Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp189-193)

Text-to-Speech Synthesis for Embedded Speech Communicators

JERNEJA ŽGANEC GROS
Alpineon Research and Development Division
Alpineon RTD
Ulica Iga Grudna 15, SI-1000 Ljubljana
SLOVENIA
http://www.alpineon.com

Abstract: - We present a scalable corpus-based concatenation text-to-speech (TTS) system, which can be used in a variety of applications, ranging from server-based systems to embedded applications. For embedded applications, limited memory and processing power consumption criteria have to be met, therefore the language resources applied in TTS, primarily lexicon and speech corpora need to be reduced. We used an implementation of a greedy subset selection algorithm that extracts a compact subset of high coverage text sentences out of a larger set of sentences. An experiment on a reference text corpus demonstrated that the subset selection algorithm yields a compact sentence subset with a low redundancy.

Key-Words: - text-to-speech synthesis, embedded speech synthesis, speech corpus design

1 Introduction

Text-to-speech synthesis (TTS) enables automatic conversion into spoken form of any available textual information. Despite its considerable promise, text-to-speech synthesis is still not being used on a wide-scale basis in public service contexts. Wider acceptance of TTS devices will depend on three factors: better quality, smaller footprints, and more attractive pricing.

With the evolution of small portable devices, porting of high quality text-to-speech engines to embedded platforms has been made possible [1], [2]. Many applications in mobile telephony and portable computing require high-quality speech synthesis systems with a very modest computational and memory footprint.

TTS systems, which are using a corpus-based concatenative approach, yield close-to-natural sounding speech [3], [4], [5]. However, the linguistic resources required to build embedded TTS modules need to be scaled down to meet the hardware specifications of the embedded devices. The major memory and processing power consuming linguistic resources that need to be reduced are lexica and speech corpora [6], [7]. The application of these reductions to Slovenian is demonstrated in the paper by utilizing efficient exception lexicon and speech corpus reductions. They were performed on the baseline full-size Alpineon TTS system [8], [9], which uses a 95 Mb read speech corpus. A new, compressed speech corpus with a reduced set of read sentences was designed, which covers the most frequent allophone sequences in the language. The sentence subset has been selected from a large phonetically transcribed text corpus. The goal was to extract a sentence subset with high phonetic coverage and small size. The quality of the synthesized speech was assessed in a listening experiment, whereby the small-footprint TTS system was compared against the baseline full-size Alpineon TTS system.

2 Concatenation-based TTS

In the Alpineon TTS system, input text is transformed into its spoken equivalent by a series of modules, as shown in Figure 1: a grapheme-to-phoneme module produces strings of phonetic symbols based on information in the written text; a prosodic generator assigns pitch and duration values to individual phones; final speech synthesis is based on speech unit concatenation, where the elemental units are selected from a prerecorded and annotated speech corpus and later concatenated using a pitch-synchronous overlap-and-add technique. The linguistic front-end speech synthesis phases used in the system are described in the following two subsections.

2.1 Grapheme-to-Allophone Conversion

Input to the TTS system is unrestricted Slovenian text. It is translated into a series of allophones in two consecutive steps. First, input text tokenization and token-to-word conversions are performed. Abbreviations are expanded to form equivalent full words using a special list of lexical entries. The text normalizer converts further special formats, such as numbers or dates, into standard grapheme strings. The rest of the text is segmented into individual words and basic punctuation marks.

Next, word pronunciations are derived based on a user-extensible pronunciation dictionary and letter-to-sound rules. We constructed a dictionary that covers over 1,400,000 Slovenian inflected word forms. When

dictionary derivation fails, words are transcribed using automatic lexical stress assignment and letter-to-sound rules. The use of rules enables the TTS system to generate a first attempt at pronunciations of neologisms and named entities.

