METHOD OF PROSODIC MODIFICATION IN THE TEXT-TO-SPEECH SYNTHESIS SYSTEMS

Maksat Kalimoldaev[1], Yedilkhan Amirgaliyev[2], Rustam Mussabayev[3]

Institute of Informatics and Control Problems, Almaty, Kazakhstan
[1]*mnk@ipic.kz*, [2]*amir_ed@mail.ru*, [3]*rmusab@gmail.com*

1. Introduction

One problem of segment concatenation is that it doesn't generalize well to contexts not included in the training process, partly because prosodic variability is very large. There are techniques that allow us to modify prosody of a unit to match the target prosody. These prosody-modification techniques degrade the quality of the synthetic speech, though the benefits are often greater than the distortion introduced by using them because of the added flexibility. The objective of prosodic modification is to change the amplitude, duration, and pitch of a speech segment. Amplitude modification can be easily accomplished by direct multiplication, but duration and pitch changes are not so straightforward.

Overlap-and-add (OLA) [1] and Synchronous Overlap and Add (SOLA) [2], two algorithms to change the duration of a speech segment. Pitch Synchronous Overlap and Add (PSOLA) [3,4,5], a variant of the above that allows for pitch modification as well. The PSOLA approach is very effective in changing the pitch and duration of a speech segment if the epochs are determined accurately. Even assuming there are no pitch tracking errors, there can be problems when concatenating different segments:

- Phase mismatches. Even if the pitch period is accurately estimated, mismatches in the positioning of the epochs in the analysis signal can cause glitches in the output.
- Pitch mismatches. These occur even if there are no pitch or phase errors during the analysis phase. If two speech segments have the same spectral envelope but different pitch, the estimated spectral envelopes are not the same, and, thus, a discontinuity occurs.
- Amplitude mismatch. A mismatch in amplitude across different units can be corrected with an appropriate amplification, but it is not straightforward to compute such a factor. more importantly, the timbre of the sound will likely change with different levels of loudness.

All listed mismatches are solved in the proposed approach.

2. The offered approach

In the given article offered the approach on realization of a prime tone frequency contour dynamic modification (F0-contour). At the same time each phoneme of the speech signal can have the set of different intonational variants. It is offered to represent the given intonational variants in the form of F0-contours set. The choice of the most suitable contour is realized with use of a competence matrix of phonemes. For each phoneme depending on its classification and phonetic environment the most suitable intonational contour gets out by the gathered points.

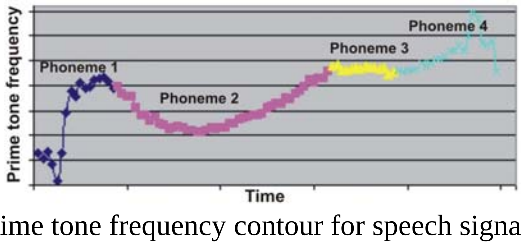


Figure 1. Example of prime tone frequency contour for speech signal with duration in a word

In Figure 1 the example of prime tone frequency contour for speech signal with duration in words is shown. It is visible, that the contour has considerably changeable, but smooth structure. Change of the given contour together with change of phonemes duration allows giving to a synthesized signal the necessary intonational and emotional structure [6, 7]. In the given article the method of F0-contour modification is offered. The primary application field of the given method is concatenative speech synthesis systems in which synthesis of the given signal is realized by methods of modification and series connection of set of initial speech signals fragments.

For qualitative modification of source speech signal by duration and intonational characteristics it is necessary at first to do segmentation by phonemes, and then by microsegment components of phonemes. Step by step applying for everyone microsegment the method of crossover mixing for its two copies with the given displacement and counter downturn of their amplitude levels it is possible to achieve the necessary modification of a prime tone frequency contour. For this purpose some function similar to function presented on Fig.1 should be set. By the given function for each microsegment the demanded length is defined, then calculated the difference between actual and demanded length, and modification of actual duration with the purpose of its reduction to demand duration is made.

