Quality Deterioration Factors in Unit Selection Speech Synthesis

Daniel Tihelka, Jindřich Matoušek, and Jiří Kala

University of West Bohemia, Faculty of Applied Sciences,
Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
{dtihelka, jmatouse, jkala}@kky.zcu.cz

Abstract. The purpose of the present paper is to examine the relationships between target and concatenation costs and the quality (with focus on naturalness) of generated speech. Several synthetic phrases were examined by listeners with the aim to find unnatural artefacts in them, and the mutual relation between the artefacts and the behaviour of features used in given unit selection algorithm was examined.

1 Introduction

The quality of synthetic speech, especially its naturalness, represents an issue which still remains open. Although the *unit selection* approach, or the modern *HMM–based synthesis*, are able to reach fairly high level of naturalness, there are still unnatural artefacts of a different nature (depending also on the synthesis method) occurring in synthesized speech when input text comes from the unlimited domain.

The paper evaluates unnatural artefacts (also called *glitches*) perceived in speech generated by unit selection incorporated in our TTS ARTIC [1,2,3], and attempts to reveal the relation between those artefacts and the behaviour of features used for target and concatenation costs computing. There have been a number of papers written, mostly focusing on the relation of human perception to spectral subcost of concatenation cost only – [4,5,6,7,8,9,10] belong among those most important. However, we aim to look at the whole unit selection scheme globally, to examine all the expected or potential factors which can cause the perceived artefacts (for the terminology defining the terms *artefact* and *factor* see [11] or short reminder in Section 2.1). The problem is the greater number of possible causes of artefacts in the unit selection, as well as usual co-incidence of several factors leading to perceived artefact.

The paper is organised as follows. Section 2 describes special evaluation methodology which has to be used for this task, covering also the related terminology. Section 3 is focused on the procedure of artefact collection, defines the set of expected causes (factors) of artefacts, and describes, how the artefacts and factors were related together. The results and the conclusions are presented in Section 4 and Section 5 respectively.

V. Matoušek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 508–515, 2007.

2 Evaluation Method

It is obvious that the outlined aims of the paper require the human evaluation of synthetic speech, i.e. listening tests. However, as the standard listening tests (e.g. MOS, CCR or different kinds of decision tests) usually focus either on overall quality (e.g. how close to natural the synthetic phrase is), or on one kind of artefact (e.g. does a particular concatenation point sound continuous or not), none of them is suitable for the proposed purposes. We, therefore, used special methodology designed in [11]. Although the methodology was first used for the evaluation of our single–instance version of our TTS, it is applicable for the larger set of evaluation types and/or system types. The disadvantage of the approach is its labour input, but to our knowledge it is the most straightforward way of analyzing the collected unnatural artefacts.

2.1 Terminology

Let us review terminology for this kind of evaluation, as established in [11]. An *artefact* is a term denoting every event or place in synthetic speech which annoys human listeners and causes distortions in synthetic speech being perceived as unnatural; it can also be distributed to several consecutive phones. Artefacts are represented by various kinds of glitches, clicks, cracks, murmuring, speech or prosody discontinuities, machine-like or artificial sounds and so on.

In [11] we also divided artefacts into *fragments*, as the artefacts determined by listeners overlapped in many ways due to the natural fuzziness of human perception. However, we do not explicitly build fragments in the present evaluation, as described in Section 3.1.

Finally, it is clear that artefacts are caused by a coincidence of what is carried out in the TTS algorithm before speech is created. Therefore, the set of actions which can influence the synthetic speech is usually explicitly defined, being called the set of *factors*.

3 Evaluation Procedure

The whole evaluation can be divided into several steps. Synthetic speech must be evaluated by the listeners to collect the set of artefacts. The set of factors which are apriori expected must also be determined from the knowledge of synthesis algorithm (although the methodology enables adding further artefacts revealed during the evaluation). Finally, the occurrences of the factors are searched in all phones assigned to the artefacts.

3.1 Artefact Collection

Five listeners did the listening to 50 synthetic phrases generated by our unit selection version (described in [2,3]), with the aim to determine artefacts which they found sounding unnatural – the length of each phrase was about 10 seconds. The set of most expected artefacts was predefined ad-hoc, however, each listener could create “new” artefact tag, when none of predefined made sense regarding his/her impression of the perceived distortion. Contrary to [11], listeners were not required to mark artefacts as

regions, but only to label the subjectively most prominent phone around which an artefact was perceived. It was decided due to the nature of unit selection approach to concatenate continuous (thus natural) sequences of units, which are whole expected to be affected by the artefact.

As the secondary aim of the evaluation is to acquire enough practical experience to tune the artefact collection process so that it is as easy as possible, the listeners included the authors and Ph.D. students focusing their research activity on speech synthesis. The future plan is to involve ordinary people (as potential users of the evaluated TTS) in the procedure.

