Text-to-speech system with acoustic processor based on the instantaneous harmonic analysis

Elias Azarov[1], Alexander Petrovsky[1], Boris Lobanov[2] and Lilia Tsirulnik[2]

1. Department of Computer Engineering,
Belarusian State University of Informatics and Radioelectronics
P.Brovky Str. 6, 220027, Minsk, Belarus
palex@bsuir.by
2. United Institute of Informatics Problems Nat. Ac. Of Sc. Of Belarus
Surganov Str. 6, 220012 Minsk, Belarus
lobanov@newman.bas-net.by

Abstract

This paper introduces a text-to-speech system with acoustic processor based on the instantaneous harmonic analysis. Speech signal is synthesized through concatenation in frequency domain. This approach is supposed to be a good alternative to the time domain concatenation methods due to its ability to smooth spectral amplitude and phase mismatches that occur at concatenation points. High quality of the synthesized signal is provided by improved harmonic analysis technique. The paper describes the general approach, essentials of the harmonic+noise model, analysis and synthesis technique along with some analysis/synthesis examples and experiments. In the experimental part of the paper the proposed system is compared with a similar text-to-speech system based on the time domain concatenation.

1. Introduction

Text-to-speech (TTS) system design is a complicated problem that is linguistic , phonetic and signal processing related at the same time. Today it is extremely important to develop new methods for speech signal synthesis that can provide both superior synthesis quality and small acoustical database. Despite the fact that many different techniques have been proposed the time domain concatenation method [1] is still the major approach for speech synthesis. A source speech recording is usually segmented into phonetic units that are assembled into speech signal during synthesis. In the present work the TTS-synthesizer structure described in [2] is used. The structure is presented in Fig.1. First orthographic text is processed through a number of successive operations carried out with the help of specialized processors. The textual processor is devised to transform the incoming orthographic text into a prosodically marked one. The processor performs the following operations:

- dividing an orthographic text into utterances;
- transforming numbers, abbreviations, etc;
- dividing an utterance into phrases;
- placing word's accents (weak and strong);
- dividing phrases into accentual units (AU);
- marking the intonation type of the phrases.

The prosodically marked text is then sending to the phonemic processor, that performs the following tasks:

- phonemic transcription of the text;
- transforming the phonemic text into allophonic;
- combining the allophones into allosyllables.

