EFFECT OF PROSODIC NATURALNESS ON SEGMENTAL ACCEPTABILITY IN SYNTHETIC SPEECH

Martti Vainio[1], *Juhani Järvikivi*[2], *Stefan Werner*[2]

University of Helsinki
[1]Department of Phonetics and
[3]Department of General Linguistics
P.O. Box 35, FIN-00014, Finland

Nicholas Volk[3], *Jarmo Välikangas*[2]

University of Joensuu
[2]Linguistics and Language Technology
P.O. Box 111, FIN-80101,
Finland

ABSTRACT

It is commonly agreed that one of the major goals in the development of modern text-to-speech synthesis is the improvement of prosody, especially intonation. Although high quality intonation is an important factor on the way to more natural synthetic speech, it is seldom scrutinized empirically whether and how this affects the relative performance of other components, such as segmental synthesis. The present paper discusses two preliminary rating experiments inquiring into the relation between the naturalness of intonation and subjective segmental quality in Finnish. Experiment 1 showed that the perception of intonation is dependent on the segmental quality. More crucially, Experiment 2 indicated that also the perceived segmental acceptability is significantly dependent on the relative naturalness of intonation. In light of the present observations, the goal of improved intonation is not only desirable for the overall quality's sake alone, but it is also shown to improve even the subjective perception of a very basic feature of synthetic speech such as segmental acceptability.

1. INTRODUCTION

Evaluation of text-to-speech synthesis systems through perception tests is commonly regarded as a necessary task both for development of new and improvement of existing systems. There is, however, no methodology available which is widely agreed-upon and proved valid over and above a few basic standardized paradigms [1]. Intelligibility ratings alone do not answer the nowadays more relevant questions of acceptability and naturalness [2]. A further caveat affecting especially the subjective assessment methods is created by the inherently multidimensional nature of perception [3]. Although, it is, more or less, accepted as a fact that by improving naturalness of synthetic speech, one also improves the intelligibility and comprehensibility (see Sproat et al in [4]), it is not self-evident that the relation between various parameters, such as segmental and prosodic quality, is linear in a similar manner.

For example, Terken and Lemeer [5] found that the segmental quality of synthetic speech affected the participants preferences of natural over flat intonation: natural intonation was preferred over the dull one in the high segmental quality cases but not in the low segmental quality utterances, where both intonations were judged as equally acceptable. The authors conclude that the expectations as to the naturalness of synthetic intonation are directly affected by the segmental quality of the synthesizer.

To our knowledge, there are no studies that inquire directly into the acceptability of the segmental side relative to the naturalness of the synthetic intonation. As it is clear that the two are interdependent, it is reasonable to assume that the quality of intonation may also influence the segmental acceptability in subjective perception. Predicting the direction of the influence is not straightforward and the possible effect may depend on the relative segmental quality of the synthesized speech, just as did the intonation preferences in the Terken and Lemer study. Given a reasonably high quality synthesizer, however, it seems intuitively reasonable to expect that increasing naturalness of the intonation would be likely to exert a positive influence on the segmental acceptability as well, whereas, ceteris paribus, a low quality segmental synthesis may well be perceived as relatively even less acceptable with a (near) natural intonation than it would with a low quality flat intonation contour.

Two perceptual rating experiments were designed to inquire into the relation between intonational naturalness and segmental acceptability in Finnish. In Experiment 1, the relative naturalness of natural and flat intonation was studied in the context of natural and synthesized segmental speech. Experiment 2, in turn, was designed to tap into the segmental acceptability of synthesized speech in relation to natural and flat synthetic intonation.

2. A HIGH-QUALITY FINNISH TEXT-TO-SPEECH SYNTHESIZER

The system examined here is a TTS system for Finnish that is currently developed within the framework of the Joint Finnish Speech Technology Project [6]. It is based on Festival speech synthesis system and development framework [7], integrating its own modules for, e.g., text pre-processing and prosody control. The tested version uses a diphone inventory of approximately 1200 elements, recorded from a single male native speaker. The signal generation method in the system is based on residual excited LPC, which is generally considered fairly good in prosody matching qualities but poor in segmental quality [8]. Nevertheless, the segmental quality is adequate for relatively high-quality text-to-speech systems. An example, which depicts the segmental differences between synthetic and natural speech, can be seen in Figure 3.

