ASSESSMENT OF VARIOUS ASPECTS OF SYNTHESIZED SPEECH QUALITY

Milan Sečujski[1], Darko Pekar[2]

[1]Faculty of Engineering, Novi Sad; [2]AlfaNum d.o.o, Novi Sad
secujski@uns.ns.ac.yu

ABSTRACT

The paper contains a description of a TTS quality assessment experiment, aimed at determining whether the listeners tend to form their opinion on the basis of intelligibility or rather prosody naturalness.

The experiment also examines the idea of using natural f_0 contour extracted from another sentence with identical accentuation configuration but different information content and syntax structure for synthesis.

1. INTRODUCTION

Synthesized speech quality assessment represents a well known problem, not yet solved in a satisfactory way, the reason for that being the fact that synthesized speech quality includes various interdependent factors. That is why the only means of evaluation are listening tests, where a number of listeners award points to synthesized utterances based on MOS (*Mean Opinion Score*) scale. In this way it is possible to compare different speech synthesizers (provided that they are examined simultaneously), as well as to estimate the contribution of various factors to the overall quality of synthesized speech.

The AlfaNum TTS synthesizer for Serbian language used in this experiment is an example of concatenative synthesizers using a large speech database, and performing segment selection at runtime. Selected segments are then concatenated, having performed some speech processing beforehand, mostly based on the TD-PSOLA speech model. The material for the quality assessment experiment was motivated by the fact that Serbian language belongs to the group of tonal languages, and thus f_0 contour variations have a more substantial lexical role than in languages with stress accent. The aim of the experiment was to establish the contribution of various aspects of synthesized speech quality to the general satisfaction of the listener, which, subjective as it may be, remains the only reliable measure of the quality of synthesis.

1.1. Synthesized speech quality

The *intelligibility* and *naturalness* of synthesized speech are said to be the most important aspects of its quality [1].

The intelligibility of synthesized speech actually represents its quality at the phone level. The more the articulation of each phone is natural and clear, the higher the intelligibility is. However, if the presence of artifacts caused by errors in the database or too excessive digital signal processing can be observed, the speech itself will be less intelligible. The listener is, fortunately, able to reconstruct damaged phones, especially in case of meaningful utterances – in such a case even the wider context can be taken into account. That is why the sets of anomalous sentences are often used for intelligibility assessment, preventing the listener from reconstructing damaged or missing elements based on the semantic context easily, albeit the sentences are syntactically correct. Some examples of sentences from standard Haskins set for intelligibility assessment for English language are:

- The great car met the milk.
- The short arm sent the cow. [2]

Beside these, semantically unpredictable sentences (SUS), based on introducing words selected at random into a predefined syntactic pattern, are also used. The result is similar to the sentences from the Haskins set, with the advantage of being more diverse and thus preventing the learning effect [2]. In a somewhat wider sense, the intelligibility of synthesized speech is also related to the ease of extracting information contained in the utterances. If listeners have to focus on combining the sounds heard into meaningful units, it is noted that they sometimes cannot say afterwards what the text was about in the first place [3]. Furthermore, the ease of extracting inforamtion from synthesized speech also determine how fast will the listeners begin to suffer from fatigue.

The naturalness of synthesized speech is most often defined as its resemblance to natural speech. In order to exclude as much elements already taken into account through intelligibility as possible, naturalness can be identified as the listener's impression on the similarity of *intonation* of the synthesized utterance with the intonation present in a naturally spoken utterance with the same content. It is, however, clear that even such a definition does not make naturalness completely independent from intelligibility, since the reconstruction of missing or damaged phones is easier if the listener can rely on the meaning of the utterance, and meaning is conveyed by natural intonation as well.

It should be noted here that the term *quality of synthesis* does not denote only the quality of synthesized speech, but also the capability of the system to deal with text other than orthographic words, such as numbers and abbreviations, in a correct way, depending on the context. It is clear that grave errors in text preprocessing, the first phase of speech synthesis, reduce the intelligibility in a wider sense, and thus this aspect of synthesized speech quality as well is not independent from the others. However, this paper will not deal with

it, and it will be supposed that the text contains orthographic words only.

1.2. The tonality of Serbian language

The Serbian language belongs to a relatively small group of *tonal languages*. It means that, unlike the languages with stress accent, where a syllable can be either stressed or unstressed, and there are no minimal pairs of words with the same phonetic content and different prosodic content, distinguishable only by different pitch movements, in Serbian language a different accent type, i.e. different pitch changes can express a difference in morphological categories:

(*rêčі* [gen.sg.]↔*réčі* [gen.pl.])

as well as convey a different lexical word:

(blâga[n.]↔blága[adj.]).

