The Design of Czech Language Formal Listening Tests for the Evaluation of TTS Systems

Daniel Tihelka, Jindřich Matoušek

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plze Czech Republic
dtihelka@kky.zcu.cz, jmatouse@kky.zcu.cz

Abstract

This paper presents an attempt to design listening tests for the Czech synthesis speech evaluation. The design is based on standardized and widely used listening tests for English; therefore, we can benefit from the advantages provided by standards. Bearing the Czech language phenomena in mind, we filled the standard frameworks of several listening tests, especially the MRT (Modified Rhyme Test) and the SUS (Semantically Unpredictable Sentences) test; the Czech National Corpus was used for this purpose. Designed tests were instantly used for real tests in which 88 people took part, a procedure which proved correct. This was the first attempt to design Czech listening tests according to given standard frameworks and it was successful.

1. Introduction

Once scientists started to engage in speech synthesis research, the aspect of generated synthetic speech quality evaluation had to be taken into consideration; it was naturally caused by the need of comparing the improvement (or deterioration, in a worse instance) of synthetic speech quality during synthesizer development as well as among different synthesizers. It is possible to say that one of the most important ways of evaluating speech quality is using listening tests, which take into consideration statistics from the subjective assessment of human listeners.

Several types of listening tests exist and are standardized for these purposes for English, but in spite of the fact that there are several institutions dealing with speech synthesis development in the Czech Republic, nobody, to our knowledge, designed or adapted these tests for use with the Czech language. This lack together with the need of the testing of our text-to-speech system ARTIC (Matoušek, Psutka, 2000) (being developed at our department) compelled us to make the first attempt at more systematic design of listening tests based on standardized tests for synthetic speech evaluation purposes, but with respect to the Czech language phenomena. We dare claim that it is the first more systematic attempt of such kind in the post-communist history of the Czech Republic.

The paper is organized as follows. Section 2. briefly describes tests standardized for English, Section 3. presents the filling of test frameworks (from Section 2.) with Czech words and sentences in more detail and shows the tricky aspects encountered. The end of this Section shortly describes the tests we carried out on the basis of our design. Finally, Section 4. contains the conclusion and outlines our future work.

This work has been supported by the Grand Agency of the Czech Republic no. 102/02/0124 and by the Ministry of Education of the Czech Republic MSM 235200004.

2. Existing listening tests

There exist a lot of listening tests, but some of them can be encountered more often than others. The tests presented in this chapter are those which can be met more frequently.

2.1. MRT

Modified Rhyme Test – MRT (Huang, Acero, Hon, 2001) belongs to intelligibility tests. It consists of 300 rhyming or similarly sounding monosyllabic words with CVC structure (consonant-vowel-consonant), divided into 50 groups with 6 words each (see sample in Table 1). The words in each group differ from each other by one (first or final) consonant only; 25 groups serve for the first consonant testing and the remaining 25 groups are for the final consonant testing. The listeners' task is to identify the overheard word out of the six possibilities shown (closed response), or to write any word which they thought they heard (opened response); both after one listening.

went	sent	bent	dent	tent	rent
pad	pat	pan	path	pack	pass
...					

Table 1: The sample of standardized words used for MRT test.

We preferred it to DTR (Diagnostic Rhyme Test using only two words in group) as it gives the listener less chance to guess the right answer if he/she does not recognize one of given words.

Responses are usually scored as the number of correctly identified or recognized words.

2.2. SUS

Semantically Unpredictable Sentences – SUS (Benoît, Grice, Hazan, 1996) tests are primarily used for intelligibility testing as well as the MRT. These tests use semantically unpredictable (i.e. meaningless), but syntactically correct sentences. This forces users to understand every word and

minimizes both their ability to estimate the sentence contents and the learning effect.

The test defines five simple syntactic structures, each of them having two levels. The first, *functional* level describes a sentence by categories such as *subject*, *adverbial*, etc (see the first row of Table 2). The second, *syntactic* level specifies the frame description by defining the *slots* corresponding to a particular word category (see the second row of Table 2). In addition, there exists a *word bank*; each word in it is tagged by the same word category as those used in the frames. Randomly selected words from the bank are then inserted into corresponding slots in frames according to their categories[1].

