Text-to-Speech Synthesis: A Complete System for the Slovenian Language

Jerneja Gros, Nikola Pavešić and France Mihelič

Faculty for Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

A text-to-speech system, capable of synthesising continuous Slovenian speech from an arbitrary input text is described. The text-to-speech system is based on the concatenation of basic speech units, diphones, using the TD-PSOLA technique, and no special hardware is required. The input text is transformed into its spoken equivalent by a series of the modules. The modules, constituting the text-to-speech system are described in detail. Special attention is paid to segmental duration determination, where the effect of speaking rate on phone duration is widely studied. Finally, the results of output speech quality assessment are given in terms of acceptability and intelligibility.

Keywords: text-to-speech synthesis, diphone concatenation, prosody modelling, grapheme-to-phoneme conversion, Slovenian language.

1. Introduction

Text-to-speech synthesis enables automatic conversion of any available textual information into its *spoken* form. For the Slovenian language, several attempts were made in the past, where different aspects of a Slovenian text-to-speech system were covered [Hribar, 1984], [Weilguny, 1993]. Nevertheless, none of them succeeded in building a complete system, providing high quality synthesised speech. In the Laboratory of Artificial Perception, University of Ljubljana, we have started on text-to-speech synthesis recently [4] [Gros et al., 1996], [8] [Gros et al., 1997]. Here we describe the current version of our Slovenian text-to-speech system, which is to serve as a reference system for future improvements.

The different phases of the synthesis task are performed by separate independent modules, operating sequentially, as shown in Figure 1. A grapheme-to-phoneme module produces strings of phonemic 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 phonemes. Final speech synthesis is performed by TD-PSOLA concatenative diphone technique [Moulines and Charpentier, 1990].

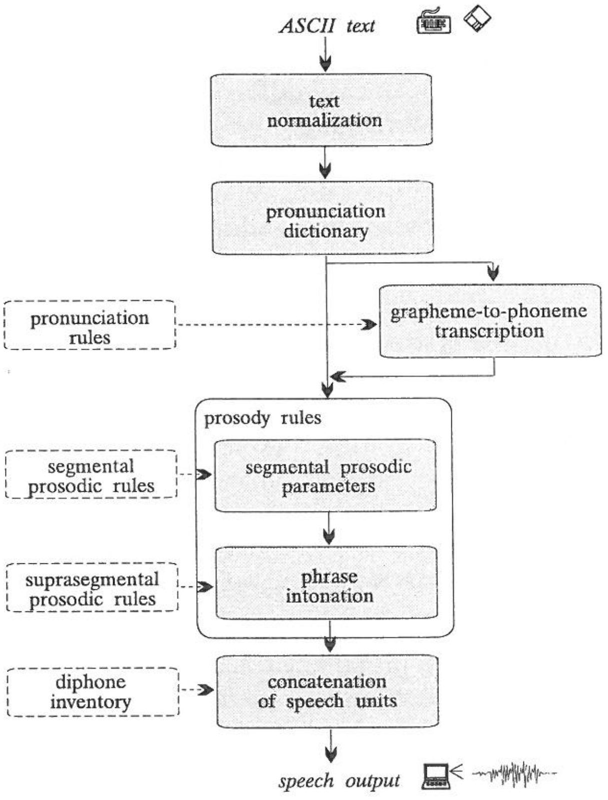

Fig. 1. Slovenian text-to-speech system architecture.

We continue with a detailed description of the modules, constituting the text-to-speech system and outline the work still needed to be done. At the end of the paper, assessment results of the text-to-speech system are given and discussed and some promising directions for future work are mentioned.

2. Grapheme-to-phoneme Conversion

Input to our text-to-speech system is unlimited text. For the time being, input text should be stored in an ASCII file; in future we intend to expand the input possibilities so that it may come from other programs or marked regions on the computer screen or databases.

Input text is translated into a *series of phonemes* (or allophones, in case different phoneme variations are differentiated) in two consecutive steps. An analysis of the Slovenian phonological system gives 8 vowel and 21 consonant phonemes. When adding allophonic variations for certain phonemes, we arrived at a total of 34.

In the first phase of our grapheme-to-phoneme conversion system, abbreviations are expanded to form equivalent full words using a special list of lexical entries. A text pre-processor (normaliser) 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 user-extensible pronunciation dictionary and letter-to-sound rules. The dictionary covers 16.000 most frequent Slovenian words along with 350 most frequent proper names.

