Slovenian Text-to-Speech Synthesis for Speech User Interfaces

Jerneja Žganec Gros, Aleš Mihelič, Nikola Pavešić, Mario Žganec, Stanislav Gruden

***Abstract*—**The paper presents the design concept of a unit-selection text-to-speech synthesis system for the Slovenian language. Due to its modular and upgradable architecture, the system can be used in a variety of speech user interface applications, ranging from server carrier-grade voice portal applications, desktop user interfaces to specialized embedded devices**.**

Since memory and processing power requirements are important factors for a possible implementation in embedded devices, lexica and speech corpora need to be reduced. We describe a simple and efficient implementation of a greedy subset selection algorithm that extracts a compact subset of high coverage text sentences. The experiment on a reference text corpus showed that the subset selection algorithm produced a compact sentence subset with a small redundancy.

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

***Keywords*—**text-to-speech synthesis, prosody modeling, speech user interface.

I. Introduction

A VITAL part of speech technology applications in modern voice application platforms is a text-to-speech engine. Text-to-speech synthesis (TTS) enables automatic conversion into spoken form of any available textual information. With the evolution of small portable devices porting of high quality text-to-speech engines to embedded platforms has been made possible [1], [2].

The initial attempts towards Slovenian TTS were mainly based on concatenation of diphones. They resulted in a few demonstration systems [3]-[5] and some first carrier-grade voice applications [6].

For the new Slovenian TTS system a corpus-based concatenative approach was chosen since it yields close-to-natural sounding speech [7]-[9]. It is based on concatenation of basic speech units, derived from a large speech corpus, instead of diphones only. Due to its modular and upgradable architecture, the TTS system can be used in a variety of speech user interface applications, ranging from server carrier-grade voice portal applications, desktop user interfaces to specialized embedded devices, where special attention has been paid to the speech corpus compression and annotation techniques.

The input text is transformed into its spoken equivalent by a series of modules (Figure 1). 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 elemental speech unit concatenation. The modules are described in more detail in the following sections.

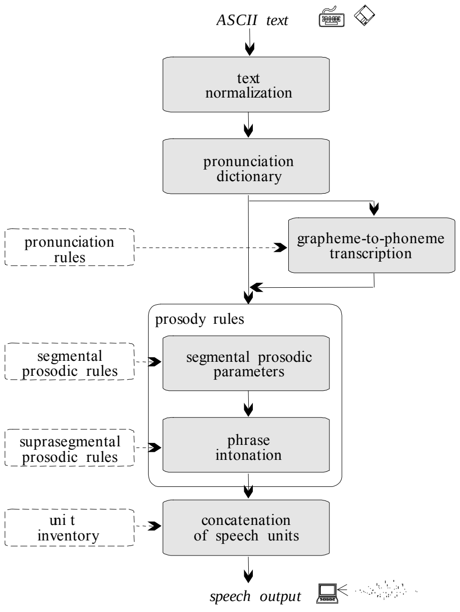

Fig. I TTS System Architecture.

Finally, quality of the synthesized speech was assessed in terms of intelligibility and naturalness of pronunciation. Various aspects of the synthetic speech production process

Manuscript received March 31, 2005. This work was partly supported by the Slovenian Research Agency, the Slovenian Ministry of Defense and the Slovenian Ministry of Information Society under grants M2-0019 and V2-0896.

Jerneja Žganec Gros was with the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. She is now with Alpineon RTD, Ljubljana, Slovenia (phone: +386 1 4239440; fax: +386 1 4239440; e-mail: jerneja.gros@alpineon.com).

Aleš Mihelič is with Alpineon RTD, Ljubljana, Slovenia (e-mail ales.mihelic@alpineon.com).

Nikola Pavešić is with the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. (e-mail: nikola.pavesic@fe.uni-lj.si).

Mario Žganec is with Alpineon RTD, Ljubljana, Slovenia (e-mail mario.zganec@alpineon.com).

Stanislav Gruden is with Alpineon RTD, Ljubljana, Slovenia (e-mail stanislav.gruden@alpineon.com).

were tested. The assessment results of the TTS system are given and discussed and some promising directions for future work are mentioned.

II. Grapheme-to-Allophone Transcription

Input to the TTS system is unrestricted text. It 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. We have built a dictionary that covers over 800.000 inflected word forms.

To reduce the memory footprint of the grapheme-to-allophone conversion module, we have compiled an exception dictionary that contains only the differences to the phonetic transcriptions obtained by applying the grapheme-to-allophone rule set. Similarly to [10], a compression factor of 10 was achieved, compared to the baseline full-lexicon representation, without sacrificing transcription accuracy.