In stress accent languages, the information carried by pitch variations is rather pragmatic, whereas in tonal languages they have a lexical function as well. It is, thus, essential to take such information into account when designing a high quality speech synthesizer. The importance of pitch as a key prosody element, as well as the fact that it is strongly influenced by accentuation, has suggested the idea of using an f_0 contour from one sentence for synthesis of another sentence with the identical accentuation, with some modifications. It was also possible to compare the sentences synthesized using this metod with sentences synthesized using an f_0 contour generated exclusively based on accentuation. The results of such comparisons are presented in the following text.

2. THE EXPERIMENT

In this section, the experiment of synthesized speech quality assessment will be presented in detail. The synthesizer used was the AlfaNumTTS synthesizer in Serbian language. The experiment was carried out in laboratory conditions, at the Faculty of Engineering in Novi Sad, in July 2004. Ten listeners took part in the experiments, and they were to listen to five pairs of sentences synthesized in nine different variants.

2.1. The assessment material

Considering the importance of pitch as a key prosody element, and the fact that in a tonal language such as Serbian it is deeply influenced by accentuation, the idea of the experiment was to find out if the synthesized speech quality can be improved by using an f_0 contour borrowed from existing sentences with appropriate accentuation, instead of synthesizing f_0 from the scratch. For example, if the speech database contains the sentence:

Nȁši kȕmovi su na vréme òtišli

then it is possible to compose another meaningful sentence with matching accentuation structure and different phonetic content and (generally) different stressed vowels:

Nȅ znam dȁ li ću da prežívi òdlazak

It is, then, possible to extract the f_0 contour from the first sentence, and to modify it in such a way that the segments of the original f_0 contour related to vowels are positioned at vowel segments of the target sentence, preserving phoneme durations in the target sentence. An example of such a time alignment is given in Figure 1. In such a way, local f_0 variations that result from accentuation are applied to the target sentence. If the f_0 contour is indeed shaped based on accentuation only, natural f_0 variations in the target sentence should have been identical. In order to avoid excessive f_0 contour distortion, an additional precaution measure was taken: if the original sentence contained two vowels with no consonant between them, the target sentence was composed in such a way that there was no consonant in a corresponding place as well.

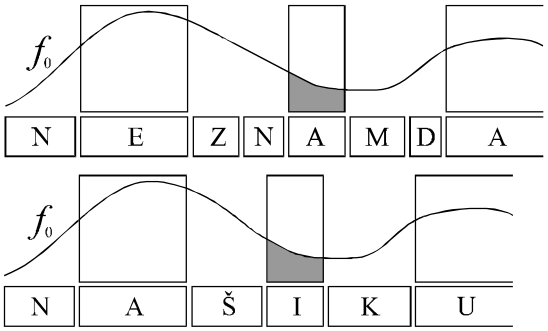

Figure 1. Time alignment of f_0 contours

Following the principles described above, 5 sentences from the speech database used by the AlfaNumTTS synthesizer were selected, and another 5 sentences with the same accentuation were composed, presented in Table 1:

Table 1. Sentence pairs used in the experiment

#	Sentences (original and composed)
1	Náčin stȉcanja zârada je uglàvnom ȉsti.
	Slúga Rȁdovan kâže da će da ìde kȕći.
2	Nȁši kȕmovi su na vréme òtišli.
	Nȅ znam dȁ li će da prežívi òdlazak.
3	Hrámovi su òpstali zahvàljujući dònātorima.
	Tréneri su zàbrinuti njègovom obèshrābrenošću.
4	Znâm jâ mnȍgo štȁ, vȉše nego što mȍžeš i zàmisliti.
	Kô znâ zȁšto bȁš zàgovornici rȁtova nè ratuju.
5	Òna se zapíljila u maglòvitu daljìnu.
	Òtpušten je návodno da se ìzbegne sramòta.

Three variants of each of these sentences were synthesized and presented to the subjects. The first variant was synthesized based on an f_0 contour generated automatically based on accentuation, and therefore identical in both sentences. In the following text, these sentences will be referred to as O_AUT and C_AUT. The second variant was based on the natural f_0 contour taken from the original sentence, and these sentences will be referred to as O_NAT and C_NAT.

The third variant was synthesized using an f_0 contour obtained as the average of the f_0 contour generated automatically and the f_0 krive borrowed from the original sentence, and these sentences will be referred to as O_MIX and C_MIX.

