A study of prosodic variability methods in a corpus-based unit selection text-to-speech system

TAMÁS GÁBOR CSAPÓ, CSABA ZAINKÓ, GÉZA NÉMETH

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics
{csapot, zainko, nemeth}@tmit.bme.hu

Keywords: prosody variation, corpus-based TTS

This paper introduces the implementation and evaluation of a method to increase the prosodic variability of synthesized speech. Different generated prosody target versions were tested in a Hungarian corpus-based unit selection Text-To-Speech (TTS) system: the baseline prosody of the synthesizer, a rule-based prosody target and the prosody of the new method.
It is based on F0 database templates which are derived from natural sentence corpora. The corpora included that of domain specific TTS and annotated radio news. The listening test validation of the new method showed that the speech quality of the corpus-based TTS was improved. Our method was tested in a Hungarian system, and it can be extended to other European languages with fixed (e.g. Finnish) and varying (e.g. English) stress.

1. Introduction

The quality of Text-To-Speech (TTS) systems is judged on the basis of how successfully the generated synthetic speech approaches the features of human speech. The intelligibility of synthetic speech is close to that of human speech in state-of-the-art TTS systems. However, there seems to be a lack of variability in most speech synthesizers: they produce deterministically the same speech output for the same textual input, when it is repeatedly given to the system. This contradicts the variability of human speech.

The variation in human speech has been addressed by Chu et al. using a database containing two repetitions of 1000 recorded sentences in Mandarin [1]. They investigated the differences of prosody (e.g. intonation, rhythm) in the paired sentences and observed the invariant and variable parts of speech. It was measured that the two repetitions had wide variations in the mean F_0 and durations of syllables, while the meaning was the same. The rhythmic organization was more stable. The results show that the variability of human speech can be as large as half of the dynamic range of a speaker, which has to be considered in speech synthesis.

The common part of the corpus-based prosody generation approaches is that they try to associate properties (e.g. F_0) of recorded speech with the text to be synthesized. However, there are some differences in the methods and element sizes that are applied. In [2], a rule-based prosody model is complemented with a corpus-based module. In the data-driven part, the F_0 templates are as small as syllables from the corpus. [3] uses a similar method. The most important difference is the length of the F_0 templates: employing flexible-sized segments allows the modeling of both macro- and microprosody. In the corpus-based approach of [4], a linear regression statistical model produces the pitch contour of a sentence, based on word-sized items. The new feature of [5] is the use of Case-Based Reasoning. They show that a data-driven model can work with stress-group units as F_0 templates reasonably well. Besides the aforementioned corpus-based methods, some superpositional corpus-based intonation generation approaches can also be found in the literature.

In the paper of [6], three levels of intonation are derived from the speech database. Sentence-, phrase- and syllable-level prosody are hierarchically separated. [7] combines decompositional modeling with corpus-based pitch contour search. The pitch contours in the corpus are typically decomposed into phrase, accent and segmental perturbation curves. [8] introduces a corpus-based synthesis system by considering several candidate intonation contours.

The method presented in this paper uses phrase-long F_0 templates without any decomposition. Our prior work concentrated on the feasibility of a method in increasing the prosodic variability in speech synthesis [9]. As the results were rather promising, this first simple approach is further developed here. In this paper, Section 2 introduces the method that tries to mimic the variable nature of human declarative sentences in TTS systems. To pair the right pitch contour with the input text, a database of recorded speech samples is used in order to find more F_0 templates to the input. Variability is ensured by the random selection from the available intonation samples. Our hypothesis is that using this kind of speech samples, variability of human speech can be modeled in artificial systems. To prove this statement, the method is applied in a Hungarian corpus-based unit selection TTS.

In Section 3, a listening test is described that was carried out to evaluate the naturalness of the generated variable synthesized speech. The results of the test are described in Section 4, which show that the method

can generate equally natural but still different versions of sentences. The last section concludes the paper.

