The Phonectic Family of Voice-Enabled Products

Jerneja Gros, Mario Žganec, Aleš Mihelič, Marko Knez, Aleš Merčun, Domen Marinčič

Masterpoint R&D
Baznikova 40, 1000 Ljubljana
Slovenia
info@masterpoint.si
http://www.masterpoint.si

Abstract

The paper presents the work of the Masterpoint Speech Technology group in the field of voice-enabled products and their implementation into robust carrier-grade solutions for telephony applications.

The core of the products is the Phonectic Voice Portal Solution, a MXML (Masterpoint XML) based voice application platform that enables rapid development of new voice applications. The platform provides support for speech recognition and text-to-speech synthesis (TTS). The platform or its parts can be also used in simple voice applications, the Phonectic SMS Reader and the Phonectic Email Reader. The Slovene mobile telephony operator Mobitel d.d. has deployed a carrier-grade solution of the system in two such services called *Glasovni SMS* and *Glasovna e-pošta*.

The Slovene Phonectic TTS system is also described. It is based on concatenation of basic speech units, derived from a large speech corpus. The input text is transformed into its spoken equivalent by several modules. Pitch modeling is based primarily on predicting the proper tonemic accent. Phone duration is predicted by a two level approach, taking into account how acceleration or slowing down affect the duration of individual phones.

Finally, a simple demonstrator system using text-to-speech synthesis is described. The application enables sending SMS messages through the mobile telephony network as well as into the PSTN network. The addressee receives the SMS message in form of a synthetically composed read message.

1. Introduction

The ability of telephone speech recognition and text-to-speech synthesis to support effective applications has been established (Meisel, 2002). The large number of successfully deployed speech technology applications has proven the technology works, cost savings or increased revenues have been achieved and speech recognition has introduced substantial improvements in the user interface over the touch-tone pad.

Telephone service providers accepted speech recognition and text-to-speech synthesis (TTS) as being of long-term strategic importance, W. Meisel continues in his resume of the major speech technology events in 2001. Major telecoms, e.g. Sprint PCS, AT&T, BellSouth and many others, launched and expanded speech driven services, and indicated plans for continuing to do so. After the period of huge investments into home internet phone devices the recent successful speech recognition implementations drive the industry back to black-phone strategies using speech technologies to introduce new value-added services, using simple traditional phones.

The Speech Technology Group at Masterpoint is active in basic research of the Slovene spoken language as well as in developing commercial speech application platforms including both speech recognition and text-to-speech synthesis covered by the Phonectic product family. The core of the products is the Phonectic Voice Portal Solution, a MXML (Masterpoint XML) based voice application platform that enables rapid development of new voice driven applications. The platform provides support for speech recognition and text-to-speech synthesis (TTS). The platform or its parts can be also used in simple voice applications, the Phonectic SMS Reader and the Phonectic Email Reader. The Slovene mobile telephony operator Mobitel d.d. has deployed a carrier-grade solution of the system in two such services called *Glasovni SMS* and *Glasovna e-pošta*.

In the following sections we describe in detail the Phonectic TTS system, a corpus based TTS system for the Slovene language.

To conclude the paper, we describe how a simple and cost-effective demonstrator system can be rapidly built for a SMS-to-Voice service. It enables sending SMS messages and receiving them in voice format. The SMS messages have to include the telephone number of the addressee as well as the message to be read. The demonstrator system described includes two mobile telephones, one for intercepting SMS messages and another one for calling the message recipients. The text message is converted to speech by a text-to-speech system.

2. Phonectic TTS System

Some initial attempts towards Slovene TTS were mainly based on concatenation of diphones, and they resulted in a few demonstration systems (Šef, et al., 1988; Gros, 2000). In (Rojc and Kačič, 2000) the authors describe the formation of only a text corpus for Slovene corpus based TTS.

The Phonectic TTS system follows similar principles as the S5 TTS system (Gros, 1997). However, it is based on concatenation of basic speech units, derived from a large speech corpus. The input text is transformed into its spoken equivalent by a series of modules, as shown in Fig. 1. A grapheme-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. Pitch modeling is based primarily on predicting the proper Slovene tonemic accent. Phone duration is predicted by a two level approach, taking into account how acceleration or slowing down affect the duration of individual phones. Final speech synthesis is based on corpus segment concatenation using the TD-PSOLA technique (Moulines and Charpentier, 1990).

2.1. Text Normalization and Grapheme-to-Allophone Conversion

Input to the Phonectic TTS system is unrestricted Slovene text. The input text is translated into an allophone sequence in two steps. First, input text normalization is performed. Abbreviations are expanded to form equivalent full words using an extensible list of lexical entries. The text preprocessor further converts 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 Phonectic TTS dictionary covers over 1.000.000 most frequent Slovene inflected word forms.

