A TEXT-TO-SPEECH SYSTEM FOR THE SLOVENIAN LANGUAGE*

Jerneja Gros, Nikola Pavešić, France Mihelič
Faculty of Electrical Engineering
Tržaška c. 25, 1001 Ljubljana, Slovenia
Tel: +386 61 1768316; fax: +386 61 1264630
e-mail: jerneja.gros@fer.uni-lj.si

ABSTRACT

A text-to-speech (TTS) system, capable of synthesising continuous Slovenian speech from an arbitrary input text is described. The TTS system is based on the concatenation of basic speech units, diphones, using the TD-PSOLA technique, and no special hardware is required. The input text is transformed into its spoken equivalent by a series of modules. These modules, constituting the TTS system are described in detail. Finally, the quality of synthesised speech is assessed in terms of acceptability and intelligibility.

1 INTRODUCTION

Text-to-speech synthesis (TTS) enables automatic conversion of any available textual information into its *spoken* form. For the Slovenian language, several attempts were made in the past, where different aspects of a Slovenian TTS system were covered [1, 2, 3]. Nevertheless, none of them succeeded in building a complete system, providing high quality synthesised speech. In the Laboratory of Artificial Perception, we started on text-to-speech synthesis one year ago [4]. Here we describe the current version of our Slovenian TTS system, which is to serve as a reference system for future improvements.

The different phases of the synthesis task are performed by several modules, operating sequentially, as shown in Figure 1. A grapheme-to-phoneme module produces strings of phonemic symbols based on information in the written text. The problems it addresses are thus typically language-dependent. So is the prosodic generator, which assigns pitch and duration values to individual phonemes. Final speech synthesis is performed by TD-PSOLA concatenative diphone technique [5]. At the end of the paper, assessment results of the TTS system are given and discussed and some promising directions for future work are mentioned.

2 GRAPHEME-TO-PHONEME CONVERSION

Unlimited input text is stored in ASCII file format. It is translated into a *series of phonemes* (or allophones, in case different phoneme variations are differentiated) in two consecutive steps. An analysis of the Slovenian phonological system

*This work was partly funded by the Commission of the European Community under COP-94 contract No 01634 (SQEL)

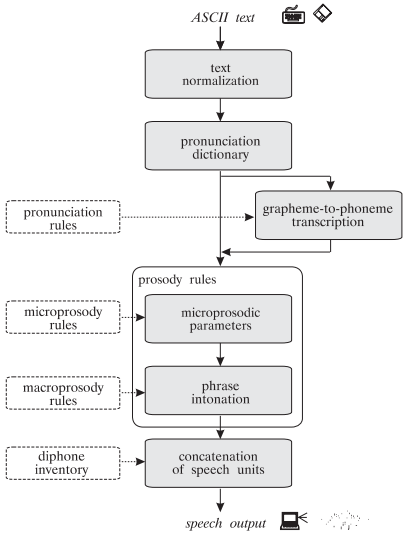

Figure 1: *Slovenian text-to-speech system architecture.*

gives 8 vowel and 21 consonant phonemes. When adding allophonic variations for certain phonemes, we arrived at a total of 34 phonemes. Initially, abbreviations are expanded to form equivalent full words using a special list of lexical entries. A text normaliser converts further special formats, like numbers or dates, into standard grapheme strings. The rest of the text is segmented into individual words and basic punctuation marks.

Next, word pronunciation is derived, based on an user extensible pronunciation dictionary and letter-to-sound rules. The dictionary covers the most frequent words in a given language and a second dictionary helps with pronouncing proper names.

In case where dictionary derivation fails, words are transcribed as a human would pronounce them upon encountering a new unknown word, using automatic lexical stress assignment and letter-to-sound rules. However, as lexical stress in the Slovenian language can be located almost arbitrarily on any syllable, errors are introduced into pronunciation of such unknown words. Nevertheless, there do exist some rules of stress assignment, based upon observations of linguists [6],

which to a certain extent determine stress position within a word. Once lexical stress has been established, a set of context-dependent letter-to-sound rules translates each word into a series of phonemes. A basic semantic analysis is also included.

3 PROSODY GENERATION

A number of studies suggest that prosody has great impact on intelligibility and naturalness of speech perception. Only the proper choice of prosodic parameters, given by phoneme duration and intonation contours, enables us to produce natural-sounding high quality synthetic speech.

3.1 Duration Modelling

First, phonemically transcribed words are syllabified by counting the number of their vowel clusters and *duration of syllables* is modelled according to the speaker's normal articulation rate, depending on the number of syllables within a word and on the word's position within a phrase.

In order to provide the synthesiser with the possibility to pronounce input text with several speaking rates, tests were made to study the impact of speaking rate on syllable duration and duration of individual phonemes and phoneme groups.

