviously it is impossible to record and store all the words of the language. Speech synthesis is applied in many domains of human life: in language education, talking books and toys, vocal monitoring, multimedia communication. High quality speech synthesis can be used in many others disciplines: in science, study, industry, aviation, etc.

Many people all over the world have damaged organ of sight. Nowadays, there are a lot of methods, which assist such persons in their daily life: tape recorders, Braille books, CD-ROM's. Unfortunately the process of preparing such materials is long and its content very quickly becomes outdated. Today, we can use computers and digital signal processors as modern tools assisting blind people. These tools are designed very often as the text analysers and speech synthesisers. As the synthesisers the specialised DSP devices can be used or popular and cheap computer sound devices.

Additionally, on the basis of human voice the sex and age of persons can be recognised. That is why, both voice analysis and synthesis are proposed and developed for several years by many investigators [2,3,4,5,6]. The widespread using of these techniques was difficult due to the associated computational complexity. Nowadays, modern digital technologies allow to overcome these difficulties.

* Division of Computer Systems, University of Silesia , ul. Bedzinska 39, 41-200 Sosnowiec, Poland

The main engine in these projects is *text-to-speech* module. The *text-to-speech* synthesiser is a computer-based system that should be able to read any text aloud, whether it was directly introduced into the computer by an operator (by cassette recorder, word processor, e-mail, etc.) or scanned and submitted to an OCR system. Considering that main users of *text-to-speech* modules are blind people, all interfaces should be prepared in user friendly technology. Additionally all keyboard actions should be supported by voice. The artificially generated computer voice should always have the characteristics of human speech, of course. Unfortunately some commercial products addressed to blind persons do not possess such conveniences.

The artificial voice generation is very hard problem, because we must take into account not only the properties of human voice but also some ethnic restrictions: abbreviations, accents, habits, phonetic transcriptions, etc. The phonetic transcription depends on processed signs sequence into phoneme sequence. For instance Polish *ch* will be replaced by *h* (word *chomik* [hamster] will be read as *homik*), on the same principles Polish word *krzak* [bush] will be read as *kszak*. Therefore during sound generation we must take so-called ADSR characteristic into consideration. This characteristic is common for all languages but some its parts can be different for national languages.

The ADSR characteristic presents Fig.1. The ADSR characteristic has four phases: attack,

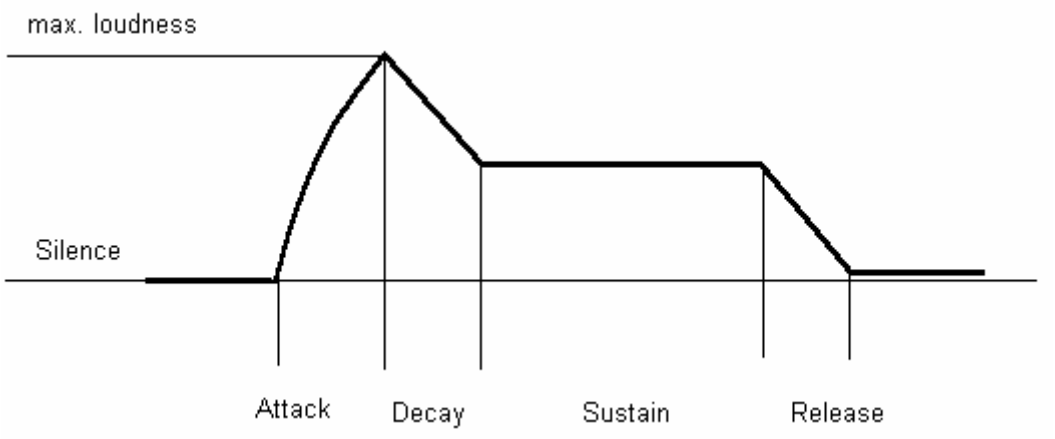

Fig.1. The ADSR (Attack/Decay/Sustain/Release) envelope

decay, sustain and release. This one presented in Fig.1 stages applied in any uttered word. More details about this problem can be found in [1].

Generally a value of stress is attached to each syllable to create a distinct stress pattern for a linguistic form such as word and phrases. When a strong stress is attached to a syllable, the duration of the syllable is longer. At the same time, a stressed syllable is pronounced with a higher pitch than unstressed syllables in the same word or phrase. Such a designation of a syllable in a word with stress is called lexical accent.

