6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

S5: THE SQEL SLOVENE SPEECH SYNTHESIS SYSTEM

N. Pave iæ, J. Gros
Artificial Perception Laboratory
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: jerneja.gros@fe.uni-lj.si

ABSTRACT

An improved version of the Slovene text-to-speech system S5 is described. S5 can be used either as a stand-alone reading system or it can be integrated into other applications.

S5 is based on concatenation of basic speech units, diphones, using the TD-PSOLA technique. The input text is transformed into its spoken equivalent by a series of modules. F_0 modeling is based primarily on predicting the appropriate tonemic accent. Phone duration is predicted by a two level approach, taking into account how acceleration or slowing down applies to the duration of individual phones.

The adequacy of the spoken output was evaluated by several subjective tests as they are recommended by the International Telecommunication Union (ITU).

1. INTRODUCTION

Text-to-speech synthesis (TTS) enables automatic conversion of any available textual information into its spoken form.

In the Laboratory of Artificial Perception at the University of Ljubljana, we started on text-to-speech synthesis in 1994. In the following year we presented the first PC-based TTS system for the Slovene language [1]. We used it as a reference system for further improvements. In 1998 another diphone based Slovene TTS system developed at the Jožef Stefan Institute in Ljubljana was presented [2].

In the recent version of our TTS system S5 we implemented a novel procedure for determination of prosodic parameters, added a pronunciation dictionary and improved text normalization [3].

The S5 TTS system has already been implemented in different applications:

- S5 as a stand-alone TTS system (reading machine),
- S5 providing the spoken output in the SQEL speech recognition and dialog system for automatic airline timetable retrieval,
- S5 integrated into a special application - HOMER, a reading system for the blind and partially sighted people, combined with optical character recognition devices and voice control,
- a TTS web server at http://sinteza.fe.uni-lj.si (under construction).

The input text is transformed into its spoken equivalent by a series of modules (Figure 1), which we describe in detail. A grapheme-to-phoneme or -to-allophone module produces strings of phonetic symbols based on information in the written text. The problems it addresses are thus typically language-dependent. A prosodic generator assigns pitch and duration values to individual phones. Final speech synthesis is based on diphone concatenation using TD-PSOLA [4].

The quality of the synthesized speech was assessed in terms of intelligibility and naturalness of pronunciation. Additionally, various aspects of the synthetic speech production process were tested. The assessment results of the TTS system are given and discussed and some promising directions for future work are mentioned.

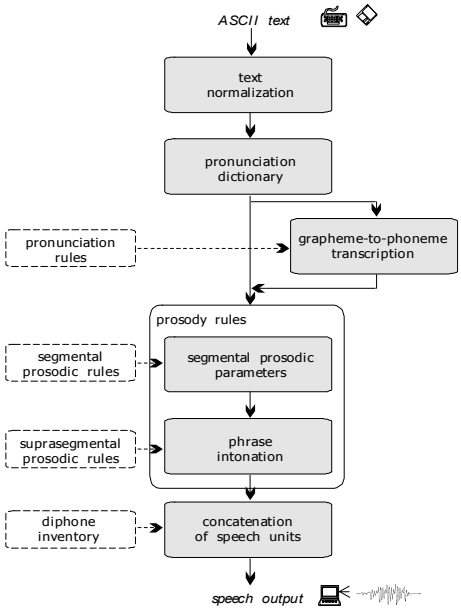

Figure1: The architecture of the S5 TTS system.

2. GRAPHEME-TO-ALLOPHONE TRANSCRIPTION

Input to the S5 system is free text. For the time being, input text should be stored in ASCII format; currently we are expanding the input possibilities so that in future it may come from other programs or marked regions on the computer screen.

Input text is translated into a series of allophones in two consecutive steps. First, input text normalization is performed. Abbreviations are expanded to form equivalent full words using a special list of lexical entries. The text normalizer 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 a user - extensible pronunciation dictionary and letter-to-sound rules. The dictionary covers over 16.000 most frequent inflected word forms.

In case where dictionary derivation fails, words are transcribed using automatic lexical stress assignment and letter-to-sound rules. However, as lexical stress in Slovene can be located almost arbitrarily on any syllable, this step can introduce errors into the pronunciation of words.

Automatic stress assignment is to a large extent determined by (un)stressable affixes, prefixes and suffixes of morphs, based upon observations of linguists [5].

For words which do not belong to these categories, the most probably stressed syllable is predicted using the results obtained by a statistical analysis of stress position depending on the number of syllables within a word.

Finally, a set of over 150 context-dependent letter-to-sound rules translate each word into a series of allophones.

3. PROSODY GENERATION

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

Prosody generation in S5 consists of four phases:

- intrinsic duration assignment,
- extrinsic duration assignment,
- modelling of the intra word F_0 contour and
- assignment of a global intonation contour.

