ISSN 2278-3091

Volume 10, No.1, January - February 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse351012021.pdf

https://doi.org/10.30534/ijatcse/2021/351012021

Syllable-Based Reading Model for Uzbek Language Speech Synthesizers

Utkir Khamdamov[1], Bakhtiyor Akmuradov[1], Djamshid Sultanov[1], Elbek Zarmasov[1]

[1]Tashkent university of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan.
utkir.hamdamov@mail.ru, b.u.akmuradov@gmail.com,sdjamshid@gmail.com, zarmasov.elbek@mail.ru

Abstract. The possibilities of modern technology, which has become a part of our lives, are increasing day by day. In particular, new areas of research are emerging on the effective use of information and communication technologies and the further simplification of solutions to everyday problems. There are also a lot of research on the recognition of human speech and the formation of artificial speech signals to simplify the use and management of existing electronic systems. This research proposes a method of generating artificial speech signals in the Uzbek language based on the syllable-based reading of textual information.

Key words: Speech, language, text, sound, speech synthesis, syllable.

1. INTRODUCTION

The issues of computer synthesis and recognition of human speech have become increasingly important since the day speech technologies entered our lives. Modern online information and electronic services require the widespread use of the achievements of natural language technologies. The main reason for this is that almost everyone has the ability to speak and understand speech. The development of natural language systems allows people to use devices such as cars and mobile phones anytime, anywhere without any additional skills.

Despite the wide range of development of speech systems, the problem of speech synthesis still exists and is currently considered to be only satisfactorily solved. It has not yet been decided which of the available approaches will give the best results and which models of speech synthesis are the most promising.

According to the interdisciplinary feature of the field, there are some difficulties in the development of the Uzbek language synthesizer. In addition to the use of modern methods and algorithms, it is necessary to take into account the phonetic features of the Uzbek language. It is possible to develop a quality system by in-depth analysis of text elements, perfect study of spelling, orthoepy and punctuation rules.

2. LITERATURE REVIEW

Text To Speech (TTS) technology, which has been known in the computer market for nearly a quarter of a century, is typically used in applications that need to convert a large number of different texts into speech signals. The main feature that distinguishes TTS from previously developed sound programs is the ability to pronounce words based on phonetic rules and a set of pre-recorded sounds [1].

Of course, this does not mean that TTS technology is the final stage. In the near future, applications based on artificial intelligence using TTS technology are expected to be able to recognize sound and "understand" the meaning of speech. Much progress has been made in this regard. For example, Younggun Lee, Taesu Kim, and Soo-Young Lee's research samples on "Voice Imitating Text-to-Speech Neural Networks" or S. Arik, C. Chrzanowski, A. Coates, and others on Deep Voice: Real-time neural text-to speech" and other research in this area. It is clear that efforts to create a new generation of improved TTS system are not in vain [2,3]. It is obvious that in the development of intelligent systems, great emphasis is placed on the use of modern neural networks. Examples include Heiga Zen, Andrew

Senior, and Mike Schuster's Statistical Parametric Speech Synthesis Using Deep Neural Networks [4].

It should be noted that speech synthesis is based on knowledge of many scientific disciplines. For example, linguistics, psychology, human physiology, computer technology and others. It is necessary to analyze the structure of the sentence, as a result of which the pronunciation, intonation of individual words and the optimal rhythm of the synthesized speech are determined taking into account the syntax and semantics [5].

3. PHONETIC UNITS OF SPEECH AND LEXICAL STRESS IN UZBEK LANGUAGE

Thoughts are expressed in words and phrases. Sentences usually consist of words. Words in a sentence and the grammatical means that connect these words are represented by certain speech sounds. It turns out that speech sounds are the material basis for the construction of words and sentences.Phonetics is a branch of linguistics that studies the sounds of speech, their formation, types, variations, accents, syllables, and tones. Phonetics is derived from the Greek word for *phone*, which literally means sound.

