Increasing Prosodic Variability of Text-To-Speech Synthesizers

Géza Németh, Márk Fék, Tamás Gábor Csapó

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
nemeth@tmit.bme.hu, fek@tmit.bme.hu, csapszi@sch.bme.hu

Abstract

The lack of prosody variation in text-to-speech systems contributes to their perceived unnaturalness when synthesizing extended passages. In this paper, we present a method to improve prosody generation in this direction. A database of natural sample sentences is searched for sentences having similar word and syllable structure to the input. One sentence is selected randomly from the similar sentences found. The prosody of the randomly selected natural sentence is used as a target to generate the prosody of the synthetic one. An experiment was conducted to determine the potential of the proposed method. The rule-based pitch contour generation of a Hungarian concatenative synthesizer was replaced by a semi-automatic implementation of the proposed method. A listening test showed that subjects preferred sentences synthesized by the proposed method over a rule-based solution.

Index Terms: speech synthesis, prosodic variability, F_0 variation, F_0 transplantation

1. Introduction

State-of-the art text-to-speech (TTS) synthesis is based on the concatenation of sub-word or longer units. These techniques can produce good quality and highly intelligible output. Recent studies showed, however, that current speech synthesis systems are still recognized as non-human when synthesizing extended passages [1]. There are a number of ways to improve naturalness. Identical or very similar pitch contours of successive sentences make the synthetic speech monotonous when synthesizing longer passages of text. This is caused by the fact, that the prosody component of TTS systems is designed to generate the intonation of formal text, and lacks the variability of natural speech. Some TTS synthesizers, including the Profivox system produce a rule-based pitch contour [2]. It is difficult to generate proper prosody resembling to human speech, thus machine-generated fundamental frequency contours are less rich than natural prosody. The goal of this work is to design a novel prosody module capable of generating more natural prosody and introducing variability over successive sentences.

It is still a question what type of individual variation could or should be implemented to obtain individually coloured forms of speech. Keller [1] concentrated on the temporal structure of an utterance. He proposed a linear prediction system that can improve the perceived naturalness of synthetic speech.

Raux and Black [3] used corpus-based approaches to F_0 modeling. They constructed the F_0 contour of an utterance by selecting flexible portions of contours from a speech corpus.

There have been other recent efforts to generate prosody that is similar to human speech. Minghui et al. [4] introduced a method which they call example-based prosody generation. The authors used natural sentences as a prosody database. The pitch contours of the database sentences were decomposed into three levels: sentence level baseline prosody, phrase level prosody pattern, and syllable level prosody. Statistical methods were applied to calculate the default values of the duration and pitch contour of a syllable. To generate phrase level prosody patterns, the database was searched based on a linguistic feature vector, produced during the setup of the prosody database. The best matching sentence was used as the prosody template for the speech to be generated. Sentence level prosody was the combination of syllable and phrase level prosody. The combined F_0 contour was copied to the generated synthetic sentence.

2. Methodology

Our final goal is to develop an improved prosody module for the Profivox synthesizer that includes higher prosodic variation. The proposed method uses a database of natural sentences, like the one described by Minghui et al. [4]. The main difference is that we search for multiple matching natural sentences during the synthesis of an input sentence. This allows us to choose among them in synthesis time, reducing the monotony of longer utterances, even in case of the repetition of the same sentence.

