FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 27, N^{o} 3, September 2014, pp. 467 - 477
DOI: 10.2298/FUEE1403467P

AUTOMATIC PROSODY GENERATION IN A TEXT-TO-SPEECH SYSTEM FOR HEBREW

Branislav Popović[1], Dragan Knežević[1], Milan Sečujski[1], Darko Pekar[2]

[1]Faculty of Technical Sciences, University of Novi Sad, Serbia
[2]AlfaNum – Speech Technologies, Novi Sad, Serbia

Abstract. *The paper presents the module for automatic prosody generation within a system for automatic synthesis of high-quality speech based on arbitrary text in Hebrew. The high quality of synthesis is due to the high accuracy of automatic prosody generation, enabling the introduction of elements of natural sentence prosody of Hebrew. Automatic morphological annotation of text is based on the application of an expert algorithm relying on transformational rules. Syntactic-prosodic parsing is also rule based, while the generation of the acoustic representation of prosodic features is based on classification and regression trees. A tree structure generated during the training phase enables accurate prediction of the acoustic representatives of prosody, namely, durations of phonetic segments as well as temporal evolution of fundamental frequency and energy. Such an approach to automatic prosody generation has lead to an improvement in the quality of synthesized speech, as confirmed by listening tests.*

Key words: *speech synthesis, speech processing, natural language processing, classification and regression trees*

1. INTRODUCTION

Explicit modeling of prosodic features of synthesized speech, as well as prediction of values of certain parameters of a model based on explicit morphological, phonetic, syntactic and other relevant rules, is considered to be a relatively poor solution in practice. This is due to an enormous number of factors that need to be considered, as well as their mutual influence, too complicated to be closely examined on reasonably large speech corpora [1]. On the other hand, inadequately determined prosodic features impair the naturalness, and in some cases even the intelligibility of synthesized speech, significantly narrowing the field of its application.

As the use of machine learning methods eliminates the need for explicit modeling of prosody, they have been widely adopted as a solution for automatic prosody generation

Received February 25, 2014; received in revised form May 21, 2014
Corresponding author: Branislav Popović
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
(e-mail: bpopovic@uns.ac.rs)

within text-to-speech systems. Furthermore, they can also provide information about the mutual influence of specific linguistic factors (e.g. masking), which is of great interest to the linguistic community.

In this paper, automatic training and subsequent prediction of prosodic features are carried out according to the methodology of classification and regression trees (CART) [2]. The idea of this methodology is to generate a tree structure through the process of automatic training based on a speech corpus of sufficient size. Such a training should identify the most relevant factors that influence the prosodic features of speech and their acoustic representatives – phone durations as well as temporal evolution of fundamental frequency and energy. The speech corpus is marked for phone boundaries as well as relevant prosodic events, such as types and levels of boundaries between adjacent intonation units, as well as levels of emphasis. Using regression trees trained on thus annotated speech corpus, the quality of synthesized speech is significantly improved compared to the quality obtained by conventional methods for prosody prediction in text-to-speech [3], [4], [5].

The paper is organized as follows. Section 2 presents the particularities of the Hebrew language, as it is well known that the properties of the target language significantly affect the development of a system for automatic speech synthesis (most notably the automatic prosody generation module). Section 3 defines the procedure of automatic part-of-speech (POS) tagging and additional morphological annotation of input text. In Section 4, prosody generation and synthesis are presented. Section 5 presents the experimental results. In Section 6, several conclusions are given.