2. Methods

2.1 Generation of variable prosody

Fig. 1 shows the steps of our prosody generation approach used in the Hungarian system. When the system is given a raw textual input, first the sentence is partitioned into prosodic phrases. Then their syllabic and stress structures are automatically determined. Intonation is assigned in a separate step.

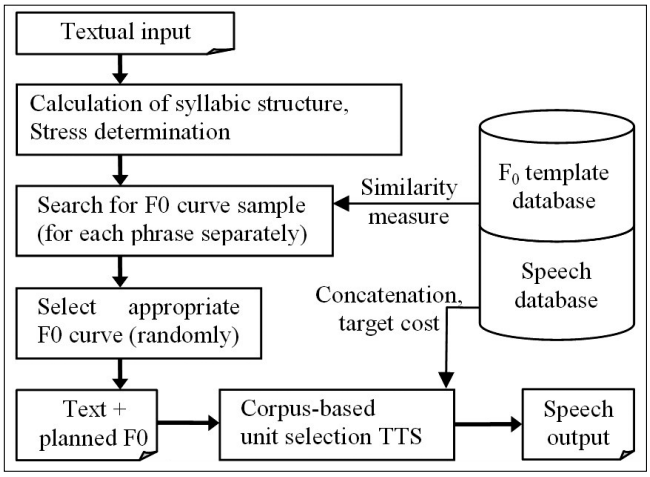

Figure 1.
Generation of variable prosody using a database of F_0 templates in a corpus-based unit selection TTS system

In this paper, the syllabic structure of a phrase is represented by the number of words in the phrase and number of syllables in each word plus their stress tags (e.g. in English a prosodic phrase might be like: "Weather warnings have been issued"; 5 words; "2+2+1+1+2" syllabic structure, plus stress marks on the words). The stress structure is specified with the rule-based method used in the Profivox TTS [10].

Based on this information target F_0 curves are searched in the database of natural sample phrases. The search is done using a similarity measure that will be discussed later. If the sample database is large and variable enough, more candidates with different F_0 curves are found. The variation of the system is realized in the next step, when one F_0 contour is selected randomly from the proposed ones. If the system is given repeatedly the same sentences or those with similar structure, this part ensures that the output speech will not always be deterministically the same, creating the desired variable prosody. The input text with the selected F_0 contour is forwarded to the corpus-selection module that tries to realize the proposed prosody during the unit selection. Note that as the system may use F_0-sample databases and speech corpora from different speakers, it is important to normalize F_0 targets, before sending them to the selection component.

2.1.1 Coverage ratio

It is not sure that an appropriate F_0 template can be found for a given input phrase. If no similar phrase is found, the rule-based prosodic model of the Profivox TTS is used. The hit rate depends on the similarity measure used and the size and variability of the database. In order to find out what degree of coverage can be reached with different similarity measures and corpora, a coverage ratio is defined. For a given input sentence, it refers to the length of the prosodic phrase for which F_0 samples are found divided by the total length of the sentence. The length is measured in number of syllables.

As an illustration in English, if the method finds a F_0 template only for the first of the two prosodic phrases in the *"A minor storm will brush the Northwest, resulting in showers."* sentence, the coverage ratio is *9/15 = 0.6*.

2.1.2 Similarity measure

Two different similarity measures are investigated in this study. They are based on the syllabic structures of the input phrases and F_0 template phrases from the database. The first one, the "exact" similarity measure means that the structures of the two phrases have to be exactly the same. The second measure, "similar" structure is less strict: the number of syllables of the longer words in the two phrases can differ in one syllable. This "loosening" in the similarity measure causes a higher coverage ratio, as discussed later. Besides the syllabic structure, the stress structure of the input and database phrases also have to be the same, in order to use the F_0 contour of the sample database.

2.1.3 Databases

The F_0 template databases were derived from several speech corpora. The textual version, phonetic transcripts, sound boundaries and measured F_0 curves were used to generate the database. The sentences were cut to phrases by the Profivox text processor [10], and for each phrase, the syllabic structures were calculated.