In case where dictionary derivation fails, words are transcribed as a human would pronounce them upon encountering a new unknown word, using automatic lexical stress assignment and letter-to-sound rules. However, as lexical stress in the Slovenian language can be located almost arbitrarily on any syllable, errors are introduced into pronunciation of such unknown words. Nevertheless, there do exist some rules of stress assignment, based upon observations of linguists [Toporišič, 1984], which to a certain extent determine the stress position within a word. Automatic stress assignment is mainly determined by (un)stressable affixes, prefixes and suffixes of morphs. Once lexical stress has been established, a set of context-dependent letter-to-sound rules translates each word into a series of phonemes. A basic semantic analysis predicting the proper inflected form for numerals is included. The vast majority of the letter-to-sound rules are context-dependent, meaning that a letter or a sequence of letters is transcribed differently according to its left and right context.

3. Prosody Generation

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

A two-level approach to duration prediction was adopted, based on [Epitropakis, 1993]. It involves an adjustment of the phoneme's intrinsic duration taking into account how stretching and sqeezing are applied to individual phonemes [7] [Gros et al., 1997].

The fundamental frequency is first predicted on a word basis – modelling the words' tonemic accent, later a global intonation countour is superponed.

3.1. Duration Modelling

In order to provide the synthesiser with the possibility to pronounce input text with several speaking rates, tests were made to study the impact of speaking rate on syllable duration and duration of individual phonemes and phoneme groups. In fact, results of perception experiments show that tempo variation contributes to the perceived naturalness of speech [Eefting and Nooteboom, 1993].

When different speaking rates are applied, it is possible to obtain different phoneme realisations that differ only in duration. In this case, the speaker is instructed to pronounce the same materials at different rates. This keeps context, stress and all other factors identical to every realisation of the sentence. As a result, pairwise comparisons of phoneme duration can be made. We opted for a relatively long meaningful text, read three times: at a normal, fast and slow rate. Reading the text took 7 minutes 32 seconds when reading at a normal rate, 12 minutes 55 seconds reading very slow and 5 minutes 45 seconds when reading as fast as possible. The speech material was initially labelled using a Hidden Markov model speech recogniser [Ipšić, 1996]. The obtained labels were manually corrected.

The effect of speaking rate on phoneme duration was studied in a number of ways. An extensive statistical analysis of lengthening and shortening of individual phonemes, phoneme groups (vowels, nasals, liquids, plosives, fricatives, etc) and phoneme components (closures, bursts) was made, the first of this kind for the Slovenian language [5] [Gros et al., 1996]. As a result of our study, intrinsic phoneme duration was determined with respect to a chosen speaking rate.

Next, the words' extrinsic duration is modelled. Words are syllabified by counting the number of their vowel clusters and the *duration of syllables* is modelled according to the speaker's normal articulation rate, depending on the number of syllables within a word and on the word's position within a phrase.

Articulation rate, expressed as the number of syllables or phones per second, excluding silences and pauses, was determined for all three speaking rates. Figure 2 shows articulation rate in number of syllables per second plotted as a function of word length in number of syllables and the word position in a sentence. The obtained values apply for normal speaking rate. In general, lexical redundancy increases with word length. Hence the need for slow and careful articulation decreases with increasing word length. Average syllable duration tends to decrease with more syllables in a word. We observed that in case enclitics are associated to their preceding words, articulation rate adopts a quasi-logarithmic contour, which can be described parametrically as in [Bakran, 1994]. Isolated words and those following a pause make an exception to this rule for words with more than four syllables, of which only a few realisations were available.