2. LANGUAGE PARTICULARITIES

The Hebrew language, one of the most widely spoken Semitic languages today, has a range of properties which drastically affect the design of a speech synthesis system. Firstly, from the orthographical point of view, it belongs to the group of so called *abjad* languages, where each symbol commonly stands for a consonant [6]. However, vowels can be indicated by (1) the use of "weak consonants" serving as vowel letters (for example, the letter *vav* indicates that the preceding vowel is either /*o*/ or /*u*/, *yodh* indicates an /*i*/, whereas *aleph* indicates an /*a*/), or (2) by using a set of diacritical symbols called *niqqud*. Another thing that should be borne in mind is that abjad languages, including Hebrew, suffer from very loose spelling rules. This means that for a number of words there can be more than one acceptable spelling, which is a very serious source of ambiguity. Namely, the revival of the Hebrew language in the late 19th century has left many unresolved issues [7]. As Hebrew speakers were almost all native speakers of European languages and thus accustomed to the Latin alphabet, it has led to the development of two parallel spelling systems: the first, where vowel indicators are used according to the historic rules, and the second, where vowel indicators are used excessively. It should also be noted that even today, a vast majority of speakers commonly makes spelling errors. Therefore, if one aims at the design of a text-to-speech system which should be able to handle arbitrary texts, spelling errors have to be accepted as a part of standard inventory. Spelling errors are thus another source of ambiguity in Hebrew, and are something that the design of a practically applicable speech synthesizer cannot dismiss.

The Hebrew alphabet has 22 letters, five of them have different forms when they are used at the end of a word. Modern Israeli Hebrew has 5 vowel phonemes. However, the meaning of a word is carried not only by its phonological content, but also by its stress, and it is not uncommon to find pairs of words containing the same string of phonemes, but pronounced differently, the only difference being the stress.

From the point of view of morphology, it should be noted that Hebrew exhibits a pattern of stems consisting typically of consonantal roots from which nouns, adjectives, and verbs are formed in various ways. Hebrew uses a range of very productive prefixes and a multitude of suffixes, dramatically increasing the number of possible morphological interpretations of each surface word form in the text.

The syntactic structure of the sentence and the word ordering in Hebrew can be considered as relatively flexible. Although particular choices in word ordering can indicate specific literary styles or genres, one commonly encounters sentences where several orders of words can be considered equivalent. This is another source of difficulty for automatic morphological annotation of text.

3. Morphological Annotation

After the text is preprocessed in order to locate sentence boundaries and reveal elements such as abbreviations, dates, punctuation, special characters, web addresses etc., it is submitted to automatic morphological annotation, aimed at assigning part-of-speech tags as well as some additional morphological information that may be of interest to any subsequent phase of automatic prosody generation.

The morphological analysis begins by assigning an empty array of "readings" to every surface word form (token) in a sentence. The term "reading" denotes a morphological interpretation of this token together with its phonological representation, i.e. a particular inflected form of a word, together with the corresponding lemma, values of part-of-speech and corresponding morphological categories, its pronunciation as well as position and type of stress. In general, it is possible to derive several hundreds of morphological forms from a single lemma in Hebrew. Ideally, the lexicon should contain entries representing each and every possible surface word form. An evaluation score will be assigned to each of the readings of a word token during the evaluation process, in order to select the reading which is most likely to be correct. The aim of morphologic analysis is, thus, to distinguish between the available readings and thus assign a correct vocalization and stress pattern to each word, which is of utmost importance for the naturalness of synthesized speech.

The novel approach to morphologic analysis described in this paper is outlined in Fig. 1 and uses a combination of active and passive methods [8]. The passive method presumes the selection of appropriate lexemes, by using the Hebrew lexicon, the lexicon of foreign words in Hebrew transcription and finally, the lexicon of frequent foreign words in Latin transcription. The active method involves an automatic morphological analysis of the input text string, as well as generation of appropriate readings by using a complex expert algorithm relying on a set of transformational rules. The use of the active method reduces the initialization time as well as the number of inflected morphological forms in the lexicon by two orders of magnitude, enabling the use of the software component within real-time applications. On the other hand, the passive methodology reduces the error rate.

