TEXT-TO-SPEECH SYNTHESIS IN

SLOVENIAN LANGUAGE

Tomaž Šef, Aleš Dobnikar, Matjaž Gams

Department of Intelligent Systems - Jožef Stefan Institute – Ljubljana, Slovenia

e-mail: {tomaz.sef, ales.dobnikar, matjaz.gams}@ijs.

 abstract

This paper presents a text-to-speech (TTS) system, capable of synthesis of continuous Slovenian speech. The system is based on the concatenation of basic speech units, diphones, using TD-PSOLA technique improved with a variable length linear interpolation process. Input text is processed by a series of modules which are described in detail. A special attention is given to modeling the F0 contour, mainly based on the so-called superpositional approach.

This system is experimentally used in an employment agent EMA that provides employment information through the Internet.

# INTRODUCTION

The aim of text-to speech systems is to convert ordinary input text (consisting of words or sentences) to understandable speech output. Any text-to-speech system can be described as a two-level structure. The first level includes the linguistic and prosodic processing, sometimes called “high level” processing, where the text is transformed into a set of prosodic and phonetic parameters. The second level is the synthesizer itself, which converts this prosodic and phonetic parameters [1]. While the first level is very language dependent, the second level can use a variety of techniques and methods.

For the Slovenian language, several attempts were made in the past [2, 3, 4]. The system for uttering isolated words [2] applied the rule based approach but it did not produce satisfactory speech quality. In a new attempt, we tried another approach to unrestricted speech synthesis [4]. We developed a system for concatenating acoustical speech units, presented in Figure 1. The different phases of the synthesis task are performed by several sequentially operating modules.

# TEXT NORMALIZATION AND GRAPHEME-TO-PHONEME CONVERSION

Unlimited input text is translated into a series of phonemes in two consecutive steps (the first two levels in Figure 1). The text normalizer first converts files in different formats into an ASCII file. Then, abbreviations are expanded to equivalent full words using a list of lexical entries. Special formats, like numbers or dates, are converted into standard graphemic strings. The rest of the text is segmented into individual words and basic punctuation marks.

Figure 1: Slovene text-to-speech system.

# PROSODY GENERATION

Prosody has great impact on intelligibility and naturalness of speech perception. The proper choice of prosodic parameters, given by phoneme duration and intonation contours, enables natural sounding high quality synthetic speech.

## Microprosodic Parameters

Microprosodic parameters determine local phoneme duration and pitch values within word. The Slovenian language is a pitch-accentuated language, meaning that pitch and stresses are strongly related [5]. Stress is marked by a pitch rise within a stressed syllable, followed by a fall, which depends on the syllable (baritone or ocsitone) and the accent (acute or circumflex).

The durations depend on the type of the phoneme, on whether the phoneme is stressed or unstressed and on the length of the whole word. Pitch depends on the surrounding phonemes and on the position in the word. Vowels and consonants were processed separately. Stress primary affects vowel duration, whereas syllable-final consonants have little stress variation.

## Segment Intonation for Slovene Language

In order to generate rules for our synthesis scheme, data was collected by analyzing the readings of ten speakers. All of them are native Slovene speakers, five males and five females. Eight of them are professional speakers on the Slovenian national radio. The largest part of the speech material consists of declarative sentences in short stories, monologues, containing sentences of various complexities and types, news, weather reports and commercial announcements. Other parts of the corpora are interrogative sentences with yes/no and wh-questions and imperative sentences. The first part of the corpora contains 500 declarative sentences, uttered by eight speakers, and the second part 100 questions and 30 imperative clauses uttered by 2 speakers [6].

The generation of intonation curves for various types of intonation consists of two main phases:

* segmentation of the text into intonation units, and
* definition of the F0 contour for specific intonation units.

In our model, the intonation unit is defined as any connected signal between two pauses, greater that 40 msec. Depending on ortographic delimiters, four phrase boundaries were introduced:

* at prefaces, between paragraphs, etc.,
* at places of rhythmical division of some clauses or prosodic phrases (before the Slovene grammatical words *in*, *ter* (and), *pa* (but), *ali* (or), ...),
* at the end of clauses, marked with a period, exclamation mark, question mark or ellipsis,
* at places of prosodic phrases inside clauses, marked with a comma, semicolon, colon, dash or parentheses.

The F0 contour is defined with the function (Figure 2) composed from [7]:

* *a global component*, related to the whole intonation unit, and
* *local components*, related to accented syllables or syntactic boundaries.



Figure 2: F0 contour based on the so-called superpositional approach.

The global component gives the baseline for the F0 contour for the whole intonation unit and often rises in the beginning of the intonation unit and slightly decreases towards the end. It depends on the type of the intonation unit (declarative, imperative, yes/no or wh-question), position of the intonation unit in a complex sentence with two or more intonation units and duration of the whole intonation unit. The local components present local movements of the F0 shape at accented syllables or syntactic boundaries. Syntactic boundaries with local ascent often indicate the final F0 shape at various types of intonation units.

Many functions were tested (linear, power, transfer, decay, exponential) for the best approximation of the natural F0 contour. In the system presented, an exponential function for the global component *G(t)* was adopted and a cosinusoidal function for accents and final boundary contours *Li(t)*. The F0 contour is thus defined by the following equation:

  (1)

where *G(t)* and *Li(t)* are defined as:

  (2)

  (3)

where the expression *(Tpi - t)* must be in the range (*-di, di*), otherwise *Li (t)* = 0.



