IMPROVEMENTS IN SLOVENE TEXT-TO-SPEECH SYNTHESIS

Tomaž Šef, Aleš Dobnikar, Matjaž Gams

Department of Intelligent Systems
Jozef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
e-mail: tomaz.sef@ijs.si

ABSTRACT

This paper presents a new text-to-speech (TTS) system that is capable of synthesising continuous Slovenian speech. Input text is processed by a series of independent modules: text normalisation, grapheme-to-phoneme conversion, prosody generation and segmental concatenation. That enables easy improvements of separate parts of the system. In order to generate rules for our synthesis scheme, data was collected by analyzing the readings of ten speakers, five males and five females. A two-level approach has been used for duration modelling and so-called superpositional approach at pitch modelling. Speech waveform is synthesized using a concatenative TD-PSOLA technique. Acoustic segments inventory consists of diphones and some frequently used polyphones.

Our system is used in several applications. It is built into an employment agent EMA that provides employment information through the Internet. Currently we are developing a system that will enable blind and partially sightless people to work in the Windows environment.

1. INTRODUCTION

While text-to-speech (TTS) systems for major world languages are quite advanced, smaller languages, like our Slovenian language, lack quality TTS synthesis. The naturalness of the synthetic speech strongly depends on the generated prosodic contours that depend on the degree of available knowledge which is for these languages rather small.

Slovenia does not have a long tradition in the development of TTS systems. In the beginning of the 90's, at the Jozef Stefan Institute, we started the first attempt at developing a rule based TTS system [1], to utter only isolated words on two commercial synthesizers: Covox speech thing and LSI phonetic synthesizers. With a lexicon and explicit rules, the accentuation module first produced a stressed form for a given word. After graphemic-phonemic conversion, the duration and pitch for each phoneme was defined. The durations depended on the type of the phoneme, on whether the phoneme is stressed or unstressed and on the length of the whole word. Pitch depended on the surrounding phonemes and on the position in the word. At the end, the LSI phonetic synthesizer and Covox speech thing converted all that information for each phoneme into synthesizers parameters and sent them to their DA ports. The quality of the results was not satisfactory, therefore another approach to unrestricted speech synthesis was used in 1995 [2]. We developed a system for concatenating acoustical speech units, using the TD-PSOLA technique.

2. SYSTEM ARCHITECTURE

The different phases of the synthesis task are performed by several sequentially operating independent modules as shown in Figure 1 [3]. This enables easy improvements of separate parts of the system (http://zlatoust.ijs.si/STTS/STTS.htm). The input into the synthesizer is unconstrained text presented in its digital form. In the first place of the hierarchical system architecture a text normalization module is located. It is followed by a grapheme-to-phoneme conversion and prosody generation module. At the end, a segmental concatenation module outputs a digital record of the synthetic speech which is produced by a computer sound card [4].

The synthesizer is controlled by a large number of parameters. They define a mode of text pre-processor (normalizer). In prosody generation module a speaking rate (normal, fast, slow or any intermediate level), pitch, pauses duration, a style of words accentuation, a type of phoneme duration modifications, etc., are set.

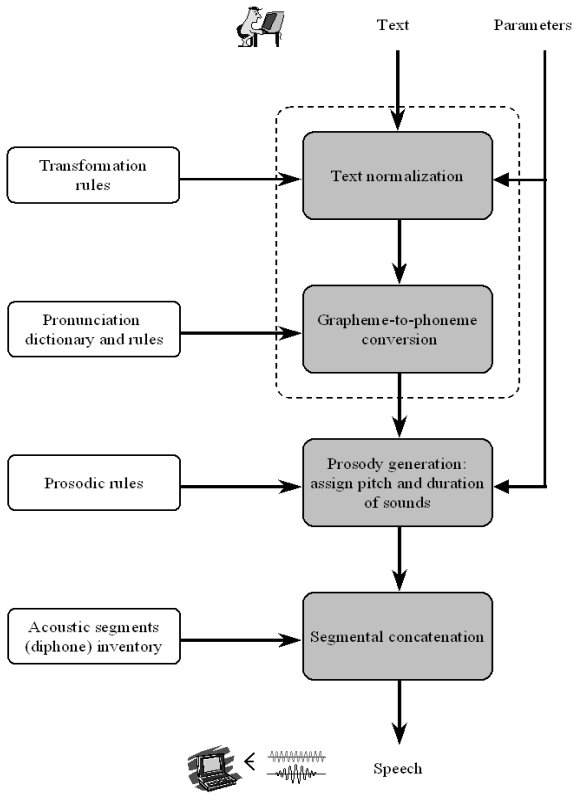

Figure 1: Slovene text-to-speech system.