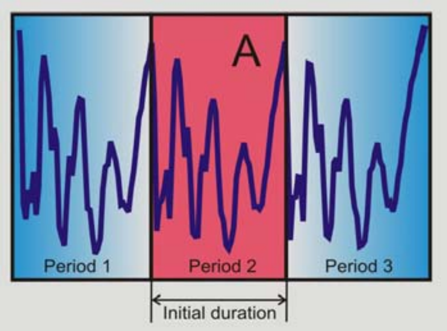


Figure 2. Initial speech signal

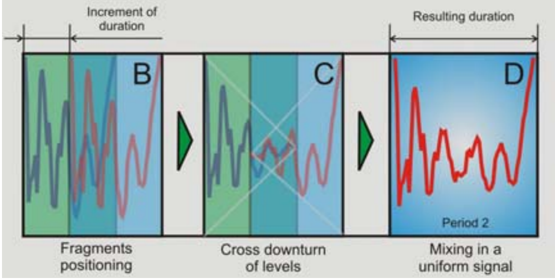


Figure 3. Illustration of crossover mixing method for two copies of signal with the given displacement and counter downturn of their amplitude levels

In Figures 2 and 3 illustrated the crossover mixing method for two copies of signal with the given displacement and counter downturn of their amplitude levels.

3. The description of algorithm

Let's consider step by step a set of operations that makes the given method:

1. Segmentation of speech signal on microsegments by change dynamics of prime tone periods: $S = (s_1; s_2; ...; s_k) = \{P; N\}$, where k - number of microsegments.
2. Then microsegments are classified and distributed between two subsets P and N. P and N - according periodic and non periodic microsegments.
3. On the input of algorithm we move one by one microsegments $p \in P$ for modification.
4. Two time copies of a modified microsegment are created p_1 and p_2.
5. The sound signal of each microsegment is represented as the ordered set of discrete samples $V = (v_1; v_2; ...; v_n)$, where n - number of discrete samples in the signal.

6. In the set $V \in p_1$ from end point executing the removal of samples subset, which number is set as follows: $q = L_1 - L_2$, where L_1 - lenght in samples of source signal of microsegment, and L_2 - is the length to which it is necessary to result L_1. Removal is made only on the stipulation that $q > 0$.
7. In the set $V \in p_2$ from the start point it is necessary to make removal subset of samples which number also is equal q. Removal is made under condition of $q > 0$.
8. In the set $V \in p_1$ operation on giving to a signal the form of smooth linear reduction of an amplitude level up to zero is made. At the given operation the modified discrete sample is set by following expression: $y_i = \left(1 - \frac{i}{n}\right) \cdot v_i$, where $i \in (0 \ldots n-1)$, $y_i \in Y$.
9. Then in the set $V \in p_2$ operation on giving to a signal the form of smooth linear increase an amplitude level from zero up to initial level is made. In the given operation the modified discrete sample is set by following expression: $u_i = \frac{s_i \cdot i}{n}$, where $i \in (0 \ldots n-1)$, $u_i \in U$.

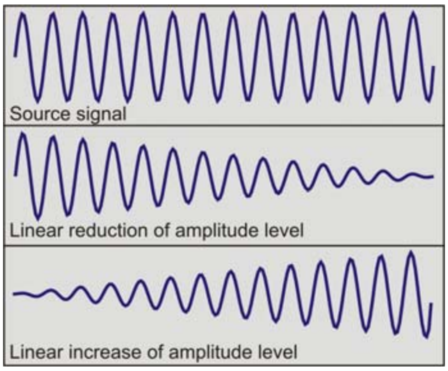


Figure 4. Illustration of operations for linear change of amplitude levels

Graphically results of the given operations are illustrated in Figure 4.