In case where dictionary derivation fails, words are transcribed using automatic lexical stress assignment and letter-to-sound rules. However, as lexical stress in Slovenian 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. For words that do not belong to these categories the most probably stressed syllable is predicted using the results obtained by a statistical analysis of lexical stress position depending on the number of syllables within a word. Finally, a set of over 190 context-dependent letter-to-sound rules translate each word into a series of allophones.

III. Prosody Generation

Corpus-based prosody modeling yields high quality and close-to-natural sounding prosody parameter prediction; however, it requires a large amount of linguistic information to rely on. Our embedded TTS system uses a compact rule-based prediction method to determine the target prosodic parameters in four phases:

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

A speech database consisting of isolated words, carefully chosen by phoneticians 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.

Another large continuous speech database was recorded to study the impact of speaking rate on syllable duration and duration of phones. 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 realization 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.

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. Pairs were first composed of normal and slow rate phones, and later of fast and normal rate phones [11].

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, 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. Read newspaper articles were processed by a pitch extractor. A manual piecewise linearization of F_0 curves into pitch contours was performed and typical prosodic segments were detected.

A. Duration Modeling

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

A 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 [12].

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.

The reliability of our two-level prediction method was evaluated on a speech corpus consisting of over 200 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 (their duration is of crucial importance to the perception of naturalness of synthetic speech).

B. Pitch Modeling

Since the Slovenian language is as a pitch accent language, special attention was paid to the prediction of tonemic accents for individual words.

First initial vowel fundamental frequencies were determined according to the parameters obtained from prior prosody measurements, 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. Finally, a linear interpolation between the defined F_0 values was performed.

A simple approach for prosody parsing and the automatic prediction of Slovenian intonational prosody which uses punctuation marks and searches for grammatical words, mainly conjunctions which introduce pauses was applied.

IV. Speech Corpus Design

For unit-selection text-to-speech synthesis a speech corpus of recorded and annotated elemental speech units is required [13]. The quality of the output synthetic speech depends crucially on the quality of the speech corpus. The longer elemental speech units are used the better and more natural-sounding synthetic speech the TTS system can yield. However, with longer elemental speech units the corpus size increases dramatically. Therefore, a compromise between the size of the speech corpus and the quality of the resulting speech has to be taken [14] that is even more pronounced for embedded TTS.

If the corpus selection method is unbalanced or random, the recorded data may lack critical phone transitions and can be full of redundances. Various corpus reduction methods have been reported, from those optimizing and reducing the contents of the prerecorded and annotated speech corpora to those that try to compress the initial text corpus to be recorded [15]-[20]. Often sentence pair exchanges are calculated using the diphone and triphone entropies. In [17], the unit coverage is maximized using prosody information. In [18], a modified greedy algorithm is applied that maximizes the hit-rate and covering-rate for sentence selection criteria. A two-stage sentence recording script design presented in [20] takes into account the balance of acoustic speech parts to provide variations in short-time speech features, while the linguistic parts provide long-time speech features, like words or frequent word sequences.

We wanted to include the most frequent allophone sequences in a given language to be represented in the final sentence subset, therefore we have implemented a greedy algorithm, similar to the one described in [17] to reduce the initial text sentence set to a compact and efficient subset.

The process of designing a speech corpus for unit-selection TTS was divided into three phases:

– representative sentence set selection,

– recording of selected texts and

– segmentation and annotation of the recorded speech material.

A. Sentence Subset Selection

Initially, we collected a large corpus of texts covering various text styles, ranging from newspaper articles to novels. All sentences shorter than 5 words were discarded from further analysis. The remaining reference text corpus contained 200.000 different sentences (25 Mbyte of text in ASCII format).

The text corpus was processed by a grapheme-to-allophone converter from the TTS system in order to obtain an allophone transcription of the text corpus.

We used this corpus to perform a statistic analysis of frequent phone sequences of allophones, diphones, triphones and quadphones. It gave us an idea on how frequently certain phone combinations occur in spoken Slovenian language. Further, the analysis has shown that just are few triphones have frequent occurrences.

Therefore it makes sense to select just the most frequent triphones to be represented in the final speech corpus. We have opted for the first 500 triphones, that represent 1% of the complete triphone set but they cover almost 50% of all triphones in the transcribed reference text corpus. In a similar way 300 most frequent quadphones were selected.

To synthesize high-quality speech, the speech corpus was required to contain a wide variety of speech parts: from collocations and words to diphones and sub-phoneme parts.

With the most frequent triphones and quadphones selected we wanted to select an optimal compact subset of corpus sentences that cover all the chosen allophone sequences, including most frequent collocations and words in a given language.