⟨*subject*⟩	⟨*verb*⟩	⟨*adverbial*⟩
⟨*det.*⟩ ⟨*noun*⟩	⟨*intr. verb*⟩	⟨*prep.*⟩ ⟨*det.*⟩ ⟨*adj.*⟩ ⟨*noun*⟩
The table	walked	through the blue truth

Table 2: The sample of the frame used for SUS test. The first row shows the well-known SUS frame definition in the *functional* level, the second row shows slots in the *syntactic* level and the last row shows one of the resulting sentences obtained following a random selection of words from the word banks.

Listeners have to write whole sentence after one listening, and the number of sentences written correctly is usually used to score the intelligibility of the synthesizers. For more information see (Benoît, Grice, Hazan, 1996).

2.3. MOS

Mean Opinion Score – MOS (ITU P.800, 1996) test is, contrary to MRT and SUS tests, used for overall quality testing. Although it was developed for the speech coding quality evaluation, the test was adapted and is widely used for synthetic speech quality testing. Listeners are asked to rate several sentences, either on a five-point scale defining *Listening Quality Scale*[2] ($1 =$ bad, …, $5 =$ excellent), or on a five-point scale defining *Listening Effort Scale* ($1 =$ no meaning understood with any feasible effort, …, $5 =$ complete relaxation, no effort required); see (ITU P.800, 1996). The scores are then averaged, resulting in a overall MOS score which is analyzed by standard statistical tests.

This kind of evaluation reflects the overall impression of speech, including intelligibility as well as naturalness, smoothness, pleasantness and other aspects.

2.4. CCR

Comparison Category Rating – CCR (ITU P.800, 1996) test is used, when the direct comparison of two synthesizers, versions of a synthesizer or methods is required (the first is marked by a, the other by b in the following text). Listeners are presented with a pair of speech samples (in succession A, B) of the same sentence and they have to score the sample B relative to the A according the 7-point table ($3 =$ Much Better, …, -3 = Much Worse), see (ITU P.800, 1996). The half of the pairs played in position A must be generated by a, the half in position A must be generated by b and vice versa. Listeners can repeat listening if they like.

The rating range can be reduced to *prefer A / roughly the same / prefer B* or *prefer A / prefer B* only, if there is such a demand. The scores are then analyzed by standard statistical tests (remember that the results from Ab correspondence must be reversed).

[1]There is direct correspondence slot↔word; one word selected from given category is inserted into a slot of the same category.

[2]In the case of comparing coded speech, the rate is evaluated by implicit reference to real human speech (Huang, Acero, Hon, 2001).

3. The design of the tests for the Czech language

By adopting the standardized test framework, we can benefit from the advantages provided by standards as well as from the ability to compare our results with other systems or across the different versions of the same system.

We selected four types of tests, the MRT, the SUS, the CCR and the MOS. These tests are well-known and very well-designed for English. What had to be done was to fill the frameworks of these tests with words and sentences designed specifically for the Czech language with regard to all phenomena in this language.

3.1. MRT

The design of this test was relatively simple, although the MRT test has not been designed for the Czech so far. The biggest complication in six-words group design was that we had to take into account the rather strong assimilation property occurring in pair consonants in consonant clusters in Czech spoken language. There is a rule that each paired voiced consonant followed by an unvoiced consonant (or pause) changes into unvoiced pair member (like $b \rightarrow p$, $d \rightarrow t$, $z \rightarrow s$, etc); if voiceness is switched, the rule holds the other way round. It causes "pod" (under) and "pot" (sweat) to have the same spoken form before any unvoiced phoneme or pause. Therefore, we had to choose such monosyllabic words as to avoid having words with identical spoken form in one group (due to assimilation). We ended up being the first to introduce the unique 40 groups of such words, some of which are shown in Table 3.

pyl	pih	pij	piš	piv	pin
lev	les	lem	lep	led	len
dub	dur	dus	duc	duch	duň
mít	mís	mír	míň	míč	mým
…					
pech	cech	mech	dech	Čech	nech
jet	zet	med	ret	let	set
val	kal	žal	dal	ťal	pal
suk	kuk	puk	luk	fuk	muk
…					

Table 3: The sample of the Czech words designed for the MRT test.

3.2. SUS

The task of the Czech SUS test design was more difficult as opposed to English and other Western languages. It was caused by both the free word-order and the quite strong inflection of words manifesting itself by different suffixes for individual persons, genders, numbers, etc.