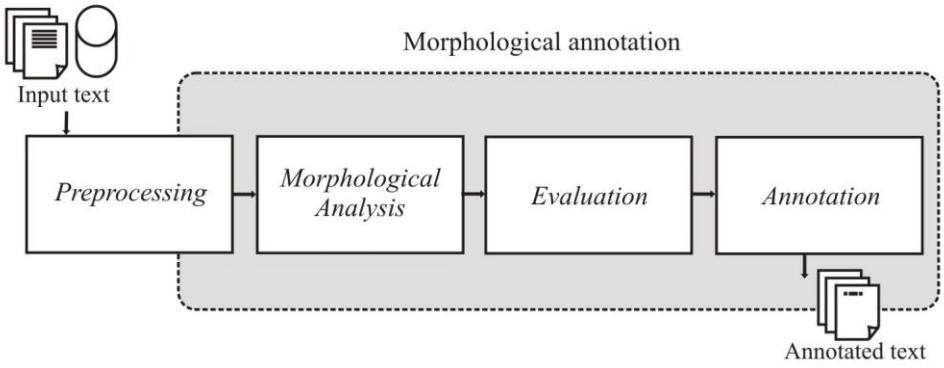

Fig. 1 Morphological annotation of input text

Transformational rules in the form of complex tree structures are applied iteratively. Branches are generated by using appropriate sets of morphological rules. Word analysis is carried out morpheme by morpheme. Every word is processed according to its left and right context. The aim is to correctly identify the surface form as a particular inflected form of a particular lemma. Currently, the system supports more than 30 part-of-speech classes with more than 3000 corresponding morphological categories.

The algorithm for the evaluation of particular readings, in order to select the most likely one, consists of a set of disambiguation tools, divided into individual scoring procedures.

The scoring of *syntactic structures* assigns syntactic indexes to words using predefined statistical algorithms, aiming at establishing the similarity between the syntactic structure of input sentence and the predefined syntactic structures. The algorithm is coupled with an accurate comparison mechanism that allows the use of existing structures in order to project on unfamiliar ones. A syntactic score indicates the level of compatibility of a certain reading to the previously tagged syntactic environment.

The scoring of *semantic structures* uses an analogous method, with only one difference: the structures represent semantic relations instead of syntactic ones. The index used is built over semantic attributes. The challenge in this process, besides building the most convenient set of indexes, is to determine the collection of a minimal number of morphological descriptors (tags) covering at the same time the maximum number of words.

Proximity scoring is the most efficient of the scoring processes. There are three types of proximity rules: generic to generic (this type of rules refers to the assignment of a relationship between linguistic items of non-specific identity, such as "there is a high probability that a verb in past tense of semantic category moving will be adjacent to a copula"; the attributes that can be used in composing these rules may be of grammatical and/or semantic nature), specific to generic (this type of rules would attach a generic rule to a specific word, e.g. "a verb in passive mood is likely to be followed by the word by") and specific to specific (this type of rules will attach two specific words, e.g. Tel is likely to be followed by Aviv). The effect of proximity scoring is clearly limited only to the words and entities for which proximity rules have been defined.

Full-niqqud scoring is a type of scoring unique to Hebrew. It determines how close a certain reading of a word is to the most commonly used spelling version. Due to the

previously mentioned lack of unique spelling standard, such a scoring procedure has to be taken into account as well.

Another scoring procedure used is *frequency scoring*, i.e. scoring readings according to their frequency in standard texts. Although such a procedure is highly inaccurate on its own (it commonly serves as a baseline for establishing the performance of more sophisticated morphological annotation techniques), it can serve as an efficient tie-breaker, i.e. it can be used in cases where other scoring procedures have assigned approximately equal scores to multiple readings. Every reading is also additionally evaluated in view of its context.

Context scores are obtained in compliance with the previously selected set of tags for the left context, as well as the set of tags for all possible readings in the right context. This is probably the most complex among all the applied scoring procedures.

Table 1 illustrates the effectiveness of the described scoring procedures, in terms of the overall accuracy of the automatic annotation process (selection of the correct reading), on the corpus of 3093 sentences (55046 words).