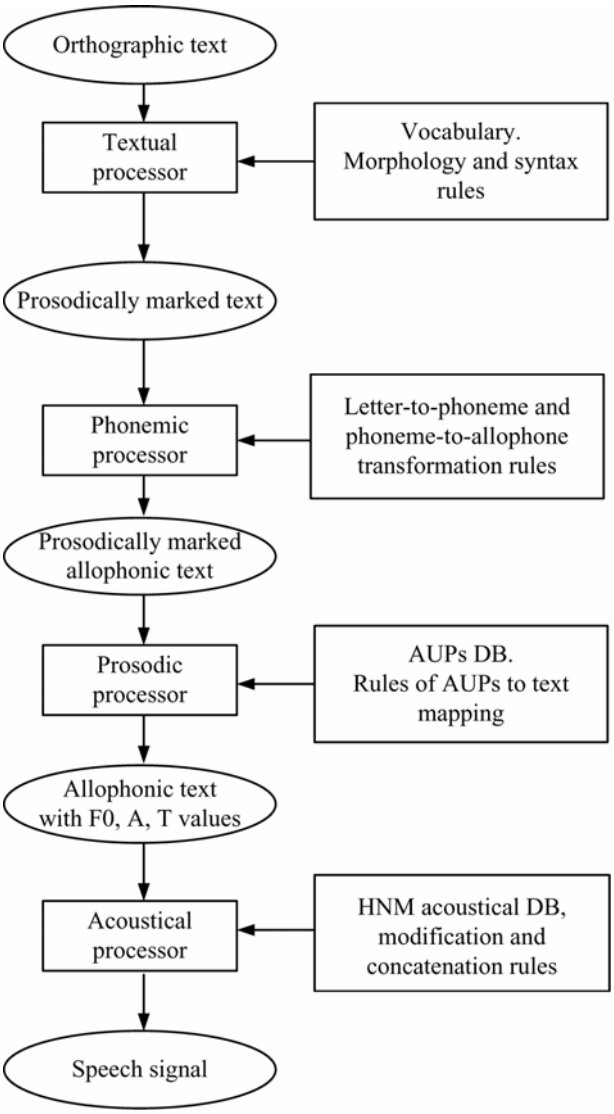


Figure 1 – General structure of the TTS-synthesizer

The prosodic processor performs the following tasks:

- splitting AU into the elements of accentual units (EAU): pre-nuclear, nuclear and post-nuclear parts;

- generating the fundamental frequency (f_0) contour as well as the amplitude (A) and phoneme duration (T) assigning values according to the accent unit portraits (AUPs) for each accent unit.

The acoustical processor uses the information that comes from the phonemic and prosodic processors to concatenate speech segments into the appropriate sequence. Segment concatenation technique is a crucial point in any TTS. The quality of segment processing to a large degree determines overall quality of the synthesized speech. The main functionality that should be implemented is listed below:

- inaudible concatenation;
- pitch shifting / matching;
- time-scale modifications;
- phase matching;

The first step is to choose an appropriate speech model for processing. In general, existent approaches can be divided into two groups: time domain concatenation and frequency domain (spectral) concatenation.

As was said before, majority of TTS systems use time domain concatenation. It means that acoustical processor modifies waveform of a signal in order to join segments in concatenation points.

Though time domain approach is usually the preferred one, it has some significant disadvantages. It is hard to change pitch or phase of a sound in time domain without audible degradation of it. Thus, it is difficult to provide phase and pitch matching between adjacent segments. Imperfect joining results in “click” effects and adds artificial inflection that hardly could be avoided.

Frequency domain approach deals with spectral representation of a signal rather than with its waveform. A time-frequency transformation is firstly carried out in order to estimate this representation. At synthesis stage the signal is generated as a sum of periodic functions which parameters are defined by target speech parameters (pitch contour, length, energy etc.). Concatenating in frequency domain is a perfect way to ensure pitch and phase matching; in addition it allows making original-like time-scale modifications.

2. Spectral acoustic processor

One of the most efficient spectral representations of speech is the harmonic model that describes speech in frequency domain as a sum of both periodic and noise parts [3]. The Harmonic+Noise (H+N) model is very flexible and allows achieving a perfect quality of the synthesis, since the model separates sounds of different nature (harmonic/noise) that can be further differently processed and synthesized. However, the major point here is adequate and accurate harmonic/noise parameters estimation, as far as synthesized speech quality entirely depends on the preceding harmonic analysis.

In this work new acoustical analyzer and processor are proposed for concatenation in frequency domain. The analysis/synthesis methods are based on (H+N) model.

The H+N model represents a speech signal as a sum of periodic and noise components [3]:

$$s(n) = \sum_{k=1}^{K} A_k(n)\cos\varphi_k(n) + h(n) \qquad (1)$$

where $A_k(n)$ - the instantaneous magnitude of the k-th harmonic component, K is the number of the harmonic components, $h(n)$ is the noise component and $\varphi_k(n)$ is the instantaneous phase of the k-th harmonic component. $\varphi_k(n)$ can be derived from the initial phase $\varphi_k(0)$ and instantaneous frequency f_k:

$$\varphi_k(n) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \varphi_k(0),$$

where F_s is the sampling frequency. The harmonic representation assumes that the next expression is true:

$$f_k(n) = kf_0(n),$$

where $f_0(n)$ is the fundamental frequency.

It is essential for this model that speech is considered to be voiced or unvoiced. The classification can be made using harmonic/noise rate (HNR):

$$HNR = 10\lg E_h \big/ E_r$$

where E_h and E_r are the energies of the harmonic and noise components respectively. Voiced frames have high HNR values, while unvoiced have them low.