2.1. Text Normalization

The text normalizer first converts files from different formats into an ASCII file. Then, all redundant symbols are removed and abbreviations are expanded to equivalent full words. Special formats, like numbers, hours or dates, are converted into standard graphemic strings [3]. Several lists of rules and lexical entries are used which can easily be expanded in case of need. A basic semantic analysis predicting the proper inflected form for numerals is also included [4].

Each sentence is processed separately. At the end, the whole text is segmented into individual words and basic punctuation marks. The text normalizer corresponds to the first level in Figure 1.

2.2. Grapheme-to-Phoneme Conversion

A grapheme-to-phoneme module (the second level in Figure 1) produces strings of phonemic symbols from the normalized text. This derivation is done in three steps [4]:

- first a pronunciation dictionary is checked for the presence of the whole word,
- if a word is found, it is transcribed into its phonetic form considering interacting influences between two adjacent words,
- for words, which are not present in the pronunciation dictionary, an automatic lexical stress assignment and letter-to-sound rules is used.

The pronunciation dictionary is divided into more components: numerals, proper nouns, collocations, common words (the most extensive), application specific words (employment, jobs, etc). It is easy to add a new component to the dictionary [4].

Automatic grapheme-to-phoneme conversion is done by a large number of context-dependent letter-to-sound rules. By lexical stress assignment, lists of stressable and unstressable affixes, prefixes and suffixes are used [5]. For other words, the most probably stressed syllable is predicted regarding the number of syllables within a word. As lexical stress in the Slovenian language can be located almost arbitrarily on any syllable, correct pronunciation is very difficult and practically impossible. This is the weakest part of our system.

2.3. Prosody Parameters Assignment

Prosody has great impact on intelligibility and naturalness of speech perception. The proper choice of prosodic parameters, given by phoneme duration and intonation contours, enables natural sounding high quality synthetic speech.

Prosodic parameters are set (third level in Figure 1) in three steps [4]:

- duration assignment,
- pitch assignment,
- insertion of pauses.

Duration Modelling

A two-level approach has been used for duration modelling. The intrinsic duration of words is determined by using factors such as phone name, phone type, phone context, segmental identity, syllable type and syllabic stress. The higher-level rhythmic and structural constraints of a phrase predict the extrinsic duration of a word. The following factors have been taken into account: the chosen speaking rate, the word's position within a phrase and the number of syllables within a word. At the end, intrinsic segment durations are modified, so that the entire word achieves its predetermined extrinsic duration [4, 7]. However, stretching and squeezing does not apply to all segments equally.

The results of the investigations made by T. Srebot-Rejec and J. Gros were helpful. The first one studied different effects on phone duration: stressed/unstressed and open/closed syllables at vowel duration [6], CC and VCV clusters at consonant duration. The second one recorded a large continuous speech database to study the impact of speaking rate on syllable duration and duration of phones [7].

Pitch Modelling

In order to generate rules for our synthesis scheme, data was collected by analyzing the readings of ten speakers. All of them are native Slovene speakers, five males and five females. Eight of them are professional speakers on the Slovenian national radio. The largest part of the speech material consists of declarative sentences in short stories, monologues, containing sentences of various complexities and types, news, weather reports and commercial announcements. Other parts of the corpora are interrogative sentences with yes/no and wh-questions and imperative sentences. The first part of the corpora contains 500 declarative sentences, uttered by eight speakers, and the second part 100 questions and 30 imperative clauses uttered by 2 speakers [8].

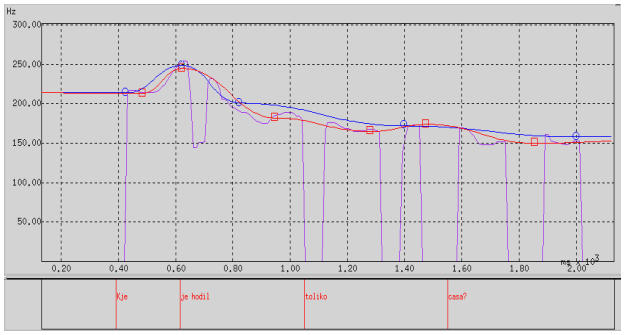

Figure 2: Example result of F0 contour modeling for the Slovene Wh-question "Kje je hodil toliko časa?" Engl.: "Where did he walk so long?".

