# Development of Text-to-Speech Synthesizer for Latvian

*Andrejs Vasiljevs (andrejs@tilde.lv)*

*Term Paper for NGSLT Speech Technology Course, Autumn 2006*

## 1 Introduction

This paper describes the development of the first text-to-speech (TTS) synthesis system for Latvian language being developed by company Tilde. The project background is briefly provided, the general approach and particular implementation aspects are described. Project is carried out by MSc. Karlis Goba with support from his colleagues. Author of the paper was one of the project initiators and participated in a preparation of the project proposal. Current role of the author is to supervise project implementation and to ensure integration of the resulting TTS in practical applications. Paper is based on project work materials, interviews with Karlis Goba, studies of literature and experience of author during project supervision.

## 2 Project Background

Latvian language is the only official language in Latvia and one of the working languages of European Union spoken by 1.6 M people. Despite important role of Latvian until now there was no text-to-speech synthesizer for this language. As a result there are no applications in use providing Latvian speech capabilities.

A population group with the most acute need for speech enabled technologies is visually impaired people. TTS is essential technology enabling them to use computer applications, browse internet and communicate with e-mail. Some of them are very efficient computer users using English TTS but majority do not have sufficient English skills. Attempts to use English TTS for reading and preparing Latvian texts were unsuccessful due to principal differences in Latvian and English pronunciation. Latvian text pronounced by English TTS is practically incomprehensible even by the most tolerant and striving users.

There was attempt to create Latvian adaptation of system WinTalker developed by Czech company RosaSoft. The pronunciation generated by pilot model was better than Latvian texts pronounced by non-Latvian TTS but still of a very low quality and barely recognizable. For this reason it was not accepted by users and was not further developed.

In the late 90th and beginning of this century some experiments in Latvian TTS were carried out by the Institute of Informatics and Mathematics at the University of Latvia. These were experiments using original approach generating speech by concatenation of individually recorded phonemes. It became clear quite soon that such approach cannot lead to human-like speech and experimental system was never completed.

Juris Grigorjevs researched different aspects of Latvian phonetics at Faculty of Philology at the University of Latvia. In this research [3] he carried out experiments on Latvian vowel generation using formant synthesis method.

In 2005 company Tilde and Association of Blind People started a project to develop Latvian TTS. Project is part of European Commission funded programme to facilitate impaired people. Description of this project is provided in this paper in more details.

# 3 System requirements

The primary purpose of Latvian TTS is to solve needs of visually impaired people working with computer in Latvian language – browsing Latvian internet, reading and creating Latvian documents, enabling e-mail and chat communication in Latvian.

At the beginning of the project requirements for the Latvian TTS have been determined in cooperation with Association of Blind People.

## 3.1. Functional requirements

- Variable speech rate and pitch
- Speech generation for arbitrary texts
- Speech generation for software interface elements in Latvian (menus, dialogs etc.)
- Fast response in text input mode - character pronunciation starts not later than 50 milliseconds after input
- Fast response in reading mode - no delays starting to read text, new part of text or word
- Possibility to modify pronunciation of particular words, part of word or characters
- Pronunciation of punctuation marks in Latvian
- Pronunciation of numbers and abbreviations in Latvian

## 3.2. Speech quality requirements

- Distinctive and correct pronunciation of separate Latvian characters during text input
- Accurate Latvian intonations and stresses on word and phrase level
- Accurate sentence level intonations for narrative, question and exclamatory sentences
- Sentence level intonation changes according to length and structure of sentence
- Noise free, emotionally neutral voice

## 3.3. Technical and compatibility requirements

- Integration with Jaws for Windows and ZoomText
- Microsoft Speech Application Programming Interface (SAPI) 4.0 and SAPI 5.0 support
- Support for both 8-bit and UNICODE text encoding

# 4 General approach

The following steps are passed in TTS:

1. Abbreviations and numericals are detected and converted to full words
2. Intonation phrases are detected and marked, punctuation marks are eliminated
3. Pronunciation of homographs is determined
4. Words are split into syllables, word-level stresses and intonation are detected and marked
5. Phonological rules are applied, character representation denoting broad transcription is transformed to narrow transcription depicting appropriate allophonic variations.
6. Prosody is modeled, F0 conture and rhythm (length) is determined for phrases, syllables and phonemes.
7. Speech waveform is synthesized by concatenating and modifying diphones.

The system architecture consists of 3 main modules.

**Text normalization module** – transformation of input text, converting numbers to textual form, expanding abbreviations, converting other nonorthographical entities of text and applying rules from speech dictionary. Output is *normalized* text containing common orthographic transcription suitable for subsequent phonetic conversion.

**Phonetic analysis module** – apply phonological rules for Latvian to convert string of letters to string of phones. Intonations, prosody and rhythm of speech are modeled and attached to phones as numerical parameters. The result is phonemic representation along with prosodic information. Process carried out in this module can be viewed as grapheme-to-phoneme conversion and prosody modeling.