We used English syntactic level frames as a basis and adjusted them to Czech; we obtained five frames without any morphological categories – let us call them *initial frames* (they consisted of similar slots as those for English, such as ⟨*noun*⟩ ⟨*verb*⟩ ⟨*prep.*⟩ ⟨*adj.*⟩ ⟨*noun*⟩). Then we used the Czech National Corpus (CNK, 2000) to find a list of real sentences corresponding to each initial frame. To be able to keep syntactic correctness of the resulting sentences, we randomly selected one pattern sentence from each list, analyzed the words in them and tagged them more precisely. The tagging included word class, case, gender, number, tense and person which is needed for unambiguous syntactic description. Then, words in sentences were replaced by their tags and thus we obtained five SUS frames consisting of *enhanced slots*. They are shown together with the original pattern sentences in the following tables.

Note that no gender is specified in the first slot of the following frame – if a particular morphological category is not specified in a slot, any of the forms distinguished within the category can be used without the loss of syntactic correctness. Individual abbreviations are explained in Table 9.

Baterie *The battery* ⟨*noun–sing.–nomin.*⟩	je *is* ⟨*intr. verb–sing.–present–*3$^{\mathrm{rd}}$⟩	pod *under* ⟨*prep.–inst.*⟩
zadním *the back* ⟨*adj.–sing.–inst.–masc./neut.*⟩	sedadlem. *seat.* ⟨*noun–sing.–inst.– masc./neut.*⟩	

Table 4: First Czech SUS frame – intransitive structure.

To keep syntactic correctness in the following frame, only auxiliary verb *to be* can be used (in any tense, however).

Starý *The old* ⟨*adj.–sing.–nomin.–masc.*⟩	pán *man* ⟨*noun–sing.–nomin.–masc.*⟩	je *is* ⟨*verb to be–sing.–masc.*⟩
totalní *a total* ⟨*adj.–sing.–nomin.–masc.*⟩	cvok. *idiot.* ⟨*noun–sing.–nomin.–masc.*⟩	

Table 5: Second Czech SUS frame – transitive structure.

We inserted an extra adjective into the following frame, since quite meaningful sentences would be generated without it even when random words were set into slots.

Vyzkoušejte *Try* ⟨*verb imperative*⟩	různé *different* ⟨*adj.–plur.–accus.–masc./fem.*⟩	odstíny *shades* ⟨*noun–plur.–accus.–masc./fem.*⟩
a *and* ⟨*conjunction*⟩	materiály. *materials.* ⟨*noun–plur.–accus.*⟩	

Table 6: Third Czech SUS frame – imperative structure.

Jak *How* ⟨*adv. of manner/reason/place*⟩	změnit *to change* ⟨*verb infinitive*⟩	systém *the system* ⟨*noun–accus.*⟩
veřejných *of public* ⟨*adj.–plur.–genit.*⟩	financí? *finances?* ⟨*noun–plur.–genit.*⟩	

Table 7: Fourth Czech SUS frame – interrogative structure.

Only reflexive pronouns *se* and *si* are used in the following frame; whenever we used more pronouns of appropriate category, resulting sentences were syntactically correct and meaningless, but they did not sound smoothly.

Vítězslav *Vitezslav* ⟨*noun–sing.–nomin.–masc.*⟩	cítil *felt* ⟨*tr. verb–sing.–masc.–present/past*⟩	potřebu *the need* ⟨*noun–accus.*⟩
se ⟨*reflexive pronoun*⟩	vsadit. *to bet.* ⟨*verb infinitive*⟩	

Table 8: Fifth Czech SUS frame – relative structure.

Substantial work had to be done on collecting words for particular slots; the Czech National corpus was used with success again. Words corresponding to nouns, verbs and adverbials in enhanced slots were looked for in the corpus, the context of following slots was used wherever it was necessary for keeping syntactic correctness. Thus we were able to obtain several thousand words for each of these slots. Words for pronouns, prepositions and conjunctions were designed by hand, as they contain only a few syntactically matching items.