A sound is the smallest unit of speech that is not phonetically indivisible, it does not express meaning, but any word is formed by sounds. So words are made up of sounds, which are made up of a number of sounds, one, two, three, four, and so on.

A phoneme is a type of sound that serves to differentiate the meaning of a word. Phonemes have three signs:

1. Acoustics (hearing)
2. Articulation (pronunciation)

3. Difference of meaning (in some literatures it is called "linguistic side")

The most important of these is the difference of the meaning sign of the phoneme. Otherwise the sound cannot be considered a phoneme.Phonetic acoustics is the study of the physical properties of speech sounds. Acoustically, any sound is a vibration of the air stream and this vibration is audible to the ear. Speech sounds are the sounds made by the vibrations of the air streams coming from the lungs and the noise made by the organs of speech.

Phonetic means are the means of distinguishing and limiting the meanings of words. These include speech sounds, accents, and melodies. The phonemes that serve to differentiate the meanings of words have been discussed above. All words and grammatical forms in dictionaries are formed by the sequential arrangement of these phonemes. Speech sounds are divided into two types: vowel sounds and consonant sounds. These are as follows:

1. In the formation of vowel sounds, the air coming out from the lungs passes through the oral cavity without any obstruction, while in the production of consonant sounds, the air coming out of the lungs encounters various obstructions in the oral cavity, resulting in noise;
2. vowels can be pronounced long, but consonants cannot be pronounced long;
3. In the pronunciation of vowels, the vocal cords vibrate, and in the pronunciation of consonants, the vocal cords may or may not vibrate. According to this feature, consonants are divided into voiced and unvoiced types [6].

In Uzbek language, letters are divided into vowels and consonants according to the fact that they are not obstructed in the pronunciation of the mouth and throat, and the symbols that represent them are called vowels and consonants, respectively. As a result, the Uzbek language spelling based on the Latin alphabet has 6 vowels and 26 consonants (Table 1).

Table 1: Grouping of vowels and consonants in the Uzbek Latin alphabet

# Type	1	2	3	4	5	6	7	8	9	10	11	12
consonants	Bb Rr	Dd Ss	Ff Tt	Gg Vv	Hh Xx	Jj Yy	Kk Zz	Ll G‘g‘	Mm Shsh	Nn Chch	Pp ng	Qq
vowels	Aa	Oo	Ii	Ee	Uu	O‘o‘						

Human speech is not a continuous stream of sounds. It contains intonation units of various sizes. Phrases, bars, phonetic words, syllables and sounds are such phonetic units.

1. A phrase is a unit of speech that is distinguished by a larger pause in the flow of speech. It is often equated with speech because of its tone and mental completeness. Phrases vary in size and can consist of one, two, three or more words.
2. Tact is part of a phrase, followed by a short pause. The word or words in it are pronounced with a single accent. The phrase is sometimes equivalent to one bar.
3. A phonetic word is a word that has its own accent or a combination of two or more words that are combined into one accent. The number of phonetic words in a beat is determined not by the total number of words in it, but by the number of accented words, that is, the more accented words in a beat, the more phonetic words ladi. Auxiliary words that do not have an independent stress have the same stress as the independent word to which they belong and form a phonetic word.
4. A syllable is a sound made by a flow to the air stream coming out of the lungs, or a single sound. A vowel is a syllable. So the more vowels in a word, the more syllables there are.
5. A sound is the smallest, phonetically indivisible unit of speech in a syllable or word structure. For example, each letter represents a sound.

Lexical stress is a part of speech or one of the words in a sentence is pronounced stronger or longer than the others. The lexical stress is usually on the vowels of the word. No matter how many syllables a word has, it has one lexical stress. Highlighting one of the parts of the sentencehas the similar features. The accented syllable in a word is called a stressed syllable, and otherwise is called an unstressed syllable. This means that the lexical stress can be on a different object - a syllable in a word, or a word in a sentence.