We can distinguish two main synthesis speech methods:

1. The synthesis by means of acoustics elements connecting. In this case a set of ethnic phonemes is used. This method allows to synthese any words in a selected language. Unfortunately these methods offer very low quality of computer pronunciation, because all phonemes sounds the same: at the beginning and at the end of a word. In reality this situation does not occur.

2. The synthesis by means of rules. In this case we connect the phonetics units, which are known and stored in memory. Besides samples, we must know sound parameters. For all phonemes frequency, amplitude and during time are fixed by means of ADSR's characteristic. This method assures computer pronunciation of a very high quality. Unfortunately accurate phonetic transcription can only be achieved if the dependency relationship between successive word is known. The rule-base synthesisers are mostly favoured with phoneticians and phonologists.

In this paper we use the second method but instead phonetics units we analyse a word with split into beginning syllables, into end syllables and into syllables which are located in the middle of the word. Additionally there are analysed the potential inter-mediate states between phonemes. It is done because in these cases the spectrum distortions can appear. In such situations audio devices can generate a noise. The suitable sound samples should be recorded taking into consideration the ADSR characteristic. The main principles of formed samples will be presented in next part of the paper. Some of speech synthesisers were designed and performed in Poland, for example the Syntalk or the Poltalk. These applications will be compared with method described below. In recent years many published works [3,4,5,6] provide information about new methods and technologies of human speech synthesis and analysis. The described problem has not found satisfactory solutions yet. The method proposed in this paper can be treated as attempt of a new view of problem.

2. THE METHOD DESCRIPTION

All sound samples used in our application are recorded and modified in Microsoft Waveform Audio File (WAV) format. The WAV files created in the program have also the structure of Resource Interchange Files Format (RIFF). During a word analysis, the samples existed in database are examined and a new WAV structure is created. This structure is a sound representation of analysing word. All samples components are stored into the Paradox database as WAV format files. The components are the fragment of Polish cluster letters, syllables and exceptions. Each word can be a single sample or a concatenation of a few samples. During a text analysis, the algorithm checks whether an encountered string can be treated as number, word, abbreviation or separator. As separators will be identified all signs which are not alphabet letters and are not digits. Some separators are always received (=,–,+,@,\ , etc.). Some separators can be omitted (sign of space, comma, colon, semicolon, etc.).
Depending on the above mentioned analysis following steps can be executed:

- If the investigated text can be treated as a number, then from database is taken immediately an appropriate sound form.
- If a word was identified, then the phonetic transcription would be carried out: $ch \mapsto h$, $ż \mapsto rz$, $ó \mapsto u$, $v \mapsto f$, $q \mapsto ku$.
- If an abbreviation were identified, the appropriate sound form would be taken from database.
- If separator was identified, the behaviour is similar. If option of separator reading is inactive, then separator signs are not analysed.

The user can switch (on/off) of spelling option. If this option is active, then:

- Numbers are uttered as single signs.
- Words are uttered only and phonetic transcription is unavailable.
- Abbreviations are treated as sequence of independents letters.

2.1. DATABASE PREPARATION

Many *text-to-speech* systems are based on phonemic decomposition. Exist systems that decompose text into constituent phonemes along with decompose with basic timing information and associated waveforms. Usually, a speaking person records each phoneme in a natural voice. The spoken phonemes are recorded as primitive waveforms.

The *text-to-speech* systems mix these waveforms together to form a continuous flow of speech (Fig.2). If all database segments have been completed, synthesis can begin itself.

The database used in application has a very simple structure and consists of two attributes. The first attribute characterises the sequence of signs of analysed text. The second attribute includes name of voice sample, which is compatible with analysing text. The pattern of database using in speech generation presents Fig.2

