ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2009
Informatyka – Zeszyt 4

Andrzej Sawicki[1], Piotr Zubrycki[1], Alexandr Petrovsky[1]

DESIGN OF TEXT TO SPEECH SYNTHESIS SYSTEM BASED ON THE HARMONIC AND NOISE MODEL

Abstract: This is a proposal of concatenative text to speech synthesizer for the Polish language, based on diphones and "Harmonics and Noise Model"(HNM). HNM has been successfully applied on a speech encoder and decoder, resulting in a high-quality of processed speech at low bit rate. Applying this model to speech synthesis system allows obtaining good quality of synthesized speech, and the small size of database parameters.
The proposed project consists of two main modules. The Natural Language Processing (NLP) is used to analyse and convert the written text for phonemes and diphones using morphological rules. NLP discovers at the same time prosodic features for later modification of synthesized speech parameters in order to obtain the stress and voice intonation. The second section is a synthesis system, derived from speech decoder, preceded by a system of adapting the parameters of speech based on prosodic rules.
The system of speech synthesis from the parameters is working in the frequency domain and uses the frequency spectrum envelope, which easily allows modifying the frequency, amplitude and duration of the signal when applying the prosodic rules. The algorithm of continuous phase designation at the speech frame borders allows concatenating portions of synthesized speech and diphones without phase distortion on the merger. Speech synthesizer operates on the diphone database, created applying fragmentation of recorded speech signal representing the pairs of phonemes. Sounds related to diphones are analyzed by speech encoder. It provides the parameters that described harmonic and noise components of speech, using the linear prediction filter LSF coefficients, resulting in a small size of diphone database.

Keywords: Speech synthesis, TTS, harmonic and noise model

1. Introduction

Most of modern Text To Speech (TTS) systems are based on unit concatenation[1]. Concatenative text-to-speech systems are designed to produce speech by concatenating small segmental units of speech, such as phonemes, diphones or triphones. TSS systems uses database of recorded, segmented and labeled utterances and words. The choice of unit size motivated by vocabulary is a key element in TTS systems for

[1] Faculty of Computer Science, Bialystok Technical University, Biaystok

improving the synthesis quality and meeting storage requirements. Good quality of produced speech, close to the original voice, assured system with huge databases of recorded speech and large concatenation speech units.

Such model however, has large memory and computing requirements, and its implementation especially in mobile systems is problematic. Small units, like diphones, gives smaller database and computational requirements, but cause several problems, such as distortions at the concatenation points. Distortions can be reduced through select suitable speech model, which provides spectral smoothing between combined parts of speech.

In this paper we propose a concatenative Text To Speech synthesiser for the Polish language, based on diphones and Harmonics and Noise Model (HNM)[8] of speech.

HNM has been successfully applied on a speech encoders and decoders, resulting in a high-quality of processed speech at low bit rate [2]. Applying this model to TTS synthesis system as essential of signal processing module, allows to get good quality of synthesised speech, and the small size of database parameters. The size of the speech database may be reduced through record only HNM speech coefficients: harmonic and noise envelopes, and pitch frequencies of diphones.

The proposed project of TTS system consists of two main modules. The Natural Language Processing module (NLP) is used to text analyse and to provide to the synthesiser necessary prosodic and phonemic information. HNM Synthesis module is used to produce synthetic speech.

The paper is organised as follows. Section 2 of document provides description of Natural Language Processing module in TTS system for Polish language. Section 3 describes in detail HNM for speech analysis and synthesis. In section 4 process of diphones database creation is introduced. In section 5 TTS synthesis system is presented. Section 6 comprises conclusions of the article.

2. Natural Language Processing

Natural Language Processing (NLP) module is responsible for text analysing and its conversion to a phonetic transcription. First the incoming text must be accurately converted to its phonemic and stress level representations. Written text and all symbols, numbers, abbreviations and non-text expressions should be converted into speakable forms. This includes determination of word boundaries, syllabic boundaries, syllabic accents, and phonemic boundaries. Then prosody properties of text should be discovered for proper intonation and stress in synthesised speech pronunciation.There are

numerous methods that have been proposed and implemented for the text processing for English, especially. For Polish language, see [5].

