INCREASEMENT OF NATURALNESS IN SYNTHETISIZED SPEECH

EINAR MEISTER
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

MART ROHTLA
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

MAIDU RAUDSEPP
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

ABSTRACT

An algorithm for synthesizing pitch contours is presented. It is shown that the relationship between the fundamental frequency and the frequency of the first formant of stressed vowel should be simulated. The problem of naturalness of synthesized speech is discussed.

INTRODUCTION

The text-to-speech systems become more and more used nowadays. The demands on the quality of the synthesized speech differ greatly. It has been found out that under certain conditions the synthesized speech has even advantages over natural speech because of its machine-like quality attracting the attention of the listener [1]. Although the systems vary in their capability and structure, for those who use them the most important thing is their intelligibility and the level of naturalness.

One of the main drawbacks of the synthesized speech is prosody, i.e. the changes of intonation, intensity and duration which are not enough controlled. According to many authors the investigation and the modelling of intonation is the key-problem in increasing the quality of the synthesized speech.

When investigating the changes in the pitch of natural speech, two characteristic components could be brought out: firstly, slow and large changes of the pitch (intonation contours) and secondly, fast and small changes of speech (the "fluctuation" of pitch). Both the components are intrinsic for the speech signal and their absence causes losses in the quality of the synthesized speech.

The movement of the pitch has been investigated by many authors already for a number of years. Various algorithms for the modelling of the intonation contours have been created [2,3,4] and the micromelody in different contexts has been studied [5,6].

When describing the intonation of Russian, the intonation contours of syntagma are widely accepted units [7]. Syntagma is the minimal prosodical unit which could still be divided into the following functional parts: (i) precentre, (ii) centre, (iii) postcentre. A special role in a syntagma is played by the so called intonation centre (by the intonation centre we mean the most important word of the syntagma) because the changes of the pitch in the centre are the most important feature in distinguishing different intonation types.

In speech the intonation is organically connected with other components of the signal. There are two types of components in a speech signal. On the one hand, there are components controlled by the articulatory program and on the other hand, components depending on the structure of articulatory organs. It may be assumed that the naturalness of human speech depends on interdependent parameters of the signal, i.e. a change in a certain parameter brings about a change in some other parameter. In other words, there exists a principle of "integral unity" between the parameters of the signal.

The aim of our research is to increase the naturalness of the synthesized speech through the modelling of the intonation contours; to point out the possible causes of machine-like sound of the synthesized speech and to introduce the connections between the parameters of a speech signal according to the principle of the "integral unity".

THE ALGORITHM FOR THE SYNTHESIS OF THE INTONATION CONTOURS

Under the synthesis of the intonation contours we mean the creating of the control parameters .for the pitch generator which are based on the duration of the synthesized message, punctuation marks, word stress and sentence stress.

Pitch generator

The control parameters for the pitch generator are the fundamental frequency and the time of transition from one frequency to another (duration). The typical range of the fundamental frequency for the male voice is 80-180 Hz. During the synthesis this range is divided into 8 levels; as for the time of transition, it is sufficient to have 4 meanings in the range of 50 up to 300 ms. It is also possible to choose the form of glottal pulse.

The models of intonation contours

There are various descriptions of intonation contours in Russian [8]. In choosing the models for the present paper we used the descriptions given in [7,9]. At the present moment the declarative, the interrogative, the nonterminal and the exclamatory models have been realized.

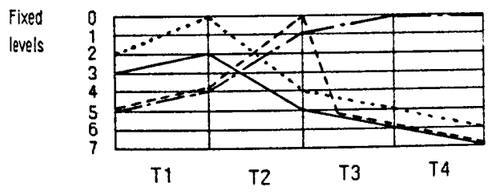

Fig. 1. The models of intonation contours.
—— a declaration,
--- an interrogation,
—·— a nonterminal,
····· an exclamation,
T1 - precentre,
T2 - centre,
T3 - postcentre,
T4 - the end of the syntagma.