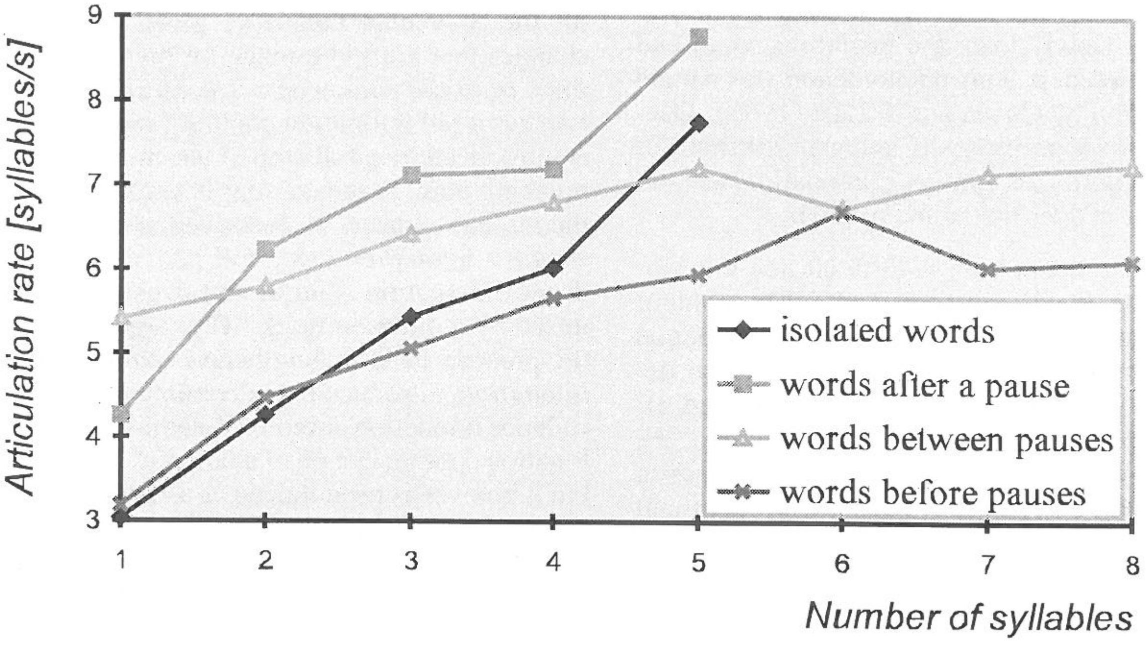

Fig. 2. Articulation rate expressed in syllables per second, depending on the number of syllables in a word is given for four different types of word positions.

Tonemic Accent

The Slovenian tonemic accent is marked by a pitch rise within a stressed syllable, followed by a fall, which depends on the syllable (baritone or ocsitone) and the accent (acute or circumflex) [Srebot, 1988].

Segmental prosodic parameters were determined for every phoneme with respect to accent position within a word, its type, syllable position and its type and syllable duration, previously determined by the articulation rate. Experimentally obtained results from measuring pitch values for a given speaker were used as initial values. They were modified according to some previous observations from phoneticians [Srebot, 1988].

Intonation Contour Determination

In the same way as the multitude of audibly different pitch changes relates to a limited number of pitch movement types, the diversity of pitch contours converges on a restricted set of melodic categories, the *intonation patterns*.

In trying to relate pitch movements to some sort of reference level, one is confronted with the most important global attribute of a pitch (F_0) curve, namely the tendency of the frequency to decrease slowly from the beginning to an end of the utterance. This phenomenon was named *declination* by Cohen and T'Hart (1965). Declination is a perceptually relevant attribute of pitch contours, as synthetic intonation without declination does not sound natural.

Intonation recognition is difficult due to problems with the F_0 contour itself. The contour is not continuous as there is no fundamental frequency in unvoiced regions. F_0 is also dependent on segmental effects - before and after stops the contour can deviate sharply. F_0 tracking is difficult and prone to errors and even when successful, there is a considerable amount of short-term variation in the F_0 contour due to pitch perturbations. Therefore, smoothing and linearisation of the F_0 contour are performed.

We used a relatively simple approach for prosody parsing and automatic prediction of Slovenian intonational prosody which makes no use of syntactic or semantic processing [Sorin et al., 1987], but rather uses punctuation marks. Sentences are parsed into prosodic groups, i.e. segments between punctuation marks or so called grammatical words (e.g. conjunctions).

Our interest was to detect typical prosodic segments by means of pitch (F_0) contours. Again, a set of measurements was made in order to define four typical intonation contours for four Slovenian basic intonation types [Toporišič, 1969]. Read newspaper articles were processed by an AMDF (*A*verage *M*agnitude *D*ifference *F*unction) pitch extractor [Ross et al., 1974], thus mimicking the *reading* type of prosody. Then, the manual linearization of F_0 curves into pitch contours was performed. The obtained pitch contours were a rather rough approximations of stylised pitch contours as IPO suggests in [Collier, 1990].

According to [Toporišič, 1984], an intonation segment consists of a prosodically irrelevant body and a prosodic head. The prosodic head plays an important role in intonation determination since it incorporates all the major changes in a intonation contour. When there is no emphatic stress in the phrase, the intonation head begins with an interrogative pronoun, otherwise it is located on the last stressed syllable within a prosodic group.

