TEXT-TO-SPEECH CONVERSION FOR MACEDONIAN AS PART OF A SYSTEM FOR SUPPORT OF HUMANS WITH DAMAGED SIGHT

Ljubomir Josifovski[1], Dragan Mihajlov[2], Dejan Gorgevik[2], Suzana Loskovska[2]
[1] Faculty of Mechanical Engineering - Skopje
Karpos II bb, 91000 Skopje, Macedonia, ljupco@ereb.mf.ukim.edu.mk
[2] Faculty of Electrotechnical Engineering - Skopje

Abstract - A subsystem for text-to-speech (TTS) conversion for Macedonian language as a part of a system for support of humans with damaged sight will be presented in this paper. The whole system includes recognition of printed Cyrillic text, archiving, TTS conversion and printing on a Braille printer. A subsystem for real-time TTS conversion from unrestricted text is under development. The whole architecture of the subsystem and some of its major modules will be presented here.

Keywords: text to speech, speech synthesis, neural networks

1. INTRODUCTION

The humans with damaged eyesight can still receive information by other senses, notably hearing and touch. Therefore written materials are recorded on audio tapes or printed in Braille writing. This is slow and costly process. Materials are restricted and not always actual. A system for helping people with damaged sight is under development (Mihajlov, 1993) as a joint collaboration between Faculty of Electrical Engineering and Department for Rehabilitation of Children and Youngsters with Damaged Sight "Dimitar Vlahov" in Skopje. The system provides automatic reading of printed Macedonian Cyrillic text, its archival, conversion to speech by text-to-speech system and printing on a Braille printer. This paper addresses the subsystem for text-to-speech conversion for Macedonian language.

2. A SYSTEM FOR SUPPORT OF HUMANS WITH DAMAGED SIGHT

Detailed description of a number of systems and devices for helping visually disabled persons can be found in (Gill, 1993). Typically, systems in the class of "reading machines" consist of scanner (or camera), Optical Character Recognition (OCR) and TTS engine. However, each of the systems is usually suited only for a particular language.

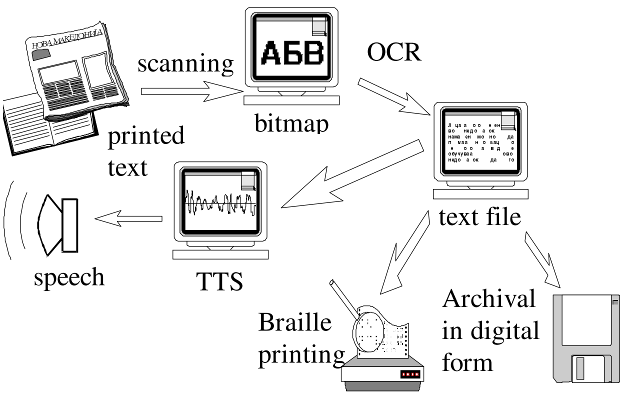

Fig. 1. A system for helping people with damaged sight

The conceptual scheme of our system is shown on Fig. 1. System includes a personal computer, a graphical scanner, a speech synthesizer and a Braille printer. Printed text found in newspapers, magazines, books etc. is scanned and enters the computer as bitmap. OCR on the Cyrillic text (Đorđević, 1995) is performed next. The result is an ASCII text file. This file can be printed on a Braille printer or feed to a TTS system. Once converted to text form, documents can be archived and reused for printing or speech synthesis.

3. AN ARCHITECTURE OF THE TTS SUBSYSTEM

Numerous research in the area of TTS conversion from unrestricted text has been undertaken in the last 30 years. As a result, number of systems such as MITalk (Allen, 1985), CNET (Stella, 1985), AT&T Bell Labs. multilingual TTS system (Sproat, 1995), TTS systems for Polish (Imiolezyk, 1994) and Japanese (Hirokawa, 1993) & (Kawai, 1994) etc., have been developed. However, ideal TTS system, indistinguishable from the human reading, is yet to be constructed.