Table 1 The overall accuracy

Scoring type	Status								
Syntactic	*on*						*on*	*on*	*on*
Semantic		*on*					*on*	*on*	*on*
Proximity			*on*				*on*		*on*
Full niqqud				*on*					*on*
Frequency					*on*				*on*
Context						*on*		*on*	*on*
Acc. [%]	92.3	85.9	44.7	45.1	32.1	46.9	99.3	99.4	99.6

Table 2 presents the correlation matrix among the different scoring procedures. A high correlation between proximity, context and full-niqqud score can be noted. Although such an analysis of the correlation between different scoring procedures is not immediately aimed at the improvement of the quality of synthetic speech, it can give an insight into the directions of the future development of the scoring system. At the same time, high correlation between particular scoring procedures, besides giving a linguistic insight into the problem, confirms the validity of the algorithms.

Table 2 The correlation matrix

Scoring type	Syntactic	Semantic	Proximity	Full niqqud	Context
Syntactic	1	0.062	0.238	0.224	0.239
Semantic	0.062	1	0.356	0.364	0.342
Proximity	0.238	0.356	1	0.945	0.982
Full niqqud	0.224	0.364	0.945	1	0.929
Context	0.239	0.342	0.982	0.929	1

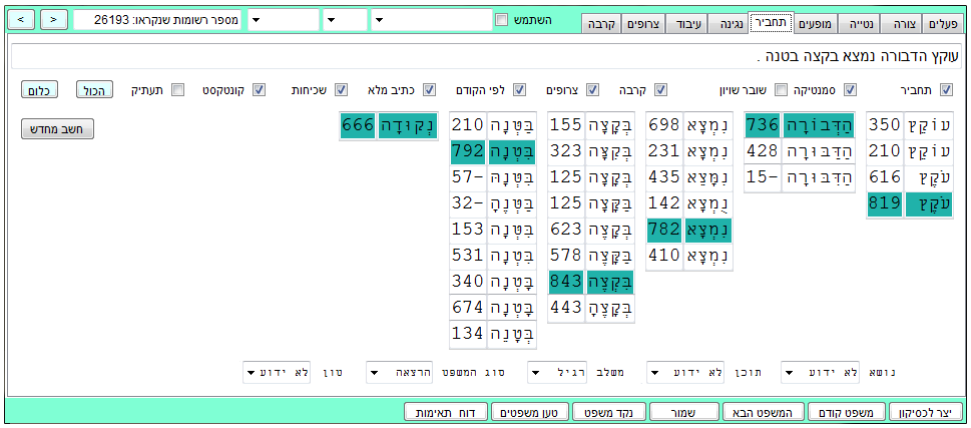

Fig. 2 Evaluation scores and manually selected readings

Evaluation scores for an example sentence are presented in Fig. 2. The sentence is given in the top right corner, and the readings with the highest scores (highlighted) match the actual correct readings.

Features recovered by automatic morphological annotation (primarily vocalization and stress pattern) constitute the symbolic representation of the prosody of a given input sentence. This representation will be used as an input to the CART prosody generator, which will, in turn, produce a corresponding sequence of values of fundamental frequency and energy, as well as phone durations.

4. PROSODY GENERATION AND SYNTHESIS

As has been mentioned before, it is well known that fully expert systems used for modeling of prosodic features are not of great practical use within speech synthesizers, mostly due to the large number of factors that influence prosody as well as their mutual effects, which are too complex to be sufficiently analyzed on speech corpora of reasonable size. Speaker inconsistence represents an additional problem. Even a single speaker can be expected to pronounce the same sentence differently on different occasions, each of the resulting utterances being equally acceptable to the listener. For all these reasons, the prediction of prosodic features is performed using machine learning, namely the methodology of classification and regression trees (CART) [9].

