F0 GENERATION IN TTS SYSTEM FOR RUSSIAN LANGUAGE

O.F.Krivnova, A.V.Babkin
MSU, Philological Faculty,
okri@philol.msu.ru

ABSTRACT

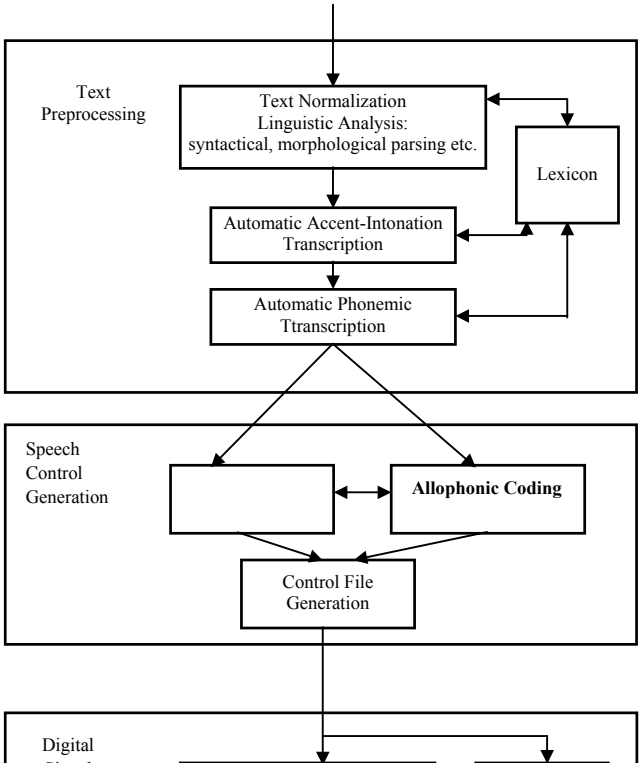
In this paper the strategy and ways of F0 contour generation in TTS system for Russian language are described. The system is developed in Lomonosov Moscow State University and based on two methods: concatenation of allophones' waveforms and prosodic rules to control pitch, duration and intensity. These rules form a part of speech control module which carries out the interface function, bridging the gap between the output of text linguistic processing and the input of speech signal generation module. As a result each segment (allophone) in a phrase being synthesized is attributed by at least two F0 values as its starting and ending points. Three and even more F0 values can be assigned to the phone if it is necessary. Signal generation is implemented according to the phrase control file, which describes the phrase as a sequence of allophones code names with assigned duration, energy and fundamental frequency values. To transform the base allophones to required prosodic values we use procedures that are close to TD PSOLA technology. All steps in development F0 modification algorithm based on TD-PSOLA technology are described and additional attention is paid to the ways of increasing naturalness of synthesized speech.

1. OVERALL ARCHITECTURE OF THE SYSTEM

The overall structure of our system is in line with the functional organization of a general TTS synthesizer. It consists of several blocks or modules, each of which has its own tasks and functions (Krivnova 1998). The structure of the system is shown on Fig.1.

2. GENERATION OF PITCH CONTOUR

The basic unit, for which the pitch contour is generated, is an intonational phrase (IP) - a coherent, grammatically organized fragment of a text to which one intonational model (abstract tune) is attributed. The type of intonational model for IP gets out as a result of the work of accent-intonation transcriptor and is fixed as an abstract prosodic marker.



1

Fig. 1. Overall structure of TTS system for Russian.

This device also determines the levels of words' prominence that is important to generate naturally sounding pitch contours. We assume that rhythm and accentuation is adjusted by two functionally different mechanisms: focus accentuation and rhythmization. The focus accents (to contrast or emphasize some words) are substantially defined by a speaker intention or by the whole information structure of a text. Frequently this structure has no evident cues to determine an accent place and its type. Therefore the formalization of focus accentuation represents the most difficult problem for TTS-systems. Our synthesizer is able to synthesize phrases with different focus accents but we have no rules to determine their localization automatically: it should be done manually. If a phrase has words with accent markers, the last of them is considered as the intonational center (nuclear) of a phrase. Otherwise the last content word of a phrase is as its intonational nuclear by default. It is the most typical situation for the narrative Russian texts, which construction is based on the use of neutral linear - accent structures with a final position of the intonational center.