The H+N representation implies harmonic parameters estimation. There are many techniques that can provide accurate results [4-6]. As soon as the parameters (amplitude, frequency and phase) are estimated the harmonic part can be synthesized and subtracted from the source signal providing the noise part.

In this work the instantaneous harmonic parameters estimation technique is used as described in [6]. The system of filters is applied to the signal, providing instantaneous harmonic parameters ($MAG(n)$ - amplitude, $f(n)$ - frequency, $\varphi(n)$ - phase):

$$MAG(n) = \sqrt{A^2(n) + B^2(n)},$$

$$f(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi} F_s + f_0(n)\cdot k, \qquad (2)$$

$$\varphi(n) = 2\pi f_0(n)kn + \alpha(n)$$

where

$$A(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi} \sin\left(\frac{\pi}{F_s} F_\Delta(n-i)\right)\cos\left(\frac{2\pi}{F_s}\varphi_k(n)\right),$$

$$B(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi} \sin\left(\frac{\pi}{F_s} F_\Delta(n-i)\right)\sin\left(\frac{2\pi}{F_s}\varphi_k(n)\right),$$

$$\alpha(n) = \arctan\left(-\frac{B(n)}{A(n)}\right),$$

$$\varphi_k(n) = (\sum_{i=0}^{n} F_0(n) - \sum_{i=0}^{N/2} F_0(n))k \cdot$$

The harmonic parameters can be calculated for every instant of time. This technique gives a good time/frequency resolution for voiced speech. In the TTS system every segment of speech corresponds to a definite phonetic sound and should be analyzed separately. The harmonic parameters estimation for the TTS is presented in Fig. 2.

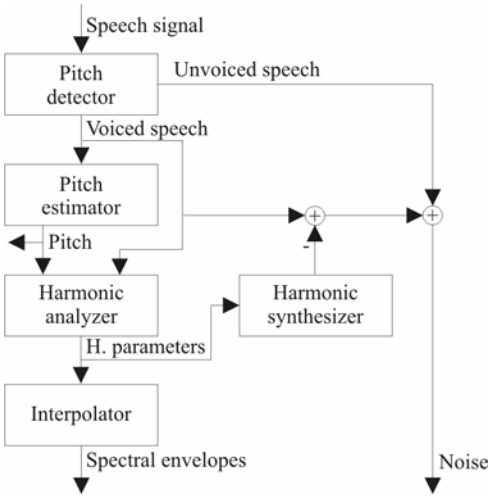


Figure 2 – Harmonic parameters estimation and harmonic/noise separation

First source speech signal is processed by pitch detector that makes voiced/unvoiced classification. Then for voiced frames the pitch contour is estimated. The pitch contour estimation technique is described in [6]. The harmonic parameters are estimated in harmonic analyzer using expression (2). Though pitch parameter $f_0(n)$ explicitly exists in the H+N model it cannot be changed directly at synthesis stage. Harmonic amplitudes should be interpolated at their new frequencies in order to perceive the original spectral envelope. Thus harmonic parameters are recalculated by the interpolator in a predefined set of frequency points to form an envelope that does not depend on pitch value. A simple piecewise-linear interpolation of the original amplitudes is used. Experiments showed that this interpolation gives a good approximation quality and does not add any perceptible envelope distortions.

The noise part of the signal consists of unvoiced frames and parts of the signal that are not described by harmonic parameters. The noise stored in the database and is not modified by pitch shifting during synthesis procedure.

In the present TTS system harmonic analysis is used at the preparation stage in acoustic database forming procedure. When the source training set is prepared and segmented every speech unit (allophone) is processed through the analysis scheme in order to obtain its spectral envelope and noise part. Since allophones can be different in length the impulse response of the analysis filter varies accordingly. Estimation becomes inaccurate close to the edge of an allophone and the edge values are interpolated from those closer to center in order to get maximum analysis accuracy. The analysis results are stored in acoustic database for future synthesis.

