SOLUTION OF PROBLEMS OF SINOOTH CONCATENATION OF SPEECH SEGMENTS IN FIELD OF SPEECH SYNTHESIS SYSTEMS REALIZATION

Rustam Musabaev

Kazakh National Technical University, Almaty, Kazakhstan
rmusab@gmail.com

There are a few common tasks of natural human speech synthesis by compilative principles: classification of sound fragments, determination of sound fragments and extraction from big audio-arrays. concatenation of small sound fragments in greater for creation of result fragment [1].

As is known, sound fluctuations are fluctuations of particles in elastic environments. Fluctuations extends in the form of longitudinal waves which frequency is in the limits of perceived by a human ear. The average values of frequencies are in a range from 16 up to 20 000 Hz [2]. In a general sense sound is a subjective perception which arises in a human ear under action of the given fluctuations.

Basically sound fluctuations are registered on analog-digital converters in the form of discrete samples. These samples are carried out at regular intervals and reflect a peak deviation of a signal from position of balance in the environment.

For example, if there is a set of fragments representing sound recordings of each separate phoneme for concrete language in the simplified kind the synthesizer of speech can be presented as consecutive connection by rules of the necessary microfragments to one single macrofragment which will be an end result of synthesis and it will be possible to submit it on an audio-output.

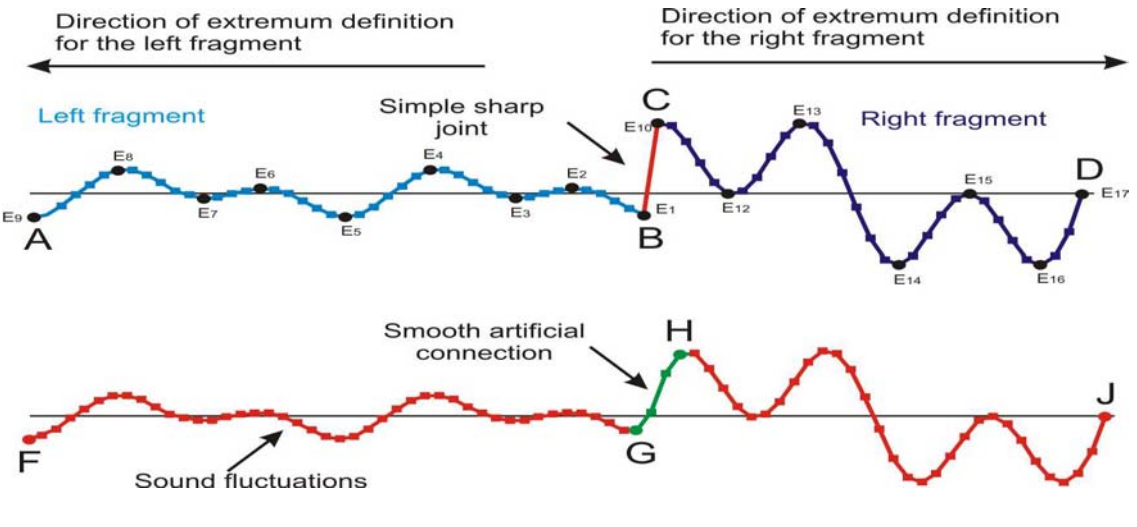
Direction of extremum definition for the left fragment
Direction of extremum definition for the right fragment
Left fragment
Simple sharp joint
C
E10
E13
Right fragment
E8
E4
E6
E2
D
E15
E17
E9
E7
E5
E3
E1
E12
A
B
E14
E16
Smooth artificial connection
H
J
F
G
Sound fluctuations

Fig.1. Examples of simple connection with a sharp defective joint and smooth artificial connection

At a stage of connection of various sound fragments there are certain difficulties. They are caused first of all by necessity of the matching of diverse amplitude and frequency components in areas of a joint. If we make connection of set of fragments without their preliminary matching then defects in the form of sharp short-term clicks on a place of each joint are clearly audible. Thus the human ear is very sensitive to similar defects.

If closely look narrowly at sound fluctuations which have a speech origin it is easy to notice their form in the shape of a sine and smooth change in time. If to consider all aforesaid then the problem of the smooth matching of two various speech signals can be reduced to a choice of a suitable fragment of a sinusoid by which they and will be connected.

Let's consider in detail given algorithm:

1) There are two various fragments of a sound wave. It is necessary to make their smooth connection in one solid fragment without any visible and heard defects at a level of their joining. We shall designate a fragment being at the left - left, and being on the right - right.

2) For each of two fragments it is defined their adjacent sides which will participate during joining:

a) For the left fragment - its extreme right side;
b) For the right fragment - its extreme left side.

3) For each of two fragments on their adjacent sides of joining it is defined the set quantity of extremums [3] with parameters accompanying them ($E_1 \dots E_2$, $E_{10} \dots E_{17}$). The quantity of extremums for both parties is defined in the form of a constant in view of that by this quantity the average statistics on duration of a half-cycle and amplitude of fluctuation will be calculated. A plenty of extrema can worsen speed of algorithm.

4) From the chosen two sets of extremums, moving aside their definitions from the parties of joining, are eliminated extremums inappropriate to conditions:

$$A_{extr} = A_{avg} \pm 25\%,$$

$$\frac{T_{extr}}{2} = \frac{T_{avg}}{2} \pm 25\%,$$

where A_{extr} is amplitude of a considered extremum, A_{avg} - average value of amplitudes on all extremums, $\frac{T_{extr}}{2}$ is half-cycle of a current extremum, $\frac{T_{avg}}{2}$ is average value of half-cycles on extremums. Elimination is realized until the first extremum corresponding given conditions will be found. That we get rid of pseudo-extremums resulted crossing a half-cycle of a sound wave.