2.1 Text preprocessing and normalisation

Text normalization encompasses all aspects of conversion from the mixture of abbreviations, symbols, numbers, dates, and other no orthographic entities of text into a appropriate orthographic transcription. Text is divided into phrases, using end of phrase punctuation marks as '.', ',', '?', '!'. Then sentences are splitted into individual tokens based on whitespaces and punctuation marks. Each token can be classified into one of the following group:

- Words
- Symbols, E.g.: =, +, -, %, $.
- Abbreviations, E.g.: mgr., n.p.
- Numbers, E.g.: 1,2,3 etc.
- Time, E.g.: 12:34
- Date, E.g.: 01/02/2008, 12.03.06, 2008 r.,
- URLs and E-mails, E.g.: somebody@domain.com

Identified tokens we have to expand to full text, using lexicon and rules, e.g. + - plus, 2 - dwa (eng. two), 12:34 - dwunasta trzydzieści cztery (eng. thirty four past twelve) , 01/02/2008 - pierwszy luty dwa tysiące osiem (eng. the first of February two thousand eight).

2.2 Grapheme to phoneme conversion

Pronunciation of a words may be determined either by a lexicon (a large list of words and their pronunciations) or by letter to sound rules. In our implementation for Polish language, grapheme to phoneme rules are the basis of the conversion process. For nonstandard and foreign language words, lexicon with direct phoneme translation is used.

In this article we use SAMPA (Speech Assessment Methods Phonetic Alphabet) symbols of phonemes, which mapping symbols of the International Phonetic Alphabet onto ASCII codes. The Polish language acoustic system comprises 37 phonemes, included eight vowels:

- Oral: a, o, u, e, I, i, pronounced [6] as in words *para, bok, paru, era, dary, lis* .
- Nasal: e , o , pronounced as in words *idę, tą*.

Consonants comprise 29 sounds:

- Plosives: p, b, t, d, k, g, pronounced as in words *pani, bak, tama, dam, kura, noga*.
- Fricatives: v, f, s, z, Z, S, z', s', x, pronounced as in words *wam, fajka, sam, koza, żona, szary, kozia, Kasia, chór*.
- Affricates: ts, tS, ts', dz, dZ, dz', pronounced as in words *koca, lecz, lać, rydza, dżem, działa*.
- Nasals: m, n, n', N, pronounced as in words *mam, len, bańka, bank*.
- Approximants: w, l, r, j, pronounced as in words *łam, pal, rak, daj*.

Our TTS system use diphones for concatenate speech synthesis. Diphones are adjacent pair of phonemes, selected from the stationary area of first phone to stationary area of second phone in recording.

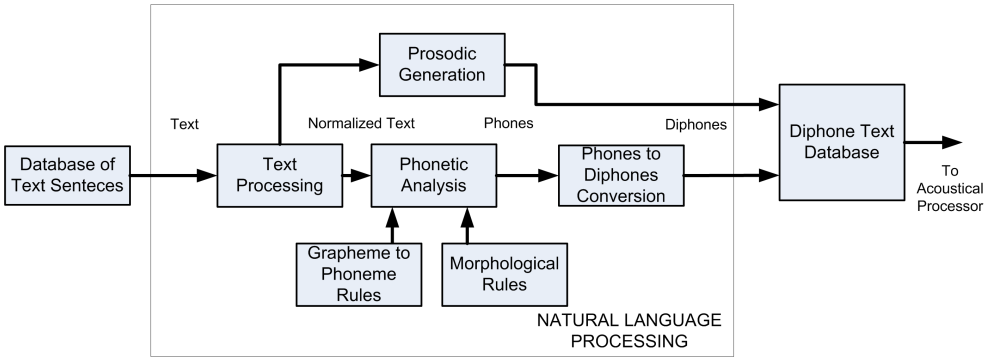

Fig. 1. Natural language processing module

Processing of natural language (NLP) is based on the model proposed by A.W. Black and P. Taylor in the Festival [4] TTS system and used in open source software speech synthesizer eSpeak [14]. Text analysis and prosody annotation for Polish language is based on the work of S. Grocholewski and G. Demenko [5]. Our system use for tests and research database of recorded texts in Polish language 'CORPORA' developed by prof. S. Grocholewski [3]. Diagram of proposed NLP module is presented in fig.1.