A result of allophone harmonic analysis is presented in Fig.3. Here the sound /a/ of a male speaker is analyzed. Sampling frequency is 8000 Hz, length of the signal – 250 ms. As can be seen the signal is non stationary within the analysis frame and every harmonic has frequency and amplitude modulations. However, estimated frequency contours are very close to those in the source signal and the estimated harmonic part of the signal (Fig.3.C) has the same structure in frequency domain. The analyzed sound has a low fundamental frequency (near 100 Hz) that is the worst case for such analysis. The fact that harmonics are located close to each other does not allow using analysis filters with wide bandwidth and requires accurate pitch contour estimation. As a result of this female speech is simpler to analyze because of higher pitch values and less number of harmonics within frequency band.

The speech synthesis procedure is shown in Fig.4. Concatenation rules contain target pitch contour, time-scale modification parameters, energy envelope. Spectral envelopes are extracted from the acoustical database, and recalculated in frequency points defined by pitch values. Derived harmonic parameters are used in the sinusoidal synthesizer along with concatenation rules to form the waveform of the periodical signal.

The sum of the periodic signal and the noise results in the synthesized speech. Phase values are interpolated in concatenation points insuring inaudible segments joining.

Such speech synthesis approach provides control over prosodic speech characteristics (like speech speed and intonation). Therefore speech prosody modifications are easily implemented using proposed model.

For noise part of the signal additional processing is needed in order to implement time scale modifications. The linear prediction coding (LPC) methods can be used for this purpose [7].

Using LPC technique the noise can be represented as a white noise generator and a spectral envelope defined by prediction coefficients.

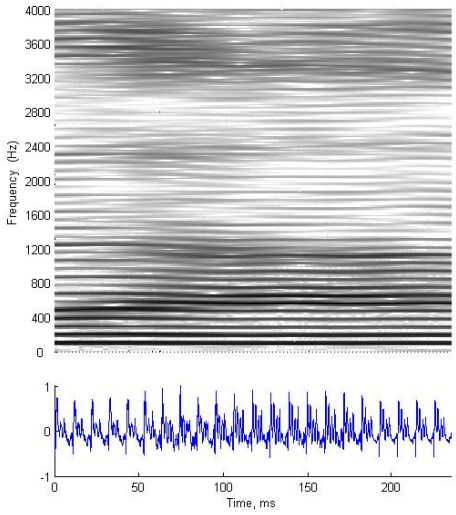
Combining harmonic and noise processing, prosodic speaker characteristics can be easily modeled providing high quality of the synthesized signal.

3. Experimental results

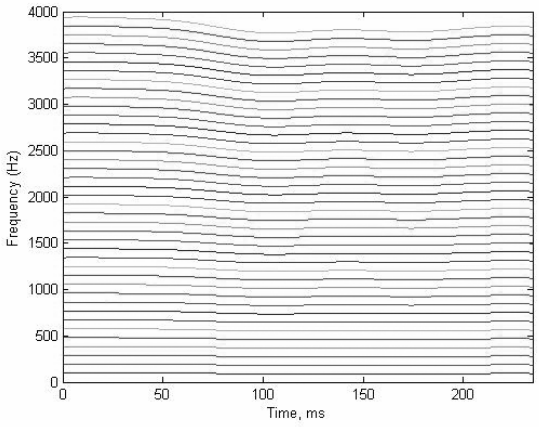
In this section an example of proposed speech synthesis is shown and compared with a time domain concatenation approach.

In order to estimate overall synthesis quality of the system for different voices two different databases were prepared – one for male and one for female speaker. Then control phrases were generated for each of them. For comparison the same phrases generated by time domain concatenation TTS system were used. For synthesis were used the same concatenation rules. The phrase consisted of 10 words in Russian pronounced with specified intonation. Speech signals were sampled at 8000 Hz.