Beside the six variants described above, another three were presented to the subjects. Both the original and the composed sentence were synthesized using the natural f_0 contour taken from the original sentence, i.e. the same as was the case with O_NAT and C_NAT, but the intelligibility of these sentences was lower, because they were synthesized using only a smaller part of the speech database, and some limitations were introduced in the process of segment selection as well. In this way, the sentences were synthesized using less convenient speech segments (generally) than in case of synthesis with no such limitations. These sentences will be referred to as O_DEGR and C_DEGR. The last variant represented the sentence synthesized using only the segments from the original sentence in the speech database, that is, with no cuts and concatenations of non-matching segments, but with an f_0 contour automatically generated based on accentuation, as in sentences O_AUT and C_AUT. This sentence will be referred to as O_NOSEG. Introduction of the last three variants was motivated by the intention to find out if the listeners prefer sentences with absolutely natural intonation, which would enable them to reconstruct intelligibility impairments that might occur, or they find intelligibility more important and rely on it more.

2.2. The experiment description

The subjects listened to the aforementioned 5 pairs of sentences within 5 slides containing a visual display of the sentences in a particular order as well as fields for entering points. The subjects listened to the sentences in silence, in identical conditions. They were asked to award points according to the MOS scale, and to grade three aspects – beside intelligibility and naturalness, a general subjects' preference was also judged to be of importance, i.e. the subjects were asked to decide if they used a speech synthesizer regularly, how fond would they be of synthesized speech that sounds just like that.

The position of different variants on all slides was the same, since some of the variants were interesting for direct comparison, and they were always placed next to each other on the slides. The subjects were allowed to listen to the sentences in an arbitrary order, but they were asked to pay special attention to adjacent pairs of sentences, and to try not to miss comparing them directly. The visual arrangement of the sentences is shown on Figure 2.

The succession effect was somewhat disregarded by having different variants always in the same place [3], but the subjects were suggested not to listen to the sentences in the same order on every slide.

2.3. The experiment results

The experiment results (the average intelligibility, naturalness and general impression grade according to the MOS scale, for every sentence variant) are shown in Table 2:

Table 2. The experiment results

	intelligibility	naturalness	impression
O_AUT	3,88	3,26	3,36
O_MIX	3,82	3,28	3,44
O_NAT	4,32	4,14	4,06
C_AUT	4,20	3,42	3,66
C_MIX	3,76	3,30	3,22
C_NAT	3,58	3,20	3,10
O_DEGR	4,20	4,06	3,96
C_DEGR	3,12	2,94	2,78
O_NOSEG	3,94	3,50	3,42

2.2. The experiment results analysis

The most striking thing was that the results were less variable than expected. This is in accordance with the subjects' comments – everyone has the impression that the marks given were unreliable, and some of the subjects had the impression that there was no difference whatsoever between some sentence variants (despite the fact that careful listening in all cases reveals significant differences).

The answer to te question whether the subjects prefer an automatically generated f_0 contour composed based on accentuation only, or a natural f_0 contour taken from a sentence with matching accentuation structure could be obtained by comparing results for C_AUT, C_MIX and C_NAT. The listeners estimate that the sentence with an f_0 contour automatically generated sounds more intelligible than a sentence with an f_0 contour taken from natural speech, from another sentence with matching accentuation (the difference in grade average being 0,62). Moreover, they estimate tat it sounds more natural (0,22), and they would prefer to use a speech sythesizer that generates prosody features automatically (0,56).

An explanation of this, seemingly paradoxical result could be given based on analysis of the results for O_AUT, O_MIX and O_NAT. If speech synthesis is carried out using an f_0 contour extracted from the *same* sentence, rather than from another sentence with the same accentuation, the situation changes. In that case, the subjects consider the variant with the original intonation not only as more natural (0,88), but as more intelligible as well (0,44), and prefer it to the other (0,70). This leads to the conclusion that the f_0 contour is

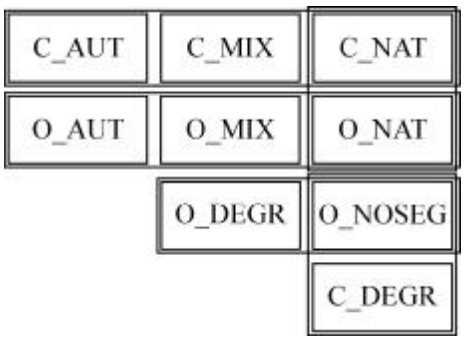

Figure 2. Slide layout

influenced by factors other than accentuation, or that applying an f_0 contour from one sentence to another required other prosody modifications, such as duration or energy adjustment.

