Uncertainty of Phone Voicing and Its Impact on Speech Synthesis

Daniel Tihelka[1(✉)], Zdeněk Hanzlíček[1], and Markéta Jůzová[2]

[1] New Technologies for the Information Society, University of West Bohemia, Pilsen, Czech Republic
{dtihelka,zhanzlic}@ntis.zcu.cz
[2] Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
juzova@kky.zcu.cz

Abstract. While unit selection speech synthesis is not at the centre of research nowadays, it shows its strengths in deployments where fast fixes and tuning possibilities are required. The key part of this method is target and concatenation costs, usually consisting of features manually designed. When there is a flaw in a feature design, the selection may behave in an unexpected way, not necessarily causing a bad quality speech output. One of such features in our systems was the requirement on the match between expected and real units voicing. Due to the flexibility of the method, we were able to narrow the behaviour of the selection algorithm without worsening the quality of synthesised speech.

Keywords: Speech synthesis · Unit selection · Target cost · Phone voicing

1 Introduction

In the past few years, the use of deep neural networks for speech synthesis become widely attractive [2,6,8,13,23,25]. Although the DNN can achieve very natural-sounding speech output, it still requires rather powerful hardware to run on. Also, since it models the speech in such a way that the model is somehow "spread" through the network weights, it is virtually impossible to make an ad-hoc "fix" when something goes wrong.

On the other hand, the unit selection approach suffers from occasional unnatural artefacts [10], causing speech perception annoyances on the otherwise very natural-sounding speech. Since it uses "raw" speech data, it closely mimics the voice style of the original speaker and thus it is not flexible in changing speaking style and/or other characteristics. On the other hand, when an artefact is perceived in the synthesised speech, the identification of its cause is rather straightforward [21] and it can be fixed much more easily. This is one of the factors why the deployment of unit selection is still considered in commercial applications, where fast fixes are desirable.

A. Karpov and R. Potapova (Eds.): SPECOM 2020, LNAI 12335, pp. 581–591, 2020.
https://doi.org/10.1007/978-3-030-60276-5_56

In the present paper we are going to show a case study of such an artefact fix in order to illustrate the flexibility of this synthesis method. And although we illustrate the problem on our particular feature handling, a linguistic/phonological attributes are almost always used in many other systems, despite the fact that the actual features are rarely revealed in the research papers (possibly due to language dependency).

2 Costs in Unit Selection

The key part of the unit selection algorithm is the computation of target and join costs [1,4,5,22]. It is also expected that when synthesising a phrase being recorded in the corpus, the sequence of units from this phrase will be selected. It is simply due to the fact that the concatenation cost $CC(c_{i-1}, c_i) = 0$ for unit candidates c_{i-1} and c_i neighbouring in the speech corpus [18,20]. Similarly, the target cost $TC(t_i, c_i) = 0$ since the features in target specification t_i must be the same as features of the candidate c_i (the target feature generator used when building unit selection database is the same as that used when synthesising input not seen before).

In some of deployments of our English version of TTS system ARTIC [19] we had reports that despite the output sounding natural, it differs from the original when synthesising a phrase from speech corpus. Closer analysis revealed that there is one feature in target cost preventing the $TC(t_i, c_i) = 0$ requirement.

2.1 Voicing Mismatch Feature

The target cost features used in our TTS system ARTIC describe prosody on deep-level [14,15], so called IFF – independent feature formulation [16]. There is one special feature, called *voiced penalty*, introduced originally to prevent the selection of units with incorrect boundaries placement in the process of automatic speech segmentation [3,11]. It checks the expected and real voicing based on phonetic properties of a unit and F_0 computed from the unit signal:

$$TC_v(t_i, c_i) = \begin{cases} 0 & \Leftrightarrow V(t_i) == V(c_i) \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where

$$V(t_i) = \begin{cases} 1 & \text{for voiced phones} \\ 0 & \text{for unvoiced phones} \end{cases} \tag{2}$$

and

$$V(c_i) = \begin{cases} 1 \Leftrightarrow \mathrm{F}_0(c_i) > 0 \\ 0 \Leftrightarrow \mathrm{F}_0(c_i) \leq 0 \end{cases} \tag{3}$$

It penalises the selection of a candidate if it should be voiced but there is no F_0 detected at it, or the other way round, if it should be unvoiced and there is F_0 detected. Let us note that this example is for phones, but our system works with diphones, where the features are supposed to be stable enough [9]. Thus, there are two independent checks of TC_v, for the left phone and for the right phone respectively. The region in which the F_0 is analysed is 5 pitch periods long, centred around the diphone boundary [20].