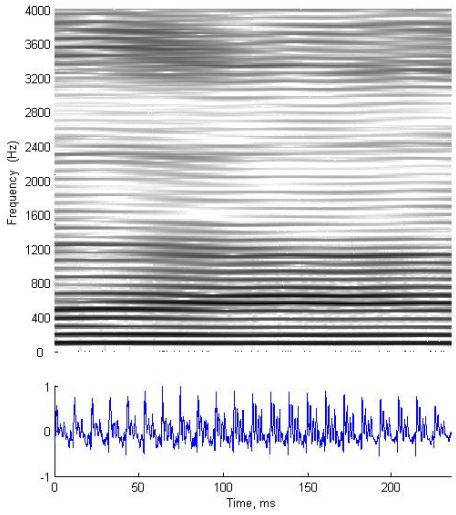
Listening tests were carried out with participation of 5 speech experts. All of them came to the conclusion that spectral based TTS system produces synthetic speech that is closer to original than the speech synthesized by time domain based TTS system.



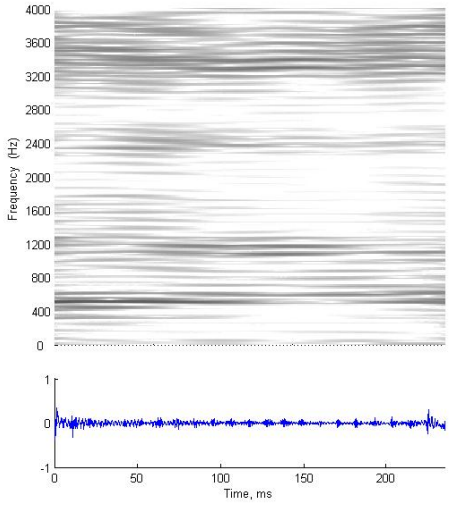
A



B



C



D

Figure 3 – Allophone harmonic/noise separation
A – source allophone, B – estimated harmonic frequency contours, C – harmonic part, D – noise part

Listening proved that the proposed approach produces sounds that are closer to original and free from concatenation artifacts; moreover intonations are more natural due to advanced prosodic modifications. Improvements were more noticeable for female speakers since time concatenation techniques are less suitable for sounds with high pitch and require special correction and accuracy.

To demonstrate the difference between two approaches in Fig.5 two signals are shown. They were synthesized for the word /al'o:/ using different methods. The concatenation rules were the same and the same allophones set was used (male speaker) for synthesis. In Fig.6 two respective spectrograms are shown that indicate differences in spectral domain representation of the signals.

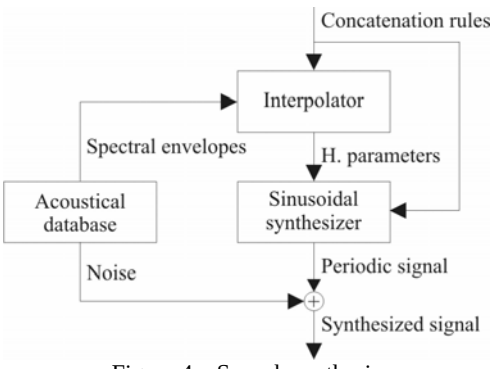


Figure 4 – Speech synthesis

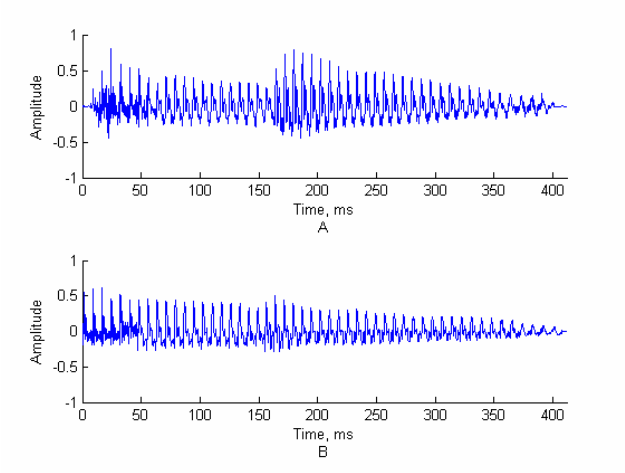


Figure 5 – Synthesized word
A – spectral concatenation
B – time domain concatenation