The first and the third phase are sometimes referred to as microprosodic parameter assignment, since they are performed on speech units smaller than a word. The second and the fourth phase are also called macroprosodic parameters determination, since they operate above the word level.

3.1 Prosody Measurements

A speech database consisting of isolated words, carefully chosen by phoneticians [6], was recorded in order to study different effects on phone duration and fundamental frequency, which operate on the segmental basis. Vowel duration and f_0 were studied in different types of syllables: stressed/unstressed, open/closed. Consonant duration was measured in CC and VCV clusters [7].

Another large continuous speech database was recorded to study the impact of speaking rate on syllable duration and duration of phones [8]. A male speaker was instructed to pronounce the same material at different speaking rates: at a normal, fast and slow rate. Thus context, stress and all other factors were kept identical to every realisation of the sentence. As a result, pair-wise comparisons of phone duration could be made.

The effect of speaking rate on phone duration was studied in a number of ways. An extensive statistical analysis of lengthening and shortening of individual phones, phone groups and phone components, like closures or bursts was performed, the first of the kind for the Slovenian language [7].

Pair-wise comparisons of phone duration were calculated. Average mean duration differences and standard deviations were computed for pairs of phones pronounced at different speaking rates. Prior to the comparison, phone duration was normalized to the corresponding normal phone duration. Pairs were first composed of normal and slow rate phones, and later of fast and normal rate phones.

The closures of plosives change but slightly and maintain almost the same duration regardless of the speaking rate. Short vowels, contrary to long vowels, increase more in duration when speaking slower than they do shorten when speaking faster. From these observations we may draw a conclusion: phones or phone components, which are considered as short by nature, except for plosive bursts, increase more in length at a slow rate than they do shorten at a fast rate. The opposite holds for affricates and long vowels.

Articulation rate expressed as the number of syllables or phones per second, excluding silences and filled pauses [9], was studied for the different speaking rates. In other studies, articulation rate is usually determined for

speech units with the length of individual words or entire phrases. We studied the articulation rate of words along with their associated cliticised words at different positions within a phrase: isolated, phrase initial, phrase final and nested within the phrase.

The articulation rate increases with longer words, as average syllable duration tends to decrease with more syllables in a word. The articulation rate immediately after pauses is higher than the one prior to pauses.

A set of measurements was made in order to define four typical intonation contours based on four Slovenian basic intonation types [10]. Read newspaper articles were processed by an AMDF pitch extractor.

Then, a manual piecewise linearization of F_0 curves into pitch contours was performed. Our interest was to detect typical prosodic segments by means of F_0 contours.

3.2 Duration Modeling

Regardless of whether the duration units are words, syllables or phonetic segments, contextual effects on duration are complex and involve multiple factors.

Similarly to [11], our two-level duration model first determines the words' intrinsic duration, taking into account factors relating to the phone segmental duration, such as: segmental identity, phone context, syllabic stress and syllable type: open or closed syllable.

Further, the extrinsic duration of a word is predicted, according to higher-level rhythmic and structural constraints of a phrase, operating on a syllable level and above. Here the following factors are considered: the chosen speaking rate, the number of syllables within a word and the word's position within a phrase, which can be isolated, phrase initial, phrase final or nested within the phrase.

Finally, intrinsic segment duration is modified, so that the entire word acquires its predetermined extrinsic duration. It is to be noted that stretching and squeezing does not apply to all segments equally. Stop consonants, for example, are much less subject to temporal modification than other types of segments, such as vowels or fricatives.

Therefore, a method for segment duration prediction was developed, which adapts a word with an intrinsic duration t_i to the determined extrinsic duration t_e, taking into account how stretching and squeezing apply to the duration of individual segments [8].

The reliability of our two-level prediction method was evaluated on a speech corpus consisting of over 150 sentences. The predicted durations were compared to those in the same position in natural speech. Natural duration variation was evaluated by averaging the duration differences for words, which occurred in the corpus several times, in the same phonetic environment and in the same type of phrase.

Standard deviation of the difference between natural and predicted duration difference is 15.4 ms for normal speaking rate, and even less for stressed phonemes the duration of which is of crucial importance to the perception of naturalness of synthetic speech.

3.3 F_0 Modeling

Since the Slovenian language has been defined as a pitch accent language [6], special attention was paid to the prediction of tonemic accents for individual words.

First initial vowel fundamental frequencies were determined according to previous measurements as suggested by [6], creating the F_0 backbone. Each stressed word was assigned one of the two tonemic accents, characteristic for the Slovenian language. The acute accent is mostly realized by a rise on the posttonic syllable, while with the circumflex the tonal peak usually occurs within the tonic. Five typical F_0 patterns were chosen from the variety of F_0 patterns described in [6]. Finally a linear interpolation between the defined F_0 values was performed.