The basic principle of CART prosody prediction will be shown on an example of predicting the durations of phonetic segments (phones). The initial and the most important step is to identify the features to be used for training. This step has some basis in expert knowledge but the rest of the procedure is completely automatic. The set of features considered to be relevant for the phone duration includes phonemic identity, primary and secondary stress (with values: *stressed*, *unstressed*; applicable to vowels only), position within the syllable and position within the intonation boundaries (expressed as number of syllables), but many others as well. The durations of phones and relevant features are known for the training set and this set is thus the basis for prediction of duration for all other phoneme instances.

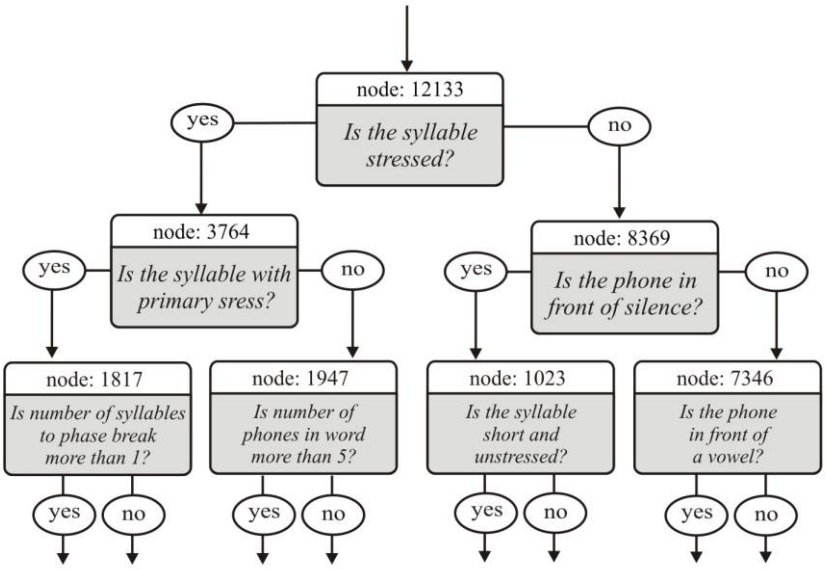

Fig. 3 The first 3 levels of the regression tree used for estimation of phone duration

The tree branching is performed as follows. All the possible YES/NO questions based on the selected features (e.g. "*Is the phone stressed?*", "*Is the distance to the nearest phrase break more than 3 syllables?*" etc.) are evaluated for each phone instance in the training set. Every question splits the starting *N* phoneme instances ("root" node) into two distinct subsets ("child" nodes) based on the answer (YES or NO), and every question generally splits the set differently. The most relevant question is the one that reduces the total diversity (in terms of duration) of both "child" nodes to the greatest possible degree. At this point, the initial node is split into two "child" nodes based on the most relevant question (e.g. "*Is the phone stressed?*"), and the procedure is recursively repeated for every descendant node, until the tree is fully branched. Every terminal node ("leaf" node) is assigned a value – the average duration of all instances assigned to that node. The final tree usually contains multiple phoneme instances assigned to each "leaf" node. Although the branching procedure is very computationally complex, the final use of the tree is exceptionally simple and fast. During the synthesis phase, the instance of the phone with known answers to all the relevant YES/NO questions is propagated through the tree – from the root node to one of the leaf nodes. The exact path to the leaf node and the final node itself depend on the answers to YES/NO questions. The estimated phone duration is the one assigned to the "leaf" node during the training phase (average duration for all the instances assigned to that node). As an illustration, Fig. 3 shows the first 3 levels of the regression tree for the prediction of phone duration. The number within the node indicates the occupancy, i.e. number of phone instances within the node.

The module for automatic prediction of prosodic features of the synthesized speech based on the regression trees for the Hebrew language is trained on the speech database which consists of approximately 4 hours of speech from one professional speaker (the same database is used for synthesis). The database is annotated for phone boundaries and

phonological content, which corresponds to the phonological inventory of modern Israeli Hebrew. Some phones are split into subphones (such as occlusions and explosions of stops and fricatives). Stress is also marked (primary and secondary). For the purposes of CART training, the database is marked for a number of prosodic events including types and levels of intonational phrase boundaries (*up*, *down*; *none*, *weak*, *medium*, *strong*, *very strong*) as well as levels of emphasis (*very weak*, *weak*, *neutral*, *strong*, *very strong*).