5) Let's designate point E_l is an last extremum from the side of a joint of the left fragment corresponding average conditions of oscillatory statistics for the fragment. And point E_r we shall designate the same, but only for the right fragment.

For a case with an example in Fig. 1: $E_l = E_1, E_r = E_{10}$.

6) We delete all discrete samples in the left fragment from the side of a joint up to a point E_l. Also we delete all discrete samples in the right fragment from the side of a joint up to a point E_r.

7) We calculate coordinates of points E_l and E_r in a simple coordinate space XY. Let's designate coordinates of a point E_l: X_1 and Y_1. And coordinates of a point E_r: X_2 and Y_2.

8) Let's check up conformity of points E_l and E_r to following conditions:

a) E_l - maximum, E_r - minimum, $Y_1 \geq Y_2$;
b) E_l - minimum, E_r - maximum, $Y_1 \leq Y_2$;

If points mismatch the above-stated conditions then it is necessary to carry out search from both sides on all range of the certain extremums with method of search of all possible combinations of points corresponding with rules (a) and (b). From the found set of combinations which correspond to rules, one combination gets out only. The total distance of its points from the sides of joining should be the least.

9) If any combination is not found, then it is necessary to stop job of algorithm and to execute connection of two fragments without any matching. However, for example, it is possible to double the set quantity of extrema and to come back to item 3. This iteration can be repeated until the set quantity of extremums will not exceed total samples count in any of two

joined fragments. Whether it is necessary to make joining of fragments on distances greater, than their middle?

10) Repeatedly we carry out item 6.

11) The basic point gets out of points E_l and E_r by a principle of a greater half-cycle. If the half-cycle of point E_l is equal to a half-cycle of point E_r, then as the basic point point E_l gets out.

12) As a result we have two opposite in directions extremums E_l and E_r which are on the ends of the adjacent sides of joining.

13) We increase the right fragment from the adjacent side by length of a half-cycle of the basic point of an extremum. Again added area it is filled by zero samples.

14) Repeatedly we carry out item 7.

15) We should define function on which smooth connection of extrema E_l and E_r will be carried out. For this purpose we check position Y_1 concerning position Y_2:

$$(Y_1 \le Y_2) \Rightarrow f(x) = \cos(x),$$
$$(Y_1 > Y_2) \Rightarrow f(x) = \cos(x + \pi)$$

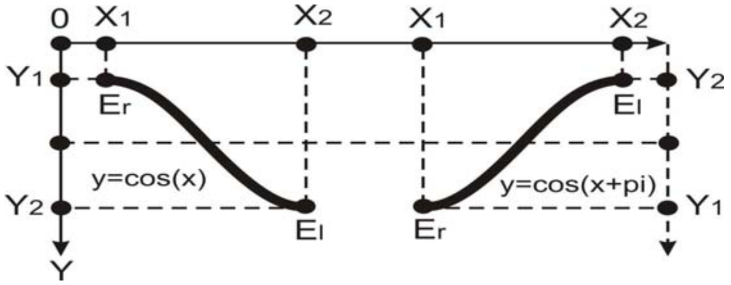

Fig.2. Kind of functions of smooth connection for two various positions of extremums rather each other.

In Fig.2 types of connection of extremums for two possible mutual positions are shown.

16) Now it is necessary to construct from E_l up to the graph certain in the previous item during a half-cycle of function set by the basic extremum. Any harmonious function of kind can be certain by the following equation:

$$y = a \cdot \cos(\frac{x}{b}) + c,$$

where **a, b** and **c** are some numbers. The problem is reduced to that, knowing coordinates of two extremums of this harmonious function (E_l and E_r), it is necessary to calculate values of parameters **a, b** and **c**.

17) It is possible to make and solve system of the equations from three unknowns **a, b** and **c**, using thus of four known parameters X_1, Y_1, X_2, Y_2:

$$\begin{cases} Y_1 = a \cdot \cos(\frac{X_1}{b}) + c \\ Y_2 = a \cdot \cos(\frac{X_2}{b}) + c \end{cases}$$

If to consider, that X_1, Y_1, X_2, Y_2 are coordinates of the nearest and opposite in a direction extremums, then:

$$a = \frac{Y_1 - Y_2}{2},$$
$$b = \frac{X_2 - X_1}{\pi},$$
$$c = X_1 - \frac{X_1 - X_2}{2}.$$

In this case it is visible, that change of parameter **a** leads to a stretching or narrowing of the graph of function on height. Change of parameter **b** leads to changes on width, and change of parameter **c** to displacement on height. Now coordinates (X;Y) any point of a connecting interval between extremum $E_l(X_1;Y_1)$ and extremum $E_r(X_2;Y_2)$ it is possible to calculate under the formula:

$$y = \frac{Y_1 - Y_2}{2} \cdot \cos(\frac{\pi x}{X_2 - X_1}) + X_1 - \frac{X_1 - X_2}{2}.$$

18) Under the deduced formula it is possible to calculate coordinates for each point on an interval from E_l up to E_r.

19) Now we have all necessary to execute smooth connection of two diverse sound fragments.

On the basis of the described algorithm its program realization in programming language Object Pascal is carried out. The given algorithm is used in development of universal speech synthesis system of English, Kazakh and Russian speech. During experiments the algorithm has shown the high efficiency, reliability and speed.

References

1. E.N. Amirgaliyev, R.R. Musabaev. Information technologies of artificial synthesis of speech // Bulletin of Kazntu, 2007, # 4 (20), pp. 26-34 (In Russian)
2. M. Libenzev A rate of physics. -M.: "Higher school", 1968 (In Russian)
3. N.S.Piskunov. Differential and integrated calculations. -M.: "Science", 1976 (In Russian)