Whenever dictionary derivation fails, words are transcribed using automatic lexical stress assignment and letter-to-sound rules.

Automatic stress assignment in the Phonectic TTS system is to a large extent determined by (un)stressable affixes, prefixes and suffixes of morphs, based on the rules described in (Gros, 1997). For words not affected by such rules, 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 and the syllable structure.

A set of over 190 context-dependent letter-to-sound rules translates each word into a series of allophones. The 150 rules described in (Gros, 1997) were used as a basic rule-set. In the Phonectic TTS system most of these rules have been updated and augmented.

2.2. Pitch and Duration Prediction

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 the Phonectic TTS system follows the proposal for Slovene prosody prediction fully described in (Gros, 1997) and consists of four phases:

- intrinsic duration assignment,
- extrinsic duration assignment along with prediction of pause duration,
- modeling of the intra word pitch 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.

The prosodic parameters for the current version of the Phonectic TTS system have been derived from extensive measurements as performed and openly published in (Gros, 1997). In this study, a speech database consisting of isolated words, carefully chosen by phoneticians (Srebot-Rejec, 1988), was recorded in order to study different effects on phone duration and fundamental frequency, which operate on the segmental basis. Vowel duration and fundamental frequency curves 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 (Gros, et al., 1997). The effect of speaking rate on phone duration was studied in a number of ways. Articulation rate expressed as the number of syllables or phones per second, excluding silences and filled pauses (O'Shaughnessy, 1995), was studied for the three different speaking rates.

Duration modeling was based on the variation of the two-level approach presented (Gros, 1997), separately modeling intrinsic and extrinsic duration and forming the final sound duration by taking into account how changes in articulation rate affect individual phone duration.

Similarly to (Gros, 1997) we used a simple approach for prosody parsing and the automatic prediction of Slovene intonation prosody which makes no use of syntactic or semantic processing (Sorin et al., 1987), but rather uses punctuation marks and searches for grammatical words, mainly conjunctions which introduce pauses. 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. We intend to introduce more sophisticated intonation models into future versions of the Phonectic TTS system.

2.3. Corpus Based Speech Synthesis

Given a sequence of phonetic symbols and prosody markers, the final step within the Phonectic TTS system 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 corpus-based approach has been used for speech synthesis for the first time in Slovene speech synthesis. First an extensive analysis of the frequency of Slovene polyphone sequences was performed. Large Slovene text corpora have been transcribed into allophone sequences and statistically processed. The texts for the spoken corpus were selected by an optimization process optimizing the number of the most frequent polyphones covered by the spoken text and a minimum amount of the text to be read by the speaker.

Given an input sequence of phonetic symbols a rather sophisticated segment selection algorithm first selects the segments to be concatenated. It takes into account a number of criteria ranging from more and less preferred allophones for gluing, the length and phonetic contexts of polyphones, spectral discontinuities, etc.

The TD-PSOLA technique (Moulines and Charpentier, 1990) was used for speech segment concatenation as it enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications without considerably affecting the quality of the synthesized speech.

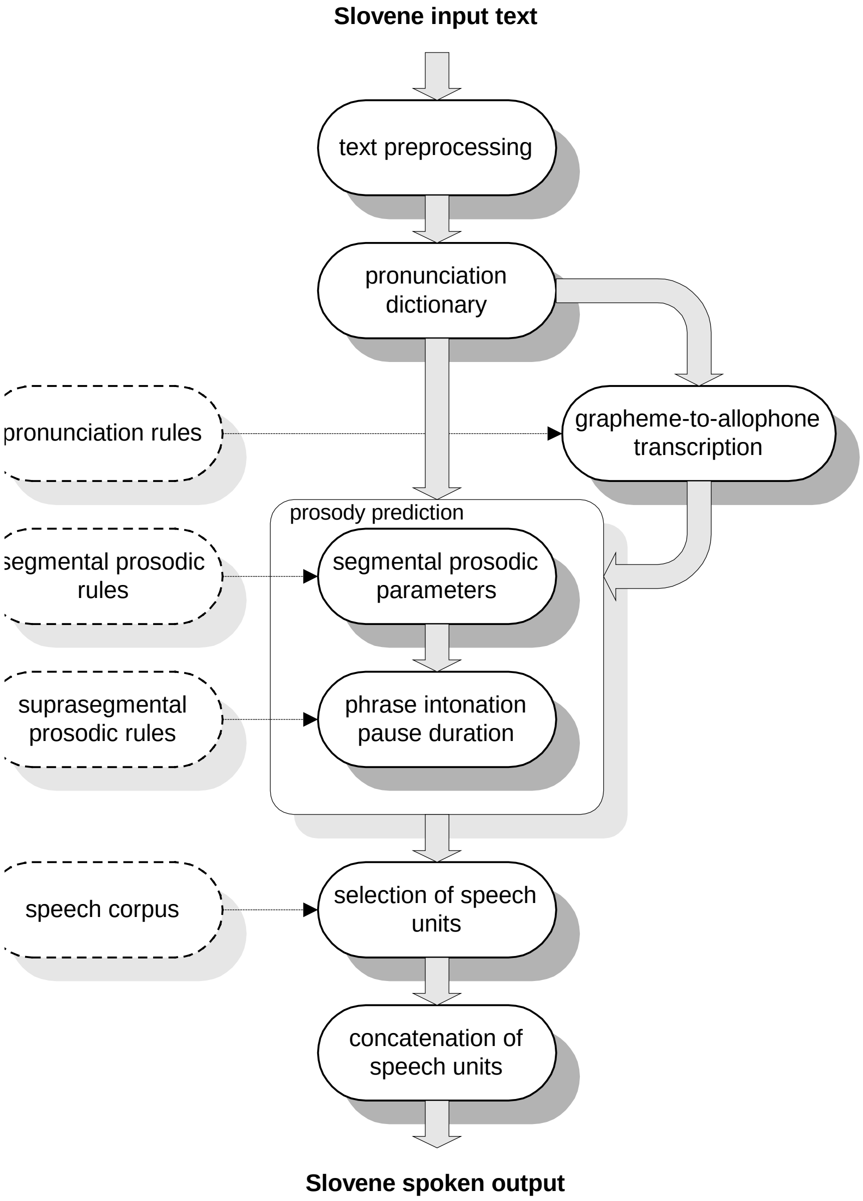
Slovene input text
text preprocessing
pronunciation dictionary
pronunciation rules
grapheme-to-allophone transcription
prosody prediction
segmental prosodic rules
segmental prosodic parameters
suprasegmental prosodic rules
phrase intonation pause duration
speech corpus
selection of speech units
concatenation of speech units
Slovene spoken output