For each syllable, the mean F_0 was calculated. It was used when selecting an F_0 contour for the input text. The stress structure of the phrases was derived from

Table 1.
Corpora and attributes of each of them used in this study

Name	Sentences	Phrases	TTS	F_0-1	F_0-2	F_0-3
Weather	5239	13803	x	x	x	
Ph. rich	1941	3146	x		x	
Numbers	205	(205)	x			
Railway	682	1291	x		x	
Prompts	672	990	x		x	
Radio news	3651	8746				x

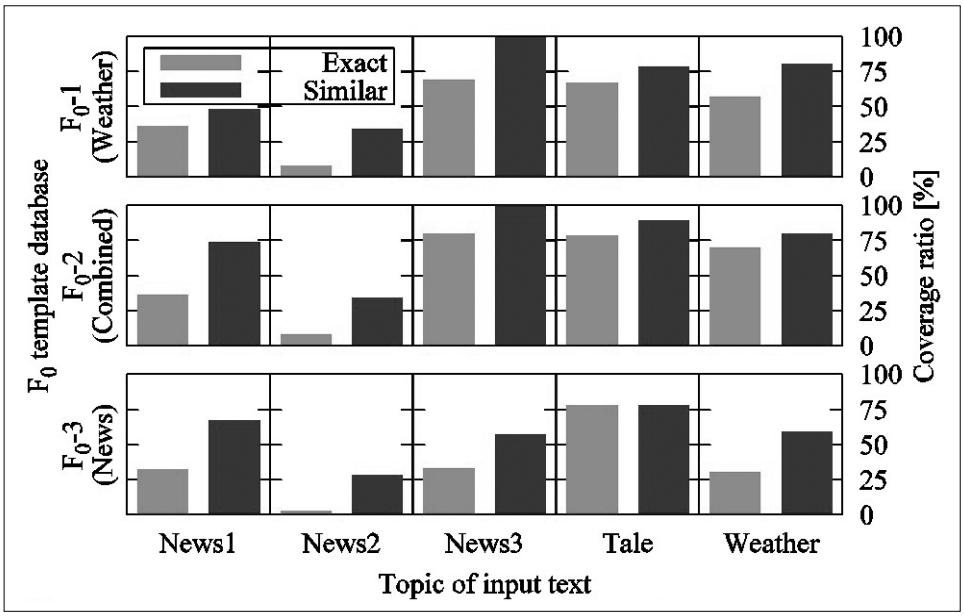

Figure 2.
Dependence of coverage ratio on similarity measure, topic of the input sentences and the domain of the F0 sample database

the textual transcripts. This solution introduces some errors because the written form does not have a one-to-one correspondence with the utterance. The development of an automatic method is in progress in order to determine the realized stress structure in the utterances of the corpora.

In this study six corpora were explored, as described in *Table 1*. The sentences of the first five corpora were read by a professional voice actress in a sound studio. The largest ("Weather") was built from weather forecast sentences [11]. The second largest corpus ("Ph. rich") contains phonetically rich and balanced sentences [12]. We also used sentences which were originally recorded for a number-to-speech synthesis application ("Numbers", [13]). The rest of it was produced for a railway station announcer system ("Railway") and other smaller, fixed inventory systems (the "Prompts" of e-mail reader systems). The last corpus ("Radio news") contains news spoken by three different announcers. Except the last one all other corpora were used in the limited domain corpus-based TTS.

Three different F_0 template databases were built for investigations of the prosody generation method: one from the "Weather" corpus itself (F_0-1), one combining four corpora of the same speaker (F_0-2) and one from the "Radio news" corpus (F_0-3).

We collected text samples in five topics: Public transport (News1), Economy (News2), Sport (News3), Tale and Weather forecasts. *Fig. 2* indicates the coverage ratios defined in 2.1.1, that can be reached with the different databases (Y axis) on input sentences from various topics (X axis). In order to illustrate which sample database is more appropriate for the variable prosody generation method, we conducted a simple test. For the different domains, we investigated up to 8 sentences to demonstrate the dependence of coverage ratios on the type of the databases.