Finally, we generated 250 (50 for each frame) syntactically correct and meaningless sentences by randomly selecting words from the bank and substituting them to the corresponding enhanced slots in the frames; if a word was used, it was removed from the word bank. You can find some of the examples in Table 10.

adj.	adjective	*adv.*	adverb
p.pron.	personal pronoun in the object form		
prep.	preposition	3^{rd}	3^{rd} person (verb)
present	present tense	*past*	past tense
intr.	intransitive verb	*tr.*	transitive verb
sing.	singular	*plur.*	plural
nomin.	nominative	*genit.*	genitive
accus.	accusative	*inst.*	instrumental
masc.	masculine	*fem.*	feminine
neut.	neuter	/	means "or"

Table 9: The explanation of abbreviations used in the frames description.

1: Jídlo visí pod teplým listem.
The food is hanging under a warm leaf.
2: Čtecí bojovník byl běžný parník.
The reading fighter was a usual steamboat.
3: Jezte lesní stoly i vejce.
Eat sylvan tables and eggs.
4: Jak léčit stavbu velkých hostí.
How to treat the construction of big guests.
5: Holič viděl korozi si povídat.
A barber saw corrosion talking to itself.

Table 10: A sample of the Czech sentences generated from the Czech SUS frames.

3.3. MOS and CCR

There is no need of special sentences or their structure design for the MOS and CCR tests. The only requirement according to (ITU P.800, 1996) recommendation is to use simple, meaningful, short and easily understandable sentences selected, for example, from non-technical literature or newspapers. The length of the sentences should be from 5 to 10 words depending on their length (they should fit from 2 to 3 seconds). That is also how we proceeded.

In the CCR test, we used only *prefer A / prefer B* scoring; however, we observed that *prefer A / roughly the same / prefer B* would be more appropriate. There were some listeners who were not able to decide, although most of the others were, and so they guessed.

3.4. The use of the tests

During our text-to-speech system ARTIC development we tried several modifications of a speech corpus segmentation process in order to increase the quality of speech generated by ARTIC, and thus we immediately used the designed tests to establish the speech quality improvement or deterioration, depending on the type of modification (in fact, the tests were originally designed exactly for this purpose).

88 listeners took part in the tests which is quite a large number. Most of the listeners were university students, all without any experience with synthetic speech and even with listening tests. The tests took place in an empty room at our university, but without any special sound isolation or adjustment. The sound was played to listeners from pregenerated wav file via high-quality sound card and high-quality headphones. The results were collected via a PHP interface specially designed to supervise the course of the tests depending on their standardized requirements. In addition, an experienced person explained to each listener his/her task and supervised the tests. The complete set of tests took the listener approximately one hour. Thanks to this, we found the best combination which will from now on be used in our system.

However, we do not intend to present what we tested nor the results of the tests – partly because we want to present the design of the listening tests and partly because the description would considerably exceed this paper.

4. Conclusion

Although it was the first attempt at such systematic design in the Czech Republic, we ended up with fully functional tests for the Czech language, along with keeping standardized framework. These tests can be used (and they indeed were) for the testing and comparing of the Czech synthetic speech.

On the other hand, we can say that there is still some work to be done. One of the tasks is to finish the MRT table by the defining of the remaining 10 six-word groups. As for the SUS test, we did not remove synonyms from word banks, as this still has to be done by hand, which is a rather laborious task. Furthermore, there is a possibility to choose real sentences from the corpus (these being used as patterns for enhanced slots tagging) with the aim of containing shared morphological categories; the word banks would then be organized as some kind of tree structure with specialized slots (described by more morphological categories) in leaf nodes and more generalized slots in superior nodes, containing all words from subordination nodes.

There is also a possibility of analyzing the theoretical distribution of the most probable Czech sentence types (if there is such distribution) and of the design of SUS frames according to this.

5. References

J. Matoušek, J. Psutka 2000. ARTIC: A New Czech Text-to-Speech System Using Statistical Approach Approach to Automatic Speech Segment Database Construction. *ICSLP2000 Proceedings*, IV:612–615.

Ch. Benoît, M. Grice, V. Hazan 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18:381–392.

ITU-T Recomendation P.800 1996. Methods for objective and subjective assessment of quality.

X. Huang, A. Acero, H-W. Hon 2001. *Spoken Language Processing. A Guide to Theory, Algorithm, and System Development*, chapter 16, pages 834–844. Prentice Hall PTR, New Jersey

CNK – The Czech National Corpus – SYN2000 2000. The Czech National Corpus Institute, Prague.