The source corpus used for the synthesis of evaluated phrases phrases contains 5,000 phonetically balanced utterances, giving approximately 12.5 hours of speech. It was read by a semi-professional female speaker (with radio news broadcasting experience) in a relatively consistent news-like style, and recorded together with glottal signal in an office at our department. The corpus was segmented automatically, using HMM-based approach [12,13]. Although the corpus was not originally designed for the unit selection approach, it was found to be suitable enough for the experiments with this approach, until a more appropriate corpus is recorded [14].

3.2 Factors in Unit Selection

In [11], where the same evaluation method was used for the single-candidate version of our TTS, the definition of factors was much easier, based on the well-known problematic parts – need of signal modification, sensitivity to segmentation inaccuracy and spectral discontinuity between candidates. On the contrary, artefacts considered in the present paper are caused almost purely by the choice of inappropriate unit in the selection algorithm (when no further signal processing is carried out, which is our case), no matter how the "suitability" is measured. In the present paper we, therefore, limit ourselves to the analysis of factors intuitively expected to decrease speech quality, and factors directly joined to target and concatenation costs features:

Segmentation inaccuracy (*SI*). In the case of unit selection with diphone units, the absolute boundary misplacements is not such a problem as it is in the single-candidate system (due to boundary shift [15], triphones are not taken into account yet [16]). Instead, we look for units which are missegmented as a whole – especially sequences as [*la*], [*ij*], [*mn*] and/or similar tend to have one of phones very short. To determine if the factor was presented in artefact, the segmentation of units had to be checked manually – there is no automatic measure available yet (we found out HMM score not to be useful very much). Once there is such, the missegmented boundaries could automatically be corrected or the selection can avoid those "bad" units, and this factor would not have to be considered.

Spectral discontinuity (*SD*). The measure of spectral smoothness is the standard part of concatenation cost. Although there were many experiments with various measures compared to human perception carried out [6,7,8,9,10], Euclidean distance between MFCC vectors is still often utilized (also in our case) thanks to its simplicity. However, as our experience is increasingly showing us that more appropriate measure of perceived smoothness must be established, we manually checked the

discontinuities of spectra around concatenation points for units related to artefact. Manual inspection is also necessary here, just due to the fact that MFCC seems not to be good predictor of perceived discontinuities.

Target features mismatch (*TFM*). This meta-factor covers the whole range of factors related to target cost, as described in [2,3]. Its aim is to reveal the relation of artefact occurrences to the mismatch of features used to measure the suitability of units to express the required prosody (or communication function). Obviously, it expects that the features used really describe measured requirements (and mismatch then causes artefact), which does not necessarily have to be true. Therefore, we work on a new approach of determining target features for unit selection. The given *TFM* factors are collected automatically during the synthesis of evaluated phrases.

Concatenation features mismatch (*CFM*). It is also a meta-factor covering all factors used to the measure of concatenation smoothness, see [2,3]. The difference from *SD* factor is that while *SD* handles the real occurrence of spectral discontinuity in synthetic speech (is/is not), those factors are supposed to provide a relation of artefact occurrences to the behaviour of features which are expected to ensure both spectral and prosodic smoothness (differences around join point). The *CFM* factors are also collected automatically.

Although listed factors are related to the evaluated TTS, we are convinced that the results will tend to display similar results for each TTS system using similar features, features with similar behaviour, or with similar range of values.

3.3 Factors Assignment

As described in Section 3.1, the listeners were asked to mark only the most prominent phone of an artefact perceived. As it can be supposed that artefacts naturally occur around concatenation points or on very short unit sequences of one or two phones (except suprasegmental artefacts like inappropriate communication function, which are not considered by the paper), and as listeners are usually unable to determine the exact phone/phones of artefact either (due to fuzzy nature of perception the place cannot usually be even defined), the sequence of phones affected by each artefact was defined aposteriori as follows. The units preceding the phone holding artefact label were examined until the continuous sequence of at least 3 units was found; similarly for the units following the phone (in the context of [11], such regions can be considered as fragments). In those regions, only the units around *sequence break* were further analysed, as illustrated by the following example:

```
...
mJ  Sentence0457
Ji  Sentence0457 >> region beginning (including the diphone)
ix  Sentence0457
xo  Sentence0457      s1
ov  Sentence4569      s1,s2
vj  Sentence1775         s2
je  Sentence1775
```

```
eC  Sentence1775      s3
Ct  Sentence2666      s3
ti  Sentence2666
iR  Sentence2666  << region end
Ri  Sentence2666
...
```

As described in Section 3.2, some of factors needed manual inspection of phones around sequence breaks – it cannot be done automatically, simply due to the lack of appropriate automatic measure, which, if it exists, could directly be used by selection algorithm to avoid the examined artefacts. Due to the laboriousness of the inspection, only 9 phrases were fully evaluated, which may seem to be a small number for significant results. However, the analysed phrases tend to display very similar tendencies, as shown in Section 4. On the other hand, the manual inspection revealed that some units around sequence breaks were chosen from phrases differing in the voice quality – it defined additional factor *VQ*.