**Sound synthesis module** is responsible for generation of speech waveforms from phonetic transcription by selecting, concatenating and modifying diphones.

### 4.1. Text normalization

Text normalization is the process of generating normalized orthography from text containing words, abbreviations, numbers, punctuation and other symbols.

Latvian language is highly inflected language. This means that creation of correct orthographical form in most cases requires not only expansion of symbols or abbreviations but also inflection in required form.

Some Latvian abbreviations have fixed representations:

| | |
|---|---|
| *u.t.t.* | *un tā tālāk* (etc.) |
| *u.tml.* | *un tamlīdzīgi* (and that sort of thing) |
| *piem.* | *piemēram* (for example) |
| *t.i.* | *tas ir* (id est, ie.) |
| *t.sk.* | *tai skaitā* (inter alia) |

Other abbreviations should be represented in inflected form corresponding to the context:

| | |
|---|---|
| *u.c.* | *un citi/un citiem/un citos* (etc.) |

Pronunciation of acronyms depends on linguistic traditions. Some acronyms traditionally are pronounced letter by letter:

| | |
|---|---|
| *ASV* | *ā es vē* (USA) |
| *PVN* | *pē vē en* (Value Added Tax, VAT) |
| *LMT* | *el em tē* (Latvian Mobile Telephone) |
| *ES* | *ē es* (European Union, EU) |
| *LTV7* | *el tē vē septiņi* (Latvian TV7) |

Some acronyms are pronounced as words:

| | |
|---|---|
| *NATO* | *nato* (NATO) |
| *VID* | *vid* (State Revenue Service) |
| *SIA* | *sia* (Limited Liability Company, Ltd.) |
| *KNAB* | *knab* (Corruption Prevention Bureau) |

There are foreign language acronyms used in Latvian that are pronounced according to pronunciation in original language, usually English,:

| | |
|---|---|
| *UNESCO* | *junesko* (UNESCO) |

Some acronyms traditionally are expanded while reading texts:

| | |
|---|---|
| *A/S* | *akciju sabiedrība* (Stock Company) |

In such cases appropriate inflectional form should be determined. The same problem is with traditional abbreviations for measurement units:

| | |
|---|---|
| *g.* | *gads* (year) |
| *g* | *grams* (gram) |
| *kg* | *kilograms* (kilogram) |
| *m* | *metrs* (meter) |
| *km* | *kilometers* (kilometer) |
| *h* | *stundas* (hours) |
| *min.* | *minūtes* (minutes) |
| *s./sek.* | *sekundes* (seconds) |

Different variations of number formats and numerical phrases should be expanded to full orthographical transcription:

| | |
|---|---|
| *2:3* | *divi pret trīs, divi, kols, trīs* |
| *40,2%* | *četrdesmit, komats, divi procenti* |
| *2.4* | *divi punkts četri* |
| *#4* | *numur četri* |

Date and time expressions should also be normalized applying appropriate Latvian inflection that depends on context:

| | |
|---|---|
| *20. janv.* | *divdesmitais janvāris* (twentieth of January) |
| *15.30* | *piecpadsmitos trīsdesmit* |

The following are normalization examples for money and currency notations:

| | |
|---|---|
| *milj.* | *miljardi* |
| *mlj.* | *miljoni* |
| *Ls/LVL* | *lats* |
| *Ls 3,5 milj.* | *trīs komats pieci miljoni latu* |

Most symbols have traditional Latvian representation:

| | |
|---|---|
| / | *slīpsvītra* |
| + | *plus* |
| & | *un* |
| [ | *atverošās kvadrātiekavas* |

Still there are symbols those oral representation in Latvian is not widely known or is used only by specific groups.

| | |
|---|---|
| @ | *et* |

User has possibility to change pronunciation of these and other symbols making appropriate entries in user dictionary.

### 4.2. Phonetic analysis

In author's view Latvian in general can be considered as a *phonetic language* – language with relatively simple relationship between orthography and phonology as defined by Huang etc. [1].

Still number of rules and exceptions should be taken into account during grapheme-to-phoneme conversion process. Some of these rules and exceptions have been identified during the project.

For example, Latvian vowels *ā, ē, ī, ū* ("long" vowels) have the same characteristics as corresponding "short" vowels except length – they are significantly longer. This effect was discovered in previous research made by Karlis Goba. It is also described by Juris Grigorjevs [3]. Author of the paper observed the same effect in analysis of his voice during first course work (see Table 1).

