The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. Section #2 "Intellectual Systems"
www.pci2008.science.az/2/28.pdf

ON AN APPROACH TO COMPUTER SYNTHESIS OF AZERBAIJAN SPEECH

Samir Rustamov[1], Aygun Saadova[2]

Institute of Cybernetic of ANAS, Baku, Azerbaijan
[1]*samir.rustamov@gmail.com,* [2]*aygun_saadova@mail.ru*

Introduction. Usage of computers is getting easier on the basis of machine-human relations in XXI century in which information society formed. Actual problems such as inputting information on the computer by means of speech and vice versa, speaking information on the computer which are the new methods of exchanging information between the user and computer.

The article is about an approach used to generate synthetic speech. The actuality of this problem is seen in its application. Text to speech technology is applied in many areas, for example, information bureaus, speaking programs of web pages (reading e-mail and fax), news agency, service enterprises.

Many scientific-research groups have been investigated the problem of text to speech for many years. For instance, MARY program (considered in German, English and Tibetan) introduced by DFKI, Free program based upon Flite derived from the Festival Speech Synthesis System (considered in English, Welsh and Spanish) from the University of Edinburgh and the FestVox project from Carnegie Mellon University, Epos program (considered in Czech and Slovak).

Problem statement. Text to speech-is converting the text and digital information to the synthetic speech whose naturalness is as close to real speech, corresponding to the pronunciation forms of special language. These systems performed this process are called text to speech systems. Input element of TTS system is text, output element is synthetic speech. Although there are different approaches in the text to speech, but still exist some problems. Such kinds of problems are connected with stress, intonation, specification of pronunciation and variety of languages. Sounding the question sentences isn't pronounced naturally even in the most developed modern programs. Words with different pronunciations aren't pronounced correctly. The Azerbaijani language has its specific features[1]. Some words aren't pronounced as its written form in Azeri [2], example, the Azeri word "ailə" like [ayilə], "müəllim" like [mə:lim]. As it is shown the sound "y" is added to the first word, the sounds "ü" and "l" aren't pronounced in the second word. Or the word "toqqa" is pronounced like [tokqa], here the first sound "q" is changed into "k".

Well known approaches. There have been three generations of speech synthesis systems. During the first generation (1962-1977) formant synthesis of phonemes was the dominant technology using rules which related the phonetic decomposition of the sentence to formant frequency contours. The intelligibility and naturalness were poor in such synthesis In the second period the diphones were represented with the LPC parametres. It was shown that good intelligibility synthetic speech could be reliably obtained from text input by concatenating the appropriate diphone units. The intelligibility improved over formant synthesis, but the naturalness of the synthetic speech remained low. The third generation of speech synthesis technology was the period from 1992 to the present, in which the method of "unit selection synthesis" was introduced and perfected, by Sagisaka at ATR Labs. in Kyoto. The resulting synthetic speech from this period had good intelligibility and naturalness that approached that of human-generated speech.

In the simplest approach to creating a speech utterance corresponding to a given text string the words can be stored as waveforms and concatenated in the correct sequence. This approach generally produces intelligible, but unnatural sounding speech, since it does not take into account the "co-articulation" effects of producing phonemes in continuous speech, the adjustment of phoneme durations or the imposition of pitch variation across the utterance.

In the word concatenation approach the vocabulary words are stored in a parametric form. A set of word concatenation rules is used to create the control signals for the synthesizer. Many words to store in a word catalog for word concatenation synthesis to be practical would have to be spoken and stored. In this method phonemes and diphones were suitable for synthesis units.

Efforts to overcome the limitations of word concatenation followed two paths. One approach was based on controlling the motions of a physical model of the speech articulators based on the sequence of phonemes from the text analysis. This requires control rules that are derived emprically. From the vocal tract shapes and sources control parametres (e.g., formants and pitch) can be derived by applying the acoustic theory of speech production used to control a synthsizer and used for speech coding. An alternative approach eschews the articulatory model and computes the control signals for a source/system model (e.g., formant parametres, LPC parametres. Pitch period, etc.). The rules for computing the control parametres are derived by emprical means.