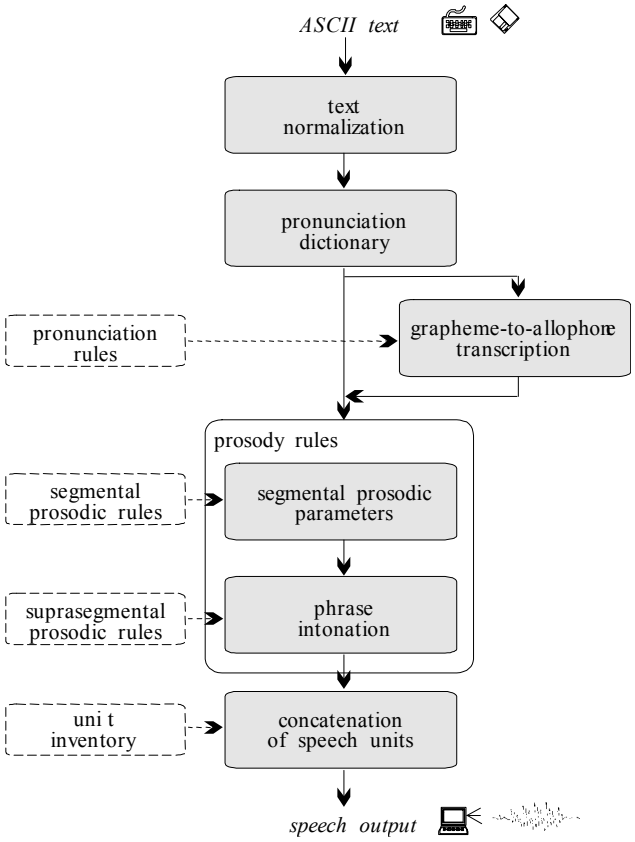

Fig.1 Modular system architecture of the TTS system.

To further reduce the memory footprint of the grapheme-to-allophone conversion module, we compiled an exception dictionary that contains only the differences from the phonetic transcriptions obtained by applying the letter-to-sound rule set. Similar to [7], a compression factor of ten was achieved, compared to the baseline full-lexicon representation, without sacrificing transcription accuracy. Further lexicon size reductions were achieved by modeling the exception lexicon in form of a decision tree, as proposed in [10].

2.2 Prosody Modeling

Corpus-based prosody modeling yields high-quality and close-to-natural sounding prosody parameter prediction; however, it requires a large amount of linguistic information upon which to rely. We used 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 F0 contour, and assignment of a global intonation contour [8].

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 at the syllable level and above.

Finally, intrinsic segment duration is modified, so that the entire word acquires its predetermined extrinsic duration. Stretching and squeezing does not apply to all segments equally. A method for segment duration prediction was used, which adapts a word with an intrinsic duration to the determined extrinsic duration, taking into account how stretching and squeezing apply to the duration of individual segments [12].

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

3 Speech Corpus Design

For unit-selection and other types of concatenation-based text-to-speech synthesis, a speech corpus of recorded and annotated elemental speech units is required [4]. 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, as do the recording and annotation costs. Therefore, a compromise between the size of the speech corpus and the quality of the resulting speech has to be taken [13] 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 may be full of redundancies. 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 [14], [15], [16], [17], [18]. Often sentence pair exchanges are calculated using diphone and triphone entropies. In [14], the unit coverage is maximized using prosody information. In [15], 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 [17] takes into account the balance of acoustic speech parts to provide variations in short-time speech features, and the linguistic parts provide long-time speech features, such as words or frequent word sequences.

We wanted the most frequent allophone sequences in a given language to be represented in the final sentence set, and therefore we implemented a greedy algorithm, similar to the one described in [15], to reduce the initial text sentence set to a compact and efficient subset. The process of designing a speech corpus for concatenation-based TTS was divided into three phases:

- Representative sentence set selection,
- Recording of selected texts, and
- Segmentation and annotation of the recorded speech material.

3.1 Sentence Subset Selection Algorithm

Initially, we collected a large corpus of texts covering various text styles, ranging from newspaper articles to fiction. All sentences shorter than five words or longer than 25 words were discarded from further analysis. The remaining reference text corpus contained 500,000 different sentences, corresponding to 50 Mb 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. A statistical analysis of frequent phone sequences of allophones, diphones, triphones and quadphones was performed on this corpus. It provided us with information about how frequently certain phone combinations occur in spoken Slovenian. In addition, the analysis has shown that only a few triphones have frequent occurrences.

Therefore, it makes sense to select only the most frequent triphones to be represented in the final speech corpus. We opted for the first 1,000 triphones: these represent 1% of the complete triphone set but cover almost 50% of all triphones in the transcribed reference text corpus. In a similar way, the 500 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 implemented 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, and the lowest to a frequent preselected triphone. In order to avoid the selection of long sentences, which 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 performed for this new set of preselected allophone sequences and a new sentence was chosen for the final text corpus. The same process was repeated in a loop until 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 299 sentences were selected out of the initial 500,000 sentences from the reference text corpus. The phonetic transcription of the selected sentence set covered all preselected most-frequent triphones and quadphones.

3.2 Recording and Segmentation

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

The recorded speech material was segmented and annotated. The final speech corpus contains read sentences with 1,993 words. For use in embedded devices, the speech corpus was compressed using standard voice compression techniques.

4 Evaluation Results

Various guidelines have been proposed for evaluating the quality of text-to-speech systems. Yet there are still no existing standards for their evaluation, although a number of different methods have been tried and it has been pointed out that the test results they yielded were often inconsistent [24].

The objective of the test was to compare the quality of the small-footprint TTS system to the baseline full-blown large-footprint unit-selection server-based TTS system [8]. The experiment was performed in laboratory conditions with 51 test subjects. It was designed according to ITU-T Recommendation P.85, describing

methods for subjective performance assessment of the quality of voice output devices.