Figure 1: The Phonectic TTS system architecture.

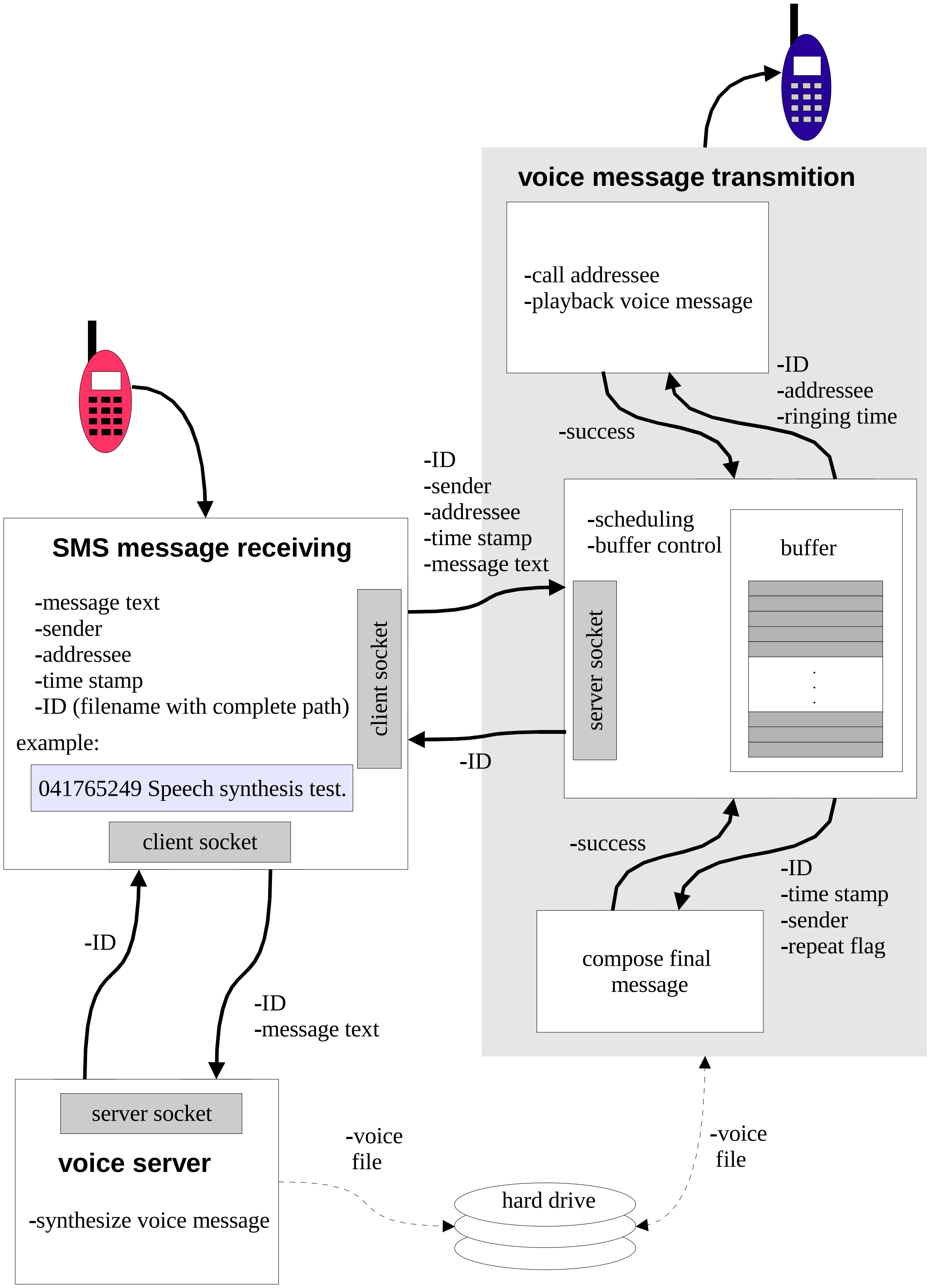

Figure 2: Architecture of a simple SMS Reader demonstrator system. A sample message sent to the system is *'041765249 Speech synthesis test'*, meaning the synthesised message 'Speech synthesis test' will be read to the owner of the telephone number 0417656249.

3. SMS Reader Demonstrator System

In this section we show how to easily build a simple demonstrator system using a TTS engine. A SMS Reader demonstrator system is a speech technology application that enables sending SMS messages and receiving them in voice format. The SMS messages have to include the telephone number of the addressee as well as the message to be read.

A simple implementation of a SMS Reader includes two mobile telephones, one for intercepting SMS messages and another one for calling the message receivers. The text message is converted to speech by a text-to-speech system. In case the majority of the voice SMS messages is expected to terminate in the fixed line telephony network a simple telephony card (analog or ISDN) can be used instead of the second mobile telephone handling outbound calls.

The simple SMS Reader demonstrator consists of 3 main modules, as shown in Fig. 2. The first module intercepts the incoming SMS message and performs the parsing of the message. The following information is sent to the next modules: message ID, message text, sender phone number, addressee phone number, time stamp and ID (file with complete path).

The text part of the message is sent to the second module, the voice server that performs the text-to-speech conversion.

The final message transmission module composes the final message, e.g. 'You have received a message from number 031765249. The message is as follows: Speech synthesis test. End of message'. Finally, the voice message is sent to the addressee by a push call.

More sophisticated and advanced SMS-to-Voice carrier grade platform solutions can be viewed as important value-added services for mobile telephony operators.

4. Conclusion

The paper describes the Phonectic family of Voice products and some of their core components, like the corpus based Phonectic TTS system, capable of synthesizing intelligible continuous speech from an arbitrary Slovene input text.