4. ANALYSIS OF TEXT ELEMENTS

It is advisable to organize the work of the sound synthesizer based on the Uzbek language in the Latin alphabet in several stages. Before developing a sound synthesizer, it is necessary to study the world experience and, based on this, to develop methods and algorithms that are compatible with Uzbek grammar and phonetics. The main function of a sound synthesizer is to convert input text into audio signals. Such synthesis systems use a number of algorithms depending on the type of text and its spelling features. Therefore, it is logical to begin the initial process of synthesizer development with the analysis of the text [7]. The following is a sequence of speech synthesizer functions.

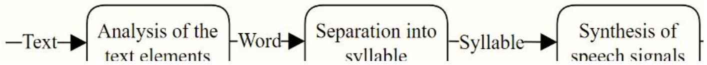

Figure 1. Step-by-step organization of the work of the sound synthesizer using syllable-based reading

In the text element analysis block, the incoming text data is analyzed in the following sequence.

1. Checking the alphabet in which the incoming text is written. This synthesizer is a special case that only synthesizes text in the Latin alphabet of Uzbek language. Therefore, if the text is written in the Latin alphabet, the process is moved to the next stage, otherwise the user is informed that the system is not able to synthesize such text;

2. Once it is determined that the incoming text is written in the Latin alphabet, it is analyzed what elements it consists of. The presence of letters, symbols, punctuation marks, informational signs, and symbols that are familiar and unfamiliar to the system in general is explored. If an unfamiliar character is observed in the system, it is stored in a special memory, and in the future the creators of the system will carry out appropriate analytical and organizational work on this character;
3. Determining how many paragraphs there are in the text information, which consists of symbols familiar to the system. Paragraphs are taken as a step in text analysis. The introduction of this rule will allow to avoid waiting in the process of synthesizing large amounts of data, as well as in the process of synthesizing continuous text data entering in real time, and the system will not fall into overload mode;
4. Determining the number of words in the text of a paragraph. A system iss introduced a word which is a conditional that is a sequence of characters separated by paragraphs, punctuation, or spaces on both sides. From this it can be concluded that a sequence of characters in a text with a grammatical error is also a word for the system. Once you have determined the number of words in a paragraph, you can move on to the next step;
5. At this stage, the words are rendered in a way that is convenient for further analysis of the system and passed to the next syllable block.

5. SEPARATION INTO SYLLABLES

The syllable block of words in the system performs the task of dividing the incoming words into syllables according to a number of rules and passing them to the next processing block. The process of breaking words into syllables can be divided into several stages:

1. The total length of a word is determined, that is, the number of letters (symbols) in a word. Ideally, words consist only of a sequence of letters. Special rules have been developed for other characters that can be found in the word structure, for example, the sign ('), the hyphen (-), the at sign (@), and in such cases work is done according to the rules;
2. Determining how many vowels a word contains. This is based on the rule that in Uzbek words, each syllable can contain only one vowel. That is, the number of vowels in a word is equal to the number of syllables in the word. Once the number of syllables is determined, the process of dividing the word into syllables begins;
3. This step is the most basic and complex stage of the system and is required to be carried out in the Uzbek language based on the rules of division and transfer of syllables. In the Uzbek Latin alphabet, one letter represents one sound. Here are the key points to keep in mind. There are two types of letters in Uzbek language: vowels and consonants. Syllables are formed by adding one or more consonant letters to a single vowel. Note that a single vowel can also represent a single syllable. However, a sequence of only one or more consonant letters is never considered a syllable.

As a result of the analysis of existing words and phrases in the current Uzbek dictionary, the possible syllable forms have been identified [8]. In pure Uzbek words, up to four sounds can be present in one syllable. Given that a syllable has a single vowel sound, there are 9 types of syllable structure models (Table 2). Conditionally, V is a vowel, C is a consonant.