It can be seen that the use of "F_0-2 (Combined)" database creates the highest coverage ratios, for all of the input sentences. The "Similar" similarity measure enhances the coverage ratio over "Exact", as expected. The low coverage value in "News2" sentences is due to the fact, that they contain extremely long phrases. As our method is based on the correspondence between the phrases of the input sentence and the database ones, F_0 patterns for these long input phrases were not easily obtainable.

2.2 Corpus-based unit selection TTS

The unit selection TTS that was used in our experiments is described in detail in [11]. The currently used speech databases contain sentences from several domains, as described in Section 2.1.3. The synthesizer can generate the prosody in two ways, depending on the type of the input sentence. If the sentence fits in the domain of the corpus, a simple prosody model is used, based on the relative position of words within a prosodic phrase. Because it is based on words, it will work properly only if most words of the input sentence are found in the corpus.

If the sentence is out of theme, there will not be enough whole words, which can determine the prosody. The prosody is undefined where whole words are missing. On those parts of sentences the F_0 values are determined only by the continuity criterion of the F_0 of the units, but this allows irregular prosody. In that case the F_0 generation method described in Section 2.1 is extremely useful, as it defines a target F_0 for the input sentence. In order to realize variability, always different but still natural F_0 curves are compared to the sentences with similar structure after each other. The obtained F_0 values are used in the target cost function of the TTS to follow the F_0 curve. Besides F_0 values the generated phoneme durations also are respected in target costs, but with less weight than F_0.

The words of the TTS corpus cover about 55% of the Hungarian texts [14]. It means that those words of a sentence that are out of domain are often missing on the

word level. Because the missing words are usually built from 4-5 syllables, in some cases as much as 60% of the synthesized sentence is determined by prosody rules for words that are concatenated from shorter (phone/diphone/triphone) units. In this case the overall prosody of the sentence is not determined by whole words of the corpus.

3. Experiments

Several sentences from different domains were collected in order to find out whether the TTS with the variable prosody generation method produces synthesized speech that sounds natural enough. A listening test was conducted to verify our hypothesis that the variability of human speech can be modeled using a TTS.

3.1 Test sentences

Only two sentences were chosen from each domain in order to decrease the duration of the listening test. The same domains were used as in the simple test of Section 2.1.3. We generated five versions for each sentence. The first is synthesized with a triphone based concatenative TTS (Profivox). It works with a rule based prosody module, and it can produce the waveform with the prescribed prosody. The second version is the original output of the corpus-based TTS, with its simple position-based prosody module (with minimal prosody modification). In that sentence the prosody is determined by the parameters of the units in the speech corpora. In the third version the corpus-based TTS uses the rule based prosody (of the Profivox algorithm) as the target F_0 curve. The last two versions are generated by the corpus-based TTS, but the target F_0s come from the variable prosody module which is described in Section 2.1.

During the collection of the different sentence versions we found that for one of the “News3” (Sport) sentences only a bad F_0 target was available. The end of that target is wrong, it contains an increasing end (in F_0) instead of a decreasing one. In spite of its incorrect prosody we inserted it in the test, in order to measure the tolerance of the audience for this type of error.

In the listening test we evaluated these sentences in two ways. Each sentence appeared in a Mean Opinion Score (MOS) test of the naturalness of prosody. To detect smaller differences between versions, we conducted a paired comparison test, too. In the paired comparison test, only sentences generated by the corpus-based TTS with different prosody modules were investigated. On the basis of a preliminary test the triphone-based concatenative TTS is definitely weaker than the other four versions of each sentence. The paired comparison test contained four versions of 10 sentences, making 60 pairs altogether. The MOS test contained five versions of the 10 sentences.

3.2 Listening test

A web based listening test was conducted to determine the naturalness and quality of synthetic sentences. 103 native speakers of Hungarian participated in the test with no known hearing loss. The results of 10 listeners were excluded from the evaluation because they either did not finish the test, or were found to respond randomly. Some of the excluded listeners reported playback difficulties. The remaining 93 listeners consisted of 67 male and 26 female testers having a mean age of 32 years. 49 listeners used head- or earphones while 44 testers listened to loudspeakers. The listening test took 38 minutes to complete, on average.