4 Results

There were 90 artefacts collected in the 9 analysed phrases (some determined by more than one listener), and 127 unique sequence breaks were found in the artefact regions. Let us note that there were 316 sequence breaks in total in the analysed phrases, and 9 artefacts did not contain any phrase break at all.

The first part of results covers factors which acquired binary values present/absent in the assignment procedure – *SI*, *SD*, *VQ*, and factors from *TFM* set. The special treatment was, however, required for "position in prosodic word" feature, which has been designed not to provide any sharp delimitation of the feature, as described in [2]. The feature was "sharpened" by splitting the position into 5 equally spaced regions (beginning, beginning-middle, middle, etc.), where match/mismatch could be determined. As the target cost is measured independently for units i and $i+1$ around each sequence break, the results are collected separately for both units; however, as expected, the results are very similar. Let us also note that context mismatch is considered if either left of right context of unit does not fit. Results show that the segmentation inaccuracy cannot still be neglected, as it was found in almost 24% of sequence breaks. Even worse

Table 1. The occurrences of factors with binary present/absent decision values in all sequence breaks (127). Features from *TFM* set are listed separately for units i and $i+1$.

factors	occurrences	percentage
SI	30	23.62
SD	42	33.07
VQ	36	28.35
context $i/i+1$	107/106	84.25/83.46
prosodeme $i/i+1$	7/ 8	5.51/ 6.30
word pos. $i/i+1$	43/ 39	33.86/30.71

situation is for spectral discontinuity, despite subcost aiming to deal with it. A similar situation is for word position mismatch, but we need to further analyse the failure cases to be able to draw a conclusion. The most frequently occurring factor is context mismatch; however, as it is the least important feature (the least weighted in selection) the precise match is yielded in favour of target features defined to be more important.

The second part of the results is focused on *CFM* factors. We compared the behaviour of individual concatenation subcosts through all sequence breaks (note that concatenation cost is 0 aside sequence breaks), shown in Table 2. Let us note that due to the fact that the concatenation features acquire continuous values, they cannot be evaluated as the previous factors. To do so, we would have to define a threshold which would split the values to correct/failure sets. However, the threshold can, in fact, be chosen randomly, each choice giving different results.

Table 2. The mean, standard deviation and median values of *CFM* factors for all sequence breaks. The F_0 values were measured only at the sequence breaks of voiced units i and $i + 1$.

features	mean value	std. dev.	median
F_0 difference [Hz]	6.10	5.23	4.71
F_0 cost	0.10	0.08	0.07
intensity cost	0.06	0.07	0.04
spectral (MFCC) cost	0.69	0.22	0.66

Table 3. The behaviour of MFCC cost in relation to the occurrences of *SD* factor and the voice of unit transitions

voice i, $i + 1$	SD occurrence	number	mean value	std. dev.	median
voiced	y	38	0.70	0.20	0.67
voiced	n	46	0.58	0.20	0.54
unvoiced	y	4	0.77	0.09	0.77
unvoiced	n	39	0.82	0.22	0.84

In addition, we compared the behaviour of spectral measure using Euclidean distance with MFCC in relation to *SD* factors, which is shown in Table 3. It can be seen that the MFCC-related subcost has the largest contribution to the concatenation cost, and it therefore significantly influences the choice of units to concatenate. However, it does not seems to be able to measure spectral discontinuity very well – the average subcosts are relatively close each other (considering also standard deviation), whether *SD* occurred or not. Moreover, there are more than 26% of distances in voiced sequence breaks without *SD* higher than the average of distances in voiced breaks without *SD*. For the unvoiced sequence breaks the numbers are even closer (which, however, was expected due to the noisy character of unvoiced sounds).

The last part of the results covers the brief comparison of the behaviour of both target and concatenation costs. While target cost acquired the mean value 0.11 (with std.dev. 0.09) for units i and 0.11 (std.dev. 0.08) for units $i + 1$ around all sequence breaks, the

concatenation cost acquired the mean value 0.82 (with std.dev. 0.24), which is almost 8-times higher. In unit selection modules using features similar to ours, it is, therefore, the concatenation cost (and MFCC cost as its sub-part) which decides which units to concatenate.