Regression trees are trained for duration, energy, the value of F0 and its derivative, log ratio of F0 values at 1/4 and 3/4 of the duration of a vowel, as well as log ratio of F0 values between two successive vowels (measured at 3/4 of the duration of the first vowel and 1/4 of the duration of the second one). Energy and durations are directly obtained, while the final F0 curve is derived from the outputs of the 4 F0-related trees.

A total of 600 different criteria (YES/NO questions) are taken into account during the process of regression trees branching. These criteria are defined based on the phonetic context, type of phoneme, phoneme position within a word, the corresponding word's position within the sentence, etc. A number of compound criteria are also used (e.g. "Is the phone vowel AND stressed?"). In this case, with a training corpus of approximately 4 hours of speech, the maximum number of levels in the trees was 11. However, it should be pointed out that this value is, in general, greatly dependent on the criterion used for stopping the branching procedure (e.g., a number of instances in the node is less than some predefined threshold, or the reduction of the impurity of the node has been reduced by branching by a value which is less than some predefined threshold).

After the trees have been built, at synthesis time, the expert systems analyze the input text and attempt to recover the correct reading for each word in it. By doing so, they recover the symbolic representation of the desired prosody for the input text, including the positions of stressed syllables as well as types and levels of intonational phrase boundaries and levels of emphasis for each word. These features exactly correspond to the features used in CART questions, and will be used for "passing" each phoneme of the input sentence down the tree, thus providing the acoustic representation of the desired prosody.

After the acoustic representatives of prosody have been generated, segments used for speech signal synthesis are selected. The basic unit on which the segment selector operates is a half-phone. Half-phones that are selected as candidates to be used for concatenation are assigned concatenation and target costs. A trellis structure is formed and the Viterbi algorithm is used to find the optimal path (half-phone sequence) through the trellis, i.e. the one with the minimal accumulated cost. The cost assignment is performed based on multiple criteria, which can be classified into two basic groups: *target criteria* and *concatenation criteria*.

The target criteria determine the mismatch between the acoustic features of the candidate half-phone and the required prosodic features, and express it through target cost, which is thus the measure of the unsuitability of the phonetic segment for being used in actual synthesis. The features taken into account for target cost are duration, F0 and its derivative, as well as energy.

On the other hand, the concatenation criteria determine the cost of concatenating any two half-phones [10]. The quality of the synthesized speech greatly depends on the frequency of concatenation points, as well as the audibility of each of them. The concatenation cost, assigned to any ordered pair of half-phones, is defined as the measure of their acoustic mismatch at concatenation points and thus their incompatibility for being

concatenated. For pairs of half-phones which are adjacent and in the same order as in the speech database this cost is equal to zero, which means that such pairs of segments will, whenever possible, be selected for concatenation. In other words, the basic units for synthesis are thus, in fact, not limited to half-phones, but can include strings of half-phones of unlimited length. In practice, the strings of half-phones selected for concatenation are mostly between 3 and 5 half-phones long.

The speech signal synthesis module performs signal concatenation. This module is based on the *Time-Domain Pitch Synchronous Overlap and Add* (TD-PSOLA) algorithm, as implemented previously in [11]. The outputs of the prosody generator module and the segment selection module are used as inputs for the concatenation module. Since it is impossible very unlikely to have the segments that ideally match the prosody requirements, it is usually necessary to additionally adjust the selected segments as regards their durations, F0 and/or energy.