The test consisted of six parts. The first and the second part were used for another unrelated study. To lessen the load of the testers and to improve attention we cut in half both the MOS and the paired comparison tests. In the third and fifth parts the listeners compared the sentence pairs. In the fourth and sixth part, the sub-

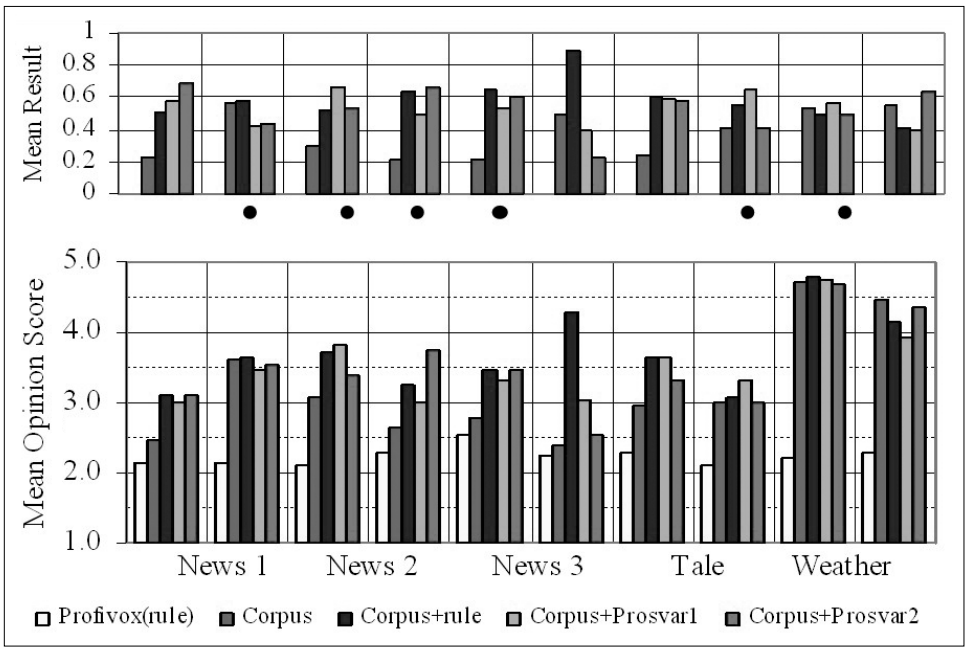

Figure 3.
Mean paired comparison results and Mean Opinion Scores of the sentences

jects used a 5-point scale to grade the quality of a sentence. The test was self-paced. The listeners had the option to replay a stimulus as many times as they wished, but they were not allowed to go back to a preceding stimulus, once they rated it. The playback order and the order of utterances in the paired comparisons were randomized individually for each listener.

4. Results

The bottom section of *Fig. 3* shows the results of the MOS test for all the 10 sentences under study. The means were calculated on the basis of the judgment of the listeners. The first column of each group shows the variants generated using a triphone rule-based TTS. It is rated between 2.1 and 2.5 MOS points. These results coincide with our earlier study [11].

The versions synthesized with the corpus-based TTS are divided into two parts: the first 8 groups were less natural than the last two groups according to the expectations. The sentences of the "News" and "Tale" area are out of domain for the synthesizer, the weather forecast sentences conform to the corpus. For "Weather" sentences changing the original prosody module of the corpus-based TTS does not show major improvements, at the second "Weather" sentence it caused rather a degradation. The worst sentences probably significantly differ from human expectations.

Except "Weather" sentences in all the other groups better results were reached with the modified, target-based prosody modules. The three versions of prosody targets (rule, Prosvar1, Prosvar2) scored nearly equally. The bars marked with dots show sentences when the "Radio-news" corpus was used as the F_0 template database for the prosody generation subsystem. These sentences were evaluated similarly to versions "Prosvar1" and "Prosvar2". In some cases, the variable prosody generation method failed to produce human-like utterances. The corpus-based TTS with the rule-based prosody method generates the best MOS score at the second sentence of the "News3" group. This sentence has a correct prosody and enough proper word units in the corpus. The "Prosvar1" and "Prosvar2" prosodies are incorrect in this case, as expected (described in Section 3.1). Subjects gave low scores, they did not accept declarative sentences with high F_0 values in the end.