Let us emphasise that this feature is not a hard-stopper for the selection – in such a case the affected unit candidates could be removed from the inventory a-priori. Instead, it rather penalises the selection of such candidates, but they can still be used for synthesis if there is no better candidate available (as regards the other target features and concatenation cost).

2.2 Voicing Mismatch Origins

Although the phones are strictly categorised into voiced and unvoiced in theory [9], there are a surprisingly large number of voicing mismatches in phone centres (diphone boundary) where the signal should be stable enough. In two of our corpora (Jan and Kateřina, see [19]) we examined, the 1.37% of 633,387 and 2.45% of 557,556 phones contain voicing mismatch as defined by Eq. 1.

Looking at the individual phones in the corpora, in Fig. 1 we show the relative number of candidates with voicing mismatch in the middle of phones. It can be seen that the majority of mismatches are for phones [P\] (in SAMPA notation [24]) for both voices, and for [Z] and [d_z] depending on the speaker. As noted before, all the statistics are related to the centres of the phones.

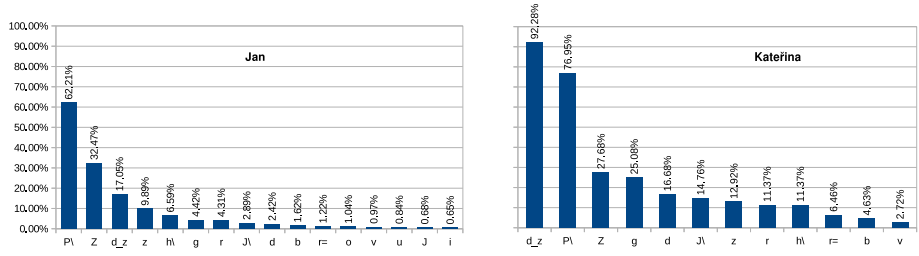

Fig. 1. Voicing mismatch occurrences (in % relative to unit count) for both examined voices as occurring in the speech corpus recordings. Only phones with mismatch >0.5% are presented.

The deeper analysis of the individual cases revealed various, but not the only, categories we have encountered:

GCI Detection Failure. Since the F_0 value is computed from glottal closure instants (pitch-marks), either from a glottal [7] or speech signals [12], the ability

to reliably determine voicing parts is crucial in these algorithms. In Fig. 2 there is rather nice signal for clearly voiced [h\] phone, but an inability to detect GCI (pitch-marks) by [12] caused the middle of the phone to be considered as unvoiced.

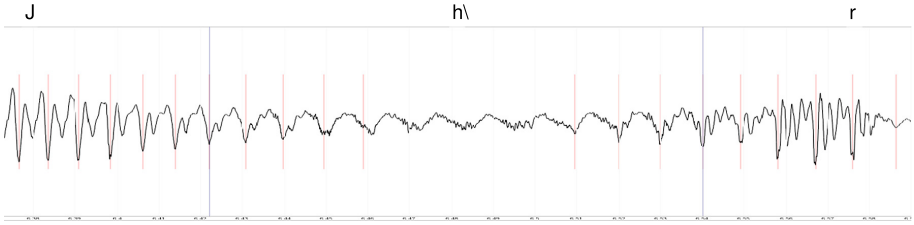

Fig. 2. Signal of [h\] from recording [. . . alespoJ h\ rubi: . . .] with clearly visible voicing structure but without GCI detected. The black vertical lines are automatically detected phone boundaries, the red vertical lines are GCI instants assigned. (Color figure online)

Devoicing. Especially in the case of paired consonants, but not only there, a devoicing may occur under some conditions [9], causing a temporal stop in GCI detection and thus no F_0 assignment, as illustrated in Fig. 3. The devoicing and GCI detection failures were the reasons of voicing mismatch in the majority of cases for units with the highest mismatch score.

Inappropriate Segmentation. When a voiced unit neighbours with unvoiced, the process of automatic segmentation [3,11] may place the boundary of a voiced unit too far into the unvoiced region. The diphone boundary may then fall in the unvoiced part of the signal, causing mismatch in the voicing comparison (Fig. 4).

Inappropriate Alignment. When automatic segmentation is carried out, several pronunciation variants of each word are examined [3,11] to increase the robustness. It seems that although the segmentation model used matches the signal more precisely, an inappropriate variant is sometimes chosen, as illustrated in Fig. 5.

Naturaly, there is often the combination of such factors. For example, the significant amount of mismatches for [d_z] and [R] is caused by the GCI detection failure.