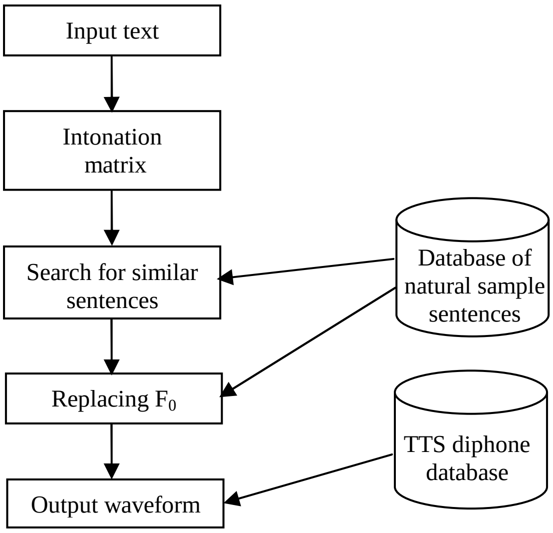

Figure 1: *Overview of the proposed synthesis method.*

Figure 1 shows an overview of the synthesis method used in our experiment. First, the Profivox synthesizer transforms the input text into a so-called intonation matrix. The intonation matrix contains the phonetic transcription and the prosodic attributes (sound durations, pitch, and intensity) of the sentence to be synthesized. In the next step, the database containing natural sample sentences is searched on text level for sentences similar to the input text. The F_0 values in the intonation matrix are replaced by F_0 values extracted from one of the selected natural sentences. The intensity contour and sound durations generated by Profivox are not changed. These parameters will be the subject of further studies. The synthetic waveform is concatenated from a diphone database. The time-domain prosody modification algorithm implemented in Profivox is used to realize the target prosody defined in the modified intonation matrix.

2.1. Database of natural sample sentences

The database consists of waveform files with their textual transcript. It contains 5200 Hungarian declarative sentences extracted from weather forecasts and weather related health news [5]. All sentences were uttered by the same female speaker. Sound and word boundaries were labelled automatically using a Hungarian speech recognition engine [6] in forced aligned mode. Pitch periods were marked using the fundamental frequency detection algorithm of Praat [7].

2.2. Analysis of sentences

The central idea of our approach is to use multiple prosody examples in the database for a sentence to be synthesized. For a given declarative input text, the method tries to find similar sentences in the database. If no similar sentence is found, the standard Profivox rule based F_0 contour is used. For the purposes of this study we apply a simplified definition of the syllabic structure of a sentence as the number of words in the sentence and the number of syllables (based on vowel nuclei) in each word. In Hungarian stress is (nearly) always realized on the first syllable of a word.

We hypothesize that similar syllabic structure indicates similar prosodic structure in this language. Therefore in our initial studies the measure of similarity between two sentences is based on their syllabic structure. The similarity measure can be improved by more detailed analysis of the input text. This is especially important for languages with highly variable word accent positions (e.g. English). After the matching sentences are found, one is selected randomly and its fundamental frequency values are copied to the sentence to be synthesized.

2.2.1. Analysis of syllabic structure

It is important to find an appropriate measure of similarity between the syllabic structures of sentences In the first experiment we used 200 sentences and the number of words as the similarity measure. That approach yielded proper variation in a few cases only. In the experiment reported here a larger database of 5200 sentences and a more complex similarity measure were used.

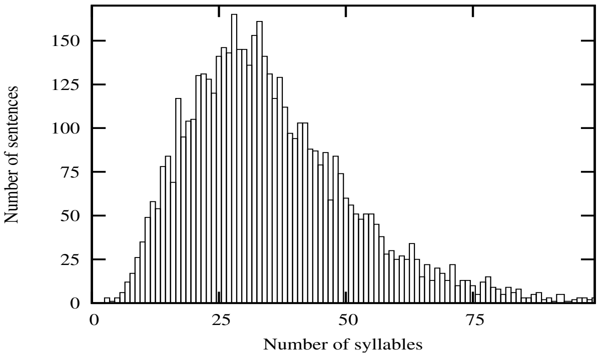

Figure 2: *Distribution of the number of syllables in the sentence database*

After analyzing the database, we found several groups of sentences which had the same syllabic structure. That is, the sentences in a group had the same number of words and the same number of syllables in all of their words. As shown in Figure 2, the database contains a large number of long sentences with over 25 syllables. The sentences were longer than usual because they were extracted from weather information sources. The maximum length of the selected sentences with matching syllabic structure was limited to 20 syllables in order to better approximate standard sentence lengths of Hungarian. Figure 3 shows an example of two sentences with matching syllabic structure. There are ten syllables, and three words. The first word is four syllable long, the second and third words have three syllables. The meaning of the sentences is completely different (3056: “Some may suffer from headache”, 3373: “There may be clouds from the evening”).