5. The Quality of Speech

It should be noted that there are several independent sources of the differences between the prosody of synthesized speech and the prosody of natural human speech. Besides the intrinsic variability of speech prosody (the fact that no speaker will pronounce the same utterance twice in the same way, and that a wide range of the values of prosodic parameters can be considered acceptable), there are two major factors that affect the accuracy of synthetic prosody. Firstly, any error in morphologic annotation (and thus stress assignment) or the assignment of some other prosodic event such as phrase break or emphasis will lead to an error at the input of CART based prosody predictor. This would inevitably result in audible prosodic errors. On the other hand, even in cases when the input to CART is quite accurate, the output still may be of inferior quality due to corpus tagging errors (largely eliminated through manual inspection), data sparsity (insufficient training corpus size), inadequately estimated feature set or simply the intrinsic inability of the CART technique to adequately cover all the peculiarities of spoken language. The errors introduced by CART are most often less audible, and the final outcome is an intonation contour characteristic of accurate, albeit somewhat emotionless speech.

The evaluation of the proposed automatic prosody generation module was carried out through the perceptual evaluation of the quality of synthesis. Within the listening tests, 10 listeners (native speakers with no background in speech processing, text-to-speech synthesis or speech prosody) rated the TTS system performance in terms of naturalness of synthesized speech on a scale from 1 (unnatural, robotic speech) to 5 (speech with apparently natural prosodic features). The listeners were presented with examples of synthesized speech using either the proposed CART-based generator or its previous version based on an expert system implementing explicit rules governing prosodic features. The utterances (a total of 20) were not marked, and their ordering was varied. The average score given to the CART-based system was 3.9, as opposed to 3.5 given to the rule-based version (the corresponding standard deviations were 0.39 and 0.41 respectively). Figure 4 shows a comparison of three fundamental frequency contours for the sentence 'תיטמוטוא הארקה תכרעמ תועצמאב תעמשומ תאז העדוה', corresponding to the utterance as rendered by the native speaker (blue), referent system [5] (grey) and

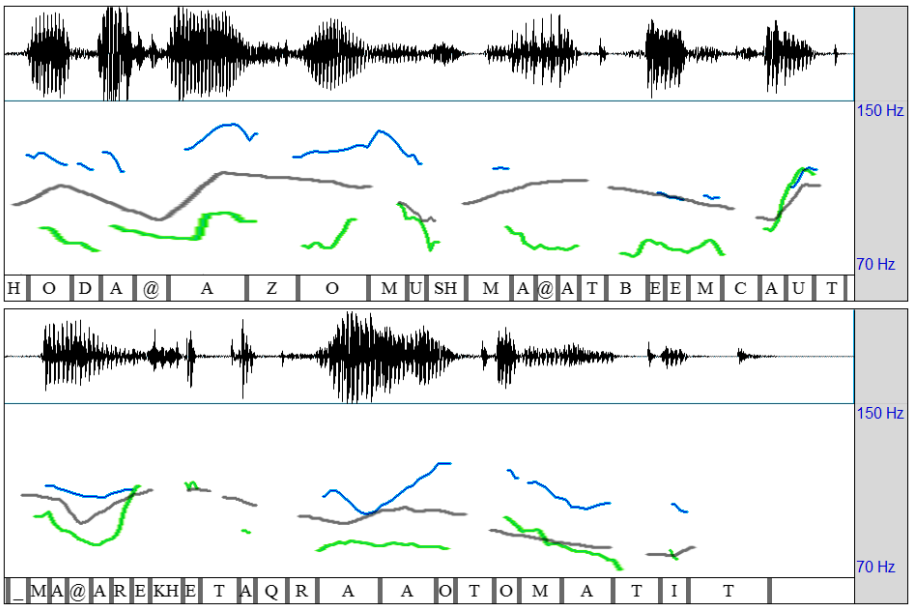

Fig. 4 Fundamental frequency contours for an example sentence, corresponding to the native speaker (blue), referent system [5] (grey), and proposed system (green).

proposed system (green). The three contours have been manually time-aligned to the utterance as rendered by the human speaker (indicated by the waveform and the phonemic labelling). It can be observed that the intonation curve as generated by the referent system seems quite regular, unlike the curves corresponding to the native speaker and the proposed system, which seem to exhibit more variation. Furthermore, it can be seen that a much greater percentage of frames in the speech signal generated by the referent system were identified as voiced, in comparison to the other two systems. This is related to the characteristic buzziness present in the speech signal generated by the referent system, which (together with a rather monotonous intonation) was one of the major drawbacks of the referent system as reported by the listeners. However, most listeners also reported that the intonation contours of both synthesizers are adequately related to the positions of stressed syllables.