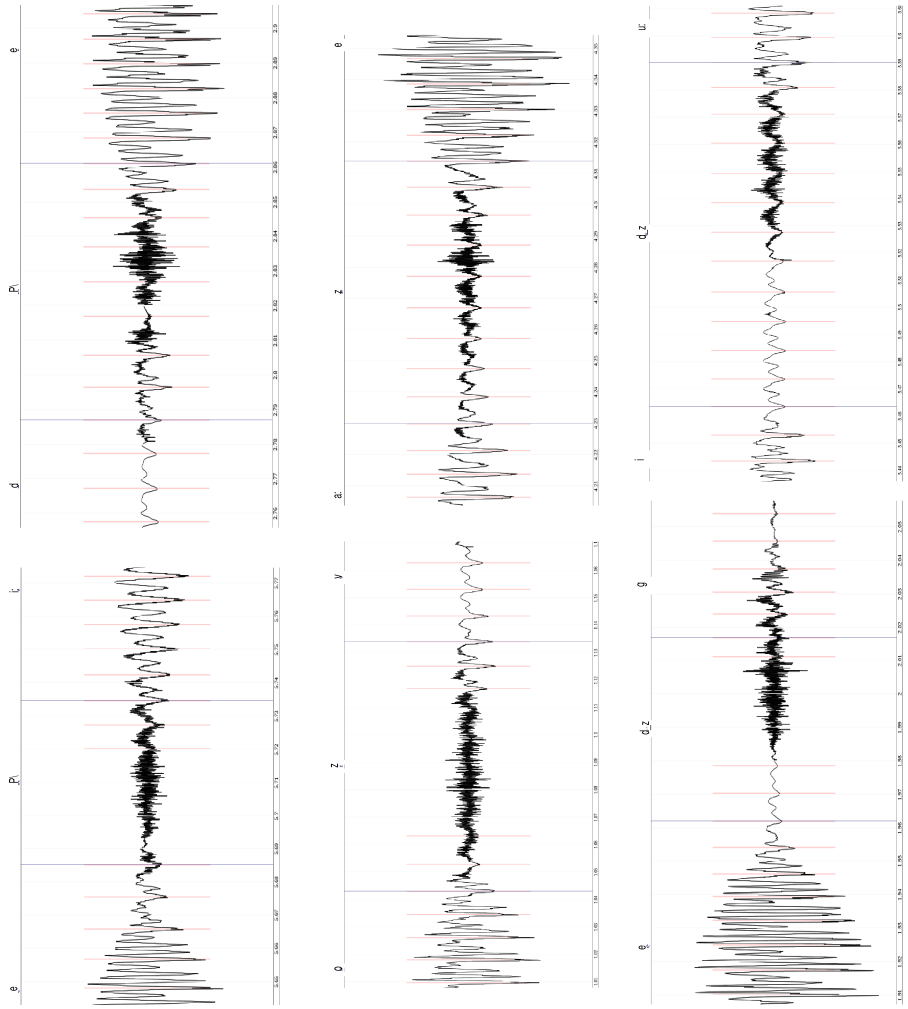

Fig. 3. Signal of phones [P\,z,d_z] with reduced voicing and thus without GCI detected at the left side. The right side shows the same phone with clear voicing structure, and thus with GCIs detected correctly.

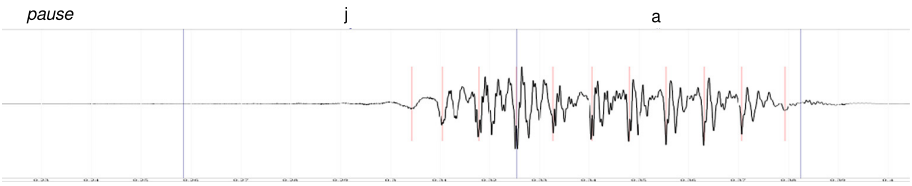

Fig. 4. Signal of [j] from recording with the left boundary placed incorrectly too far into the preceding pause. The black vertical lines are automatically detected phone boundaries, the red vertical lines are GCI instants assigned. (Color figure online)

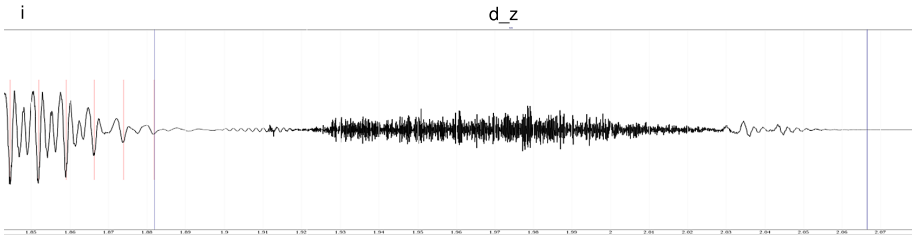

Fig. 5. Signal of [t_s] from recording [. . . ut_SebJit_s *pause*] where inappropriate pronunciation variant [ut_SebJid_z] was used. It is clearly visible that there is no voicing part in the signal and [t_s] is also audible; the model of [d_z] matched the signal more precisely, though.