All of the mentioned systems share the notion that the process of TTS conversion is to be divided into two major steps (Fig. 2). In the first step, input text is converted into some form of linguistic representation. Such representation includes information on phonemes to be produced, the duration, pitch and power contours. In the second step this information is converted into speech waveform - the final output of the whole TTS subsystem.

Text preprocessing includes expansion of abbreviations, conversion of numbers into words, dates, times, telephone numbers, etc., special handling of formulas, punctuation marks, hyphens. Grapheme (i.e. letters) to phoneme conversion methods include morphological analysis (Kawai, 1994), context sensitive letter-to-sound rules (Stella, 1985), or their combination (Allen, 1985). This conversion requires vast number of rules and exception dictionary for English, French, German. It is feasible in a number of languages (Spanish, Italian, Russian) with moderate effort and virtually no error. For Macedonian it's mostly trivial.

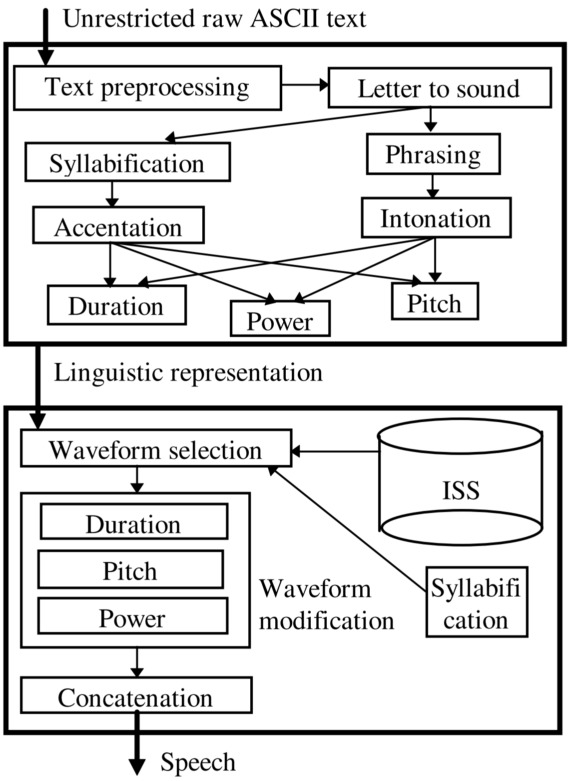

Fig. 2. Architecture of our TTS system

4. PROSODY GENERATION

Prosody is mostly comprised of accentuation and intonation. Prosodic models, as described in (Bruce, 1993) (Rossi, 1993), identify prosodic categories and map them to their phonetic correlates: duration, fundamental frequency (F_0) and intensity.

Lexical stress is property of the morphemes and affects all phonetic categories (F_0, duration, intensity). The detection is often rule-based (Allen, 1985), and carried out during the grapheme to phoneme conversion. For Macedonian, stress placement is trivial after the word syllabification (see Section 6).

Prosodic phrases are basic intonational units primarily related to F_0 (Bruce, 1993) (Imiolezyk, 1994), but can affect duration and power (Rossi, 1993). Punctuation marks, lists of functional words, decision tree trained on annotated corpus of text (Sproat, 1995) or relation to the word accent (Bruce, 1993) can be used to indicate good places to break.

Duration models (Bruce, 1993) (Imiolezyk, 1994) or employment of general purpose statistical methods (Hirokawa, 1993), both refined on large corpora of recorded and labeled speech, can be used for segments duration prediction. The more general sum-of-products model (Van Santen, 1994) predicts segment duration by summing the products of scaled factors. Each factor is some computable property of the segment.

Fundamental frequency contours are typically specified in hierarchical fashion. In (Allen, 1985) rule based approach is applied. In (Sproat, 1995) F_0/time pairs values are computed from various phrasal parameters using topline/baseline limit for speaker's pitch range. In (Imiolezyk, 1994) F_0 values are computed using functions with parameters extracted by fitting to the contours of read newspaper text.