A greedy sentence selection algorithm was designed for this purpose. Each sentence in the reference text corpus was equipped with a cost attribute, based on the amount of the preselected frequent allophone sequences they contained. The highest cost value was attributed to a rare preselected quadphone (or collocation), the lowest to a frequent preselected triphone. In order to avoid the selection of long sentences (that contain more allophone sequences than shorter sentences) the cost value was normalized by the total number of allophones within the sentence.

The sentence with the highest score was selected for the final text corpus. The preselected allophone sequences covered by this sentence were eliminated from the list. Then the cost derivation and sentence selection process was preformed for this new set of preselected allophone sequences and a new sentence was chosen for the final text corpus. The same process repeated in a loop until the all of the initial preselected allophone sequences were covered in the resulting corpus of selected sentences.

The sentence selection algorithm was capable of selecting a rather modest subset of sentences out of the reference text corpus that cover the most frequent collocations, words, quadphones and triphones in the given language. A total of 297 sentences were selected out of the initial 200.000 sentences from the reference text corpus. The phonetic transcription of the selected sentence set covered all preselected most frequent triphones and quadphones.

B. *Recording and Segmentation*

These selected sentence subset was recorded along with logatoms containing all phonetically possible diphone combinations for the spoken Slovenian language. The speaker was instructed to read the phonetically transcribed sentences and logatoms in supervised recording sessions.

The recorded speech material had to be segmented and annotated. For segmentation of elemental speech units a semi-automatic procedure was used. A dynamic time warping acoustic alignment procedure between the synthesized voice and the recordings [21] was used to obtain preliminary phone boundaries since it performed better on detecting consonant segment boundaries than the HMM approach [22].

Manual corrections were needed on consonant boundaries within consonant clusters. The performance of the acoustical clustering plus dynamic time warping method is being tested along with boundary specific correction by means of a decision tree, as recently proposed by [23].

The final speech corpus contains 297 read sentences containing 1814 words. Additionally, 1668 logatoms were recorded. For the usage in embedded devices the final speech corpus had been compressed – a 1:14 corpus size reduction was achieved without significantly degrading the quality of the output speech signal. The resulting footprint of the compressed speech corpus was under 2MB. It was distributed into several files and it is used.

V. Evaluation

The adequacy of the TTS system was evaluated in terms of acceptability and in terms of intelligibility. The experiment was performed in laboratory conditions with 32 test subjects. It was conceived according various ITU-T Recommendations, describing methods for subjective performance assessment of the quality of speech voice output devices [24], [25].

The test was divided into three parts, as in [26]. 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, slightly above the score that the S5 system had received in a similar experiment [26]. Slovenian Eurom1 texts pronounced by a male professional radio announcer from the Multext-East corpus were used as reference speech. We encourage other groups working on Slovenian TTS to use the same reference speech when evaluating 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 the TTS system to be pleasant and quite natural sounding, sufficiently rapid and not over-articulated.

VI. Conclusion

We have presented the design of a corpus-based text-to-speech system, capable of synthesizing intelligible continuous speech from an arbitrary input text. Further improvements of intelligibility and naturalness depend in particular on proper lexical stress assignment and more sophisticated generation of prosodic parameters.

Additionally, major footprint reduction considerations for embedded TTS implementation are discussed. We concentrated on shrinking the speech corpus and yet keeping a high coverage of the frequent allophone sequences in a given language: our goal was to extract a sentence subset with high coverage and small size. The sentence subset was selected out of a large phonetically transcribed text corpus.

The greedy sentence selection algorithm implementation described in the paper was capable of selecting a rather modest subset of sentences out of the reference text corpus that cover the most frequent collocations, words, quadphones and triphones in the given language. A total of 297 sentences were selected out of the initial 200.000 sentences from the initial text corpus. The phonetic transcription of the selected sentence set covered all preselected most frequent triphones and quadphones, words and collocations.

An implementation of the algorithm for the Slovene language has resulted in a small footprint TTS system yielding intelligible and quite natural sounding speech.

References

[1] A.W. Black and K.A. Lenzo, "Flite: a small fast run-time speech synthesis engine," In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, 2001, pp. 204-207.

[2] M.L. Tomokoyo, W.A. Black and K.A. Lenzo, "Arabic in my hand: small footprint synthesis of Egyptian Arabic," In *Proceedings of the Eurospeech'03*, Geneva, Switzerland, 2003, pp. 2049-2052.

[3] T. Šef and M. Gams, "Speaker (GOVOREC): a complete Slovenian text-to speech system," *International journal on speech technologies*, vol.6, 2003, pp. 277-287.

[4] N. Pavešić, J. Gros, S. Dobrišek and F. Mihelič, "Homer II - man - machine interface to internet for blind and visually impaired people,". *Computer communications*, 2003, vol. 26, pp. 438-443.