Four basic intonation contours are described for the Slovenian language, given as relative changes that are to be made on previously defined prosodic parameters. *Declarative intonation* has a fall intonation contour (cadence) and is introduced by a full stop or a semicolon. The breaking point in the contour is associated with the intonation head, as described above. *Interrogative intonation* has a fall and rise (anticadence) intonation contour and is usually introduced by a question mark. The target vowel at the prosody head is lengthened. *Non-terminal intonation*, introduced by a comma, has a semi-cadence intonation contour. A semi-cadence intonation contour can be of a fall or of a rise type, but it is never as pronounced as a cadence or anticadence. Without a profound syntactic analysis the correct type cannot be chosen, therefore a rise type was chosen as it predominantly appears in the Slovenian language. *Exclamatory intonation* is introduced by an exclamation mark and receives an extended cadence or anticadence intonation contour, depending on the semantic meaning of the prosodic group. The cadence version was chosen as its usage is more frequent. The target vowel at the prosodic head is lengthened. Graphical schemes for the different types of intonation contours used are given in [6] [Gros et al., 1996].

Finally, the length and relative location of prosodic groups determine the insertion of pauses according to the type of grammatical categories. The basic idea is to introduce pauses after long prosodic groups in order to simulate breathing pauses and to reduce the mental load of the listener. However, the location of such pauses should be prosodically plausible.

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 analysis. A more sophisticated intonation model needs to be developed, like the Dutch stylisation model, the Fujisaki model or the rise and fall model [Taylor and Isard, 1992].

4. Diphone Concatenation

Once appropriate phonetic symbols and prosody markers are determined, the final step within a text-to-speech system is to produce audible speech by assembling elemental speech units. This is achieved by taking into account computed pitch and duration contours, and synthesising a speech waveform.

A concatenative synthesis technique was used. The TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) scheme enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications [Moulines and Charpentier, 1990] without considerably affecting the quality of synthesised speech.

Samples of synthetic speech produced by our text-to-speech system are available on the WWW on the address "`http://luz.fer.uni-lj.si/english/SQEL/synthesis-eng.html`".

Diphones were chosen for concatenative speech units. A diphone can be defined as a speech fragment which runs roughly from half-way one phoneme to half-way the next phoneme. In this way the transition between two consecutive speech sounds is encapsulated in the diphone and need not be calculated. Considering a diphone as a building block of spoken language, in order to generate a word, one simply concatenates the appropriate diphones.

Preparation of the Diphone Inventory

For every possible phoneme combination in a given language one diphone is required. A Slovenian diphone inventory comprising 955 pitch-labelled diphones was recorded, hand-segmented and hand-labelled in order to provide optimal coupling at concatenation points. Design and recording of the diphone inventory were given special attention. Diphone speech segments were extracted from nonsense plurisyllabic sequences (logatoms). The target diphone was always found within an unaccented syllable and the speaking rate was relatively slow, opting for a high intelligibility of the text-to-speech system. The diphones were placed in the middle of isolated plurisyllabic nonsense words, logatoms, pronounced with a steady intonation, except for the cases when the silence phone was part of the required pair: there the diphone was word-initial or word-final. Speech signals were recorded by a close talking microphone using a sampling rate of 16 kHz and 16 bit linear A/D conversion.

After the recording phase, logatoms were hand-segmented and the centre of the transition between the phonemes was marked, using information from both temporal and spectral representation of the speech signal. Finally, pitch markers were manually set for voiced parts of the corresponding speech signal. Figure 3 gives an example of the diphone *am* along with its spectrum.

While concatenating diphones into words it suddenly turned out that there was a large discrepancy between the duration of allophones, as suggested by the prosody module, and the actual corresponding diphone duration stored in the diphone inventory. This occured due to an exaggerated eagerness of the speaker trying to pronounce the meaningless logatoms in a correct and clear way. As a result, the quality of the synthetic speech was considerably affected and we have to record another diphone inventory.