3. Harmonic and noise model for speech analysis and synthesis

3.1 Speech analysis

High quality speech coding at low bit-rates is major interest of speech coding methods. Speech signal modelling based on HNM was presented in [8]. Speech is divided into two subbands by the maximum voiced frequency, lower band is considered fully voiced and upper band fully unvoiced. From the speech production point of view it is clear, that both voiced and unvoiced components are present in whole speech band. This idea was used by Yegnanarayana et. all [9] in speech decomposition method into voiced and noise components. Decomposition of speech is performed on excitation signal approximated with use of inverse linear prediction filter. Idea of work is to use an iterative algorithm based on Discrete Fourier Transform (DFT)/Inverse Discrete Fourier Transform (IDFT) pairs for noise component estimation. Voiced component of excitation is obtained by subtracting noise component from LP residual. These approaches were designed without taking into account time-varying fundamental frequency and harmonic amplitudes.

In this paper we present another approach to speech decomposition into voiced and noise components and its application to speech synthesis system. As the methods presented in [9],[10] our method considers voiced and noise components present in whole speech band. Pitch-Tracking modification is applied to standard DFT in order to provide spectral analysis in harmonic domain rather than frequency domain. Voiced component parameters (i.e. harmonics amplitudes and phases) are estimated in harmonic domain. Estimation is done every 16ms. Amplitudes and phases of harmonics are interpolated between points of estimation. Voiced component is generated with time-varying frequency and harmonics amplitudes. After voiced component generation decomposition of speech signal is done in time domain. An iterative algorithm is used for decomposition in order to obtain exact components separation. Time-domain speech components separation and voiced component modelling method is sensitive to pitch estimation errors, thus precise and accurate pitch detection algorithm is needed. A robust pitch detection method based on tuning pitch frequency to its harmonics presented in [11],[2] is used.

Pitch estimation and tracking. Pitch tracking algorithm operates both in time and frequency domain. Preliminary pitch estimation is taken every 8ms using autocorrelation method. This estimate is used to refine pitch frequency using algorithm working in spectral domain similar to the one proposed in [12]. In order to prevent Gross Pitch Errors (GPE) and pitch track distortions a tracking algorithm is used. Scheme of pitch

estimation and tracking algorithm is shown on fig. 2. First, the autocorrelation vector is computed. In order to improve robustness of the algorithm low-pass filtering and signal clipping operation are performed according to Sondhi [13]. Autocorrelation vector is computed in interval corresponding to possible fundamental frequency values (typical from 60 to 500Hz) using formula:

$$R(k) = \sum_{n=0}^{N-1} s(n)s(n+k), k = -l..l, \tag{1}$$

where l is maximum lag corresponding to minimal fundamental frequency.

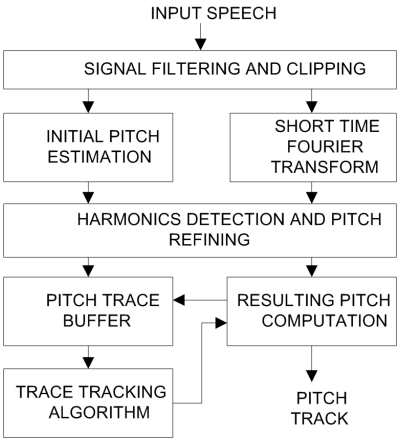

Fig. 2. Pitch estimation algorithm

Length of autocorrelation window is 32ms and 16ms before and after the frame is needed, thus algorithm operates with delay of 16ms. This approach to autocorrelation computation enables transient frame detection, because resulting vector is not symmetric. If the speech segment is transient maximum of autocorrelation is available only at one side of vector, and the side depend on the frame is beginning or ending of voiced segment. Initial pitch estimation is computed as weighted mean of maximums of autocorrelation sequence on left and right side.