The Phonectic TTS system, along with a Slovene speech recognition module, is the vital component of the Phonectic product family for voice application carrier grade platforms, including the Phonectic SMS2Voice Platform, the Phonectic E-mail Reader Platform (using an efficient language identification module developed by Masterpoint) and the Phonectic Voice Portal as part of the broader Phonectic MVAP platform (Multi-protocol Voice Application Platform).

Finally, the architecture of a simple demonstrator system using text-to-speech synthesis is described. The application enables sending SMS messages in voice format through the mobile telephony network as well as into the PSTN network. The addressee receives the SMS message in form of a synthetically composed read message.

5. References

Gros, J. (1997). *Samodejno pretvarjanje besedil v govor*. PhD Thesis. University of Ljubljana, Slovenia.

Gros, J., Pavešić, N. and Mihelič, F. (1997). *Syllable and segment duration at different speaking rates for the Slovenian language.* Proceedings of the EUROSPEECH'97. Rhodes, Greece.

Gros, J., Pavešić, N. and Mihelič, F. (1997). *Speech timing in Slovenian TTS.* Proceedings of the EUROSPEECH'97. Rhodes, Greece, pp. 323-326.

Gros, J. (2000). *Samodejno tvorjenje govora iz besedil.* Linguistica et Philologica. Založba ZRC, Ljubljana, Slovenia.

Meisel, W. (2002). *Looking back at 2001 and forward to 2002*. Speech recognition update. 103:2-3.

Moulines, E., Charpentier, F. (1990). *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones.* Speech Communication, 9:453-467.

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

Rojc, M. and Kačič, Z. (1990). *Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system. Proceedings of the Second international conference on language resources and evaluation.* Athens, Greece, pp. 321-325.

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

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

Šef, T., Dobnikar, A., Gams, M. and Grobelnik, M. (1998). *Slovenski govor na internetu. Proceedings of the Conference Language Technologies for the Slovene Language ISJT'98.* Ljubljana, Slovenia, pp. 60-64.

Toporišič, J. (1984). *Slovenska slovnica.* Maribor, Slovenia, Založba Obzorja.