As far as rhythmization is concerned, we distinguish three degrees of vowel prominence within a word (stressed, strong unstressed, weak unstressed) and four degrees for lexically stressed vowels (1 for full clitics, 2 for functional words, 3 for nonnuclear content words, 4 for nuclear content word). It should be noted that in Russian the prominence markers are very important not only for adequate pitch generation but also to determine correctly the duration of sounds.

In our system we use 7 abstract intonational models: 1 model of finality; 1 - non-finality; 3 - interrogative models (general, special, comparative questions); 1 – for exclamation (or command). For all models the possibility of a different position of the intonational center is taken into account. The formation of F0 contours for concrete phrases within the same intonational model is carried out in the separate submodule.

The strategy of pitch generation in each intonational submodule is as follows. The contour of the synthesized IP is formed as a result of concatenation of two types of tonal objects - tonal accents the main of which are nuclear and nonnuclear accents, and tonal plateaus. Each intonational model is considered as a cluster of these tonal events with the possibility of various phonetic realization determined by the rhythmical and sound structure of the IP.

Tonal accents are aligned with lexically stressed syllables if their prominence level is not less than 3 and if they are not considered atonic in the chosen intonational model. The main control parameters for pitch accents are the type of pitch movement (tonal figure), the realization time domain (part of a phrase to which the accent is phonetically anchored, stressed syllable including), the localization of pitch target points of the accent in a speaker pitch range and in realization time domain. We recognize that in Russian pitch movements forming the accent (and their targets) are very closely correlated with the boundaries of sound segments.

The tonal plateaus are aligned with unstressed and atonal stressed syllables in the beginning and end of IP and also in the intervals between pitch accent realization domains. The controllable parameters in this case are F0 values at the margins of intonational phrases and an interval of pitch change.

The temporal alignment and amplitude of tonal events are controlled by rules taking into account the intonation model itself, the rhythmical pattern of IP and its segmental make-up. To make it possible the preliminary coding of syllables in IP is carried out which fixes such features as accent status of a syllable, its prominence level according to the IP rhythmical structure, position in the IP and sound make-up. All pitch rules are hand-written and based on phonetic and acoustic analysis of read-aloud texts.

The calculation of F0 curves is implemented in two steps: at first in a semi-tone scale with respect to the average pitch (reference line) of a speaker, then these values are transformed into Hz. The calculated curve settles down in a working area of the speaker voice range, the boundaries of which are typical for realizations of the chosen intonational model.

3. PROSODY MODIFICATION ALGORITHM FOR RUSSIAN TTS

One of the most popular approaches in the creation of the high quality TTS system is the synthesis by concatenation. Formation of the synthesized speech signal is implemented in this case by means of concatenation of the acoustic waveform samples which are called elements of concatenation. The elements of concatenation are formed from the original samples of the speech signal, storing in the system acoustical database, by means of modification of their prosodic characteristics (such as duration, fundamental frequency and energy) in accordance with the requirements of the speech control file, generated for the IP being synthesized.

The theoretical base for the developing our methods of forming the required prosodic characteristics of the speech signal is TD-PSOLA technology (Babkin 1998). The main idea of TD-PSOLA consists in the following: the original database allophone is multiplied by a sequence of time windows synchronized with its pitch periods. The received sequence of acoustic segments, which are preliminary shifted about each other in time, is summed up, thus making the modified allophone with required sequence of pitch periods. To change the duration of the allophone the technology of repetition or elimination of some acoustic segments is used. In the traditional realization of this algorithm, in the case of noticeable increase of the duration of speech signal, and caused by this many-timed repetition of some identical segments, a particular unnaturalness is observed in perception of the resultant speech. To make the signal more natural in sounding we have built special algorithms based on random repetition and making some changes in the sequence of the identical acoustic segments. The described algorithms are realized in the module of signal processing (Fig.2)

In our Russian speech synthesis system the elements of concatenation, in the majority of cases, have the phonemic size and, thus, are allophonic realizations of the traditional phonemes. The structure of the module that is modifying the prosodic characteristics of the vocal allophones is given on Fig 2. (In this paper we do not discuss the prosody modification algorithms for unvocal allophones. In this case only duration and energy are needed to be changed because of this the modification methods are not so complicated as for vocal allophones.