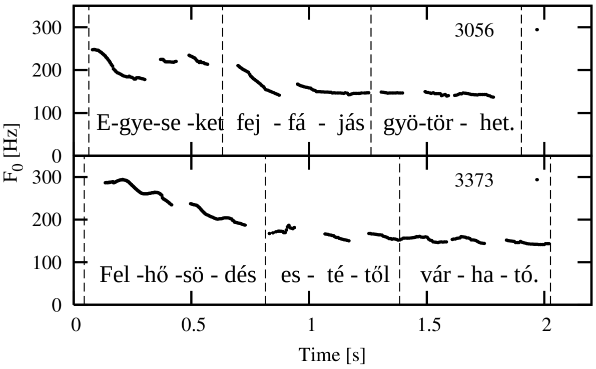

Figure 3: *Pitch contour of two similar sentences with matching syllabic structure. The vertical dashed lines indicate word boundaries.*

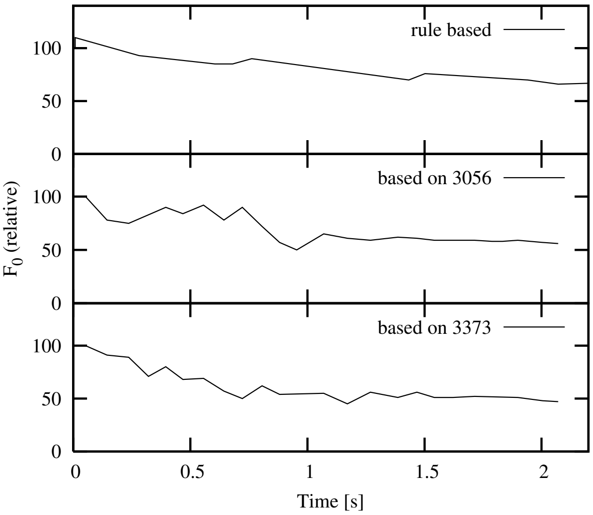

Figure 4: *Pitch contour of sentence 3056 in three styles.*

2.2.2. Analysis of pitch contour

First, we synthesized all sentences in the selected sentence groups with the rule-based Profivox pitch contour. The pitch values in the intonation matrix were replaced, without changing the generated intensity and duration values. F_0 modification was based on the pitch contours of the natural sentences in the selected group. The syllables in the sentences were time-aligned based on the center of vowel nuclei. The F_0 curves were stylized by creating breakpoints in the middle of each vowel. The F_0 value at a breakpoint was calculated as the mean normalized F_0 of the corresponding vowel. The basis of normalization was the pitch at the beginning of each utterance.

The mean F_0 values were written into the intonation matrix. That is, for the first sentence in Figure 3, we obtained three synthetic sentences: one with rule based F_0 contour, one with an F_0 contour copied from its natural version, and one with an F_0 contour copied from the other sentence. The same synthetic temporal structure was set for all sentences, thus they only differ in their pitch contours.

2.3. Pitch contour generation

In the Profivox TTS system, a fundamental frequency curve is defined by one breakpoint on each sound. Linear interpolation is used between the breakpoints. The position of a breakpoint within a sound is not fixed, it can be put on any sample within the sound. Figure 4 shows the three synthetic F_0 curves generated for sentence “3056” in Figure 3 by a male voice of the Hungarian version of Profivox.

3. Experiments

Six groups with matching syllabic structure were selected for feasibility tests. There was one group that consisted of three sentences, and five groups with two sentences. For sentences in the 2-sentence groups we obtained four variants as described in section 2.2.2. In the 3-sentence group we synthesized four variants for each sentence: one with a rule-based F_0 contour, one with an F_0 contour copied from its natural version, and two with F_0 contours copied from the other two sentences in the group.

3.1. Test environment

The goal of the test was to compare the F_0 copying method to the standard rule-based solution and to evaluate different natural-based F_0 samples. We paired the sentence variants in each group, so that the 3-sentence group had six pairs and the 2-sentence groups had three pairs per sentence. Altogether 48 sentence pairs (18 in the 3-sentence group, 30 in the five 2-sentence groups) were created.. The paired stimuli were separated by an interval of 0.5 seconds, and the actual sentence order was cross-balanced in the experiment.

To lessen the load of the testers, a listener had to evaluate only a part of the pairs. The subjects listened to 8 sentence pairs in random order. They had the option to replay a stimulus as many times as they wished but they were not allowed to go back to previous stimuli. Listeners were asked to select the better sounding variant from the sentence pair, or to mark them as equal. The test was performed using a web-interface via the Internet which made the participation of numerous listeners possible.

On the average the whole test took 18 minutes to complete. It consisted of six parts. The listeners evaluated four sentence pairs in both of the second and fifth parts. The first and fourth parts were unrelated to this experiment. The third and sixth parts were included to filter out random clickers. In these two parts, the listeners used a 5-point scale to grade the quality of a sentence in one natural and three synthesized versions. Those providing inconsistent quality judgements were excluded from the evaluation.