3. FINNISH PROSODY

In relation to this study, it may be relevant to note, that in Finnish, intonation does not play as important a role as in many other languages. For instance, intonation is not systematically used to mark questions (which are marked with lexical means by interrogative

0-7803-7395-2/02/$17.00 ©2002 IEEE

particles) [9]. The basic intonation shape in Finnish is falling with all content words receiving an accent [10]. All of the utterances in this study follow this shape.

Finnish is also a so called *quantity language* with two degrees of quantity (short and long) for all sounds. The prosodic parameter responsible for the perception of quantity is segmental duration; the long sounds are, on the average, twice as long as the short ones. Nevertheless, the actual duration of a sound depends on a multitude of factors and, for instance, a short vowel in a stressed syllable may be twice as long as another one within the same word (see [11]). Long consonants are always geminates and have a syllable boundary within them, which renders them ambisyllabic.

Finnish has a fairly free word-order, rich morphology with suffixation and enclisis. The number of grammatical cases is fairly large (15). Suffixation, enclisis and free word order has lead to the situation where the lexical morphemes are situated at the beginning of words and, invariably, receive the lexical stress – i.e., the lexical stress is always on the first syllable of the word. The tendency for suffixation can be seen on the level of the utterance as the basic falling shape of the intonation contour, which usually ends with a creaky or whispery voice. Creak is used in Finnish as a turn-yielding construct and it usually has a fairly abrupt onset [12]. All of the utterances in this study exhibit a creaky ending. The creaky voice has some consequences for the results; see Section 5.

4. EXPERIMENTS AND RESULTS

4.1. Experiment 1

The first experiment was carried out with two main objectives in mind: First to replicate the earlier observations of the prosodic naturalness relative to segmental quality, and second, to establish a relative perceptual difference between the three intonational conditions.

Materials. The materials consisted of 15 phonetically balanced sentences with an average duration of approximately 3 seconds, which were recorded by a male, native speaker of Finnish (*natural* stimuli) in a relatively sound-proof room at the Department of Phonetics onto a digital audio tape (Tascam DA-P1) with a noise reducing condenser microphone (AKG HSC 200 SR) using the tape recorder's AD converter and microphone pre-amplifier. The digital recordings were transferred to a computer where they were further processed by removing unwanted environmental noise by gating and down-sampled to 22050 Hz to match the synthetic stimuli.

The recorded utterances were manually segmented by a trained phonetician and the F_0 contours were extracted with an auto-correlation based pitch detector. The segmentation and F_0 contour were used to produce two sets of synthetic stimuli: a set with both superimposed segmental durations and an F_0 contour (*semi-synth* stimuli) as well a set of synthetic stimuli with superimposed segmental durations and flat F_0 contour (*synth-flat* stimuli). Both synthetic sets were produced with the Festival system described in Section 2. The *natural* stimuli were scaled so that the average intensity of the stimuli matched with the *semi-synth* stimuli within one dB. The same speaker who was responsible for the original diphone inventory of the synthesizer was used for the recording.

Subjects. Nineteen students from the University of Joensuu participated in Experiment 1. All were native speakers of Finnish and reported no hearing problems.

Procedure.The 45 sentences were presented to the listeners over a professional-grade PA system installed in a university auditorium. A computer program was used to select and play the stimuli, add pauses of constant length (4 s) between them, and simultaneously display their number through a data projector. The subjects were instructed to evaluate on a scale from 1 to 7 how natural they thought the intonation of each sentence was on a scale from 1, extremely unnatural, to 7, extremely natural. The subjects responded by checking the appropriate box on a printed form. Before the experimental trials, the subjects responded to six practice trials.

Results. Figure 1 summarizes the average mean opinion scores from the three conditions. As expected the scores from the *semi-synth* condition(4.2) fell between the *natural* (6.7) and *synth-flat* (1.8) conditions (SDs 0.7, 1.4 and 1.2, respectively). An analysis of variance confirmed that the observed difference between the three conditions was also statistically significant (F (2,36) = 253.21, p < .0001). Further pair-wise comparison (a dependent groups t-test) revealed a significant difference between the *natural* and *semi-synth* conditions (t (18) = 16.02, p < .0001) as well as the *semi-synth* and *synth-flat* conditions (t (18) = 9.96, p < .0001).