Initial data

When developing the algorithm for the generation of intonation contours the following initial data was taken into consideration:
- the input text should be supplied with word-stress marks (') and with sentence-stress marks ('');
- in order to distinguish between different types of intonation the following punctuation marks are used at the end of a syntagma:
  [.] - a declaration,
  [,] - a nonterminal,
  [?] - an interrogation,
  [!] - an exclamation;
- the input data for the algorithm is a sequence of elements where each byte carries information about the duration of the element, about the stress and about the end of the syntagma;
- the possibility of changing the existing models and of adding other models of intonation contours corresponding to other punctuation marks;
- when describing the intonation contours no knowledge about programming is needed;
- the minimizing of the working time of the algorithm and the capacity of the memory;
- the realization of the algorithm using a 8-bit microprocessor I8080.

The description of the algorithm

The algorithm functions in 3 steps:
(i) the punctuation mark determines the type of the model of intonation contours of the syntagma and the durations T1, T2, T3, T4 are computed;
(ii) depending on the stress-mark the duration of the segments is determined during which the fundamental frequency changes with a constant rise;
(iii) the approximation of the fundamental frequency with linear cuts, i.e. the control parameters of the F0 generator are computed.

The algorithm needs about 1 KByte memory and works in real time.

Auditory estimation

In order to estimate the efectiveness of the algorithm, sentences consisting of one and two syntagmas with all types of intonation contours were synthesized. The type of intonation was in most cases distinguished correctly by the listeners and the intonation contours were said to correspond satisfactorily to those of human speech. At the same time the synthesized speech as a whole still had a machine-like sound.

Thus, in order to increase the naturalness of the synthesized speech it is not enough to control only one of the parameters.

THE PROBLEM OF NATURALNESS

The text-to-speech systems used by us /10/ and many other authors contain two main blocks - the model of vocal tract and the block of control on the basis of a microcomputer.

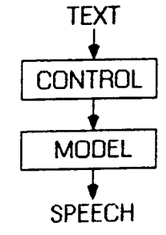

Fig. 2. The structure of the text-to-speech systems.

In case of such a structure of the system, the optimal correspondence between the control program and the technical aids has been found. The more exactly vocal tract is modelled, the harder it is to control it; on the other hand, a too simple model cannot provide the sufficient quality of speech.

We have used the classical model of formant synthesis:

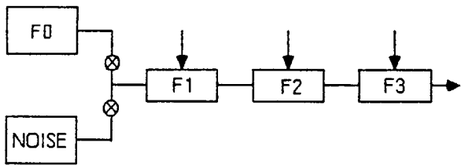

Fig. 3. The model of the vocal tract.

The control parameters of the model are established on the basis of a text which is provided with punctuation marks and stress-marks. This process consists of the following steps:

- transforming the orthographic text into a phonematic text;
- establishing the time structure;
- computing the intonation contour;
- taking into account the phenomena of coarticulation.

The model is characterized by the possibility of controlling all the parameters separately, i.e. the parameters of the model are not interdependent. Thus, incorrect control is possible.

The control errors can be eliminated in the process of determining the control parameters when the quantity of the knowledge and the rules of controlling the synthesis, both able to take the influence of the phonetic context into account, are big enough. But this brings about the enlarging of the memory of the system and the demands on the speed of operating the information.

Another way of eliminating the errors is to set uncontrollable by the program connections between the parameters of the vocal tract model. The existence of such connections is proved by many authors (more thoroughly are considered the relations between the intrinsic pitch of the vowels and the openness of the vocal tract; the changes of the pitch frequency in vowels of CVC context; the changes of the shape of the glottal pulse in the process of articulation). Such phenomena are rooted in the peculiarities of the vocal tract, and are not controlled by the program of articulation. People are used to such phenomena and regard them as compulsory. It may be assumed that not taking these connections into account is one of the reasons of the machine-like sound of the synthesized speech.

This can be proved by introducing into the vocal tract model the additional connections which imitate the above mentioned phenomena.

As the frequency of the first formant is the best determinant of the openness of the vocal tract, the connection between the frequency of the first formant and the pitch frequency is introduced. The experiments carried out by using the formant synthesizer showed that the machine-like sound diminishes if a dependence in accordance with

$$F0 = f0 * K/F1$$

is introduced, where F0 is the frequency of the pitch generator, F1 is the frequency of the first formant of a vowel, f0 is the computed frequency of the pitch and K is the coefficient of connection (K=500).

The dependence should be introduced for stressed vowels, for in all the other cases F0=f0.