We used a relatively simple approach for prosody parsing and the automatic prediction of Slovenian intonational prosody which makes no use of syntactic or semantic processing [12], but rather uses punctuation marks and searches for grammatical words, mainly conjunctions which introduce pauses. We considered it more important to predict the word F_0 contour modeling the tonemic accent as reliably as possible than to explore sentence intonation.

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 or even semantic analysis.

4. DIPHONE CONCATENATION

Once appropriate phonetic symbols and prosody markers are determined, the final step within S5 is to produce audible speech by assembling elemental speech units. This is achieved by taking into account computed pitch and duration contours, and synthesizing 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 [4] without

considerably affecting the quality of the synthesized speech.

Diphones were chosen for concatenative speech units as a compromise between the size of the unit inventory, the complexity of the concatenation rules and the resulting speech quality.

5. EVALUATION

The adequacy of the S5 system was evaluated in terms of acceptability and in terms of intelligibility. The experiment was performed in laboratory conditions with 21 subjects within the age span between 19 and 45 years, ten of them being female. It was conceived according various ITU-T Recommendations, describing methods for subjective performance assessment of the quality of speech voice output devices.

The test was divided into three parts [13]. The first part was to evaluate whether the quality of the synthetic speech was sufficiently high for a real application of the system in an automatic information retrieval system. The subjects were asked to fill in different templates related to the chosen application domain based on the information they heard.

The second part of the test served to compare several features describing the synthetic voice quality to those describing the quality of natural speech distorted with different levels of gaussian noise. The synthetic speech received a mean opinion score, which was between distorted natural speech with a SNR ratio of 5dB and 10dB. Slovene EUROM1 texts pronounced by a male professional radio announcer, acquired in scope of the MULTEXT-EAST Copernicus project, were used as reference speech [14]. Hereby we encourage other groups working on TTS in Slovenia to use the same reference speech when evaluating the quality of their TTS systems.

In the third part of the test, different methods for prosody assignment were evaluated. The major part of the subjects estimated the synthetic speech produced by S5 to be pleasant and quite natural sounding, sufficiently rapid and not over-articulated.

6. CONCLUSION

S5 is a text-to-speech system for the Slovene language, capable of synthesizing intelligible continuous speech from an arbitrary Slovene input text. Further improvement of intelligibility and naturalness depend in particular on proper lexical stress assignment and a more sophisticated generation of f_0 prosodic parameters.

REFERENCES

[1] Gros, J., Pavešiæ, N. and Miheliè, F. (1996), A text-to-speech system for the Slovenian language. *Proceedings of the EUSIPCO'96*, Trieste, pp. 1043-1046

[2] J. Gros, N. Pavešiæ, F. Miheliè, Text-to-speech synthesis: A complete system for the Slovenian language, *Journal of Computing and Information Technology*, Vol. CIT-5, No. 1, pp. 11-19, 1997.

[3] T. Šef, A. Dobnikar, M. Gams, M. Grobelnik, Slovenski govor na internetu, *Proceedings of the Conference Language Technologies for the Slovene Language ISJT 98*, pp. 60-64, Ljubljana, Slovenia, 1998.

[4] J. Gros, *Samodejno pretvarjanje besedil v govor*, PhD Thesis, University of Ljubljana, 1997.

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

[6] J. Toporiši è, *Slovenska slovnica*, Založba Obzorja, Maribor, 1984.

[7] T. Srebot-Rejec, *Word Accent and Vowel Duration in Standard Slovene: An Acoustic and Linguistic Investigation*, Slawistische Beiträge, Band 226, Verlag Otto Sagner, München, 1988.

[8] J. Gros, N. Pavešiæ, F. Miheliè, Syllable and segment duration at different speaking rates for the Slovenian language, *Proceedings of the EUROSPEECH'97*, Rhodes, Greece, 1997.

[9] J. Gros, N. Pavešiæ, F. Miheliè, Speech timing in Slovenian TTS, *Proceedings of the EUROSPEECH'97*, pp. 323-326, Rhodes, Greece, 1997.

[10] D. O'Shaughnessy, Timing patterns in fluent and disfluent spontaneous speech, *Proceedings ICASSP'95*, pp. 600-603, Detroit, USA, 1995.

[11] J. Toporiši è, Slovenska stavèna intonacija, *V. seminar slovenskega jezika, literature in kulture,* 1969.

[12] G. Epitropakis, D. Tambakas, N. Fakotakis, G. Kokkinakis, Duration modelling for the Greek language, *Proceedings of the EUROSPEECH'93*, pp. 1995-1998, Berlin, Germany, 1993.

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

[14] J. Gros, H. Miheliè, N. Pavešiæ, Speech Quality Evaluation in Slovenian TTS, *Proceedings of the LREC 98*, Granada, Spain, 1998.

[15] *East meets West: A Compendium of Multilingual Resources*, Eds. T. Erjavec, A. Lawson. L. Romary, CD-ROM, produced & distributed by TELRI, 1998.