In presented spectrograms (Fig.6) it is seen that time domain concatenation approach produces audible artifacts in concatenation points. They are caused by phase and pitch differences, that cannot be effectively handled. The proposed spectral concatenation provides inaudible phase and pitch matching, that does not distort spectral and formant structure of the signal. However, some fricative sounds like Russian 'r' require special adjustments of filter parameters in analysis procedure, because of complex structure of these sounds.

4. Discussion

Segment concatenation is still the main approach in TTS systems and widely used in different applications. Concatenation technique in time domain forces TTS systems designers into significant complication of the system in order to eliminate artificial accent from the synthesized speech. There is a tendency towards database extension, and storing combinations of sounds as well as single allophones. This solution can provide necessary results, however requires many efforts at database preparation stage.

The spectral concatenation approach can provide much more natural quality with much smaller acoustic database. It is possible due to smooth concatenation and efficient prosody modifications that can be easily implemented using H+N model. Final quality of speech synthesis however entirely depends on analysis accuracy (harmonic parameters estimation and periodic/noise separation). Sufficient accuracy is achievable by means of instantaneous speech harmonic analysis technique that is used in this work. The experiments have shown that this technique is good enough even for short and fricative segments.

5. Conclusions

In this paper a TTS system with spectral acoustic processor has been proposed. The proposed concatenation technique is based on the instantaneous harmonic analysis, that concatenates segments in frequency domain. The experiments that were carried out showed that this approach provides

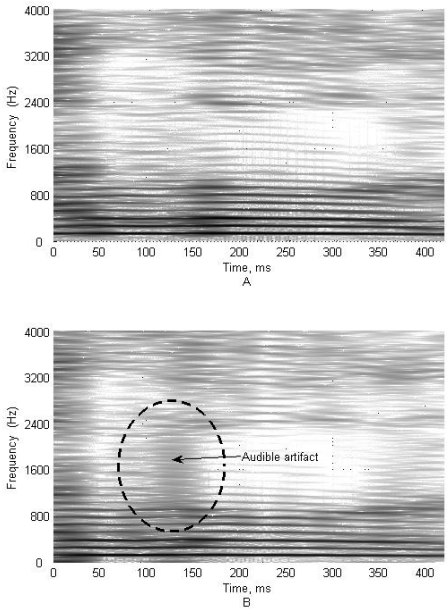


Figure 6 – Synthesized word spectral representation
A – spectral concatenation
B – time domain concatenation

better speech quality and can be used in various TTS applications.

6. Acknowledgements

The authors are grateful to a Byelorussian fund of fundamental researches (the grants No. F08R-016 and No. T08MC-040) for a partial financial support of the present research.

7. References

[1] T. Dutoit. "An Introduction to text-to-speech synthesis"// Kluwer Academic Publishers, 1997. – 286 p.

[2] B. Lobanov, L. Tsirulnik. "Competer Speech Synthesis and Cloning" // Belarussian science, Minsk, 2008, 337 p. (in Russian)

[3] P. Zubrycki, A. Petrovsky. "Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform" // in *Proc. of the 15th European Signal Process. Conf., (EUSIPCO-2007)*, Poznan, 2007, pp.2336-2340.

[4] George E.B., Smith M.J.T. „Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model", *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 5, pp. 389-406, 1997.

[5] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. ICASSP*, 1995, pp. 756–759.

[6] E. Azarov, A. Petrovsky, M. Parfieniuk. "Estimation of the instantaneous harmonic parameters of speech" // in *Proc. of the 16th European Signal Process. Conf., (EUSIPCO-2008)*, Lausanne, 2008, CD-ROM.

[7] X. Huang, A. Acero, H.W. Hon, "Spoken language processing" Prentice Hall, New Jersey 2001, 980 p.