We opted for a relatively long text, read three times: at a normal, fast and slow rate. Reading the text took 7 minutes 32 seconds when reading at a normal rate, 12 minutes 55 seconds reading very slow and 5 minutes 45 seconds when reading as fast as possible.

The effect of speaking rate on phoneme duration was studied in a number of ways. An extensive statistical analysis of lengthening and shortenings of individual phonemes, phoneme groups (wovels, nasals, liquids, plosives, fricatives, etc) and phoneme components (closures, bursts) was performed. As a result of our study, different inherent values for initial phoneme duration when calculating microprosodic parameters were determined with respect to a chosen speaking rate.

As another part of the study, articulation rate, expressed as the number of syllables or phones per second, excluding silences and pauses, was determined for all three speaking rates. Figure 2 shows articulation rate in number of syllables per second plotted as a function of word length in number of syllables and the word position in a sentence. The obtained values apply for normal speaking rate. In general, lexical redundancy increases with word length. As expected, we can find evidence of polysyllabic shortening: average syllable duration tends to decrease with more syllables in a word.

3.2 Microprosodic Parameters

Microprosodic parameters determine local *phoneme duration and pitch values* within a word. The Slovenian language is a pitch accentuated language, meaning that pitch and stress are strongly related [6]. 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). Microprosodic parameters were determined for every phoneme with respect to accent position within a word, its

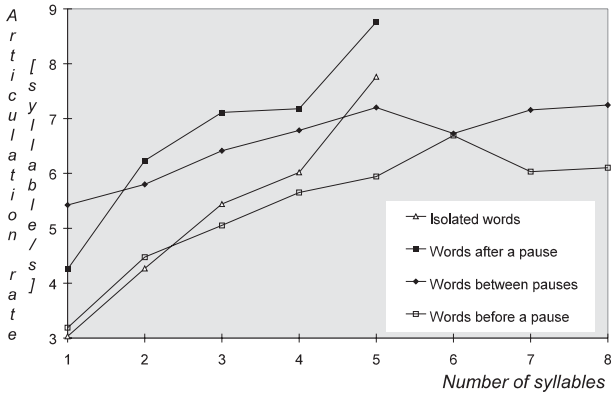

Figure 2: *Articulation rate, expressed in syllables per second, is given for four different types of word positions.*

type, syllable position and its type and syllable duration, previously determined by the articulation rate. Experimentally obtained results on measuring microprosodic parameters for a given speaker were used as initial values. Vowels and consonants were processed separately, vowels being responsible for major audible pitch variations within a word. Stress primarily affects vowel duration, whereas syllable-final consonants have little stress variation.

3.3 Macroprosodic Parameters

Sentential intonation recognition is difficult due to problems with the F_0 contour itself. The contour is not continuous as there is no fundamental frequency in unvoiced regions. F_0 is also dependent on segmental effects - before and after stops the contour can deviate sharply. F_0 tracking is difficult and prone to errors and even when successful, there is a considerable amount of short-term variations in the F_0 contour due to pitch perturbations. Therefore, smoothing and linearisation of the F_0 contour are performed.

We used a relatively simple approach for prosody parsing and automatic prediction of Slovenian intonational prosody which makes no use of syntactic or semantic processing [7], but rather uses punctuation marks and searches for grammatical words.

A lexicon of grammatical words was chosen so that they have a relatively stable F_0 contour, that acts as a trampling before the higher initial pitch of the following lexical word (e.g. conjunction or complementizer). Once grammatical words are determined, sentences can be parsed into prosodic groups, i.e. segments between punctuation marks or grammatical words.

A set of measurements was made in order to define four typical intonation contours for the four Slovenian basic intonation types [8]. Read newspaper articles were processed by an AMDF (*A*verage *M*agnitude *D*ifference *F*unction) pitch extractor [9], thus mimicking the *reading* reading type of prosody. Then, the manual linearisation of F_0 curves into pitch contours was performed. The obtained pitch contours

were a rather rough approximation of stylised pitch contours, as suggested by IPO [10]. The prosodic head plays an important role in accentuation. The intonation head can begin with an interrogative pronoun, otherwise it is located on the last stressed syllable within a prosodic group.

Finally, the length and relative location of prosodic groups determine the insertion of pauses according to the type of grammatical categories. The basic idea is to introduce pauses after long prosodic groups in order to simulate breathing pauses and to reduce the mental load of the listener. However, the location of such pauses should be prosodically plausible.

The drawbacks of such a syntactically independent prosodic parser are important, as in many cases prosodic parameters are determined by the syntactic structure of a phrase and cannot be reliably estimated without a deep syntactic analysis. Therefore, a more sophisticated intonation model needs to be developed.