One of the main requirements which essentially increase quality of the synthesized speech is the minimization of the distortions in acoustic characteristics of the transitional parts of the allophone. Within the framework of this requirement the modification of the fundamental frequency (via pitch periods) is realized along the whole length of the original allophone; the change of the duration of the allophone occurs only on its specially calculated part called stationary section. The calculation of the stationary part can be accomplished on the stage of speech database construction thus increasing the speed of synthesis process. But in our system it is performing in the signal processing module, because only at this stage of synthesis it is known to what degree original allophone has to be changed thus giving the possibility to estimate the length of the stationary part.

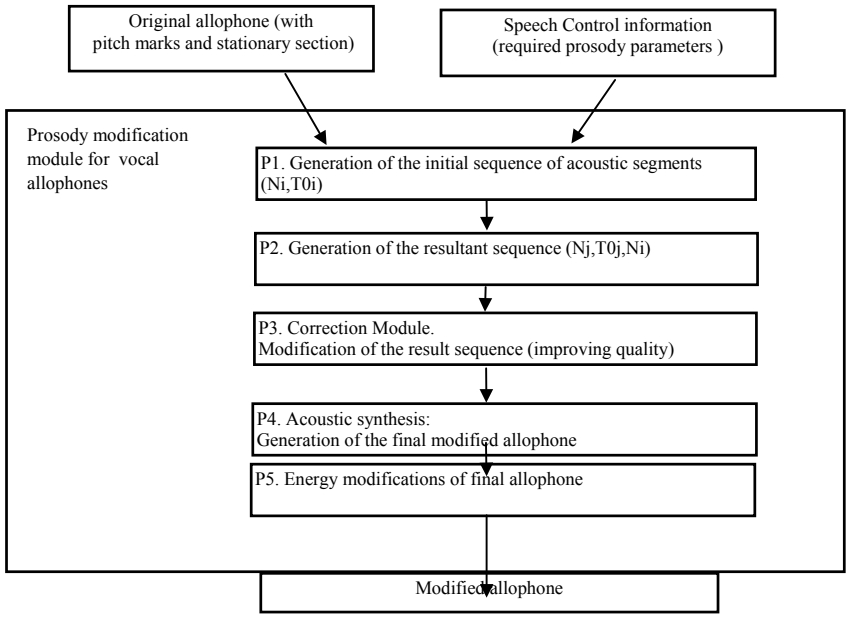


Fig. 2. The structure of the prosody modification module.

Now let us discuss all steps of generation of the modified allophone. The prosody modification module receives the original allophone with pitch period marks from the system database and creates the initial sequence of acoustic segments (step P1). Each segment has it own number and duration witch is defined in the speech database. It is calculated during the database creation. At the next step (P2) the requirements, that are specified in the speech control file, are analyzed and the resultant sequence of acoustic segments are generated. Each segment in this sequence has the reference to the initial element and the new duration of the segment is calculated. To avoid some speech unnaturalness the algorithm realized at this step makes some changes in the sequence of elements that has the reference to the same initial segment.

In the process of the F0 contour generation each acoustic element of the resultant period sequence receives duration that is calculated by linear way between the values in the 'start' and 'end' points of the pitch movement. It brings some shade of the unnaturalness because it does not reflect natural fluctuation of the fundamental frequency – and such a signal is perceived by a listener as a 'computer voice'. It occurs with the essential increase of the duration of the allophone as for example in the synthesis of the 'singing voice' – in which the fundamental frequency becomes fixed on the same value. In real speech F0 changes occasionally in certain limits around the given value.