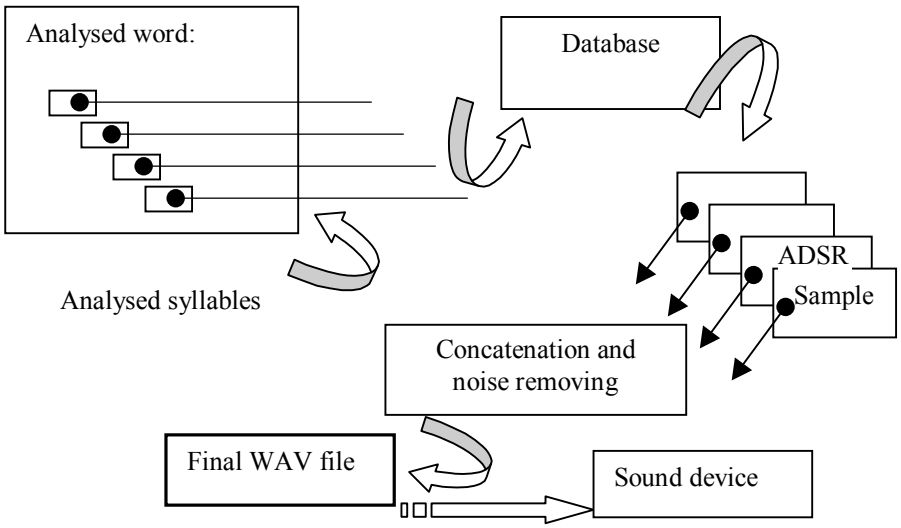

Fig.2. The principles of syllables association with database data

Program has been written as Builder C++/ver.5.0 application.

Main algorithm of text analysis has been presented in Fig.3. In this algorithm the REST and the PRESENT are signs chain variables. The mentioned variables are indexed. In Fig. 3 the PRESENT:=PRESENT-PRESENT[1] notation describes an operation where the chain of the PRESENT is cut at one sign. The TEMP variable describes a working file. The FINISHED variable pointed final, processed file, where sound representation of analysed word is stored. Special, additional sign '_' is joined for the sake of programming needs. This is recommend because the same sounds are generated differently depend on the occurrence in analysed word. This means that sound stresses in words can be different. Therefore the artificial generated speech is similar to human voice.