The role of intensity in signaling prosody is discussed in (Grenstroem, 1992). The slope of the spectrum and relative level of fundamental change globally while speaking from weak to strong voice.

5. SPEECH SYNTHESIS FOR MACEDONIAN LANGUAGE

Waveform concatenation of syllables as a method for speech synthesis is chosen in our system. The aim is inventory of speech segments (ISS) with variable length stored with their phonemic context and their alteration entirely in time domain (Fig. 2).

Formant synthesis (Allen, 1985) (Imiolezyk, 1994) was rejected because it's impossible to acquire data and numerous rules necessary to operate such synthesizer for Macedonian at present time. Linear Predictive Coding (LPC) (Sproat, 1995) (Stella, 1985) synthesis was dropped as less natural and computationally more expensive compared to time-domain based methods (Dutoit, 1994) (Hirokawa, 1993).

Phonemes (Hirokawa, 1993), diphones (Stella, 1985) (Sproat, 1995), demmysyllables (Kraft, 1992) or multiple instances (with different F_0, duration and phonemic context) of one to several phones (Kawai, 1994) were possible inventory units (IUs) of choice. The need to eliminate the coarticulation effects on the IUs in greatest extent (because of lack of research in this area for Macedonian) was the main factor that influenced our decision.

The ISS consists of speech segments and labels about their preceding and succeeding environment (Kawai, 1994), pitch marks and duration. IUs are extracted from recording of a text uttered by a professional speaker (Kraft, 1992). Syllabic environment in carrier words ensures that syllable in question doesn't carry lexical stress, and can be easily modified both ways (stressed and unstressed). Integrated tool for segmentation and labeling aid is under development.

In the process of translation from a phonemic text to a list of IUs longest match is found (Sproat, 1995). Among several IUs, the one with with F_0, duration and phonemic context closest to the desired is chosen. Guides (Kawai, 1994) about the allowable F_0 and duration modification are also obeyed.

The chosen IU is modified according to F_0 and duration contours. In (Moulines, 1995) there is a thorough review of the non-parametric techniques for pitch and time scale modification both in frequency and time domain.. We applied classic time domain pitch synchronous overlap-add (TD-PSOLA). The waveform is passed through a 246 Hz low-pass filter (Hirokawa, 1993) and pitch marks are set at the local peaks. Original waveform is decomposed into a stream of short-time analysis signals by multiplying with two pitch periods wide time-translated Hanning windows, centered around the pitch marks.

Modification of the duration is performed using the model proposed in (Kubin, 1994) (see Section 7). Power control is trivial. Samples are simply multiplied by the desired factor.

After the proper modification, IUs are glued together to produce the final waveform. Pitch-synchronous cross fading is applied in order to smooth the overlapped junction. The exact beginning of the region is determined using fast algorithm to find the minimum absolute error (MAE) (Lin, 1995) between the apposite regions.

6. SYLLABIFICATION IN MACEDONIAN

Syllabification in Macedonian is necessary for lexical accent placement, and in the process of IU's selection from the ISS. Syllabification is considered as (probably) the first step in the phonetic analysis performed by humans (Korubin, 1955). Even kids have "feeling" for and can syllabify words. Considering the nature of grapheme-to-phoneme conversion in Macedonian, syllabification of graphemes is treated.

Two approaches to syllabification are proposed in (Apostolovska, 1987). The usage of complete dictionary of presyllabified words is discarded as unrealistic for the time being. The second method employs number of rules, together with small exception dictionary. We estimated that significant effort and time are needed for rule adaptation and evaluation. Therefore, we consider yet another approach: a neural network (NN) based estimator of the probability that there is a syllable break after a given letter (or phone - almost the same in Macedonian).