After initial estimation input speech signal is weighted in 256-point time window and STFT is computed. Initial pitch value is tuned to all present pitch harmonics [2]. In case of inability to identify at least two out of four leading pitch frequency harmonics, the segment is considered unvoiced. Refined pitch value F_r for each weighting window is identified with the harmonic factor, which can be understood as

adequacy of the estimation:

$$h_f = \frac{n_h}{n_{max}} \tag{2}$$

where n_h is number of present harmonics, n_{max} is number of all possible harmonics with given pitch.

False pitch frequency estimations got during speech flow pauses are discarded on the base of analysis of the weighting factors of pitch frequency estimations and analysis of the values of the input signal level, of the speech and silence levels. In order to prevent gross errors and provide better quality, pitch estimation is performed with a delay of two analysis windows. Estimations of the pitch frequency are included in the current track in case the difference between neighbouring windows does not exceed the allowed one. Trace tracking estimation of pitch frequency is calculated using linear approximation of current trace according to the least-square method. The condition determining end of the trace tracking is checked by availability of preliminary estimations to the right of the window being analysed and by harmonic factors. Resulting pitch frequency is determined as:

$$F_0 = h_f F_r + (1 - h_f) F_t \tag{3}$$

where F_r is refined pitch; F_t is trace tracking estimation.

Pitch-tracking modified DFT. Pitch-Tracking modified DFT transform providing analysis in spectral domain is given by:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi n k F_0}{F_s}}, k = 0..K \tag{4}$$

where $X(k)$ is k-th spectral component corresponding to k-th harmonic, $x(n)$ is input signal, N is transformation length, F_s is sampling frequency, F_0 is fundamental frequency. Kernel of transformation has to be modified in case tracking analysis. Argument of exponential can be written as follows:

$$\varphi(n,k) = \sum_{i=1}^{n} \frac{2\pi k (F_0(i) - F_0(i-1)}{2F_s}, n \neq 0 \tag{5}$$

$\varphi(n,k) = 0$, `for`$: n = 0$, and $F0(i)$ is fundamental frequency at the time specified by i. Transformation providing tracking harmonic analysis is given as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\varphi(n,k)}, k = 0..K \tag{6}$$

Non-orthogonal transformation kernel can cause energy leakage to neighbouring spectral lines. Time-Varying window is used in order to deal with leakage phenomenon. Idea of this solution is to design a spectral window, which follows fundamental frequency changes. Window length is chosen to contain at least two pitch periods. Experiments showed, that window length should be approximately 2.5 pitch period which is a trade-off between spectral resolution and computational efficiency. Good results could be achieved when Kaiser window is used as a prototype [11]:

$$w(n) = \frac{I_0(\beta\sqrt{1-(\frac{2x(n)-(N-1)}{(N-1)})^2})}{I_0(\beta)}, n = 0..N-1 \tag{7}$$

where N is window length, β is window parameter, $I_0(x)$ is zero order Bessel function, $x(n)$ is a function enabling time-varying feature, given as:

$$x(n) = \frac{\varphi(n-1)}{2\pi\bar{F}_0\frac{N}{F_s}} \tag{8}$$

where $\phi(n,1)$ is computed using formula (5), $\bar{F}_0$ is average fundamental frequency in analysis frame.

Decomposition algorithm. In this paper we present solution which is based on continuous harmonic component generation. Continuous generation of voiced component is performed with a delay of 16ms which is necessary to iterative algorithm. Proposed method performs decomposition in whole speech band, which leads to more accurate representation of speech. Synthesised signal sounds more natural. Speech signal decomposition scheme is shown on figure 3.

First step of decomposition is pitch tracking. This information is passed to iterative decomposition algorithm. It performs decomposition every 16ms. First step of decomposition is speech windowing with time-varying window. Centre of time window is set every 16ms. In order to reduce leakage length of window is chosen as integer multiple of pitch periods. Pitch-Tracking Modified DFT every 16ms gives an information about instantaneous amplitudes and initial phases of the harmonics. For synthesis of the harmonic component a set of time-varying oscillators can be used:

$$h(n) = \sum_{k=0}^{K} A_k(n)cos(\varphi(n,k) + \Phi_k) \tag{9}$$

where phase $\varphi(n,k)$ is determined using formula (5). While pitch harmonics amplitudes estimation is performed every 16ms instantaneous amplitudes of harmonics have