The key idea of a concatenative TTS system, using unit selection methods, is to use synthesis segments. The word concatenation method is the simplest embodiment of this idea. The units for unit selection can be as large as words and as small as phoneme units.

Before any synthesis can be done, it is necessary to prepare an inventory of units. These units are coded for efficient storage. At the final synthesis stage, the units are decoded into waveforms for final merging, duration adjustment and pitch modification.

The approach proposed. The two fundamental processes are performed by all TTS systems: text anlysis and speech synthesis. The text analysis must determine from the input text following features:

1. *Pronouncation of the text string:* the text analysis process must decide on the set of phonemes, the degree of stress in speaking, the intonation of the speech, and the duration of each of the sounds in the utterance;
2. *Syntactic structure of the sentence to be spoken:* the text analysis process must determine where to place pauses, what rate of speaking is most appropriate for the text and how much emphasis should be given to individual words and phrases within the speech;
3. *Semantic focus and ambiguity resolution:* the text analysis process must resolve homographs and also must use rules to determine word etymology to decide on how best to pronounce foreign words and phrases.

The input data for the analysis is Azerbaijan text. The first stage of processing does text processing operations, including detecting the structure of the document containing the text , normalizinig the text and performing a linguistic analysis. The text processing benefits from an online dictionary of word pronunciations along with rules for getermioning word etymology. The output of the basic text processing step is tagged text, where the tags denote the linguistic properties of the words of the input text string.

The document structure detection module determines the location of all punctuation marks in the text, and to decide their significance with regard to the sentence of the input text. For example, an end of sentence marker is usually a period, ., a question mark, ?, or an exclaimation point, !. However this not always the case as in the sentence, "Her weight is 55.5 kilogram, height is 159.50 metre." Where there are two periods, neither of which denote the end of the sentence.

Text normalization methods handle the problems, including abbreviations and acronyms:

Example 1: "I live in Sh. I. Khatai street."

Example 2: "Azerbaijan Republic is the member of UNO."

In Example 1, the text "Sh. I" is pronounced as "Shah Ismail" , and in Example 2 the acronym UNO can be pronounced as either the word "uno" or "United Nations Organization", but it is never pronounced as the letter sequence "U N O".

Let's look at another example. The string $25.30 should be pronounced as "twenty five dollars and thirty cents" rather than as a sequence of characters.

One other text normalization problem concerns the pronunciation of proper names of foreign languages.

The third step in the basic text processing block is a linguistic analysis of the input text, with the goal of determining the following linguistic properties:

- The part of speech of the word
- The sence in which each word is used in the current context
- The location where a pause in speaking might be appropriate
- The word (or words) on which emphasis are to be placed, for prominence in the sentence
- The style of speaking, e.g., irate, emotional, relaxed, etc.

Ultimately, the tagged text obtained from the basic text processing block of a TTS system has to be converted to a sequence of tagged phones. The phonetic analysis block enables the TTS system to perform this conversion, with the help of a pronunciation dictionary. That is why the following operations are performed.

The homograph disambiguation operation must resolve the correct pronunciation of each word in the input string that ha smore than one pronunciation. In the Azerbaijan phrase "qırmızı alma" the word "alma" is a noun and the accent is on the second syllable, in the Azerbaijan phrase "bu kitabı alma" the word "alma" is a verb and the accent is on the first syllable.

The second step of phonetic analysis is the process of grapheme-to-phoneme conversion, namely conversion from the text to speech sounds. Although there are a variety of ways of performing this analysis, the most straightforward method is to rely on a standart pronunciation dictionary, along with a set of letter-to-sound rules for words outside the dictionary.