A classic feed forward NN with three layers, as shown on Fig. 3 is used. Units in apposite layers are fully connected, and the ones in the same layer are not. First layer has 217 (7 x 31) units. Each one of the 7 groups (with 31 units) corresponds to a single letter. As there are 31 letters in

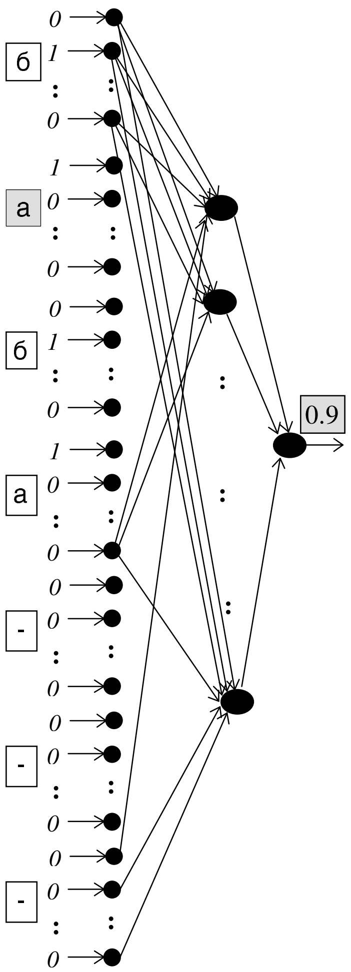

Fig. 3. The net architecture

Macedonian, the presence of a particular letter is modeled by turning the corresponding unit on. The other 30 units in the group are turned off. Syllable breaks depend on the letter together with it's context. The context is modeled by taking into account one letter before, and five letters after the particular letter. It's the smallest one yielding to consistent training and evaluation sets.

The hidden layer consists of 12 to 31 units. Training of nets with less then 12 units in hidden layer showed unsuccessful in reasonable time. Nets with as few units in the hidden layer as possible are preferred because of the insufficient number of elements in the training set (compared to the number of weights in the NN). The output layer consists of 1 unit, giving the probability for syllable break after a particular letter.

During the working phase, the letter and it's context are fed to the input of the network. There are 31 units reserved for each letter: the letter before, the letter in question, and five successive letters after. For each of seven letters, only the unit correspondent to letter value is turned on (set to 1). Other 30 units are turned off (set to 0). If there are no letters before or after, all corresponding 31 units are turned off. The activation function is the classic sigmoid $\frac{1}{1+e^{-x}}$. The output from the single unit in the last layer is in (0,1) range, 0 meaning absence, and 1 presence of syllable break. Fig. 3 shows an example of how the probability of syllable break existence is determined for the second letter ("a") in the word "baba".

Classic back propagation (BP) of error, and modified BP with variable step length (Jacobs, 1988), are used in the process of training. We considered other improvements to speed up the training process (Alpsan, 1995). The usage of *tanh* instead sigmoid transfer function significantly increased the learning speed. However, the learning process failed occasionally.

Weights are updated after the presentation of each training set pattern (pattern learning). The learning phase is performed only on missclassified patterns. The training set is extracted from a random text and contains 1953 words with 4617 syllable breaks. It carries 417 different contexts, i.e. training patterns.

We evaluated two trained nets, with 31 (net1) and 12 (net2) units in the hidden layer. Evaluation set is also extracted from random text and contains 1180 words with 3215 syllable breaks. Words in the evaluation set are different from the ones contained in the training set. We found 274 different contexts in the evaluation set, 209 of them being new (not found in the training set).

The nets performances are shown on Fig. 4. We counted number of insertion and deletation errors in syllable breaks placing. Given that 76.28% of the contexts in the test set were different from the ones in the training set, the nets show high generalization ability. That was our primary concern: the ability of the net to deduct and generalize the syllabification rules from the examples presented in the training set, and apply them in unknown contexts.

Two nets with different number of hidden units are trained on the same training set in order to investigate the relation with their generalization ability. Better performance of the smaller net on the same evaluation set indicate that both nets are undertrained. We couldn't find strict rules about the number of training patterns vs. number of weights ratio in the literature. But common approach is that training set should contain at least twice the number of weights elements. That means at least 5232 different contexts (instead of 417) in the training set. However, the purpose was to investigate if the neural network based approach to the syllabification problem leads to satisfactory results. We believe that the first results show so.