This means that prolonging of sound duration can be used to generate "long" vowels from the same base phonemes as used for the "short" vowels.

|  | Duration (s) | F1 frequency (Hertz)<br>Mean/(Standard deviation) |
|---|---|---|
| **a** | 0.061 | **544 (sd 60)** |
| **ā[a:]<br>(māja)** | 0.168 | **539 (sd 23)** |
| **e** | 0.049 | **413 (sd 23)** |
| **ē** | 0.151 | **428 (sd 16)** |
| **e<br>(desa)** | 0.13 | **520 (sd 18)** |
| **ē<br>(tēvs)** | 0.1525 | **572 (sd 18)** |
| **i** | 0.127 | **402 (sd 19)** |
| **ī** | 0.155 | **381 (sd 28)** |
| **u** | 0.039 | **346 (sd 29)** |
| **ū** | 0.128 | **374 (sd 18)** |

**Table 1 Duration and formant frequency for unstressed Latvian vowels. Results of experiments during Assignment 1.**

Latvian unvoiced stops *p, t, k, ķ* may have regular and extended length depending on context. For example, *k* has regular length in word *maks* but extended length in word *aka.*

In Latvian there are number of homographs with different pronunciation for different senses of homograph. For example, word *robots* as noun (/r o b o t s/) and *robots* as adjective (/r u o b u o t s/). Homograph disambiguation and morphological analysis can be applied to determine word sense and corresponding pronunciation.

### 4.3. Prosody modeling

Prosody is acoustic properties of sentence pronunciation to express attitude, assumptions and attention which are not described by the sequence of phones derived from the text. Prosody is expressed by pauses, pitch, rate or relative duration and loudness, with pitch being the most expressive phenomena [1].

Latvian TTS deals only with the basic aspects of prosody such as syllable stress and phrase intonation. General rule for Latvian is that syllable stress is on the first syllable. There are number of exceptions to this rule that are stored in a word-list, for example, such frequent words as *paldíes (thank you), labrít (good morning*) and superlatives like *visgudrākais (the smartest).*

For Latvian TTS prosody is modeled by variations of pitch and relative duration of speech elements. Stress is modeled as extruded pitch and/or lengthened duration. In such a way stressed and unstressed syllables are modeled on word level.

Phrase level stress is modeled as well to improve prosodic quality were possible. Prosodic phrases are determined in simplified way by using text punctuation.

To synthesize F0 contour Fujisaki pitch model is used [6] that superimpose both word level and phrase level prosody modulations.

### 4.4. Sound synthesis

In final phase of TTS output speech waveforms are generated from phonetic transcription created in previous phases.

According to Morais and Violaro [4] in this decade corpus-based synthesis approach is a dominant trend in speech synthesis that provides high naturalness, accuracy and intelligibility.

In corpus-based synthesis prerecorded speech units are concatenated and transformed to produce speech. At runtime, appropriate acoustic patterns and prosody of a sentence are superimposed during concatenation by means of digital signal processing techniques.

The drawback of corpus-based synthesis is high development costs and relatively high memory and processing requirements for running system. Corpus-based synthesis that use large number of larger recorded units like sentences, words, phrases and morphemes, may produce higher quality speech but requires lot of processing power and memory for unit storage.

Similar to corpus-based synthesis is diphone synthesis. In this method speech units – building blocks for speech generation – are diphones. Diphones are small units of speech that extend from the middle of one region of steady-state sound to the middle of the next; thus, they represent transitions between speech sounds.

There are usually 700 – 3000 diphones for a language so diphone preparation process is not very costly and their storage is memory effective. From other side the quality of the resulting speech is generally worse than that of corpus-based systems.

For Latvian TTS, a combination of corpus-based and diphone synthesis is used. Very frequent words are used as single speech units, e.g., weekdays, months, frequent country and city names etc. Other words are synthesized from diphones.

In traditional diphone synthesis only one example of each diphone is contained in the speech database. For Latvian TTS it was decided to keep several examples of some diphones to improve speech quality and to include contextual variations of vowels and consonants. Also diphones that were consecutive in the original speech recording were included, thus effectively increasing the unit length. Such combined approach is a good compromise between speed and effectiveness of speech synthesis, and quality of produced speech.

For speech data creation text to be recorded was prepared consisting of phonetically and prosodically balanced sentences. Special attention was made to include necessary diphones in different accent, stress and intonation combinations. Several professional speakers were tested to select the most appropriate voice. Association of Blind People was involved in speaker selection. Their preference was to use male voice with balanced and articulated voice. The recording was made in professional studio by recording company experienced in preparation of voice notifications and messages. Recorded sentences were phonetically segmented and manual selection of diphones was made from recorded material.

### 4.5. LPC for speech representation and synthesis

For conversion of phonetic transcription to the speech waveforms the most appropriate diphones should be selected from a set of corresponding recorded diphones.

In the Latvian TTS system speech units are coded using linear predictive coding (LPC) method [1]. LPC is used in speech processing for representing the spectral envelope of a digital signal of speech using the information of a linear predictive model by predicting the current sample as a linear combination of its past samples.