Table 2: Types of syllables in Uzbek language

#	Syllable	Type	Example
1.	V	vowel	o-na, o-pa, a-ka;
2.	CV	consonant + vowel	bo-bo, to-za, ki-tob;
3.	CCV	doubleconsonant + vowel	bri-ga-da, tri-mó, stu-dent;
4.	VC	vowel + consonant	or-zu, os-mon, ep-chil;
5.	VCC	vowel + double consonant	umr, ost, ishq, aql-li;
6.	CVC	consonant + vowel + consonant	mak-tab, bog‘-bon, ki-tob;
7.	CVCC	consonant + vowel + double consonant	qarz, dars, sharq;
8.	CCVC	double consonant + vowel + consonant	stul, kran, trak-tor;
9.	CCVCC	double consonant + vowel + double consonant	sport, start, shtamp;

Some foreign words may have up to five sounds. Research based on the developed vocabulary shows that words and terms from other languages do not correspond to the above platforms. With this in mind, it is advisable to add additional types of syllables (Table 3).

Table 3: Syllable types

# of sounds Cases	1	2	3	4	5
Case 1	V	VC	VCC		
Case 2		CV	CVC	CVCC	CVCCC
Case 3			CCV	CCVC	CCVCC
Case 4					CCCVC

The table above shows that there are a total of 11 types of syllables in the current Uzbek language. Separation of syllable means the division of one or more of these types of syllables into parts according to a logical sequence. These types of syllables are used only to divide words that occur in the Uzbek language into syllables. However, in the words used today, which come from different foreign languages, there may be syllables that do not correspond to these types of syllables. Exceptions will be considered for such special cases.

The order of words into syllables is from the end of the word to the beginning. In particular, if we consider the order in which the word "MUKAMMALLIK" (perfection)is divided into syllables, it looks like this:

Syllables	**MU**	**KAM**	**MAL**	**LIK**
Types	CV	CVC	CVC	CVC

The division of words into syllables is determined by the norms of Uzbek literary language, spelling and grammar. The rules of syllable translation in the Uzbek language also play an important role in this regard [5]. At the end of the process of separating words into syllables, a hyphen is placed between each syllable.

6. BASES OF SYLLABLE FRAGMENTS

If we pay attention to the words and terms that exist in Uzbek language, a single syllable can be used in several words and, in combination with different types of syllables and can form words that have different meanings. For example, if we analyze the syllable "KAM", we can give the following examples given in Table 4.

Table 4: Forms of the syllable

№	In Uzbek	Meaning in English
1.	KAM	Little (Small)
2.	KAM - TAR - LIK	Humility
3.	KAM - YOB	Rare
4.	MU - KAM - MAL - LIK	Perfection
5.	HA - KAM - LAR	Judges
6.	MAH - KAM	Firmly (hard)

As can be seen from the Table 4, a single syllable is expressed in the same way in a text, regardless of how many syllables it contains. In Uzbek, some words may consist of a single syllable. Even such words can be used as a link to another word.

However, the role of the syllable in pronunciation is important. As mentioned above, emphasis is placed on the semantic ordering of speech. Given that an emphasis can fall on only one syllable of a word, the syllable in question may or may not be stressed in any word. The lexical stress is also linked to the meaning of the text and the punctuation. With this in mind, in the process of dividing into syllables, each syllable in the word structure is indexed according to its position:

MU	KAM	MAL	LIK
(1)	(2)	(2)	(3)

There are 3 types of indexes - 1, 2, 3 and they define the place of the syllable in the word. For example:

(1) MU – Syllable is at the beginning of the word;
(2) KAM – Syllable is at the middle of the word;
(2) MAL – Syllable is at the middle of the word;
(3) LIK – Syllable is at the end of the word.

Indexing of syllables determines the position of the same syllable in the word structure. However, the same posision may come with a different index in another word. One syllable can be at the beginning of a word, as amiddle syllable of another word, and can form the last syllable of a word. In Uzbek, prepositions, possessive suffixes, word-formative and modifier suffixes are used to form sentences. Therefore, in a single text or in a single paragraph, a single syllable can be repeated multiple times and with different indexes. Indexing, that is, the position of syllables in a word, determines the height of their pronunciation, the form of stress. Given the fact that the pronunciation of single-syllable words also has its own peculiarities, single-syllable words or syllables with independent meanings are also different. It can be said that the pronunciation form exists. In this case, the following 4 forms of the KAM joint selected above are formed:

KAM (0) – Syllable as an independent word;
KAM (1) – Syllable is at the beginning of the word;
KAM (2) – Syllable is at the middle of the word;
KAM (3) – Syllable is at the end of the word.