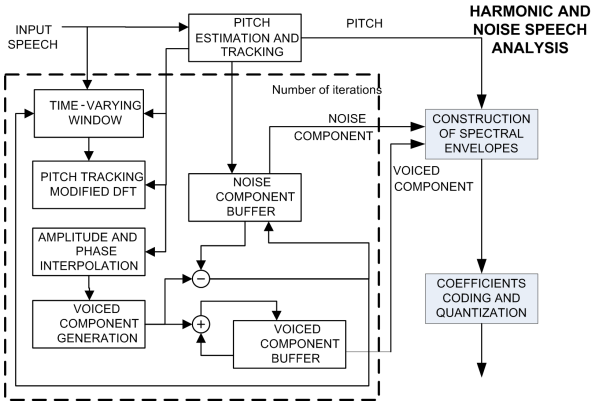

Fig. 3. Harmonic and noise speech analysis.

to be computed using interpolation algorithm. Piecewise Cubic Hermite Interpolation method is used, as it can be easily implemented in real-time applications. Having harmonic component, noise component is defined as a difference between input speech signal and harmonic component:

$$r(n) = s(n) - h(n) \tag{10}$$

Voiced component is stored in voiced component buffer and noise component is stored in noise component buffer. After first iteration noise component still consists of some voiced component. This is mainly due inability of perfect reconstruction of harmonic amplitudes tracks imposed by amplitudes variations. In the next iterations noise component is processed in the same way as original speech in first iteration. Remaining in noise voiced component is subtracted from noise component. Output voiced component is sum of harmonic components from all iterations and noise component is residual from last iteration.

3.2 Speech synthesis

The periodic spectrum is created by a set of sinusoidal generators working at variable frequency changing linearly within time window, using equation:

$$H(n) = \sum_{k=0}^{K} A_k(n) cos(\varphi(n,k) + \Phi_l) \tag{11}$$

Correctly calculated phases of pitch harmonics avoid any reverberation in the synthesised speech, especially in diphone concatenation points. In proposed model

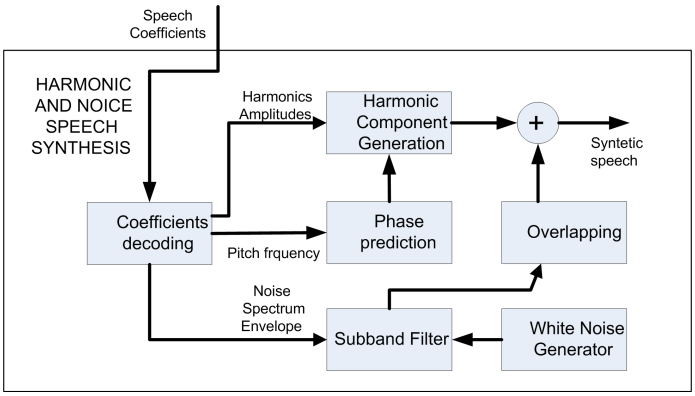

Fig. 4. Harmonic and noise speech synthesis

phases are modelled by a parabolic temporal model of instantaneous phase witch take changes on pitch from frame to frame Phase of k+1'th frame is predicted from k'th frame. New phase is calculated on time-varying contribution of the frequency trajectory:

$$\Phi_l(t) = \frac{\omega_l^{k+1} - \omega_l^k}{2T} t^2 + \omega_l^k t + \phi_l^k \tag{12}$$

where T is frame length, and ϕ_l^k is phase offset at the begin of k'th frame. An initial phases for voiced groups of speech frames are chosen randomly. The unvoiced speech is synthesised by the bandpass filtering of white noise. The voiced and the unvoiced components are added. Detailed view of presented speech synthesiser can be found in [2].

4. Diphone Database Creation.

For diphone database creation we used database of Polish Speech CORPORA [3] , projected by prof. Grocholewski. Database consists 45x365 segmented and labelled utterances. Files in database were recorded using 12 bits resolution and frequency 16 kHz. In Polish language there are 37 phonemes. We use additional symbol sil (#) in order to indicate silence at beginning and end of the words, thus the optional number of phoneme to phoneme connections is about 1400. In [3] it is suggested, that not all combination of neighbouring phonemes occur in polish language, or such diphones happen very rarely. Finally, our diphone database consist of 1270 diphones.