	Evaluation set		Words with syll. errors		Misplaced syll. breaks	
	No. words	No. syll. breaks	No.	%	No.	%
Net1	1180	3215	112	9.49	137	4.26
Net2	1180	3215	93	7.88	96	2.99

Fig. 4. Nets performances

7. TIME - SCALE MODIFICATION

The duration is modified according to the model and technique introduced in (Kubin, 1994). Speech is modeled as a nonlinear oscillator. Short-time stationary intervals which characterize vowels are interpreted as attractors of the underlying system. Noise-like waveform consonant patterns are transitions from one attractor to another.

The oscillator is defined as nonlinear feedback system using state space representation. The N-dimensional state vector X(n) is obtained from N delayed samples: X(n) = [x(n), x(n-M), ... x(n-(N-1)M)]. The delay parameter M defines the "subsampling" factor and is not necessary equal to 1. Product NM defines the duration of the state vector.

The state transition function a(X) is a nonlinear mapping from X(n-1) to X(n). Actually, only the most recent component x(n) of X(n) is produced. Other (N-1) components are shifted components of X(n-1), as shown on Fig. 5. The unknown component x(n) of X(n) is estimated using the "nearest-neighbor" method. First, a table of all state transitions observed in given waveform is constructed. Known the current value of the state vector, the output value is the table entry nearest to the state vector. Oscillator stability is guaranteed, at a price of lack of any noise suppression mechanism.

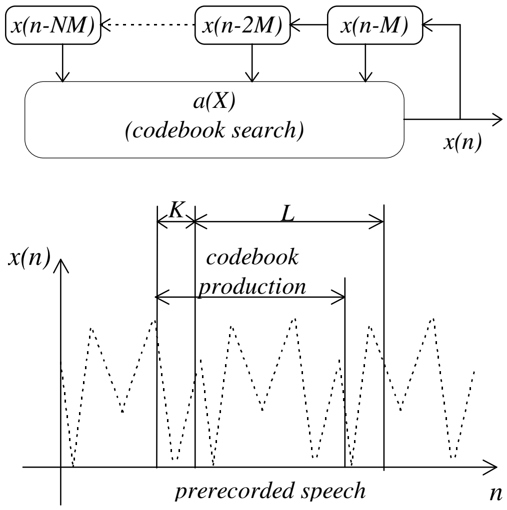

Fig. 5. Speech modeled as nonlinear oscillator

The model is adapted to the short-time nature of speech by adaptation of the state-transition codebook approximately every 5 ms (see Fig. 5). Speech is divided in frames of length L (20 to 40 ms). Next frame advances over the previous with step K (5 ms or less). All samples in a frame are used to build the state-transition codebook of size L x (N+1) samples. Each of the L codebook entries consists of a short waveform segment of length N and an immediately following sample. Once the codebook is built, signals of arbitrary length can be synthesized from free-running oscillator. For time-scale modification, after the frame is advanced over previous and new codebook is built, free-running oscillator produces RK samples (R is time-scaling factor).

Codebook search computational cost can be greatly reduced if performed only once every J samples. Once the best match is found, J subsequent samples are copied from the original waveform. But clearly audible discontinuity appears in the output waveform on the place where sample is obtained by codebook search. In (Kubin, 1994), linear predictive smoothing is used as an alternative to costly time-domain interpolation of the original waveform. However, linear predictive smoothing is itself a costly procedure if the speech coder is not LPC based. Instead, we implemented computationally less expensive cross-fading with satisfactory results.

8. CONCLUSION

We presented a subsystem for real-time TTS conversion for Macedonian in this paper. It's part of a system for support of humans with damaged sight beside TTS conversion includes recognition of printed Cyrillic text, it's archival and printing on a Braille printer. The architecture of the subsystem as a whole and the modules for syllabification and time scale modification were presented. Lack of research for Macedonian language in the area of TTS conversion influenced proposed solutions for old and well known problems in this area in greatest extent.