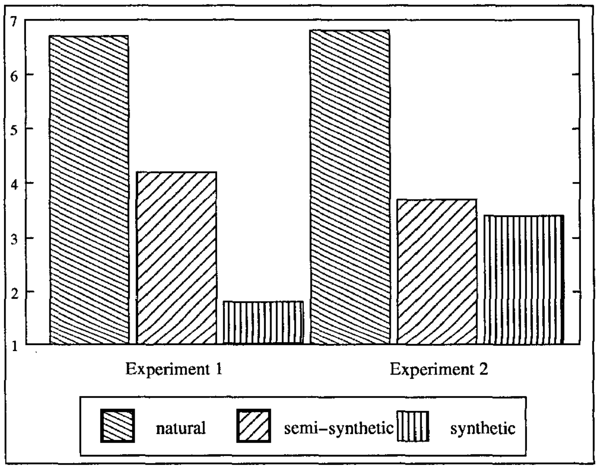

Fig. 1. Mean Opinion Scores (MOS) in Experiments 1 and 2.

The results show a comparable effect to that observed by Terken and Lemeer [5] for Dutch, in that the segmental quality of the voice significantly affects the perception of the relative naturalness of the intonation contour. Furthermore, it is notable that this applies even to the natural intonation contour employed for the sentences in both the *natural* and *semi-synth* conditions.

Having established relative perceptual difference between the three intonation conditions, we proceeded to inquire into whether the perception of segmental quality would be affected by the relative naturalness of the intonation to a similar degree. In other words, here the critical contrast is between the identical segmental qualities of the synthetic conditions, differing only in the natural intonation of the *semi-synth* and *synth-flat* intonation of the synthetic conditions.

4.2. Experiment 2

Materials. Materials were the same that were used in Experiment 1.

Subjects. Eighteen of the subjects from Experiment 1 participated in Experiment 2.

Procedure. The procedure was identical to that in Experiment 1 in all other respects, except the subjects were instructed to consider how acceptable they thought the segmental quality of each sentence was, and were asked to respond on a scale from 1, highly unacceptable, to 7, highly acceptable by checking the appropriate box on the test form.

Results. The results are again summarized in Figure 1. As expected the top line condition (natural) received the highest mean opinion scores (MOS 6.8 ; SD 0.6) of the three, the semi-synth and synth-flat conditions differing only slightly from each other (MOS 3.7 and 3.4; SDs 1.57 and 1.65, respectively). An analyses of variance for the rating data revealed a reliable main effect of condition (F (2,34) = 101.3, $p < .0001$). However, in a pairwise comparison the difference between the two critical conditions was only marginally significant (t (17) = 1.51, p = .07). To assess the generalizability of the observed marginally significant difference, a further t-test was carried out by experimental stimuli in the two conditions. This time a significant difference was observed between the subjective segmental acceptability of the two synthetic conditions (t2 (28) = 2.06, $p < .03$). Thus, the results suggest that the perceptual segmental acceptability of synthetic speech is affected by the relative naturalness of intonation to a signifigant degree.

5. SUMMARY AND DISCUSSION

The present study is a preliminary step towards evaluating the relative influence of various parameters on the perceived subjective acceptability and naturalness of text-to-speech synthesis. The first experiment showed that the naturalness of intonation in Finnish is affected by segmental quality to a significant degree. Similar effects have also been observed earlier by Terken and Lemeer [5] for Dutch. More crucially, the results from Experiment 2 indicate that the mutual interdependence of the two experimental parameters work also the other way around. Thus, exactly the same segmental synthesis was judged as significantly more acceptable when it was accompanied with a natural intonation countour than when a flat countour was used. As the interpretation of the results depends on the assumption that the two conditions differed only with respect to the factor on intonation, a closer look at the experimental stimuli is in order to make sure no other factors intervened in a signifigant degree.