The responses of the listeners in the paired comparison test were summarized in the top section of *Fig. 3*. For the two utterances in a pair, the results were calculated for each answer as follows: the more natural sentence was given a 1.0 score, the less natural one received a 0.0 score. If a listener could not hear any difference between them, a score of 0.5 were given for both variants in the pair. The averages of these values were calculated in each sentence group. The top section of *Fig. 3* shows the mean values for each variant. ANOVA tests were run for each sentence, Tukey-HSD post hoc tests showed the significant differences. In six cases of the eight non-weather sentences, the versions with prosody generated by the external module were significantly better ($p<0.05$).

Fig. 4 shows the summarized results of the paired comparison tests, averages for each F_0 generation method. The utterances generated by the corpus-based TTS without external prosody information (left column) are significantly ($p<0.05$) less natural than the other three cases, in which a target F_0 curve was given to the synthesizer. The mean results for the three different approaches of the variable prosody generation subsystem (right three columns) are not significantly different.

5. Summary and conclusions

We successfully integrated a new prosody generation method to a Hungarian corpus-based unit selection TTS system. It can provide variable prosody while increasing the quality of synthesized sentences when the input is outside the corpus domain. A listening test showed that in most cases, versions of the same sentence with different intonations were evaluated as equally natural, indicating that the variability of human speech can be applied into speech synthesis. The method can also be applied for increasing prosody variation even when the TTS works in a closed domain, but some quality degradation may occur then. Rule-based target prosody and the alternative sample-based prosody gave similar MOS values. It was found, that the method can

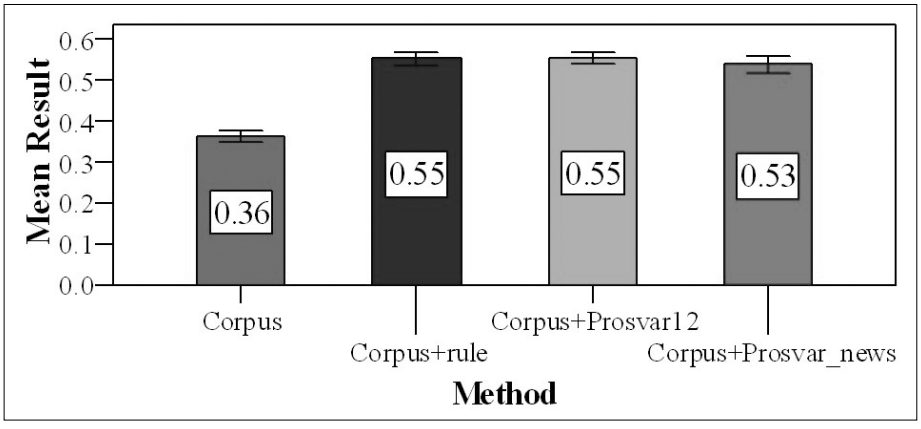

Figure 4.
Mean and 95% confidence interval of paired comparisons

work with various F_0 template databases. The use of F_0 targets from the "Radio news" did not significantly decrease the quality of synthesized speech over the "Combined" corpus.

Future work should address the construction and analysis of several F_0 template databases, in order to analyze the relationship of the realized variance in synthesized speech and the size, type and domain of speech databases. Our variable speech generation method can be extended to other languages. Languages with fixed stress (e.g. Hungarian, Finnish) are easier to handle in this system. The method can be used for languages with varying stress (e.g. English) as well, if we can determine the stress structure of the sentences based on textual input. Using other technologies (e.g. HMM) requires a different similarity measure. The results can be applied in improving the acceptability of long synthesized texts such as synthesized talking books.