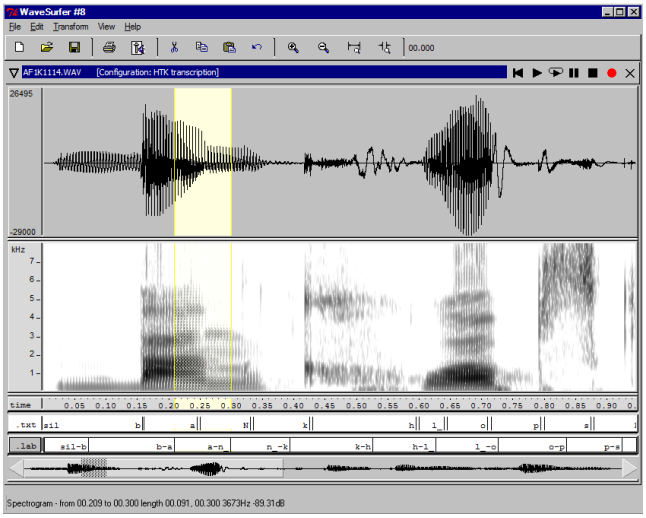

Fig. 5. Waveform diphone labelling using WaveSurfer

Text of database sentences was converted to diphones using NLP module. Preliminary waveform marker description (e.g. beginnings and endings of the particular phoneme) is given in the Corpora database. Proposed system uses diphones as a basic units, thus accurate determination of the exact boundary locations of a diphone in waveform should be performed manually, as shown in figure 5. We use open source program WaveSurfer [7] for waveform labelling. Based on labelled waveform, diphone database for speech synthesis is prepared, using harmonic and noise speech analysis module.

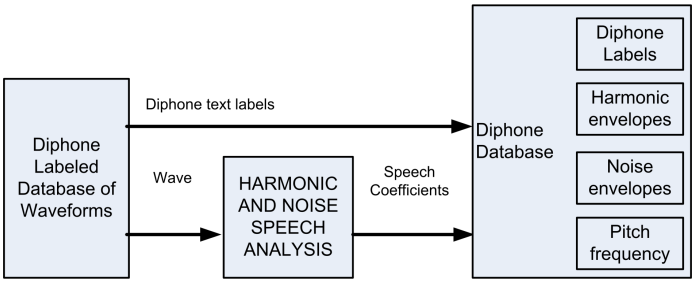

Fig. 6. Diphone database creation.

Each diphone in database is described by set of 16 milliseconds frames included coefficients of speech model:

1. harmonic envelope
2. harmonic gain
3. noise envelope
4. noise envelope gain
5. pitch frequency

For each diphone additional information about its duration is stored in the database. The duration is determined by waveform labels.

5. Text To Speech synthesis using Harmonic and Noise Model.

TTS system consist of 3 main components described above: NLP section, HNM synthesis module and diphone database. The diphone database is created for one speaker. All utterances are analysed using method described in section 3.1 and the database is created as described in section 4. As a result of text processing in NLP module, we get diphone labels and prosody description. The prosody description gives information about stress of the syllable and necessary modifications of the parameters i.e. relative changes of a F_0, duration and gain for particular diphone. Using diphone text labels, a set of frames for actual diphone is selected from database. Speech model coefficients for each frame are transformed in accordance with prosody descriptions. It is worth notice, that using harmonic model makes diphone transformation process easy. Gain and pitch modification is simply straightforward, modification of the diphone duration is done by changing the time offset between frame coefficients (i.e. single frame time-scaling). Converted coefficients are passed to the signal processing module, where synthetic speech is generated in accordance with the HNM, as was described in section 3.2.

For example for given text:

"Artykuł przedstawia projekt konkatenacyjnego syntezatora mowy z tekstu dla języka polskiego."

phonetical description, using presented above phoneme notation, with word stress (appointed as: ‘) is following:

" art‘Ikuw pSetst‘avja pr‘ojekt koNkatenatsIjn‘ego sIntezat‘ora m‘ovI s t’ekstu dla je z‘yka polski‘ego".