5 Conclusion

Although the speech generated by our TTS has been evaluated as "close to natural" [3], the present paper showed that there are still a number of issues to focus on. First of all is the use of different measure for concatenation smoothness which would follow human perception much more closely (to avoid *SD* factors). There is a also need to reduce the mismatch of target features, e.g. the scaling of context features which is required to express the perceived defect of substitution instead of the difference of phone labels. Moreover, we work on a new method of target features design based on the analysis-by-synthesis approach; we expect that it will be able to provide us with the set of features observed in both natural speech and its synthetic variant. The new costs will also need to be better balanced, and finally, the work on segmentation precision will continue. Further analysis of features aside artefacts and/or sequence breaks will yet also be carried out, to get behaviour of the features in speech regions with and without perceived distortions.

There is also the need to tune the evaluation procedure, as we plan to carry out such tests involving ordinary people. Moreover, the assignment procedure needs be simplified not to be so laborious. The evaluation will also be very important for the new unit selection corpus which we are preparing [14].

Acknowledgements

This research was supported by the GAČR 102/06/P205, and by the EU 6th Framework Program no. IST-034434. Our thanks are also due to Zdeněk Hanzlíček and Milan Legát, who participated on the artefact assessments.

References

1. Matoušek, J., Romportl, J., Tihelka, D., Tychtl, Z.: Recent Improvements on ARTIC: Czech Text-to-Speech System. In: Proceedings of Interspeech 2004 - ICSLP, Jeju Island, Korea, vol. III, pp. 1933–1936 (2004)
2. Tihelka, D., Matoušek, J.: Unit selection and its relation to symbolic prosody: a new approach. In: Proceedings of Interspeech 2006 – ICSLP. Pittsburgh, USA, pp. 2042–2045 (2006)
3. Tihelka, D.: Symbolic Prosody Driven Unit Selection for Highly Natural Synthetic Speech. In: Proceedings of Interspeech 2005 – Eurospeech. Lisbon, Portugal, pp. 2525–2528 (2005)
4. Bellegarda, J.R.: A Novel Discontinuity Metric for Unit Selection Text-to-Speech Synthesis. In: Proceeding of 5th ISCA Speech Synthesis Workshop. Pittsburgh, pp. 133–138 (2004)
5. Syrdal, A.K., Conkie, A.D.: Data-Driven Perceptually Based Join Costs. In: Proceeding of 5th ISCA Speech Synthesis Workshop. Pittsburgh, pp. 49–54 (2004)

6. Vepa, J., King, S.: Join Cost for Unit Selection Speech Synthesis. In: Text to Speech Synthesis: new Paradigms and Advances, pp. 35–62. Prentice Hall PTR, New Jersey (2004)
7. Vepa, J., King, S.: Kalman Filter-Based Join Cost For Unit-Selection Speech Synthesis. In: Proceedings of the 8th European Conference on Speech Communication and Technology Interspeech 2003 – Eurospeech. Geneva, Switzerland, pp. 293–296 (2003)
8. Kavai, H., Tsuzaki, M.: Acoustic Measures vs. Phonetic Features as Predictors of Audible Discontinuity in Concatenative Speech Synthesis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP '04, Quebec, Canada, vol. 1, pp. 657–660 (2004)
9. Donovan, R.E.: A new Distance Measure for Costing Spectral Discontinuitie In Concatenative Speech Synthesis. In: Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Scotland (2001)
10. Klabbers, E., Veldhuis, R.: On the Reduction of Concatenation Artefacts in Diphone Synthesis. In: Proceedings of International Conference on Spoken Language Processing ICSLP 98, Sydney, Australia vol. 6, pp. 2759–2762 (1998)
11. Tihelka, D., Matoušek, J.: Revealing the most Significant Deterioration Factors in Single Candidate Synthetic Speech. In: Proceedings of SPECOM 2005. Greece, pp. 171–174 (2005)
12. Matoušek, J., Tihelka, D., Psutka, J.: Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction. In: Proceedings of the 8th European Conference on Speech Communication and Technology Interspeech 2003 – Eurospeech. Geneva, pp. 301–304 (2003)
13. Matoušek, J., Tihelka, D., Psutka, J.: Experiments with Automatic Segmentation for Czech Speech Synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 287–294. Springer, Heidelberg (2003)
14. Matoušek, J., Romportl, J.: Recording and Annotation of Speech Corpus for Czech Text-to-Speech Synthesis. LNCS (LNAI). Springer, Heidelberg (2007)
15. Clark, R.A.J., Richmond, K., King, S.: Festival 2 – Build Your Own General Purpose Unit Selection Speech Synthesizer. In: Proceedings of ISCA Speech Synthesis Workshop. Pittsburgh, pp. 173–178 (2004)
16. Tihelka, D., Matoušek, J.: Diphones vs. Triphones in Czech Unit Selection TTS. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 531–538. Springer, Heidelberg (2006)