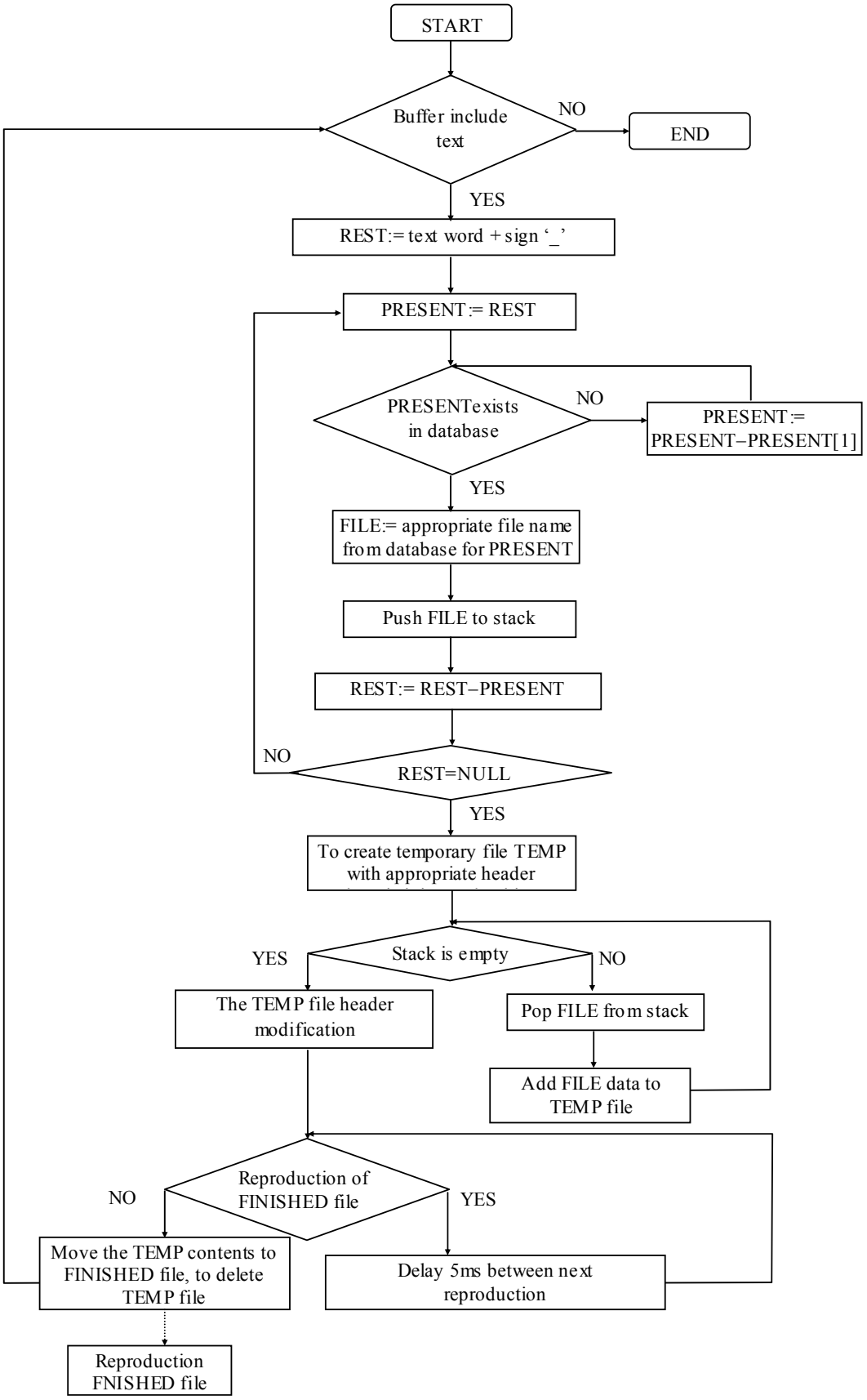

Fig.3. The algorithm of the text analysis

3. EXPERIMENTAL RESULTS

A few programs dedicated for blind persons can be found on Polish market. Some of them are produced in Poland. Unfortunately their prices are relatively high or very high for Polish users. In Table 1 the most popular products from this trade have been collected.

Name of product	Price
RealSpeak	1000 € (EUR)
Kubuś Lektor – polish language	2850 PLN
SMP-4 – polish language	1750 PLN
WinDOTS	4852 PLN
VisulexLP-DOS	9105 PLN
Jaws for Windows NT	5584 PLN
Hal and Lunar for Windows 95	2494 PLN

Tab.1. Prices of some speech synthesisers.

After talking with blind students and basing on Polish market analysis we have estimated, that the most popular products are the Syntalk (Neurosoft) and the RealSpeak (ScanSoft). It is necessary to emphasise that RealSpek synthesiser was presented for the first time in 1998. In 1999 this product received "The Best Speech Technology" title. For this reason we compared three text-to-speech synthesisers: the SynTalk our program and the RealSpeak. All of them are program synthesisers.

The tests have been implemented on a personal computer with the Celeron procesor 800Mhz and Soundblaster card. All procedures have been started from CD-ROM drive (32×). The operation system MS Windows98 has been used.

The time characteristic (synthesis) of some Polish words has been showed and compared bellow.

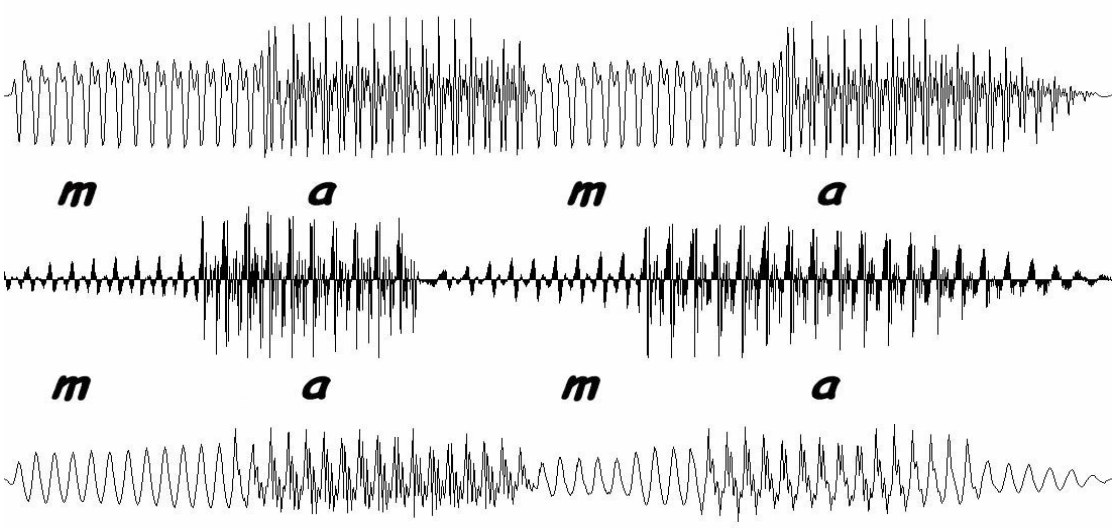

Fig.4. The Polish word synthesis "mama" [mum]. Program SynTalk (top), our program (middle) and program RealSpeak (bottom)

From the above-presented results it is obvious that the SynTalk synthesiser generates the voice more “mechanical” than voice generated by our method. Moreover, from the presented examples we can observe, that in our method the syllables have the different pronunciation depending on place occurrence. By analysing the waves from Figs 4,5,6 and 7 follows that our program fluently has accented sound in a word. The SynTalk program reads all syllables in the same way.