4 DIPHONE CONCATENATION

Once appropriate phonetic symbols and prosody markers are determined, the final step within a TTS system is to produce audible speech. This is achieved by assembling elemental speech units by taking into account computed pitch and duration contours, and synthesising a speech waveform.

A concatenative synthesis technique was used. The TD-PSOLA scheme enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications [5] without considerably affecting the quality of synthesised speech [1].

Diphones were chosen for concatenative speech units. A diphone can be defined as a speech fragment which runs roughly from half-way one phoneme to half-way the next phoneme. In this way the transition between two consecutive speech sounds is encapsulated in the diphone and needs not be calculated.

4.1 Preparation of the Diphone Inventory

A diphone inventory requires one diphone for every possible phoneme combination in a given language. A Slovenian diphone inventory comprising 955 pitch-labelled diphones was recorded, hand-segmented and hand-labelled in order to provide optimal coupling at concatenation points [11]. The design and recording of the diphone inventory were given special attention. The target diphones were always found within unaccented nonsense words, called logatoms, pronounced with a steady intonation. The speaking rate was relatively slow, opting for a high intelligibility of the TTS system. Speech signals were recorded by a close talking microphone using a sampling rate of 16 kHz and 16 bit linear A/D conversion.

After the recording phase, logatoms were hand-segmented and the center of the transition between the phonemes was marked, using information from both temporal and spectral representation of the speech signal. Finally, pitch markers were manually set for voiced parts of the corresponding speech signal. Figure 3 gives an example of diphone *am* along with its spectrum.

[1]Samples of synthetic speech produced by our TTS system are available on the WWW on the address"`http://luz.fer.uni-lj.si/-english/SQEL/synthesis-eng.html`".

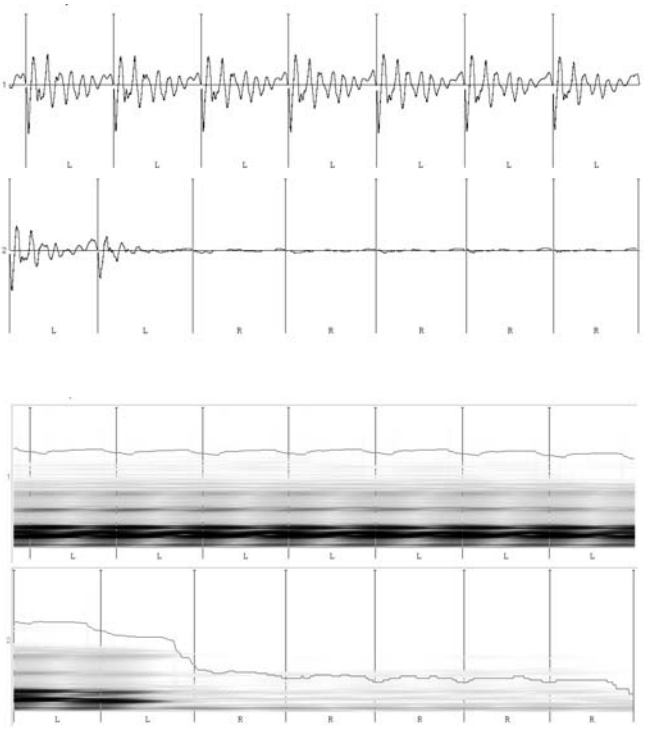

Figure 3: *Waveform* (above) *and spectral* (below) *representation of the diphone* am. *Markers* L *and* R *are set at the pitch periods of the left part of the diphone and of the right part, respectively.*

5 PERFORMANCE ASSESSMENT

Here we give some results of the Slovenian TTS system performance assessment.

The adequacy of the system was tested in two ways: in terms of acceptability and in terms of intelligibility [12]. The synthesis output was directed to a Sound Blaster audio card. The experiment was performed in laboratory conditions with 10 subjects within the age span between 24 and 43 years, three of them being female.

In our first experiment, intelligibility of synthesised speech was evaluated on three levels: segmental level, word level and phrase level. Subjects, participating in the test were asked to write down everything they heard. Figure 4 gives the percentage of correctly understood syllables and words, with word intelligibility rate being close to 80%. On phrase level, different types of texts (daily newspaper, fiction, scientific paper, poems) were chosen and the ratio of word insertions, deletions and substitutions with respect to the number of all words within the text was computed.