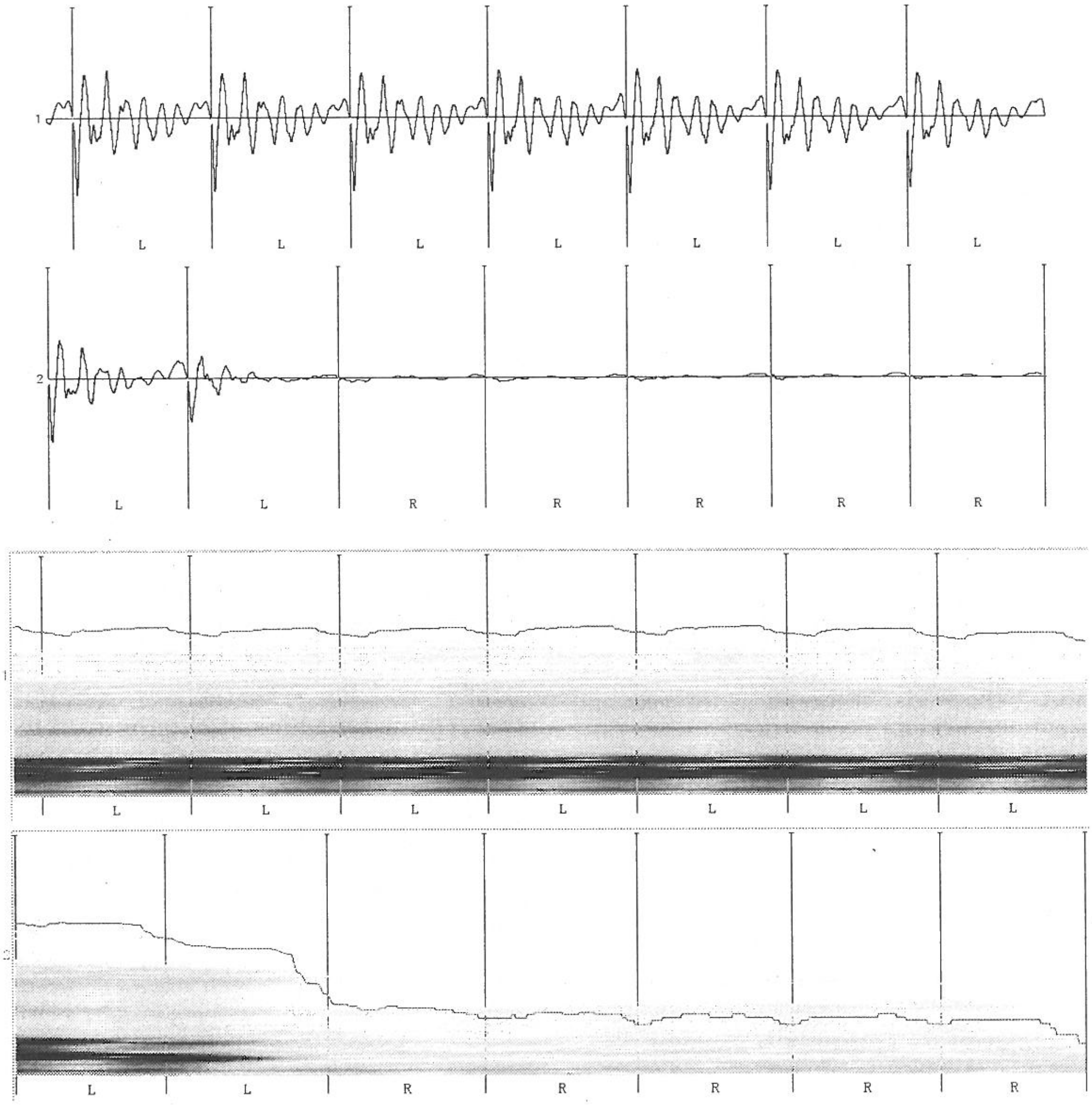

Fig. 3. Waveform *(above)* and spectral *(below)* representation of the diphone *am*. Markers *L* and *R* are set at the pitch periods of the left part of the diphone and of the right part respectively.

Automatic Diphone Segmentation

The preparation of the diphone inventory was rather laborious and time-consuming, since the whole process of extracting a diphone from a logatom was done manually. An automatic procedure for segmenting and pitch labelling of diphones should result in considerable reduction in preparation time of a new diphone inventory. It also provides a powerful tool for including new synthetic voices and for updating and supplementing existing diphone libraries. They can at least outline diphone boundaries, the refinement of which can be performed by a human expert. Therefore, in order to be able to synthesise speech in a variety of different voices, we decided to use procedures for automatic segmentation and pitch marking of spoken logatoms [5] [Gros et al., 1996].

First, logatoms from which diphones have to be excised, are segmented. A speech recogniser, based on Hidden Markov Models in forced segmentation mode is used to determine phone boundaries within spoken logatoms [Ipšić, 1996]. A statistical evaluation of manual and automatic segmentation discrepancies was performed on a much larger speech database than the logatom inventory itself, as proposed in [Schmidt, 1993]. The discrepancies are considerable. Most of the problems arise when detecting bursts of plosives as the automatic procedure tends to shorten their closures considerably. The situation improves when plosives are taken as a whole, closures and bursts together.