And diphone transcription is following: *"#-a a-r r-t t-I I-k k-u u-w w-# #-p p-S S-e e-t t-s s-t t-a a-v v-j j-a a-# #-p p-r r-o o-j j-e e-k k-t t-# #-k k-o o-N N-k k-a a-t t-e e-n n-a a-ts ts-I I-j j-n n-e e-g g-o o-# #-s s-I I-n n-t t-e e-z z-a a-t t-o o-r r-a a# #-m*

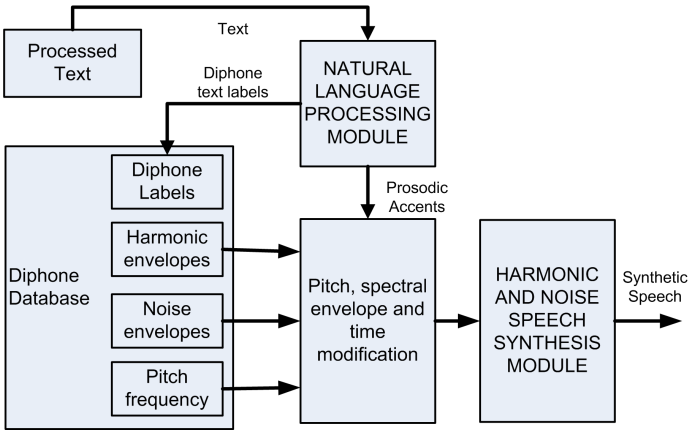

Fig. 7. TTS Synthesis

m-o o-v v-I I-# #-s s-# #-t t-e e-k k-s s-t t-u u-# #-d d-l l-a a-# #-j j-e e -z z-I I-k k-a a# #-p p-o o-l l-s s-k k-i i-e e-g g-o o-#"

Such list of diphones is synthesised using HNM model described in section 3.

6. Conclusions

In this article we propose new approach to the topic of concatenative text to speech synthesiser for the Polish language. System based on Harmonics and Noise Model of speech allow to accurate represent speech signal for diphones database building. Signal processing in frequency domain and instantaneous interpolation of harmonic amplitudes and phases allow to smooth concatenation of diphones. The HNM-based speech analysis/synthesis system proved its ability to produce high quality of synthetic speech [12], almost indistinguishable from the original one. The performance of the proposed TTS system mostly depends on NLP module. Time and frequency modification of diphones can be done separately and without any loose of quality. Presented approach can be successfully used in high-quality speech synthesis applications for TTS system.

Our TTS system is still in development. Especially prosody parsing and diphones modification in accordance with prosody annotations have need of improvement. Current version of NLP module gives information about stress in syllables, phrase intonation description function is under development. While voiced component modification is easy task, unvoiced component in current version of the system can change only its duration which can cause artifacts. Future works include improvement of

voiced component modification method. Applying effective methods of speech compression using vector quantization for database size reduction is also under investigation.

Small database size and effective algorithms of speech generation allow to implement presented system in embedded devices.

References

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

[2] Petrowsky A., Zubrycki P., Sawicki A.: Tonal and Noise Components Separation Based on a Pitch Synchronous DFT Analyzer as a Speech Coding Method, Proceedings of ECCTD, 2003, Vol. III, pp. 169-172.

[3] Grocholewski S.: Założenia akustycznej bazy danych dla języka polskiego na nośniku CD-ROM, Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław 1995, s. 177-180

[4] A. Black and P. Taylor: Festival Speech Synthesis System: system documentation (1.1.1), Human Communication Research Centre Technical Report HCRC/TR-83, 1997.

[5] Demenko, G. Grocholewski, S. Wagner, A. Szymanski M.: Prosody annotation for corpus based speech synthesis. [in:] Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology, New Zealand. Auckland, 2006.

[6] M. Wiśniewski: Zarys fonetyki i fonologii współczesnego języka polskiego, wyd. Uniwersytetu Mikołaja Kopernika, Toruń, 2007.