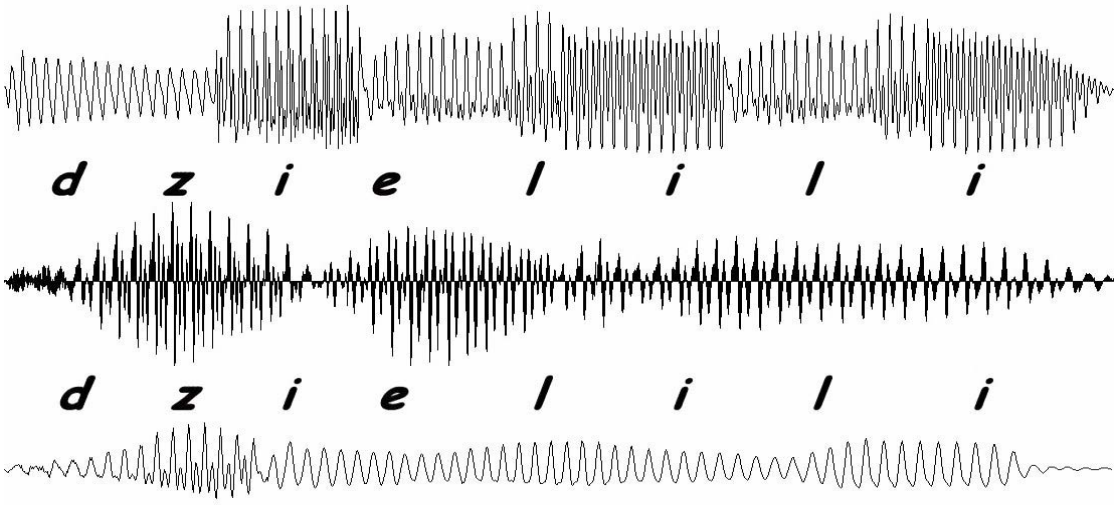

Fig.5. The Polish word synthesis "dzielili” [divided]. Program SynTalk (top), our program (middle) and program RealSpeak (bottom)

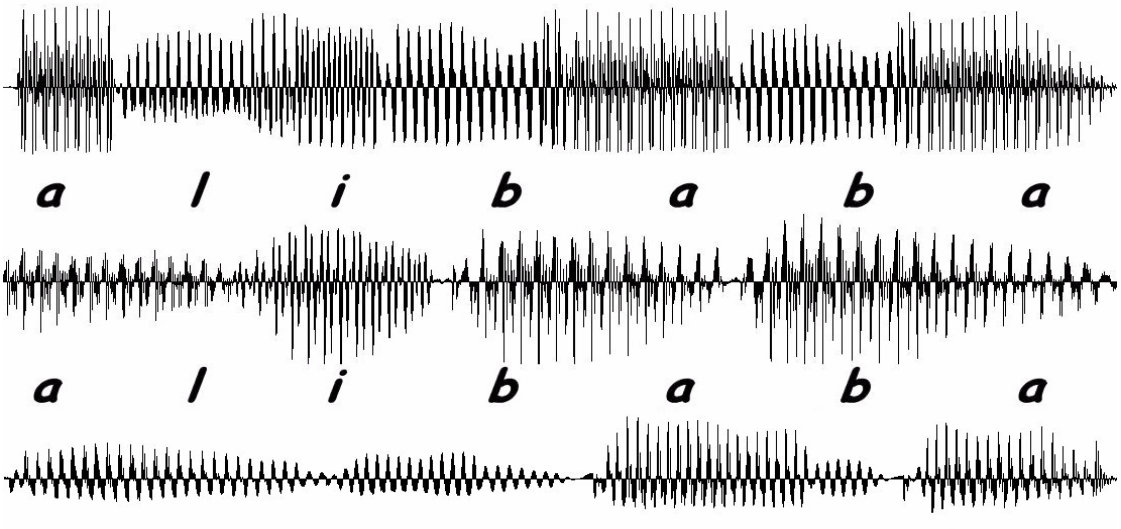

Fig.6. The word synthesis "alibaba”. Program SynTalk (top), our program (middle) and program RealSpeak (bottom)

The RealSpeak program in speech synthesis uses a woman voice, therefore its evaluation is a little difficult. From the described results follow, that the RealSpeak generates speech similarly as our program. In practice the RealSpeak application generates speech of a very high quality – but notice that this product is a very expensive and that is why it is used by few persons. In blind users

opinion our algorithm is better than the SynTalk synthesiser and in many cases it has better parameters than the RealSpeak synthesiser. All blind users did not have problems with understanding the speech generated by the proposed new method. Moreover, our application in their opinion has many additional options, which do not have others products (even RealSpeak).