Figure 3: Example result of F0 contour modeling for the Slovene Wh-question “Kje je hodil toliko časa?” Engl.: “Where did he walk so long?”.

The symbols in these equations denote:

* *Fk*: asymptotic final value of F0 in the intonation unit
* *Az*: parameter for the onset F0 value in the intonation unit
* *α*: parameter for F0 shape control
* *Tpi*: time of i-th accent
* *Api*: magnitude of i-th accent
* *di*: shape duration of i-th accent.

The parameters *Fk*, *Az*, *α* and *Api* change during the synthesis process according to the analysis results of the F0 contour. The parameter *di* models the microprosodic duration of accented syllables.

Figure 3 illustrates the results obtained. The sentence for comparison is uttered by a female speaker. The parameters for the synthesized F0 are the same for the whole sentence. The panel displays the original F0 contour modeled with the INTSINT (**IN**ternational **T**ranscription **S**ystem for **INT**onation) system [8], indicated by squares, and the synthesized F0 contour, generated with the equations presented, indicated by circles. The prosody module corresponds to the third module in Figure 1.

# SEGMENTAL CONCATENATION

Once the appropriate phonetic symbols and prosody markers are determined, the final step is to produce audible speech by assembling elemental speech units, computing pitch and duration contours, and synthesizing the speech waveform (the last module in Figure 1).

## Slovenian Diphone Inventory

In concatenation systems, both the choice and the proper segmentation of the units to be concatenated play an important key role. Acoustic differences between stored and requested segments, as well as acoustic discontinuities at the boundaries between adjacent segments have to be minimized. Diphones are the most commonly adopted compromise between the size of the unit inventory, the complexity of the concatenation rules and the resulting quality of synthetic speech. A diphone can be defined as a speech fragment, which runs roughly from halfway one phoneme to halfway the next phoneme. So the transition between two phones is preserved and does not need to be calculated.

An analysis of the Slovenian phonological system gives 8 vowel and 21 consonant phonemes. When adding allophonic variations for certain phonemes, we arrived at a total of 34 phones. One diphone for every allophone combination possible in a given language is required. Our diphone inventory contains 1224 pitch-labeled diphones. They were hand-segmented and hand-labeled in order to enable as good as possible coupling at concatenation points. In order to guarantee high synthesis quality, the diphones were recorded by a professional speaker and placed in the middle of logatoms, pronounced with a steady intonation.

## Synthesis Technique

We used a concatenative TD-PSOLA technique [9] improved with a variable length linear interpolation process. This algorithm enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications without considerably affecting thequality of synthesized speech. In contrast to the pure TD-PSOLA algorithm it also supports spectral interpolation.



Figure 4: The variable-length linear interpolation process.

# EMPLOYMENT AGENT EMA

Our text-to-speech system is used in an employment agent EMA (<http://www-ai.ijs.si/~ema/>). EMA is an intelligent agent for employment tasks on the Internet designed in cooperation with the National Employment Office. Ema's basic task is to help people searching for employment, jobs, employees, scholarships and any other form of employment related tasks. For example, 90% of all available jobs in Slovenia are presented through EMA. In addition, EMA can perform a variety of tasks, such as sending mails whenever any new information on related Internet files/sites occurs. In the last year, EMA was visited by around one third of Slovenian population with access to the Internet.

# CONCLUSION

We developed a new complete text-to-speech system for the Slovenian language based on the acoustical units concatenation. The system is capable of synthesizing continuous Slovenian speech from an arbitrary input text.

The paper also describes an attempt to model the F0 contour for Slovene intonation units by rules, generated through the analysis of a large set of utterances.

The experiment showed that the system is an appropriate tool for generating audiable speech from text in the Slovenian language (the word intelligibility rate was around 90%).

References

1. C. Sorin. Towards High-Quality Multilingual Text-to-Speech. Proc. CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology. München. 1994.
2. S. Weilguny. Grafemsko-fonemski modul za sintezo izoliranih besed za sintezo slovenskega jezika. MSc Thesis. Faculty of electrical engineering and computer science. University of Ljubljana. 1993. In Slovenian.
3. J. Gros, N. Pavešić, S. Dobrišek, M.Erpič, B. Gorenc, A. Rakar, T.Šef, V. Vračar, F. Mihelič. Sistem za sintezo slovenskega govora. Proc. ERK’95-Slovenia Section IEEE Conference, Portorož, Slovenia, 1995, pp. 265-268. In Slovenian.
4. A. Dobnikar, J. Bakran. A new approach for Slovene text-to-speech synthesis. In Proc. MIPRO’95. pp. 265-268. Opatija. Croatia. 1995.
5. J. Toporišič. Slovenska slovnica. Založba Obzorja. Maribor. 1984. In Slovenian.
6. A. Dobnikar. Določevanje stavčne intonacije pri sintezi slovenskega govora. PhD Thesis. Faculty of Computer and Information Science. University of Ljubljana. 1997. In Slovenian.
7. H. Fujisaki, S. Ohno. Analysis and Modeling of Fundemental Frequency Contour of English Utterances. Proc. EUROSPEECH’95. Vol. 2. pp. 634-637. Philadelphia. 1996.
8. D. J. Hirst. Prosodic labelling tools. MULTEXT LRE Project 62-050 Report. Centre National de la Recherche Scientifique, Université de Provance, Aix-en-Provence. 1994.
9. E. Moulines, F. Charpentier. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication 9. pp. 453-467. 1990.
10. T. Dutoit, H. Leich. MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database. Speech Communication 13. pp. 435-440. 1993.