The fact that the improvement of the perceived segmental acceptability of the semi-synthetic stimuli in Experiment 2 was not more significant may be explained by a series of hypotheses which relate to the actual segmental quality of the synthetic stimuli. That is, there are reasons to believe that the process of transferring the F_0 values from the natural utterances to the semi-synthetic stimuli may, in fact, degrade their segmental quality. The following factors may have a negative influence on the actual segmental quality of the semi-synthetic signals:

1. Asynchronous micro-prosodic variation due to imperfect transfer of natural prosody and pitch detection of the original signal.
2. Possible shape mismatch between the intensity and F_0 contours, which is liable to produce perceivable discontinuities on the segmental level.
3. Lack of creaky and whispery voice at the end of the utterances.

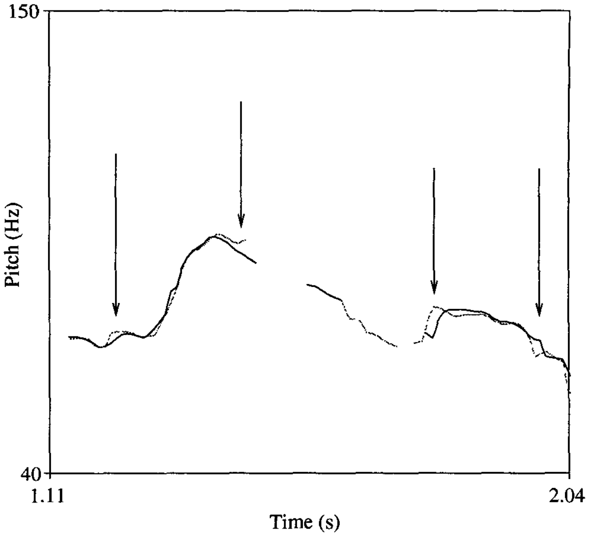

Fig. 2. Incomplete and erroneous transfer of the F_0 contour from the natural stimuli to the semi-synthetic ones: the arrows show the most obvious differences. The solid line depicts the actual values and the dotted line the transferred values. Both F_0 contours were produced with an auto-correlation based pitch detector from the actual stimuli.

The mapping of the natural F_0 contour on the synthetic stimuli was not accurate on the segmental level and micro-prosodic variation of the original signals was not correctly copied. In fact, there were obviously problems with the mapping and the micro-prosody in the synthetic stimuli could, in fact, be erroneous. The unsatisfactory mapping of micro-prosody on the semi-synthetic stimuli is clearly observable in Figure 2.

Perhaps the biggest problem concerning the relationship between prosody and segmental quality in Finnish synthetic speech, is the lack of proper models for the creaky voice, which permeates Finnish speech to the degree that it's absence is bound to have a degrading effect on the synthesis. Figure 3 shows a comparison of two utterance final words in natural and synthetic utterances. The creaky voice can be easily seen in the upper panel. Moreover, the synthetic utterance has excessive energy during the final segments, which is bound to lead to an unnaturally loud ending.

For future experimentation we need to produce semi-synthetic stimuli which minimize the problems concerning segmental quality. This can be done by avoiding micro-prosodic variation altogether by using a parametrized intonation contour (e.g., Fujisaki's model or a straight line stylization). The other two degrading conditions are not easily repaired since the concatenation method used does not allow for controlling either loudness or voice quality.

A baseline for a further version of Experiment 2 can be produced by constructing a set of synthetic stimuli which bear no