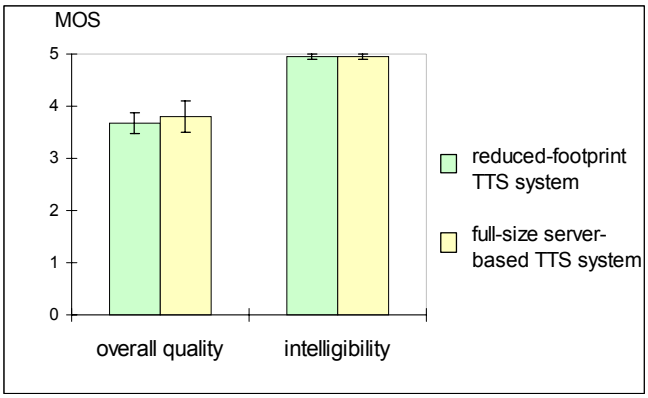

Fig.2 Evaluation results of the listening tests, given as MOS ratings for overall quality and intelligibility.

In the experiment, the performance of both TTS systems was evaluated by the listeners with grades on a five-point Mean Opinion Score or MOS scale. For the test, the sentences were synthesized by both TTS systems and presented to the listeners in random order. The listeners were asked to evaluate the overall quality and intelligibility. The results are provided in Fig.2. The majority of the test subjects evaluated the reduced-footprint, as well as the large-footprint, synthetic speech produced by the TTS system as pleasant and quite natural-sounding, sufficiently rapid and not over-articulated.

5 Conclusion

The memory and computational resources in TTS applications on embedded portable devices are inherently limited. 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 covers the most frequent collocations, words, quadphones, and triphones in a given language.

An implementation of the proposed sentence subset selection method for Slovenian has resulted in a small-footprint TTS system yielding intelligible and sufficiently natural-sounding speech, so that the system is ready for deployment in embedded applications. Listening experiments proved that the TTS system gives satisfactory performance in phonetization and speech concatenation quality with considerably reduced memory resources.

6 Acknowledgements

The authors wish to thank the Slovenian Ministry of Defense and the Slovenian Research Agency for co-funding this work under contract no. M2-0019.

References:

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

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

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

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

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

[6] Tian, J., Nurminen, J. and Kiss, I., Optimal subset selection from text databases, Proceedings of the ICASSP'05, PA, USA, 2005.

[7] Meron, J. and Veprek, P., Compression of exception lexicons for small footprint grapheme-to-phoneme conversion, Proceedings of the ICASSP'05, PA, USA, 2005.

[8] Žganec Gros, J., Mihelič, A., Pavešić, N., Žganec, M., Gruden, S., AlpSynth - concatenation-based speech synthesis for the Slovenian language, Proceedings of ELMAR'05, Zadar, Croatia, 2005, pp. 213-216.

[9] Žganec Gros, J., Human language technologies, *Public Service Review, European Union*, Vol. 10, 2005, pp. 126-127.

[10] Šef, T., An employment agent with a speech module, *WSEAS Trans.Comput.*, Vol. 2, 2003, pp. 680-685.

[11] Mohri, M., *On some applications of finite/state automata theory to natural language processing, Natural Language Engineering I*, Cambridge University Press, 1996.

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

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

[14] Kawai, H., Yamamoto and Shimizu, T., A design method of speech corpus for text-to-speech synthesis

taking into account prosody, Proceedings of the ICSLP'00, 2000, pp. 420-425.

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

[16] Bozkurt, B., Ozturk, O. and Dutoit, T., Text design for TTS speech corpus building using a modified greedy selection, Proceedings of the Eurospeech'05, Geneva, Switzerland, 2003, pp. 277-180.

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

[18] Rojc, M. and Kačič, Z., Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system, Proceedings of the LREC'00, Athens, Greece, 2000, pp. 321-325.

[20] Mihelič, F., Gros, J., Dobrišek, S., Žibert, J. and Pavešić, N., Spoken language resources at LUKS of the University of Ljubljana, *Int. Journal on Speech Technologies*, Vol. 6, No. 3, 2003, pp. 221-232.

[21] Xydas, G. and Kouroupetroglou, G., An intonation model for embedded devices based on natural F0 samples, Proc. Interspeech'04, 2004, pp. 801-804.

[22] Hoffmann, J., Jokisch, O., Hirschfeld, D., Strecha, G., Kruschke, G., Kordon, U. and Koloska, U., A multilingual TTS system with less than 1 Mbyte footprint for embedded applications, Proceedings of the ICASSP'03, Hong Kong, 2003.

[23] Alvarez, Y. and Huckvale, M., The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems, Proceedings of the ICSLP'02, Denver, CO, 2002, pp. 329-332.