REFERENCES

Mihajlov D. Djordjevik D. Kotevska N. (1993), "Computer System for Support of Humans with Damaged Sight", ETAI, Ohrid

Gill J. M. Peuleve C. A. (1993), Research Information Handbook of Assistive Technology for Visually Disabled Persons, The Tiresias Consortium

Đorđević D. Mihajlov D. Josifovski Lj. (1995), "Kompjuterski sistem za pomoć osobama sa oštećenim vidom: Podsistem za optičko prepoznavanje štampanog ćiriličnog teksta", YU Info, Brezovica

Allen J. (1985), "Speech Synthesis From Unrestricted Text", In Computer Speech Processing, ed. by F. Fallside and A. W. Woods, Prentice-Hall Int, London, pp. 461-477

Stella M. (1985), “Speech Synthesis”, In Computer Speech Processing, ed. by F. Fallside and W. A. Woods, Prentice-Hall International, London, pp. 421-460

Sproat R. W. Olive J. P. (1995), “Text-to-Speech Synthesis”, AT&T Technical Journal, Vol. 74, No. 2, pp 35-44

Imiolczyk J. Nowak I. Demenko G. (1994), “High Intelligibility TTS Synthesis for Polish”, Archives of Acoustics, Vol. 19, No. 2, pp 161-172, 1994

Hirokawa T. Itoh K. Sato H. (1993), “High Quality System Based on Waveform Concatenation of Phoneme Segment”, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. 76, No. 11, pp 1964-1970

Kawai H. et al. (1994), “Development of a TTS for Japanese Based on Wave Form Splicing”, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc, Adelaide, Vol. 1, pp. 569-572

Bruce G. Granstroem B. (1993), "Prosodic Modeling in Swedish Speech Synthesis", Speech Communication, Vol. 13, No. 1/2, pp 63-73

Rossi M. (1993), "A Model for Predicting the Prosody of Spontaneous Speech (PPSS Model)", Speech Communication, Vol. 13, No. 1/2, pp 87-107

Van Santen J. P. H. (1994), "Assignment of Segmental Duration In Text-to-Speech Synthesis", Computer Speech & Language, Vol. 8, No. 2, pp 95-128

Grenstroem B. Nord L. (1992), "Neglected Dimensions in Speech Synthesis", Speech Communication Vol. 11 No. 4/5, pp 459-462

Dutoit T. (1994), “High Quality TTS: A comparison of Four Candidate Algorithms”, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc, Adelaide, Vol. 1, pp 565-568

Kraft V. Andrews J. R. (1992), "Design, Evaluation and Acquisition of a Speech Database for German Synthesis-by-Concatenation", Proc. of 4th Australian Int. Conf. on Speech Science and Technology, Brisbane, pp 724-729

Moulines E. Laroche J. (1995), “Non-parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech”, Speech Communication, Vol. 16, No. 2, pp 175-205

Kubin G. Kleijn W. B. (1994), “Time-Scale Modification of Speech Based on a Nonlinear Oscillator Model”, Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Proc, Adelaide, pp 453-456

Lin G.-J. Chen S.-G. Wu T. (1995), “High Quality and Low Complexity Pitch Modification of Acoustic Signals”, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 5, pp 2987-2990

Korubin B. (1955), „Pogled na slogovnoto delewe na Makedonskiot jazik", Makedonski jazik, god. VI, kniga 1-2, str. 22-43, Institut za Makedonski jazik, 1955, Skopje

Apostolovska S. (1987), „Statisti~ka analiza na slogovi vo Makedonskiot jazik", Diplomska rabota, Skopje, 1987

Jacobs A. R. (1988), "Increased Rates of Convergence Through Learning Rate Adaptation", Neural Networks, Vol. 1, pp 295-307

Alpsan D. et al. (1995), "Efficacy of Modified Backpropagation and Optimization Methods on a Real-world Medical Problem", Neural Networks, Vol. 8, No. 6, pp. 945-962