In order to check the validity of this introduction, an experiment was carried out with a group of naitive listeners. Sentences consisting of one and two syntagmas and with different intonation types (10 sentences per each type) were synthesized using the formant synthesizer. The sentences were synthesized first without the connections between F0 and F1 (sentences A) and then with introducing the connection (sentences B). The listeners heard the sentences in pairs of optional order (AB or BA). The task of the listeners was to estimate which variant in a pair was more natural.

The results of the experiment prove firmly that the intorduction of the above mentioned connection is justified, for in all the cases the listeners chose variant B as more natural of the two.

The effectiveness of the introduction depends on the choice of the value of coefficient K. From the point of view of naturalness, the optimal value of K appeared to be K=500. When K<<500 the effect was not noticed, when K>>500 the structure of an intonation contour is violated and an effect of deep emotional excitement could be detected.

DISCUSSION

Although the obtained results demand to be studied thoroughly, already at the present stage the structure and the principles of encoding the speech signal can be detected. The fact that the introduction of the connections between the pitch and the first formant affect the quality and the naturalness of the synthesized speech was to be expected. There are a number of indirect data ponting to the fact that all living organisms and the signals they send out are subject to a general principle of "integral unity".

According to this principle information is encoded in a definte number of signs which are arranged in an order according to the importance and reliability and where the previous sign determines the region and corrects the strategy for searching for the following signs in the order of importance. All the signs can change within the corridor which width and the trajectories of the movement are determined by previous signs [11].

The introduced connections between the first formant and the pitch correspond to the principle of "integral unity" and can be explained by the fact that the frequency and the impulses of the pitch depend on the tension of the vocal cords, on subglottal pressure and pressure over the vocal cords. The tension of the vocal cords and subglottal pressure directly control the muscles of the larynx and the diaphragm. The pressure over the vocal cords coordinates the maxilla, the tongue and the lips. The less opened the vocal tract is, the higher is the pressure over the vocal cords and the fundamental frequency. In case of fluent speech the openness of the vocal tract changes quickly and the real fundamental frequency fluctuates around the intonation contours, depending on the position of the maxilla, tongue and lips.

CONCLUSION

Intonation is an essential factor when trying to increase the quality and the naturalness of the synthesized speech but in order to get rid of the machine-like sound it is not enough to model the intonation correctly. In a speech signal all the signs are distributed in the whole frequency-duration range and the machine-like effect is caused not only by the monotony of the changes of the pitch but are also characterized by the form of the glottal pulses, intnsity, duration structure etc. In order to completely free the synthesized speech from the machine-like sound it is necessary to control not only the changes of pitch around the intonation contour but also all the other signs. But for that it is inevitable to study the connections between the different parameters so that the optimal strategy of control can be worked out, i.e. which of the parameters are controlled by a program and to which the connections according to the principle of "integral unity" can be applied.

REFERENCES

[1] G.Kaplan, E.J.Lerner "Realism in synthetic speech" IEEE Spectrum, April, 1985, p.32-37.

[2] J.Pierrehumbert "Synthesizing intonation" J. Acoust. Soc. Am. 70(4), Oct. 1981, p.985-995.

[3] D.O'Shaughnessy "Linguistic Features in Fundamental Frequency Patterns" J. Phonet. 7, 1979, p.119-145.

[4] K.Hirose, H.Fujisaki "Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences" IEEE, 1982, p.950-953.

[5] C.H.Shadle "Intrinsic fundamental frequency of vowels in sentence context" J.Acoust. Soc. Am. 78(5) Nov. 1985, p.1562-1567.

[6] B.Lyberg "Some fundamental frequency perturbations in a sentence context" J. Phonet. 12, 1984, p.307-317.

[7] Е.А.Брызгунова Звуки и интонация русской речи" Москва, 1969.

[8] Н.Д. Светозарова "Интонационная система русского языка" Ленинград, 1984.

[9] N.D.Svetozarova "The Inner Structure of Intonation Contours in Russian", - Auditory analysis and perception of speech.- London, New York, San Francisco, Academic Press, 1975, p.499-510.

[10] О.Кюннап, А.Отт "Управляемый микропроцессором синтезатор речи" В кн.: Автоматическое распознавание слуховых образов - 12, Киев, 1982, стр. 410-411.

[11] M.Rohtla, K.Lindvere "Extraction of features from acoustic signals" EPP, Tallin, 1977, p.89-91.