Acknowledgements

The work was partially supported by the Hungarian National Office for Research and Technology (Teleauto project, OM-00102/2007) and by ETOCOM project (TÁMOP-4.2.2-08/1/KMR-2008-0007) through the Hungarian National Development Agency in the framework of Social Renewal Operative Programme supported by EU and co-financed by the European Social Fund.

Authors

CSAPÓ TAMÁS GÁBOR (1985): He obtained his MSc degree in Computer Science, major in Next Generation Networks at the Faculty of Electrical Engineering and Informatics of BME in 2008. In 2007, he was awarded with 1st prize of the National Conference of Scientific Student's Associations. Since 2008, he is a PhD student at the Speech Technology Laboratory of BME TMIT. Research fields: speech technology, speech synthesis.

ZAINKÓ CSABA (1976): obtained his MSc in Technical Informatics at the Faculty of Electrical Engineering and Informatics of in 1999. He is a member of Speech Technology Laboratory of BME TMIT. Research areas include speech synthesis, text processing, human computer interaction, multimodal information systems, user interfaces.

NÉMETH GÉZA (1959): He obtained his MSc in Electrical Engineering, major in Telecommunications at the Faculty of Electrical Engineering of BME in 1983. Also at BME: dr. univ, 1987, PhD 1997. He is the Head of the Speech Technology Laboratory of BME TMIT. Research fields: speech technology, service automation, multilingual speech and multimodal information systems, mobile user interfaces and applications.

References

[1] Chu, M., Zhao, Y., Chang, E.,
"Modeling stylized invariance and local variability of prosody in text-to-speech synthesis",
Speech Communication, Vol. 48, pp.716–726, 2006.

[2] Meron, J.,
"Prosodic unit selection using an imitation speech database", SSW4-2001, p.113, 2001.

[3] Raux A., Black, A.W.,
"A unit selection approach to F0 modeling and its application to emphasis", ASRU, pp.700–705, 2003.

[4] Saito, T.,
"Generating F0 contours by statistical manipulation of natural F0 shapes",
IEICE – Transactions on Information and Systems, Vol. E89-D, No. 3, pp.1100–1106, March 2006.

[5] Iriondo, I., Socoro, J.C., Alias, F.,
"Prosody modelling of Spanish for expressive speech synthesis", ICASSP, Vol. 4, pp.821–824, 2007.

[6] Dong, M., Lua, K.-T.,
"An example-based approach for prosody generation in Chinese speech synthesis", ICSLP, pp.303–307, 2000.

[7] van Santen, J., Kain, A., Klabbers, E., Mishra, T.,
"Synthesis of prosody using multi-level unit sequences", Speech Communication,
Vol. 46, No. 3-4, pp.365–375, 2005.

[8] Díaz, F.C., Banga, E.R.,
"A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems", Speech Communication,
Vol. 48, No. 8, pp.941–956, 2006.

[9] Németh, G., Fék, M., Csapó, T.G.,
"Increasing prosodic variability of text-to-speech synthesizers", Interspeech, pp.474–477, 2007.

[10] Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Gordos, G.,
"PROFIVOX – a Hungarian professional TTS system for telecommunications applications".
International Journal of Speech Technology,
No. 3/4, pp.201–216, 2000.

[11] Fék, M., Pesti, P., Németh, G., Zainkó, Cs., Olaszy, G.,
Corpus-Based Unit Selection TTS for Hungarian.
Proc. of Text, Speech and Dialogue, Brno, 2006.
pp.367–374.

[12] Hungarian Speech Database for Creation of Voice Driven Teleservices: Technical report EU INCO Copernicus Project, No. 977017.
European Commission Brussels, 2000.

[13] Olaszy, G., Németh, G.,
"IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method"
In: Gardner-Bonneau D. (ed.): Human Factors and Voice Interactive Systems. Kluwer, pp.237–256, 1999.

[14] Németh, G., Zainkó, Cs.,
"Multilingual Statistical Text Analysis,
Zipf's Law and Hungarian Speech Generation",
Acta Linguistica Hung. 49. (3-4), pp.385–405, 2002.