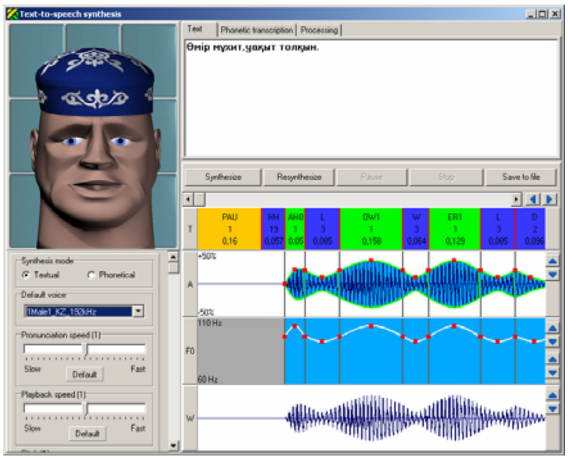
10. It is necessary to define an increment of length of an individual segment during its modification: $\Delta L = L_2 - L_1$, where L_1 and L_2 accordingly lengths of the initial and modified microsegments. If the frequency of the prime tone falling an individual microsegment inversely proportional to its length: $F0 = \frac{1}{L}$, where L - length of the microsegment, hence the increment of frequency of the prime tone during its modification has following dependence: $\Delta F0 = -\left(\frac{1}{L} - \frac{1}{L + \Delta L}\right)$. Also there is useful to use an average parameter of the prime tone frequency: $avg(F0) = \frac{1}{m}\sum_{i=1}^{i=m} F0_i$, where $F0_i$ - is a prime tone frequency for the $p_i \in P$.
11. Ordered sets Y - from a final position, and U - from an initial are supplemented with a subset of zero samples $Z = (z_1 = 0; z_2 = 0; \ldots; z_h = 0)$, where $h = \begin{cases} \Delta L, & if \quad \Delta L > 0 \\ 0, & in\ ather\ case \end{cases}$.

   Each element of the resulting set R is defined as follows: $r_i = u_i + y_i$, where $i \in (1 \ldots n + h)$, $r_i \in R$. Using the given method it is necessary to consider, that change of a F0-contour leads to change of duration of a modified signal. So the increment of a signal duration as a whole at modification a F0-contour is the sum of lengths increments of all microsegments making it: $\Delta L_{\text{сигн}} = \sum_{i=1}^{i=k} \Delta L_i$,

where ΔL_i - increment of microsegment length $s_i \in S$.

4. Conclusion

The considered algorithm also has its program realization in structure of Text-To-Speech synthesis system.



Text-to-speech synthesis

Text | Phonetic transcription | Processing

Өмір мухит.уақыт толқын.

Synthesis mode: ● Textual ○ Phonetical

Default voice: 1Male1_KZ_192kHz

Pronunciation speed (1)

Playback speed (1)

Figure 5. Graphical user interface of the TTS synthesis system

On the basis of the given realization and other methods [8,9] it is possible to make practical experiments of speech signal synthesis with the set of different F0-contours (Figure 5). There are two ways to make this kind of experiments:

1. Using of free defined F0-contours.
2. Using the contours extracted from the natural human speech.

The experiments in practice have shown full suitability of the given algorithm for application in the systems of speech synthesis.

References

[1] Crochiere, R., "A Weighted Overlap-Add Method of Short Time Fourier Analysis/Synthesis," IEEE Trans. on Acoustics, Speech and Signal Processing, 1980, 28(2), pp. 99-102.

[2] Roucos, S. and A. Wilgus, "High Quality Time-Scale Modification of Speech," Int. Conf. on Acoustics, Speech and Signal Processing, 1985, pp. 493-496.

[3] Moulines, E. and F. Charpentier, "Pich-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," Speech Communication, 1990, 9(5), pp. 453-467.

[4] Moulines, E. and W. Verhelst, "Prosodic Modification of Speech" in Speech Coding and Synthesis, W.B.Kleijn and K.K.Paliwal, eds. 1995, pp. 519-555, Elsevier.

[5] Dutoit, T. An Introduction to Text-to-Speech Synthesis, 1997, Kluwer Academic Publishers.

[6] P. Taylor, *Text to Speech Synthesis*, University of Cambridge, 2007.

[7] X. Huang, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, 2001.

[8] Y. Amirgaliyev, R. Musabaev, "Algorithms of phonemes classification in field of compilative speech synthesis systems realization", *PCI'2008 - "Problems of cybernetics and informatics"*, Baku, Azerbaijan, Volume I, pp. 108-111.

[9] R. Musabaev, "Solution of problems of smooth concatenation of speech segments in field of speech synthesis systems realization", *PCI'2008 - "Problems of cybernetics and informatics"*, Baku, Azerbaijan, Volume II, pp. 279-282.