LPC is based on the simplified assumption (source-filter model) that a speech signal is produced by a glottis created buzzer at the end of a vocal tube (voiced sounds) with occasional added hissing and popping sounds (sibilants and plosive sounds) generated by the action of the tongue, lips and throat. The resonances of vocal tube are characterized by formants. The buzz is characterized by its intensity (loudness) and frequency (pitch).

LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue. LPC coding uses numerical coefficients which describe the intensity and frequency of the buzz, the formants, and the residue signal as the main characteristics of speech signal.

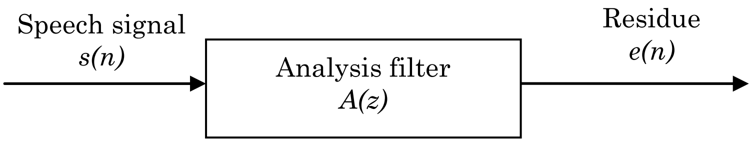

**Fig. 1 LPC analysis model**

In Latvian TTS LPC is used as effective way of compressing and storing speech data as well as used for diphone comparison and speech modifications.

LPC is used also for synthesis of the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter (which represents the vocal tube), and run the source through the filter, resulting in speech.

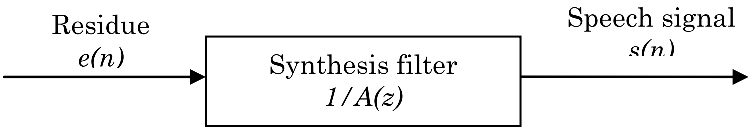

**Fig. 2 LPC synthesis model**

### 4.6. Diphone selection

In traditional approach only one recorded speech unit for each diphone is stored in the diphone synthesis system. But for quality improvements in Latvian TTS it was decided to have several different utterances for a number of diphones. During speech synthesis the most appropriate diphones are selected from the possible variations.

Selection of a pair of adjacent diphones is based on *connection-cost* algorithm. To evaluate compatibility of diphones their LPC coefficients are compared. Comparison is made at the borders of diphones to be concatenated, e.g., end of the first diphone and beginning of the second diphone. The closer are LPC coefficients of two diphones the less *cost* is assigned to the connection of these diphones.

In the ideal case if two diphones were adjacent in the recorded text used for speech unit data preparation, then LPC coefficients at their adjacent borders should be very close and connection-cost should be the least possible.

To map phones in phonetical transcription to the most appropriate diphones the *least expensive path* is determined using dynamic programming method.

### 4.7. Prosodic modifications of generated speech

The purpose of prosodic modification is to change the amplitude, duration and pitch of a speech segment. Amplitude adjustment is carried out by directly amplifying the speech waveform. For adjustment of pitch and duration of speech segments (compression and expansion) a version of LP-PSOLA is used – a method for manipulating the pitch and duration of an acoustic speech signal. In Latvian TTS residual signal and LPC coefficients of concatenated diphones are used as entry for LP-PSOLA modifications. Signal is divided into separate but overlapping smaller signals or windowing segments. These segments are being either multiplied or left out thus modifying the duration of the signal. Resulting segments are recombined through overlapping and adding to get modified speech signal.

## 5 Conclusions

Result of the project is the first Latvian TTS system currently in beta testing. Feedback from the first users is very positive. They characterize generated speech as natural-sounding, with correct Latvian pronunciation of majority of phrases and efficient work even on relatively old systems (200 MHz Pentium processor). Project demonstrates applicability of diphone synthesis in combination with such improvements as usage of multiple diphone variations and LP-PSOLA method for speech quality improvements.

Further work will concentrate on evaluation of resulting system, its integration into different application scenarios and detection of areas for further improvements.

## Acknowledgements

I would like to thank Karlis Goba for the excellent work he does developing Latvian TTS and for valuable consultations he gave me during preparation of this term paper.

# References

[1] Spoken Language Processing. A Guide to Theory, Algorithm, and System Development, Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Prentice Hall, 2001

[2] Speech and Language Processing, Daniel Jurafsky, James H. Martin, Prentice Hall, 2000

[3] Acoustic and auditory characteristics of the Latvian vowel system: synopsys of the doctoral thesis, Juris Grigorjevs, Rīga, 2005

[4] Data-Driven Text-to-Speech Synthesis, Edmilson Morais and Fábio Violaro, XXII SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, CAMPINAS, 2005

[5] Changing Pitch with PSOLA for Voice Conversion, Gina Upperman, http://cnx.org/content/m12474/latest/

[6] Fujisaki, H. and Ohno, S., Comparison and assessment of models in the study of fundamental frequency contours of speech. In ESCA workshop on Intonation:Theory Models and Applications, 1997.