3.2. Listeners

A total of 208 subjects participated in the test. 14 of them were excluded from the assessment because they reported playback difficulties, failed to finish the test or they were regarded as random-clickers. The results from 159 male and 35 female listeners were evaluated. Their age was between 18 and 40 years, with a mean of 23 years.

All were native speakers of Hungarian with no known hearing loss. Most of them were motivated undergraduate computer science students. 83 listeners used head- or earphones while 111 applied loudspeakers. 178 testers reported average quality audio equipment, 10 mentioned professional, and 6 used poor quality devices. 172 subjects took the test in a quiet environment while 22 testers reported noisy surroundings. To lessen the load on the listeners, not all of them tested all sentence pairs. The assignment of listeners to sentence pairs was uneven, but each sentence pair was evaluated by at least 21 listeners.

4. Results

Results show that fundamental frequency copying improved the prosody of the synthesized sentences. Users accepted the intonation curve of semantically different but structurally similar sentences. Rows in Table 1 and Figure 5 describe the same comparison: the choices by listeners for the variants of sentences 3056 and 3373. In case of rows 1-3 the carrier sentence is 3056 while rows 4-6 describe results with 3373 as the carrier sentence.

Table 1. *Results for a group containing 10 syllable long sentences (3056 and 3373)*

	sentence pair			choices of listeners		
	number	1. variant	2. variant	1. is better	equal	2. is better
1	3056	3056	3373	8	8	6
2		rule	3056	3	3	17
3		rule	3373	3	4	16
4	3373	3373	3056	5	10	6
5		rule	3056	4	0	17
6		rule	3373	2	2	19

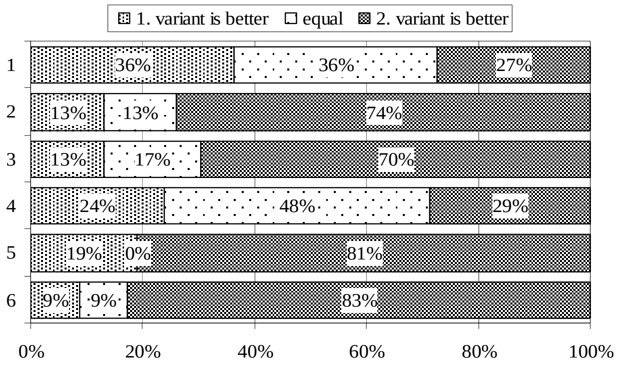

Figure 5: *Results for a group containing 10 syllable long sentences (3056 and 3373)*

The “1. variant” and “2. variant” columns refer to the prosody target of the carrier sentence. For example, rows 1 and 4 show that the prosodic structure of sentences 3056 and

3373 matches very well. Listeners judged them as nearly equal in both synthesized versions. Row 3 demonstrates that 70% of the listeners regarded sample based prosody better than the rule-based solution, while in sentence 3373 this ratio is 81%, as shown in row 5. Rows 2 and 6 show the control case for natural prosody, which should be better than the rule-based one.

Most listeners preferred the F_0 copied variants of semantically different sentences against the rule-based solution, as shown in Table 2 and Figure 6. Rows in Table 2 correspond to the horizontal axis of Figure 6. Regarding the whole test, F_0 copying was preferred in 7 cases. In 6 cases listeners regarded the two variants as basically equal. The standard rule-based F_0 contour was found to be substantially better only for one sentence pair. It turned out that the signal processing algorithm in Profivox introduced important quality degradation because of the relatively large pitch variations in that case. In pairs where the F_0 was copied from two different natural sentences, results show that their quality is almost equal. The experiment demonstrates that the new synthesis approach opens a way for increased prosody variation without sacrificing quality.