It follows that the shape of a single syllable with different indices is pronounced differently. It therefore allows an approximation to natural pronunciation in the synthesis process by implementing an indexing algorithm.

More than 31000 words in the Uzbek dictionary have been included to form a database of sounds in the system. At the same time, as a result of globalization, there are more than 7,000 popular international words also covered, such as place and region names, scientific names, names of countries, mountains, rivers, seas and lakes, capitals and major cities, etc. Comparing all the word bases, we can see that a single syllable can occur several dozen times in different words. There are even more than a thousand recurrences of such joints. That is, a single syllable can occur in more than a thousand words in a database. Here are some examples of such syllable:

Table 5: The number of repetitions of syllables in the word base

Syllable	Number
MA	1830
LI	2178
LIK	2424
LA	2453
MOQ	3678

No matter how many syllables a word contains, it is enough to pronounce it once. Each time clicking on the syllable, a single audio file is played. With this in mind, a database of more than 2,800 syllables has been formed, which can fully express the base of words formed. It should be noted that the pronunciation of the syllables varies depending on the place in the word. According to the above indexing rule in the database of syllables, there are 4 different forms of syllables respectively, and we can conditionally say that each syllable has the following shape:

KAM (0). wav
KAM (1). wav
KAM (2). wav
KAM (3). wav

Each form is radically different from the other phonetically. For this reason, it is advisable to synthesize using the indexing system in order to form the correct pronunciation of the syllables according to their place in the word. The corresponding sound fragments of more than 2,800 syllables were formed in wav format and grouped into a database of common syllable fragments.

7. ORGANIZING SYLLABLE-BASED READING ON THE FROUND OF SOUND FRAGMENTS

Given the fact that words in Uzbek are pronounced in syllables, it is advisable to carry out the process of synthesizing speech signals by organizing syllable-based reading. In this model, the text elements are broken down into syllables, and the corresponding sound fragments are retrieved from the pre-recorded database. Based on the sequence of incoming syllable, sound fragments are also sorted and transferred to the playback of audio files. In view of the above, the model of the system that converts words in Latin letters into sound signals can be conditionally given as given in Figure 2.

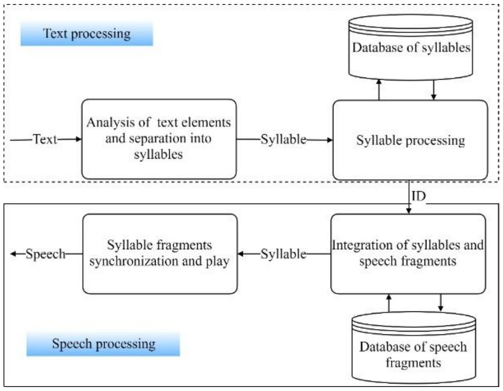
Text processing
Database of syllables
Text
Analysis of text elements and separation into syllables
Syllable
Syllable processing
ID
Speech
Syllable fragments synchronization and play
Syllable
Integration of syllables and speech fragments
Speech processing
Database of speech fragments

Figure 2: Schematic diagram of a synthesis system that converts text information into sound signals

The system generally consists of two main parts. The first part deals mainly with the processing of textual data, that is, the grammatical analysis of textual information entered into the system, the division of text into syntactic parts. The process includes following steps: dividing the words in the text into syllables and identifing them with the appropriate symbols (ID), determining their position in the text and word structure. This part of the system completes its task by sequentially passing the ID of the syllable to the next step. Regardless of the amount of text data, the system will be able to analyze and line up text lines in real time.

The second part of the system performs the task of searching the corresponding sound signals from the database of sounds on the basis of the ID identified in the first stage and determining the sequence number they hold during pronunciation. The most important part of the synthesis system is the sound base. The greater the number of nodes available in the database, the greater the system synthesis coverage.