As a result, a fully automatic segmentation of speech segments is hardly conceivable in the context of concatenation synthesis. As most phonological units originate via phonological considerations rather than on acoustic grounds, isolating them requires a deep prior knowledge of their specific features. Unsupervised segmentation, i.e. segmentation on acoustic principles only, often results in segments and subsegments boundaries being misplaced, or just missing, while undefined ones appear. However, it can be used as a segmentation outline, the refinement of which has to be performed by a human expert.

Next, diphone boundaries need to be determined. As the concatenation point of the diphones corresponds to the centre of the phone, it is somewhere in the steady region of the phone. By studying the distances from the signal to the target values, [Ottesen, 1991] claims that minimal distances tend to be found just before the middle of a phoneme. This being a general trend, we decided to divide each phoneme duration in a fixed ratio, 40 and 60 %. Unvoiced plosives are an exception to this rule: they are divided just in front of the opening burst. An automatic diphone boundary detection algorithm, minimising spectral discontinuities at concatenation points [Taylor and Isard, 1992] may also be investigated.

Finally, pitch markers are determined for voiced parts of the signal. We applied the SRPD (Super Resolution Pitch Determination) algorithm as it allows precise pitch determination [Medan et al., 1991].

We expect the whole process of creating a new voice to be semi-automatic (with manual correction of stop-consonant boundaries), allowing the synthesiser to be retrained on a new voice in less than 3 days.

5. Performance assessment

Here we give some preliminary results of the Slovenian text-to-speech system performance assessment.

Adequacy of the synthesis system was tested in two ways: in terms of acceptability and in terms of intelligibility [Pols, 1994]. The synthesis output was directed to a Sound Blaster audio card. The experiment was performed in laboratory conditions with 10 subjects within the age span between 24 and 43, three of them being female. Another test based on ITU Reccomendations involving 20 test subjects is being prepared.

In our first experiment, intelligibility of synthesised speech was evaluated on three levels: segmental level, word level and phrase level. Subjects, participating in the test were asked to write down everything they heard. Figure 4 gives the percentage of correctly understood syllables and words, with word intelligibility rate being close to 80%. On the phrase level,

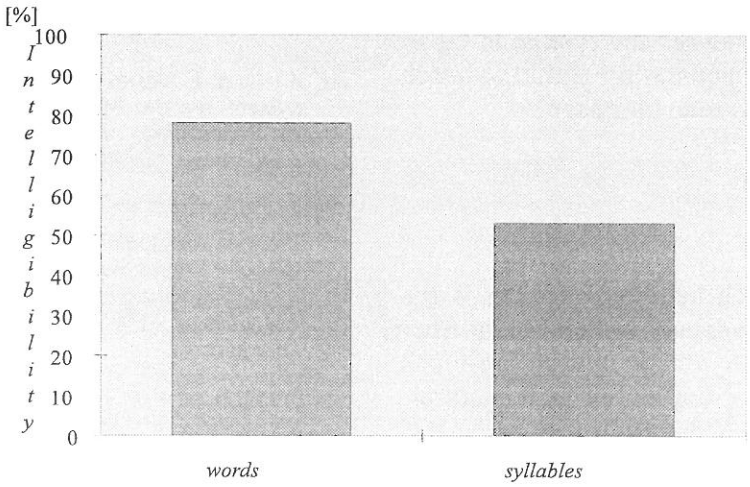

Fig. 4. Intelligibility test. Percentage of correctly understood syllables and words.

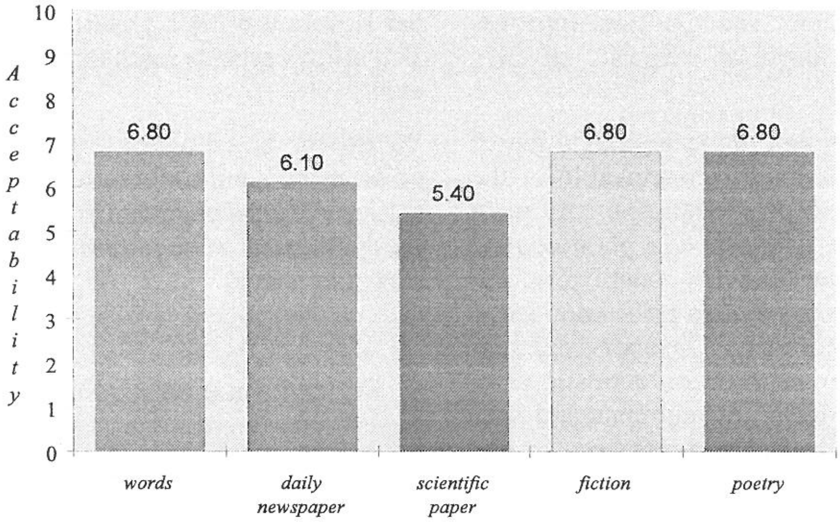

Fig. 5. Degree of naturalness of pronunciation for individual words and four different text styles.