[7] Sjolander, Kyre / Beskow, Jonas: Wavesurfer - an open source speech tool, In ICSLP-2000, vol.4, 464-467

[8] Y. Stylianou, Applying the Harmonic Plus Noise Mode in Concatenative Speech Synthesis, IEEE Trans. on Speech and Audio Processing, vol. 9, no 1., 2001.

[9] B. Yegnanarayana, C. d'Alessandro, V. Darsions An Iterative Algorithm for Decomposiiton of Speech Signals into Voiced and Noise Components, IEEE Trans. on Speech and Audio Coding, vol. 6, no. 1, pp. 1-11, 1998.

[10] P.J.B. Jackson, C.H. Shadle, Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence-Noise Components in Speech, IEEE Trans. on Speech and Audio Processing, vol. 9, no. 7, pp. 713-726, Oct. 2001

[11] V. Sercov, A. Petrovsky, An Improved Speech Model with Allowance for Time-Varying Pitch Harmonic Amplitudes and Frequencies in Low Bit-Rate MBE Coders, in Proc. of the 6ht European ?onf. on Speech Communication and Technology EUROSPEECH'99, Budapest, Hungary, 1999, pp. 1479-1482.

[12] P. Zubrycki, A. Petrovsky Analysis/synthesis speech model based on the pitch-tracking periodic-aperiodic decomposition, in Information processing and security systems (Khalid Saeed, Jerzy Peja eds.) Springer Verlag, Heidelberg 2005, pp. 33-42

[13] M.M. Sondhi, New Methods of Pitch Extraction, IEEE Trans. on Audio and Electroacoustics, vol. AU-16, no. 2, pp. 262-266, 1968.

[14] Espeak, eSpeak text to speech, http://espeak.sourceforge.net/[viewed 15/09/2009]

KONCEPCJA UKŁADU SYNTEZY MOWY Z TEKSTU OPARTEGO NA MODELU HARMONICZNE I SZUM

Streszczenie: Artykuł przedstawia projekt konkatenacyjnego syntezatora mowy z tekstu dla języka polskiego, opartego na difonach i modelu Harmoniczne i Szum. Model Harmoniczne i Szum został z powodzeniem zastosowany w układzie kodera i dekodera mowy, dając w rezultacie dobrą jakość przetwarzanej mowy przy niskiej przepływności bitowej. Zastosowanie tego modelu do układu syntezy mowy pozwala na uzyskanie dobrej jakości syntezowanej mowy, oraz niewielki rozmiar bazy parametrów.

Układ składa się z dwch głównych modułów. Moduł Naturalnego Przetwarzania Języka służy do analizy i zamiany tekstu pisanego na fonemy oraz difony, przy wykorzystaniu reguł morfologicznych. Procesor tekstu wyznacza jednocześnie warunki prozodii związane z późniejszą modyfikacją parametrów syntezowanego głosu w celu uzyskania akcentowania i intonacji. Drugim układem jest moduł syntezy, oparty na dekoderze mowy poprzedzonym systemem adaptacji parametrów mowy w oparciu o wyznaczone wcześniej reguły prozodyczne.

Układ syntezy mowy z parametrw działa w dziedzinie czstotliwości i bazuje na obwiedni spektrum, co w prosty sposób pozwala na modyfikację czstotliwości, amplitudy i czasu trwania sygnału przy stosowaniu reguł prozodycznych. Algorytm wyznaczania ciągłej fazy na granicach ramek sygnału mowy pozwala na łączenie fragmentów syntezowanej mowy oraz poszczególnych difonów bez zniekształceń fazowych na połączeniu.

Syntezator mowy operuje na bazie difonów, stworzonej na podstawie fragmentaryzacji nagranego sygnału mowy na części, reprezentujące połączenia par fonemów. Dźwięki odpowiadające difonom są analizowane przez moduł analizy mowy. Dostarcza on ciąg parametrów reprezentujących harmoniczne i szumowe komponenty sygnału mowy, opisane za pomocą filtrów liniowej predykcji i współczynników LSF, dając w rezultacie niewielkiej wielkości bazę difonów.

Słowa kluczowe: Synteza mowy, model harmoniczne i szum

Artykuł zrealizowano w ramach pracy badawczej W/WI/6/09.