The following example illustrates how the system works. Given that each syllable has 3 different forms according to the indexing rule, as well as 4 different forms of joints with independent meanings, we can express the process of synthesis of the word "MUKAMMALLIK" (perfection) will be as follows:

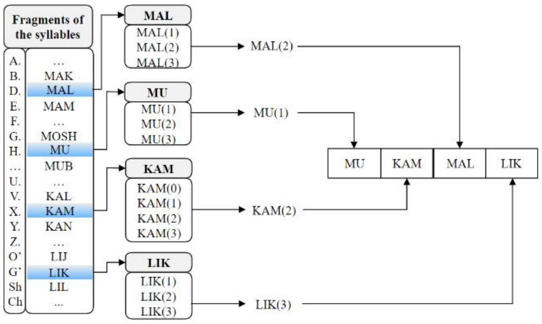

Figure 3: The order in which words are formed from syllable fragments

The process of synthesizing words by syllables allows to get closer to the real pronunciation of words by organizing them according to the above steps. In the process of synthesizing words by syllables, special rules have been developed for spaces and punctuation between words. Punctuation is a key element in sentence formation and word stress. Symbols, abbreviations, and some mathematical operations are required to be expressed in words. For such characters, a special section is allocated in the sound database, and the problem can be solved by preserving the sound form of such additional characters.

8. CONCLUSION

When words are pronounced in Uzbek, the word stress often falls on the last syllable. The articulation of words determines the intonation and content of speech. Based on the considered model, it is possible to preserve the intonational features of speech by synthesizing speech signals.

The phonology and accent rules of the Uzbek language discussed above are one of the main factors in the formation of artificial speech signals. It is impossible to generate quality speech signals in any language without following such rules. For this reason, the considered literary language norms are a key factor in the synthesis of speech signals based on syllables.

There are several types of speech signal synthesis systems and they use many algorithms and methods depending on the characteristics of the defined language. The database of syllables used in the Uzbek language, developed using the proposed model, and the database of sound fragments of syllables formed on this basis are the main components that determine the effectiveness of the method used.

It is possible to develop a relatively perfect speech synthesizer in the Uzbek language by organizing the reading of texts in the proposed way. Of course, this will be done using additional technologies to address the possible shortcomings. Typically, any speech synthesizer is improved by the wider application of established literary language norms, the emergence of new words and terms, the popularity of modern techniques and technologies, and other factors.

REFERENCES

1. TabetYoucef, BoughaziMohamed, AffifiSadek, “A Tutorial on Speech Synthesis Models” Procedia Computer Science,Volume 73, 2015, Pages 48-55

2. Younggun Lee, Taesu Kim, Soo-Young Lee "Voice Imitating Text-to-Speech Neural Networks" . arXiv:1806.00927, 2018.
3. S. Arik, C. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li et al., "Deep Voice: Real-time neural text-tospeech," in Proc. ICML, 2017, pp. 195–204.
4. Zen, H.; Senior, A.; Schuster, M.: Statistical parametric speech synthesis using deep neural networks, in Proc. ICASSP, 2013, 7962– 7966.
5. Richard Sproat, "Linguistic Processing for Speech Synthesis" Springer Handbook of Speech Processing (2008) pp 457-470.
6. D.Lutfullayeva, R.Davlatova "O'zbektiliningamaliygrammatikasi" Toshkent, Yangiasravlodi - 2010.
7. MentorHamiti, AgniDika, "Learning opportunities through generating speech from written texts" Procedia - Social and Behavioral SciencesVolume 2, Issue 2, 2010, Pages 4319-4324
8. E. Begmatov, A. Madvaliyev, N. Mahkamov, N. To'xtayev, E. Umarov, D. Xudoyberganova, A. Hojiyevlar, T. Mirzayev "O'zbektiliningizohlilug'ati", "O'zbekistonmilliyensiklopediyasi" Davlatilmiynashriyoti, 2006-2008.