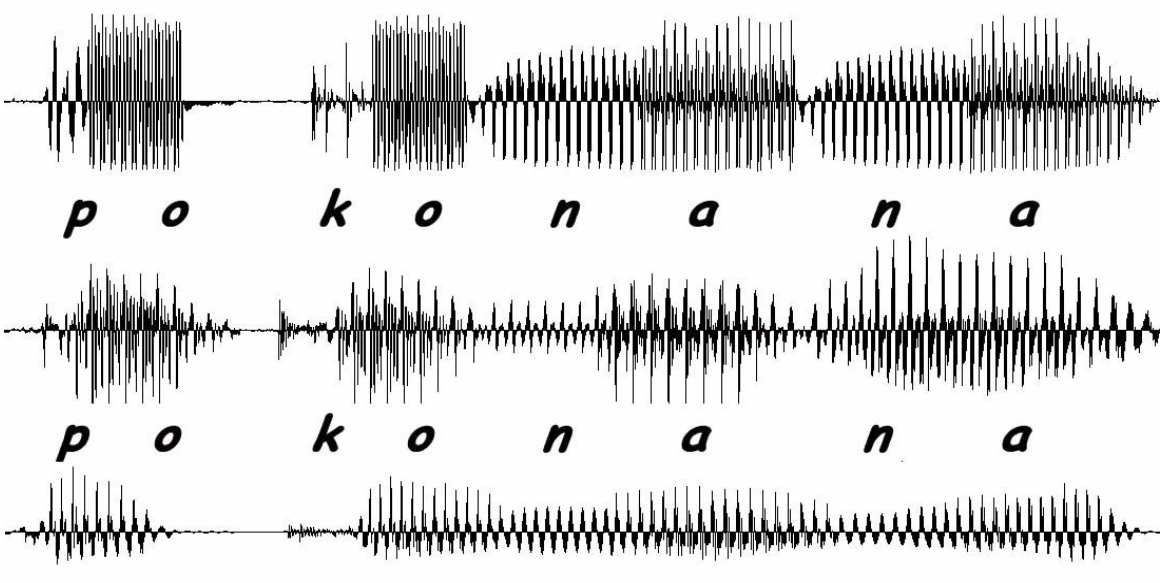

Fig.7. The Polish word synthesis "pokonana" [beats]. Program SynTalk (top), our program (middle) and program RealSpeak (bottom)

4. FINAL REMARKS

In this paper, we have developed a new efficient method for *text-to-speech* program synthesiser. The presented application was compared with others products. The final version has taken all suggestions submitted by blind persons into consideration.

All program procedures and databases can be started and executed immediately form CD-ROM. This is very convenient for blind persons, because the program can be easy carried and used. In practice hardware requirements are very low. The database includes 1466 suitably prepared of human voice WAV samples, what makes possible to generate any Polish sentence. Blind persons have tested the presented method – especially our university blind students. Its users have given the good opinion regarding the utility the of program. Some of them use this product till now.

All menu actions and messages are generated by a female voice – it is convenient for beginners. The main text is read by a male voice.

The pronunciation module provides pronunciations for most ordinary words and morphological derivatives as well as proper names. In the program the default strategies exist for pronouncing words not recognised by the database dictionary. Other components, such as the prosodic phrasing, word accentuation, sentence intonation have been analysed.

BIBLIOGRAPHY

[1] CZYŻEWSKI A., Dźwięk cyfrowy- Wybrane zagadniena teoretyczne, technologia, zastosowanie. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.
[2] NORBERT K. Programowanie kart dźwiękowych w Pascalu. Wyd. Lynx-SFT, Warszawa 1995
[3] TADEUSIEWICZ R., Sygnał mowy. WKŁ Warszawa, 1983.
[4] O'MALLEY M.H., Text-to-speech Conversion Technology. Computer. August 1990, pp. 17-23.
[5] WITTEN I.H., Principles of Computer Speech. Academic Press, 1992.
[6] D. DEW, P.J. JENSEN, Phonetic processing: the dynamics of speech. Merrill Publishing Company, Columbus, Ohio, 1977