In (Klatt and Klatt 1990) it is offered the simple formula which describes the occasional fluctuation of fundamental frequency in speech:

$$\Delta F_0 = \frac{F_0}{100}(\sin(12.7\pi t) + \sin(7.1\pi t) + \sin(4.7\pi t))/3 \qquad (1)$$

This additional fluctuation of F0 enhances the naturalness of the synthesized speech. In our TTS system this formula was converted to more complex variant with two parameters:

$$\Delta T_{0j} = A\frac{T_{0j}}{100}(\sin(12.7\pi K n_j) + \sin(7.1\pi K n_j) + \sin(4.7\pi K n_j))/3 \qquad (2)$$

where A = characterizes the degree of fluctuation of the period of the fundamental frequency and its range of values is between 0 and 100. K – the degree of casualty or quasi-periodicity. The fluctuation value (ΔT) is calculated for each element and is added to the value of pitch period (T) of this element. This is realized at a step P3. The choice of variant (2) of formula (1) is motivated first and foremost by the model which we use for prosody modification. The usage of parameters gives the possibility to enhance or to reduce the influence of this formula (and F0 fluctuations) on the synthesized speech. When A=0 the fluctuation is absent. According to the tests (Babkin,Zakharov 1999), the most 'natural' speech sounding is achieved when:

A=4 K=0.00005 (3)

These values are used as default in our TTS-system. In the course of further increase of the parameter A, for example when A=40, the effect of "sob" is observed – it could be explained by significant vibration of fundamental frequency.

At the next and almost final step (P4) the new modified allophone is generated using the information, which has been calculated at the previous steps. The final modified allophone is formed from the sequence of resultant acoustical segments by means of OLA (overlap and add) technology. In systems based on TD-PSOLA technology the type and size of window function has special significance. They are chosen to achieve the most exact spectral accordance between synthesized and real speech. Also great importance has timeline location of the window function against signal period. So we can talk about the problem of choosing the 'start point' of the period. There exists several variants of choice of these parameters and due to their small noticeable difference in perception of synthesized speech we have implemented several of these choices. They differ by window function and the localization of window within the signal period. We have conducted several tests and found that it is difficult to choose the best of them and in our system we decided to leave some and a user can switch between them.

The last step (P5) is the energy modification of the final allophone. After implementing any PSOLA algorithms the energy of the resultant acoustic signal is changed and we need to normalize it to some value. The normalization algorithm is done at this step. In our system we can choose the way of normalization. The resultant allophone can be normalized to the average energy or its energy can be increased or reduced to some value. In real speech the average energy of each period realizes not only the given energetic contour but is modified according to the casual law around the local average energetic value. We may assume that in order to improve the quality of synthesized speech it is needed to take into consideration this particular low or to talk about its mathematical realization. We haven't yet investigated this problem but it is known that any additional modification will cause certain tangible effect on the synthesized speech. For example if we take some kind of sinus periodical formula thus in some value of the period for this formula we receive the acoustic effect which is called the 'amplitude vibrato'. In the current version of synthesizer we have already reserved the place for this inquiry.

All the algorithms and methods mentioned in this paper have passed the special tests (Babkin, Zakharov1999) and are realized as a computer program, which makes part of the Russian text-to speech system being developed at MSU.

REFERENCES

Babkin A. V., Zakharov L.M., 1999: Testing of "Text-to-Speech" System Developed in MSU // *International Workshop "Speech and Computer" SPECOM99., Moscow, 1998.*

Babkin A. V., 1998: Automatic synthesis of speech — problems and methods of speech signal generation // *Proceedings of the International Workshop "Dialogue98" (Computational Linguistics and its Applications), Kazan', 1998.*

Klatt D.H., Klatt L.C., 1990 : "Analysis, synthesis and perception of voice quality (variations among female and male talkers) " // *Journal of the Acoustical Society of America. V.87, 1990.*

Krivnova O.F., 1998: TTS synthesis for Russian language (second version for female voice) // *Proceedings of the International Workshop "Dialogue98" (Computational Linguistics and its Applications), Kazan', 1998.*