The F0 contour is defined with the function composed from [9]:

- a global component, related to the whole intonation unit,
- local components, related to accented syllables or syntactic boundaries.

The global component gives the baseline for the F0 contour for the whole intonation unit. It depends on the type of the intonation unit (declerative, imperative, yes/no or wh-question), position of the intonation unit in a complex sentence with two or more intonation units and duration of the whole intonation unit. The local components present local movements of the F0 shape at accented syllables or syntactic boundaries. Tonemic accent in Slovenian language is signed by a pitch rise within a stressed syllable, followed by a fall, which depends on the accent (acute or circumflex) and syllable (baritone or ocsitone).

Many functions were tested (linear, power, transfer, decay, exponential) for the best approximation of the natural F0 contour. In the system presented, an exponential function for the global component and a cosinusoidal function for accents and final boundary contours were adopted.

Figure 2 illustrates the results obtained. The sentence for comparison is uttered by a female speaker. The parameters for the synthesized F0 are the same for the whole sentence. The panel displays the original F0 contour modeled with the INTSINT (INternational Transcription System for INTonation) system [10], indicated by squares, and the synthesized F0 contour indicated by circles.

Pauses

Pauses have a very important role in the intelligibility of speech. Normal conversations typically half of the time consist of pauses; in the analyzed readings they represent 18 % of the whole database duration. The obtained results show that pause duration is almost independent of the duration of the intonation unit before the pause, only duration of the pauses greatly depends on whether there is breathing in or not.

Pauses, the classical points of boundaries between successive intonation units, are classified into four groups with respect to types and durations:

- at prefaces, new topics of readings and new paragraphs, noted without classic orthographic delimiters (always represent the longest pauses, always connected with breathing in),
- at the end of sentences, marked with a period, exclamation mark, question mark or ellipsis,
- at places of prosodic phrases in the sentences, marked with a comma, semicolon, colon, dash, parentheses or quotation marks,
- in sentences at places of rhythmical divisions of the clause, often before the conjunction of Slovene words *in*, *ter* (and), *pa* (but), *ali* (or), ...

The stochastic variance in the range of pause durations prevents the synthetic, discrete nature of pauses in synthetic speech.

2.4. Segmental Concatenation

Once the phonetic symbols and prosody markers are determined, the final step is to produce audible speech by assembling elemental speech units, computing pitch and duration contours, and synthesizing the speech waveform (the last level in Figure 1).

Acoustic Segments Inventory

An analysis of the Slovenian phonological system gives 8 vowel and 21 consonant phonemes. When adding allophonic variations for some phonemes, we arrived at a total of 33 phones. One diphone for every allophone combination possible in a given language is required. Our diphone inventory contains 1155 pitch-labeled diphones [3]. They were hand-segmented and hand-labeled in order to enable as good as possible coupling at concatenation points. In order to guarantee high synthesis quality, the diphones were recorded by a professional speaker and placed in the middle of logatoms, pronounced with a steady intonation [4].

Currently we are adding to the existent diphone inventory some larger units: polyphones and frequently used short words. This way, the final quality of the synthetic speech is much better.

Segmental Concatenation

We used a concatenative TD-PSOLA technique improved with a variable length linear interpolation process [11]. This algorithm enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications without considerably affecting the quality of synthesized speech. In contrast to the pure TD-PSOLA algorithm it also supports spectral interpolation between voiced parts of segments. This improvement demands no excessive increase in the computational load [11].

3. APPLICATIONS

3.1. Employment Agent EMA

Our text-to-speech system is used in an employment agent EMA (http://www-ai.ijs.si/~ema/). EMA is an intelligent agent for employment tasks on the Internet designed in cooperation with the National Employment Office. Ema's basic task is to help people searching for employment, jobs, employees, scholarships and any other form of employment related tasks. For example, 90% of all available jobs in Slovenia are presented through EMA. In addition, EMA can perform a variety of tasks, such as sending mails whenever any new information on related Internet files/sites occurs. In the last year, EMA was visited by around one third of Slovenian population with access to the Internet and is the most often visited and used intelligent system in Slovenia [12]. With this system, Slovenia was among the first European countries to offer this type of information through the Internet.

The TTS system is applied to the most often visited module of the system that provides information about available jobs in Slovenia [4, 12]. The text is based on a combination of a limited (e.g. job positions) and practically unlimited input (e.g. location, desires). When a user chooses an interesting job offer and the speech option, the system sends a WAV file through the Internet to user's machine. A typical job offer has about 100-150 chars, and typical WAV file has about 250 KB.