The fig.1 shows the schematic description of the algorithm for a simple dictionary search for pronunciation. Each individual word in the text string is searched separately. If the word exists, in its entirety, in the word dictionary, the conversion to sounds is straightforward and the dictionary search begins on the next word. If not the word is separate to the "root form" and affixes and the base search attempts to find both of them. If the "root form" or affixes are not present in the dictionary, a set of letter-to-sound rules is used to determine the best pronunciation of the root form or affixes of the word, again followed by reattachment of stripped out affixes.

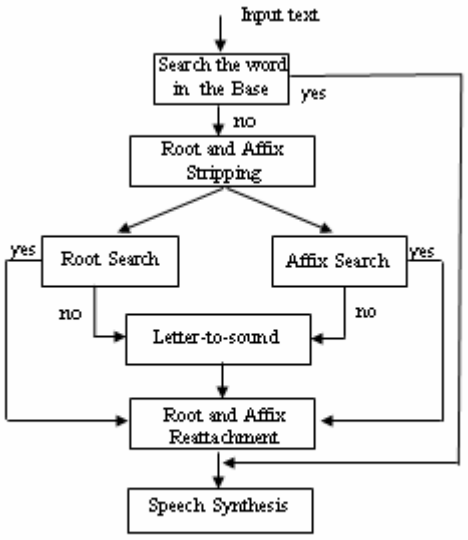

Fig.1. The scheme of the algorithm of speech synthesis.

On the paper the edit distance method is applied for finding the roots of the words. Let's explain main essence of this method briefly.

Our goal is to find the edit distance between two strings $x[1,...,m]$ and $y[1,...,n]$. Let's denote by $E(i,j)$ the edit distance between x and y words. For finding the edit distance between these words it is necessary to compare the sequence of symbols in each sets separately. For comparing $x[1,...,i]$ and $y[1,...,j]$, it can only be three cases:

$$1)\ \begin{matrix} x[i] \\ - \end{matrix};\qquad 2)\ \begin{matrix} - \\ y[j] \end{matrix};\qquad 3)\ \begin{matrix} x[i] \\ y[j] \end{matrix}.$$

The first case incurs a cost of 1 for this particular column, and it remains to align $x[1,...,i-1]$ with $y[1,...,j]$. But this is exactly the subproblem $E(i-1,j)$.

In the second case, also with cost 1, we still need to align $x[1,...,i]$ with $y[1,...,j-1]$. This is again another subproblem, $E(i,j-1)$. And in the final case, which either costs 1 (if $x[i] \neq y[j]$) or 0 (if $x[i] = y[j]$), what's left is the subproblem $E(i-1,j-1)$. In short, we have expressed $E(i,j)$ in terms of three smaller subproblems $E(i-1,j)$, $E(i,j-1)$, $E(i-1,j-1)$. We have no idea which of them is the right one, so we need to try them all and pick the best:

$$E(i,j) = \min\{1 + E(i-1,j), 1 + E(i,j-1), diff(i,j) + E(i-1,j-1)\} \quad (1)$$

where for convience $diff(i,j)$ is defined to be 0 if $x[i] = y[j]$ and 1 otherwise.

Consequently, $E(i,j)$ is computed for all $i \in [1,...,m]$, $j \in [1,...,n]$ and by means of its values the table of subproblems is constructed. Then optimal way is found based on the table of subproblem and (1) condition [3].

The last step in the text anlaysis is prosodic analysis where the sequence of speech sounds is mainly performed by the phonetic analysis. The assignment of duration and pitch contpours is done by a set of pitch and duration rules for assigning stress and determinig where appropriate pauses.

References

1. Mammadov N. The theoretical principles of Azerbaijan linguistics. Baki: Maarif, 1971, 366 p.
2. Akhundov A. The phonetics of Azerbaijan language. Baki: Maarif, 1984, 392 p.
3. S. Dasgupta, C. H. Papadimitriou, U. V. Vazirani. Algorithms 2006. 318 p. http://beust.com/algorithms.pdf
4. P. Taylor. Text to speech synthesis. svr-www.eng.cam.ac.uk/~pat40/ttsbook_draft_2.pdf