different types of texts (daily newspaper, fiction, scientific paper, poems) were chosen and the ratio of word insertions, deletions and substitutions with respect to the number of all words within the text was computed.

In our second experiment the degree of acceptability of the synthesised speech was assessed, again on word and phrase levels. Subjects were asked to mark naturalness of pronunciation from 1 to 10, 10 being the highest mark. The results obtained are shown in Figure 5.

Despite a rather good intelligibility of the synthetic speech, utterances often lack more naturalness and fluency.

Majority of the subjects estimated the synthetic speech to be pleasant and quite natural sounding, sufficiently rapid and not over-articulated. All ten of them considered the system to be an appropriate tool for generating audible speech from text in the Slovenian language.

6. Conclusion

The described speech synthesis system is the first complete text-to-speech system for the Slovenian language. The synthetic speech produced by the system is intelligible, but lacks naturalness. Improvement of intelligibility and naturalness depend in particular on proper lexical stress assignment and on a more sophisticated generation of prosodic parameters. The first attempts at developing a diphone-based synthesis system for the Slovenian language are promising, so that further work on improving individual parts of the system is encouraged.

References

[1] J. BAKRAN, Model vremenske organizacije hrvatskog standardnog govora. PhD Thesis, University of Zagreb, 1994.

[2] R. COLLIER, On the perceptional analysis of intonation. *Speech Communication*, **9** (1990), 443–451.

[3] W. EEFTING, S. G. NOTEBOOM, Accentuation, information value and word duration: Effects on speech production, naturalness and sentence processing. In *Analysis and Synthesis of Speech* (V. J. Heuven and L. C. W. Pols, Eds.), (1993). Mouton de Gruyter.

[4] J. GROS, F. MIHELIČ, N. PAVEŠIĆ, A text-to-speech system for the Slovenian language. Presented at the *Proceedings of the European Signal Processing Conference*, (1996), pp. 1043–1046, Trieste, Italy.

[5] J. GROS, I. IPŠIĆ, S. DOBRIŠEK, F. MIHELIČ, N. PAVEŠIĆ, Segmentation and labelling of Slovenian diphone inventories. Presented at the *Proceedings of the International Conference on Computational Linguistics*, (1996), pp. 298–303, Copenhagen, Denmark.

[6] J. GROS, N. PAVEŠIĆ, F. MIHELIČ, S. DOBRIŠEK, A text-to-speech system for the Slovenian language. Presented at the *Proceedings of the 3rd Slovenian-German and 2nd SDRV workshop on Speech and Image Understanding*, (1996), pp. 47–56, Ljubljana, Slovenia.

[7] J. GROS, N. PAVEŠIĆ, F. MIHELIČ, Duration modelling in Slovenian TTS. Accepted for presentation at the *EUROSPEECH'97*, (1997), Rhodes, Greece.

[8] J. GROS,, Samodejno pretvarjanje besedil v govor. Ph.D. Thesis in preparation, University of Ljubljana, Ljubljana, Slovenia, 1997.

[9] J. HRIBAR, Sinteza umetnega govora iz teksta. MSc Thesis, University of Ljubljana, Slovenia, 1984.

[10] I. IPŠIĆ, Razpoznavanje besed v vezanem govoru. PhD Thesis, University of Ljubljana, 1996.

[11] Y. MEDAN, E. YAIR, D. CHAZAN, Super resolution pitch determination of speech signals. *IEEE Transactions on Signal Processing*, **39** (1991).

[12] E. MOULINES, F. CHARPENTIER, Pitch — Synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, **9** (1990). 453–467.

[13] G. E. OTTESEN, An automatic diphone segmentation system. Presented at the *Proceedings of the European Conference on Speech Technologies*, 713–716 (1993), Berlin, Germany.