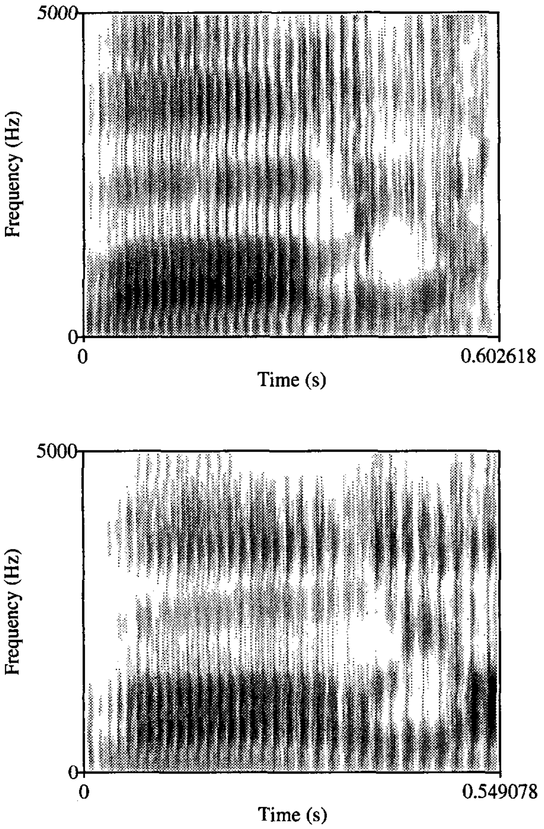

Fig. 3. A comparison of natural (upper panel) and synthetic speech (lower panel) – word "vaalia" (to shelter) from the end of the utterance "Joitakin perinteitä on mukava vaalia." (It is nice to retain certain traditions.). The creaky voice of the natural token can easily be seen as a lack of predictable structure within the pitch pulses.

prosodic relation to the natural utterances. This can be done by using randomly fluctuating segmental durations in addition to a flat, or virtually flat, F_0 contour.

It is widely agreed that the emphasis on the naturalness development of text-to-speech synthesis should be directed on its contemporarily weakest aspect, i.e., prosody, and intonation especially. It is taken as granted that increasing the naturalness of intonation will have a positive influence on the overall quality of synthetic speech. However, not much research has been devoted to actually determining experimentally whether this has any direct influence on the perception of other individual aspects, such as segmental acceptability. The present results indicate that this is indeed the case. Thus, a highly natural intonation seems to bring along a proverbial free lunch in terms of improved segmental acceptability judgments. That this is indeed the case even with a relatively high quality segmental synthesis further motivates the need for high quality prosody development for text-to-speech systems.

Whether the problems discussed above had a considerable effect on the perceptual data would deserve a study of its own. All in all, however, the most important point to note with regard to the present results is that any such effect would have worked to diminish the observed difference between the segmental acceptability judgments in the two synthetic conditions. Therefore, the more careful controlling of the stimuli in these respects can be expected to clarify the difference even further.

This work has been financed by the Finnish National Technology Agency (TEKES) under grant 1387/401/00.

6. REFERENCES

[1] Dafydd Gibbon, Roger Moore, and Richard Winski, Eds., *Spoken Language System and Corpus Design*, vol. 1, Mouton de Gruyter, 1998.

[2] K. Silverman, S. Basson, and S. Levas, "Evaluating synthesiser performance: is segmental intelligibility enough?," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1990)*, 1990, pp. 981–984.

[3] Jan van Santen, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, chapter Evaluation, pp. 229–244, Kluwer Academic Publishers, 1998.

[4] Richard Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1998.

[5] J. Terken and G. Lemeer, "Effects of Segmental Quality and Intonation on Quality judgements for texts and utterances," *Journal of Phonetics*, vol. 16, pp. 453–457, 1988.

[6] Martti Vainio, Stefan Werner, Nicholas Volk, Jarmo Välikangas, and Juhani Järvikivi, "Finnish Speech Technology: A Multidisciplinary Project," Unofficial web page at `http://www.ling.helsinki.fi/suopuhe/`.

[7] Alan W. Black, Paul Taylor, and Richard Caley, "The Festival Speech Synthesis System system," URL: `www.cstr.ed.ac.uk/projects/festival.html`.

[8] Thierry Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.

[9] Antti Iivonen, "Intonation in Finnish," in *Intonation systems – A survey of twenty languages*, Daniel Hirst and Albert Di Cristo, Eds., pp. 311–327. Cambridge University Press, Cambridge, 1998.

[10] Hansjörg Mixdorff, Martti Vainio, Stefan Werner, and Juhani Järvikivi, "The Manifestation of Linguistic Information in Prosodic Features of Finnish," in *In Proceedings of Prosody 2002*.

[11] Martti Vainio, *Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis*, Number 43 in Publications of the Department of Phonetics, University of Helsinki. Yliopistopaino, 2001.

[12] Richard Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, June 2001.