This conclusion is further confirmed by a big difference in grading of O_NAT and C_NAT (0,94 for preference), especially big on slides 4. and 5. This should not be surprising, since both sentences on slide 4. are emotionally charged – the first one is in the first person, even the word order is not standard, but the verb is at the first position, while the other expresses a personal opinion of the speaker and contains a rhetorical question. The slide 5. used a sentence from the speech database having a problematic f_0 contour in the first place, which is particularly apparent if it is applied to a sentence with different information content. The situation is the same with results for O_DEGR and C_DEGR, where the difference is even bigger (1,08 for intelligibility, 1,12 for naturalness and as much as 1,18 for general impression, all in favour of O_DEGR). On slide 4. the difference in naturalness is 1,7, and in general impression as much as 2,1 in favour of O_DEGR. This suggests that naturalness becomes increasingly important in adverse conditions, when intelligibility is lower, and especially in case the sentence is emotionally charged.

As to the question if a user of a speech synthesizer relies on intelligibility or naturalness more, it is well known that it varies from person to person, but that it is possible that variations of one of the two can have greater impact on overall speech quality in case a particular speech synthesizer is used. When the subjects compared O_DEGR and O_NOSEG directly, they decided in favour of O_DEGR more often (the difference in intelligibility being 0,24, in naturalness 0,56, and in general impression 0,54). Such a difference in general impression may seem illogical, since O_NOSEG should be ideally intelligible, since it was synthesited using the original sentence segments. Beside the already established dependence of f_0 contour and information content, the reason for this could be the fact that f_0 contour modification ultimately damages the signal. That is why O_NOSEG does not sound ideal – its intelligibility is impaired as well as its naturalness. Although TD-PSOLA is considered a technique that essentially preserves the integrity of the speech signal when modifying prosodic features, it is not that ideal when the changes are not uniform throughout the signal. In other words, if the f_0 contour of the entire sentence is to be raised or lowered by 10%, the resulting sentence will sound very intelligible and natural. However, if the f_0 contour should at some segments be raised and in others lowered, and if those segments are adjacent, it may result in a signal that does not sound natural any more, and the artifacts of such synthesis may be attributed to poor intelligibility. This fact makes the task of synthesized speech quality assessment even harder, since it is an additional factor of interdependence between intelligibility and naturalness.

In order to get a definite answer to the question of influence of intelligibility and naturalness to overall quality, the results for five variants of the original sentence were compared (O_AUT, O_MIX, O_NAT, O_DEGR and O_NOSEG) and the correlation between the general impression grade and intelligibility and naturalness grades was computed. For all five variants the answer was the same – that the contribution of naturalness is more significant (the Euclidean distance from general impression grade to the naturalness grade was 3,81 times smaller than the distance from general impression grade to the intelligibility grade). However, the problem of the automatic generation of prosody features that would reflect not only accentuation but syntactic structure and information content is still far from solution.

3. CONCLUSION

In this paper an experiment of assessment of various aspects of synthesized speech quality was described. The experiment included the analysis of the contribution of some factors to the overall speech quality. The speech was synthesized using the AlfaNumTTS speech synthesizer. Several problems of synthesis quality evaluation were pointed out, and the results obtained represent a contribution to speech synthesis in Serbian, and they will be used in improving the AlfaNumTTS synthesizer quality as well.

The presumption that the impact of accentuation on f_0 contour, and the entire prosody, was partially confirmed, considering the fact that an f_0 contour generated based on accentuation only was still quite satisfactory for most listeners. However, it turned out that this influence is still not such that would allow applying f_0 contours extracted from one sentence to another sentence with the same accentuation structure, not taking into account syntactic and other differences between them.

LITERATURE

[1] T. Dutoit: *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht/ Boston/London, 1997.

[2] S. Lemmetty: *Review of Speech Synthesis Technology*, M.Sc.E.E. Thesis, Helsinki University of Technology, 1999.

[3] *Assessing Text-to-Speech System Quality*, White Paper, SpeechWorks International.

[4] I. Lehiste, P. Ivić: *Word and Sentence Prosody in Serbocroatian*, The Massachusetts Institute of Technology, 1986.

[5] M. Sečujski, R. Obradović, D. Pekar, Lj. Jovanov, V. Delić: AlfaNum System for Speech Synthesis in Serbian Language, *TSD 2002*, Brno, Czech Republic, 2002.