[14] L. C. W. POLS, Synthesis performance assessment. Presented at the *Proceedings of the CRIM / FORWISS Workshop on Progress and Prospects of Speech Research and Technology*, (1994), Munich, Germany.

[15] M. J. ROSS, H. L. SCHAFFER, A. COHEN, R. FREUDBERG, H. J. MANLEY, Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics Speech and Signal Processing*, **5** (1977), 565–571.

[16] M. S. SCHMIDT, G. S. WATSON, The evaluation and optimization of automatic speech segmentation. Presented at the *Proceedings of the European Conference on Speech Technologies*, 701–704 (1993), Berlin, Germany.

[17] T. SREBOT REJEC, Word accent and vowel duration in standard slovene: an acoustic and linguistic investigation. Slawistische Beiträge, **226**, Verlag Otto Sagner, (1988), München.

[18] C. SORIN, D. LAURRUER, R. LLORCA, A Rhythm-Based Prosodic Parser for Text-to-Speech Systems in French. Presented at the *Proceedings of the XIth International Conference on Phonetic Sciences*, (1987), pp. 125–128, Tallin, Estonia.

[19] P. A. TAYLOR, S. D. ISARD, Automatic diphone segmentation. Presented at the *Proceedings of the European Conference on Speech Technologies*, 701–704 (1991), Genova, Italy.

[20] P. A. TAYLOR, S. D. ISARD, A new model of intonation for use with speech synthesis and recognition. Presented at the *Proceedings of the International Conference on Spoken Language Processing*, (1992), Banff, Canada.

[21] J. TOPORIŠIČ, Slovenska stavčna intonacija. *V. seminar slovenskega jezika, literature in kulture*. 1969.

[22] J. TOPORIŠIČ, *Slovenska slovnica*. Založba Obzorja. Maribor. 1984.

[23] S. WEILGUNY, Grafemsko-fonemski modul za sintezo izoliranih besed za sintezo slovenskega jezika. MSc Thesis. University of Ljubljana, Slovenia, 1993.

Received: December, 1996
Accepted: April, 1997

Contact address:
Jerneja Gros, Nikola Pavešić, France Mihelič
Faculty for Electrical Engineering
University of Ljubljana
Tržaška cesta 25, SI-1001 Ljubljana
tel: +386 61 17 68 316
fax: +386 61 12 64 630
e-mail: jerneja.gros@fe.uni-lj.si

JERNEJA GROS was born in 1968. She received her B.Sc. and M.Sc. degrees in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, in 1990 and 1994, respectively. In 1991 she became a research staff member at the Laboratory of Artificial Perception, Faculty of Electrical Engineering, University of Ljubljana, where she is currently working towards her Ph.D. in electrical engineering. Her research interests include text-to-speech synthesis, automatic speech recognition and understanding, digital signal processing and pattern recognition. Jerneja Gros is a member of the IEEE, ESCA, EURASIP and The Slovenian Pattern Recognition Society.

FRANCE MIHELIČ studied at the Faculty of Natural Sciences, Faculty of Economics and Faculty of Electrical Engineering and Computer Science, all at the University of Ljubljana. He received his B.Sc. degree in technical mathematics, the M.Sc. in operational research and the Ph.D. degree in electrical sciences in 1976, 1979 and 1991, respectively. Since 1978 he has been a staff member at the Faculty of Electrical Engineering in Ljubljana, where he is currently assistant professor. His research interests include pattern recognition, speech recognition and understanding and signal processing.

NIKOLA PAVEŠIĆ was born in Rijeka in 1946. He studied at the Faculty of Electrical Engineering, University of Ljubljana. There he received the B.Sc. degree in electronics, the M.Sc. degree in automatics and the Ph.D. degree in electrical engineering in 1970, 1973 and 1976, respectively. He was recipient of the Mario Osana Award in 1974, the Vratislav Bedjanić award in 1976, and the Boris Kidrič fund award in 1982. Since 1970 he has been a staff member of the Faculty of Electrical Engineering and Computer Science in Ljubljana, currently occupying the positions of the professor of systems, automatics and cybernetics and head of the Laboratory of Artificial Perception. His research interests include pattern recognition and image processing, speech recognition and understanding and information theory. He has authored and co-authored more than 100 papers and 3 books addressing several aspects of the above areas. Dr. Nikola Pavešić is a member of IEEE, the Slovene Association of Electrical Engineers and Technicians (Meritous Member), The Slovenian Pattern Recognition Society and the Slovenian Society for Medical and Biological Engineering. He is also a member of editorial boards of several technical journals.