[5] B. Vesnicer and F. Mihelič, "Evaluation of the Slovenian HMM-based speech synthesis system," Proc. TSD'04, *Lecture notes in computer science*, vol. 1692, Berlin, Springer Verlag, 2004, pp. 513-520.

[6] J. Gros, F. Mihelič, N. Pavešić, M. Žganec, A. Mihelič, M. Knez, A. Merčun and D. Škerl, "The phonectic SMS reader," Proc. TSD'01, *Lecture notes in computer science*, vol. 1692, Springer Verlag, Berlin, 2001, pp. 334-340.

[7] N. Campbell, "CHATR: a high-definition speech resequencing system," In *Proceedings of the 3rd ASA/ASJ Joint Meeting*, 1996, pp. 1223-1228.

[8] M. Beutnagel, A. Conkie, J. Schroeter and Y. Stylianou, "The AT&T Next-Gen TTS System," in *Proceedings of the 137th Meeting of the Acoustic Society of America*, 2000.

[9] B. Möbius, "The Bell Labs German text-to-speech system," *Computer Speech and Language*, vol. 13, 1999, pp. 319-358.

[10] J. Meron and P. Veprek, "Compression of exception lexicons for small footprint grapheme-to-phoneme conversion," In *Proceedings of the ICASSP'05*, Philadelphia, USA, March 18-23, 2005.

[11] J. Gros, N. Pavešić and F. Mihelič, "Syllable and segment duration at different speaking rates for the Slovenian language," in *Proceedings of the Eurospeech'97*, Rhodes, Greece, 1997, pp. 1-4.

[12] J. Gros, N. Pavešić and F. Mihelič, "Speech timing in Slovenian TTS", in *Proceedings of the Eurospeech'97*, Rhodes, Greece, 1997, pp. 323-326.

[13] A. Conkie, "Robust unit selection system for speech synthesis," in *Proceedings of the Eurospeech'99*, Budapest, Hungary, 1999.

[14] M. Beutnagel, R. Mohri and M. Riley, "Rapid unit selection from a large speech corpus for concatenative speech synthesis," in *Proceedings of the Eurospeech '99*, Budapest, Hungary, 1999.

[15] J. Tian, J. Nurminen and I. Kiss, "Optimal subset selection from text databases," In *Proceedings of the ICASSP'05*, Philadelphia, USA, March 18-23, 2005.

[16] J.P.H. Van Santen, "Methods for optimal text selection," In *Proceedings of the Eurospeech'97*, Rhodes, Greece, 1997, pp. 553-556.

[17] H. Kawai, S. Yamamoto and T. Shimizu, "A design method of speech corpus for text-to-speech synthesis taking into account prosody," in *Proceedings of the ICSLP'00*, 2000, pp. 420-425.

[18] C. Kuo and J. Huang, "Efficient and scalable methods for text script generation in corpus-based TTS design," in *Proceedings of the ICSLP'02*, 2002, pp. 121-124.

[19] B. Bozkurt, O. Ozturk and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection," in *Proceedings of the Eurospeech'03*, Geneva, Switzerland, 2003, pp. 277-180.

[20] M. Isogai, M. Mizuno and K. Mano, "Recording script design for corpus-based TTS system based on coverage of various phonetic elements," In *Proceedings of the ICASSP'05*, Philadelphia, USA, March 18-23, 2005.

[21] F. Malfrère and T. Dutoit, "High quality speech synthesis for phonetic speech segmentation," In *Proceedings of the Eurospeech'97*, Rhodes, Greece, 1997, pp. 2631-2634.

[22] F. Mihelič, J. Gros, S. Dobrišek, J. Žibert and N. Pavešić, "Spoken language resources at LUKS of the University of Ljubljana," *International Journal on Speech Technologies*, vol. 6, no. 3, 2003, pp. 221-232.

[23] G. Xydas and G. Kouroupetroglou, "An intonation model for embedded devices based on natural F0 samples," In *Proceedings of the Interspeech'04*, Korea, 2004, pp. 801-804.

[24] ITU, "A method for subjective performance assessment of the quality of speech voice output devices," *ITU-T Recommendation P.85*, ITU, 1994.

[25] ITU, "Telephone transmission quality subjective opinion tests - Modulated noise reference unit," *ITU-T Recommendation P.81*, ITU, Blue Book, (5), pp. 1-5, 1993.

[26] J. Gros, F. Mihelič and N. Pavešić, "Slovene interactive text-to-speech evaluation site – SITES," Proc. TSD'99, *Lecture notes in computer science*, vol. 1692, Berlin, Springer Verlag, 1999, pp. 223-228.