Table 2. *All comparisons between rule based and semantically different F_0 copied variants of sentences*

	sentence pair			choices of listeners		
	number	1. variant	2. variant	1. is better	equal	2. is better
1	3053	rule	3614	11	10	21
2	3053	rule	3855	5	11	6
3	3614	rule	3053	17	19	20
4	3614	rule	3855	6	14	1
5	3855	rule	3053	9	4	10
6	3855	rule	3614	20	13	22
7	3056	rule	3373	3	4	16
8	3373	rule	3056	4	0	17
9	1773	rule	2565	12	1	9
10	2565	rule	1773	5	3	15
11	3517	rule	3953	3	1	18
12	3953	rule	3517	1	7	14
13	2551	rule	3966	0	3	19
14	3966	rule	2551	5	13	5

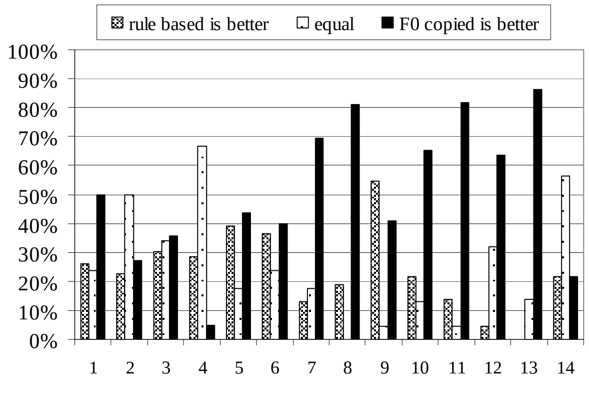

Figure 6: *All comparisons between rule based and semantically different F_0 copied variants of sentences*

5. Discussion

In this initial study we have experimented with introducing prosodic variability by F_0 generation in a Hungarian diphone TTS environment. Our method generates the F_0 contour for a sentence to be synthesized from a database of natural sample sentences. The quality of the generated F_0 contour was evaluated using a listening test. Subjects tested 48 sentence pairs. Figure 7 shows the sum of choices in the rightmost three columns of Table 2. In half of all cases subjects preferred the F_0 copied variants over versions with rule-based pitch contour. F_0 copied variants of the same sentence category were otherwise regarded to be of equal quality with only two exceptions. Given these positive results, we hope to extend this work with the timing and intensity features of prosody with the final goal of increasing prosodic variability of the whole TTS system. We suppose that by refining the similarity measure results can be further improved.

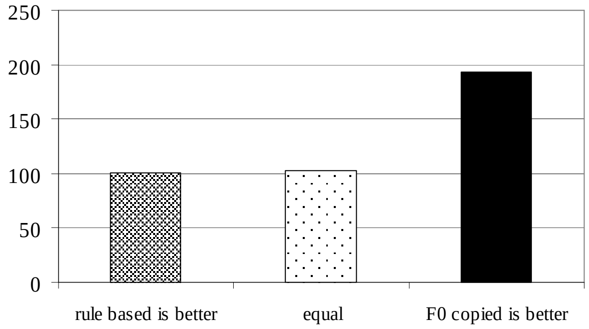

Figure 7: *Comparisons between rule based and semantically different F_0 copied variants of sentences*

6. Acknowledgements

We thank all listeners for participating in the subjective tests. The research presented in the paper was partly supported by the Hungarian National Office for Research and Technology (NKFP 2/034/2004 and NAP 00736/2005)

7. References

[1] Keller, E. (in press). “Beats for individual timing variation”, in A. Esposito et al. (eds.), The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue, IOS Press.

[2] Olaszy, G., Németh, G., Olaszi, P., “Automatic Prosody Generation - a Model for Hungarian”, in Proc. of Eurospeech 2001, Vol. 1., 2001, pp. 525–528.

[3] Raux, A., Black, A., “A Unit Selection Approach to F0 Modeling and its Application to Emphasis”, ASRU 2003.pp.700-705.

[4] Dong Minghui, Lua Kim Teng, “An Example-based Approach for Prosody Generation in Chinese Speech Synthesis”, International Symposium on Chinese Spoken Language Processing, Beijing, 2000, pp. 303-307.

[5] Fék, M., Pesti, P., Németh, G., Zainkó, Cs., Olaszy, G.: “Corpus-Based Unit Selection TTS for Hungarian.” Proc. of TSD 2006, pp. 367-374.

[6] Mihajlik P., Révész T., Tatai P., “Phonetic Transcription in Automatic Speech Recognition”, Acta Linguistica Hungarica, Vol. 49. (3-4), 2002, pp. 407–425.

[7] Boersma, P. & Weenink, D., Praat: doing phonetics by computer, (Version 4.4.34) [Computer program], http://www.praat.org/, 2006.