3.2. SoundHint System for Blind and Partially Sightless People

The TTS system is available as a program library which can be included in any application. Currently, the SoundHint system that will enable blind and partially sightless people to work in the Windows environment is under development (http://zlatoust.ijs.si/SoundHint/SoundHint.htm).

The considerably uniform structure of Graphical User Interfaces (GUI) in most of applications allow distinction between separate constructions, respectively components of the interface, like windows, buttons, menus, fields, scroll bars, etc. This property enables a development of the all-purpose sound interface for interpretation of any GUI.

In the background, the SoundHint inspects all occurrences inside Windows environment. It is able to describe the active window which enables the user to imagine what is in front of him and what is he expected to do. The changes, like crossing to the next field, menu or button are simultaneously passed to the user. Individual actions, like renewed description of active window, text reading, reading of the program list, etc. are enabled by pressing a predefined key combination on the keyboard.

Such an approach cannot take the full advantage of program specific functions. Therefore, applications that are especially interesting for blind and partially sightless people will be supported specially. The access to the Internet is already one of the fields that deserve that.

4. TESTS

The adequacy of the synthesis system was tested in terms of acceptability [3]. The experiment was performed with 11 subjects within the age of 22 and 53 years, three of them being female. Most of them have high education. Eight subjects did not hear the synthesizer beforehand.

The word intelligibility rate was 94,6 %. The listeners had more problems with short words. Usually they mixed up only one letter in the word with another closely related letter.

All of the subjects considered that the synthetic speech is pleasant, enough understandable and quite natural sounding. To their opinion the system is an appropriate tool for generating audible speech from text in Slovenian language.

5. CONCLUSION

We developed a new complete text-to-speech system for Slovenian language based on the acoustical units concatenation. The system is capable of synthesizing continuous Slovenian speech from an arbitrary input text. The modular architecture enables easy improvements of separate parts of the system.

In order to generate rules for our synthesis scheme, data was collected by analyzing the readings of ten speakers. A two-level approach has been used for duration modelling and so-called superpositional approach at pitch modelling. Speech waveform is synthesized using a concatenative TD-PSOLA technique.

The experiment showed that the system is an appropriate tool for generating audible speech from text in the Slovenian language. The advances in technology and new methods have already enriched practical usefulness. The quality of the system is still inferior to human speech, but the improvements are noticeable, compared to older systems.

6. REFERENCES

1. S. Weilguny, *Grapheme-to-Phoneme Conversion for the System for Uttering Isolated Words*, MSc Thesis, Faculty of electrical engineering and computer science, University of Ljubljana, 1993.
2. A. Dobnikar, J. Bakran, "A new approach for Slovene text-to-speech synthesis", *Proc. MIPRO'95*, pp. 265-268, Croatia, 1995.
3. T. Šef, A. Dobnikar, M. Gams, "Text-to-Speech Synthesis in Slovenian Language", *Proc. of the IX European Signal Processing Conference (EUSIPCO' 98)*, 1998, (to apear).
4. T. Šef, *System for speech mediation of employment opportunities*, M. Sc. Thesis, Faculty of Computer and Information Science, University of Ljubljana, 1998.
5. J. Toporišič, *Slovene grammar*, Založba Obzorja, Maribor, 1984.
6. T. Srebot Rejec, *Word accent and vowel duration in standard Slovene: an acoustic and linguistic investigation*, Slawistische Beiträge, 226, Vewrlag Otto Sagner, München, 1988.
7. J. Gros, *Automatic text-to-speech conversion*, PhD Thesis, Faculty of Computer and Information Science, University of Ljubljana, 1997.
8. A. Dobnikar, *Modeling Segment Intonation for Slovene Text-to-Speech System*, PhD Thesis, Faculty of Computer and Information Science, University of Ljubljana, 1997.
9. H. Fujisaki, S. Ohno, "Analysis and Modeling of Fundemental Frequency Contour of English Utterances", *Proc. EUROSPEECH'95*, Vol. 2, pp. 985-988, Madrid, 1995.
10. D. J. Hirst, "Prosodic labelling tools", *MULTEXT LRE Project 62-050 Report*, Centre National de la Recherche Scientifique, Université de Provance, Aix-en-Provence, 1994.
11. T. Dutoit, H. Leich, "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database", *Speech Communication 13*, pp. 435-440, 1993
12. M. Gams, V. Križman, T. Šef, "An Employment Agent with a NL Interface", *Proceedings of the International Conference on Systems, Signals, Control, Computers (SSCC'98)*, South Africa, 1998, (to apear).