6. Conclusion

By using the expert system in combination with CART the quality of synthesized speech is considerably increased. Based on the results of the listening tests, the system described in the paper provided much more natural-sounding speech when compared to the previous version of the system, in which the prosody was estimated using the expert system. An additional benefit of automated prosody generation is in the fact that such an automated system can be adapted to different dialects of the Hebrew language much more easily and in much less time than the expert system. Namely, covering a different dialect

of Hebrew would require that a new speech corpus be recorded and tagged, and that the automatic training procedure be repeated, which is still widely considered to be far simpler than discovering new sets of expert rules related to prosody. The quality of synthesized speech could be further improved by widening the set of relevant questions as well as by improving the segment selection and signal concatenation modules.

Acknowledgement: *This research work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and it has been realized as a part of the research project TR 32035.*

REFERENCES

[1] J.P.H. van Santen, "Contextual Effects on Vowel Duration", *Speech Commun.*, 1992, vol. 11, no. 6, pp. 513-546.

[2] M. Sečujski, N. Jakovljević and D. Pekar, "Automatic Prosody Generation for Serbo-Croatian Speech Synthesis Based on Regression Trees", In Proceedings of the 12th Annual Conference of the International Speech Communication Association, 2011, Florence, Italy, pp. 3157-3160.

[3] Ö. Öztürk and T. Çiloğlu, "Segmental Duration Modelling in Turkish", In Proceedings of the 9th International Conference on Text, Speech and Dialogue, Brno, Czech Republic, *Lect. Notes Comput. Sc.*, Springer, 2006, vol. 4188, pp. 669-676.

[4] A. Lazaridis, P. Zervas, N. Fakotakis and G. Kokkinakis, "A CART Approach for Duration Modeling of Greek Phonemes", In Proceedings of the 12th International Conference on Speech and Computer, 2007, Moscow, Russia, pp. 287-292.

[5] D. Kamir, N. Soreq and Y. Neeman, "A comprehensive NLP system for modern standard Arabic and modern Hebrew", In Proceedings of SEMITIC'02, the ACL-02 workshop on Computational approaches to Semitic languages, 2002, ACL, Stroudsburg, PA, USA, pp 1-9.

[6] N. Chomsky, Morphophonemics in Modern Hebrew. Routledge, 2012.

[7] J. Fellman, "Concerning the "Revival" of the Hebrew Language", *Anthropol. Linguist.*, May 1973, vol. 15, no. 5, pp. 250-257.

[8] B. Popović, M. Sečujski, V. Delić, M. Janev and I. Stanković, "Automatic Morphological Annotation in a Text-to-Speech System for Hebrew", in Proceedings of the 15th International Conference on Speech and Computer, Pilsen, Czech Republic, *Lect. Notes Comput. Sc.*, Springer, 2013, vol. 8113, pp. 319-326.

[9] L. Breiman, J.H. Friedman, C.J. Stone and R.A. Olsen, *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, London, New York, Washington D.C., 1984.

[10] A. Black and N. Campbell, "Optimising Selection of Units from Speech Databases for Concatenative Synthesis", In Proceedings of the 4th European Conference on Speech Communication and Technology, 1995, Madrid, Spain, pp. 581-584.

[11] V. Delić, M. Sečujski, N. Jakovljević, M. Janev, R. Obradović and D. Pekar, "Speech Technologies for Serbian and Kindred South Slavic Languages", *Adv. Speech Recognition*, Chapter 9, 2010.