In our second experiment, degree of acceptability of the synthesised speech was assessed, again on word and phrase

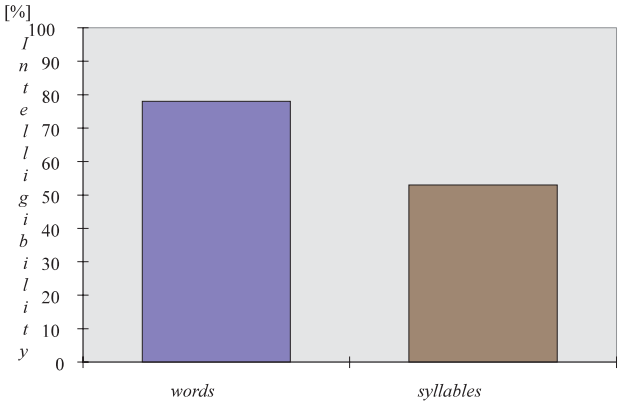

Figure 4: *Intelligibility test. Percentage of correctly understood syllables and words.*

level. Subjects were asked to mark naturalness of pronunciation from 1 to 10, with 10 being the highest mark. The results obtained are shown in Figure 5.

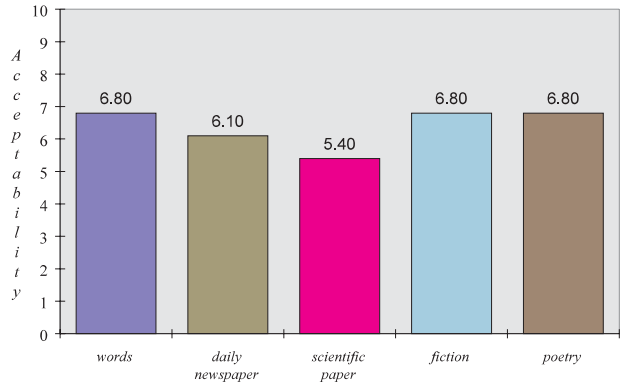

Figure 5: *Degree of naturalness of pronunciation for individual words and four different text styles.*

Despite a rather good intelligibility of synthetic speech, utterances sometimes suffer a lack from naturalness and fluency.

The major part of the subjects estimated the synthetic speech to be pleasant and quite natural sounding, sufficiently rapid and not over-articulated. All ten of them considered the system to be an appropriate tool for generating audible speech from text in the Slovenian language.

6 CONCLUSION

The described speech synthesis system is the first complete TTS system for the Slovenian language. The synthetic speech produced by the system is intelligible, but lacks more naturalness. Improvement of intelligibility and naturalness depend in particular on proper lexical stress assignment and a more sophisticated generation of prosodic parameters.

Preparation of the diphone inventory was rather laborious and time-consuming, since the whole process of extracting diphones from logatoms was done manually. An automatic procedure for segmenting and pitch labelling of diphones should result in considerable reduction in preparation time of a new diphone inventory. It also provides a powerful tool for including new synthetic voices and for updating and supplementing existing diphone libraries.

The first attempts at developing a diphone-based synthesis system for the Slovenian language are promising, so that further work on improving individual parts of the system is encouraged.

Acknowledgement

The authors wish to thank Tomaž Erjavec for proof-reading of the text and his useful comments on the article.

References

[1] J. Hribar. Sinteza umetnega govora iz teksta. *MSc Thesis*, Faculty of Electrical Engineering and Computer Science. University of Ljubljana. 1984. In Slovenian.

[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] A. Dobnikar, J. Bakran. A new approach for Slovene text-to-speech synthesis. In *Proc. MIPRO95*. pp. 265–268. Opatija. Croatia.

[4] J. Gros, N. Pavešić, F. Mihelič, S. Dobrišek. Slovenian text-to-speech synthesis. In *Proc. 3rd Slovenian-German and 2nd SDRV Workshop on Speech and Image Understanding*. Ljubljana. Apr. 1996.

[5] E. Moulines, F. Charpentier. Pitch - Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication* 9. pp. 453–467. 1990.

[6] J. Toporišič. *Slovenska slovnica*, Založba Obzorja. Maribor. 1984. In Slovenian.

[7] C. Sorin, D. Laurreur, R. Llorca. A Rhythm-Based Prosodic Parser for Text-to-Speech Systems in French. *Proc. XIth ICPhS*. pp. 125–128. Tallin. Estonia. 1987.

[8] J. Toporišič. Slovenska stavčna intonacija, *V. seminar slovenskega jezika, literature in kulture*. 1969. In Slovenian.

[9] M. J. Ross, H. L. Schaffer, A. Cohen, R. Freudberg, H. J. Manley. Average Magnitude Difference Function Pitch Extractor. *IEEE Trans. ASSP-25*, str. 565–571, 1977.

[10] R. Collier. On the perceptional analysis of intonation. *Speech Communication 9*. North-Holland. pp. 443–451. 1990.

[11] J. Gros, I. Ipšić, S. Dobrišek, F. Mihelič, N. Pavešić. Segmentation and labelling of Slovenian diphone inventories. COLING96. Denmark. 1996. accepted for presentation.

[12] L.C.W. Pols. Synthesis Performance Assessment. *Proc. CRIM / FORWISS Workshop on Progress and Prospects of Speech Research and Technology*. München. 1994.