3 Replacing the Cost

As mentioned above, we had reports that when synthesising an exact phrase occurring in the speech corpus, the result does not sound the same as the original phrase. Even when the synthetic variant did not show unnatural artefacts, it was not desirable by the user of our TTS.

The most straightforward way of dealing with this behaviour was to remove the sub-cost from the target cost computation. However, to examine the real effect of the voicing cost, with the aim of removing it, we carried out the following experiments:

- to remove the cost computation completely. Thus, we ensure that when synthesising a speech corpus phrase, the unit from that phrase will be selected since the other features depend on text only at the cost of selecting units with expected/detected voicing mismatch to the synthetic output;
- to substitute the cost by a higher F_0 penalty in the concatenation cost, preventing the selection of units with voicing mismatch at the boundary of concatenated candidates:

$$CC'_{F_0}(c_{i-1}, c_i) = \begin{cases} CC_{F_0}(c_{i-1}, c_i) & \Leftrightarrow V(c_{i-1}) == V(c_i) \\ 1000 & \text{otherwise} \end{cases} \tag{4}$$

  where $CC_{F_0}(c_{i-1}, c_i)$ is the original F_0 cost computation as described in [20]. Let us note that this is not a substitute of the original TC_v cost since that penalised selection of mismatching candidates while the current prevents the selection of candidates with mutual voicing mismatch. To mimic the original behaviour while ensuring cost = 0 for in-corpus phrase is laborious to set, since the target and concatenation costs behave and are weighted slightly differently.

Then, we have synthesised nearly 150,000 sentences and logged each usage of unit where the voicing mismatch occurred. In the following text, the *baseline* denotes the original implementation of target cost computation, taking the voicing mismatch into account and trying to avoid it, although it still does not have to

be avoided when there is no better candidate available (i.e. candidate with voicing mismatch $TC_v(t_i, c_i) > 0$ will be preferred to candidate with $TC_v(t_i, c_i) = 0$ if the first has better match of the other target features than the latter). The *no-VC* will denote the version when the voicing match-mismatch in not examined at all (though, it is still logged), and F_0-*VC* will denote the selection with modified concatenation cost, as defined by Eq. 4. Let us emphasise, that both modifications ensure the selection of the original phrase in the required case.

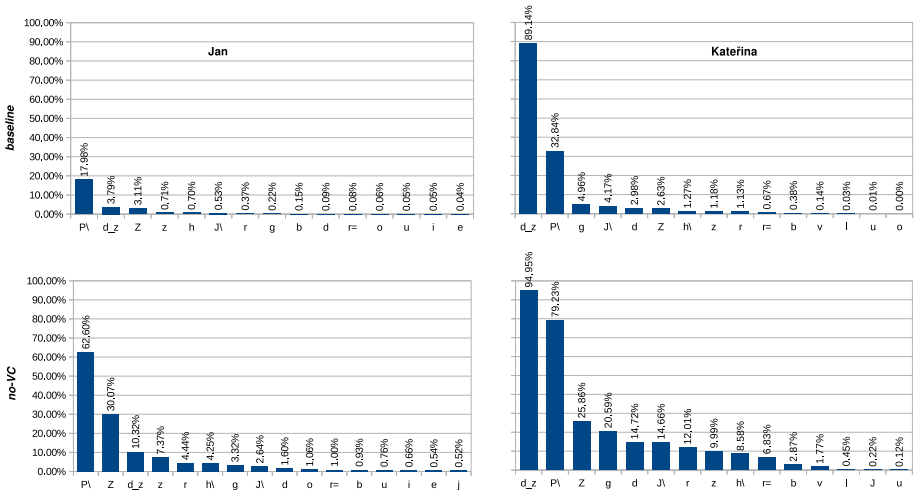

Fig. 6. Voicing mismatch occurrences (in % relative to unit count) for both examined voices as occurred in the synthesised output of the individual baseline and *no-VC* system (F_0-*VC* system is omitted since it looks very close to *no-VC*). Only the first 15 phones are presented.

It can be seen in Fig. 6 that the voicing mismatch is reduced in the *baseline* system, as expected. On the contrary, the mismatches in *no-VC* and F_0-*VC* are roughly the same. It means that the original $CC_{\mathrm{F_0}}(c_{i-1}, c_i)$ was capable enough in preventing concatenation boundary voicing mismatches.

4 Evaluation of Quality Impact

Both the proposed cost computation modifications ensure the selection of the whole phrases from the corpus, when they appear at the TTS input. However, it needs to be answered if, and how much, the modifications affect the overall quality of the generated speech, while the expectation is that omitting voicing mismatch evaluation will not perform worse than the baseline system.

To do so, we examined the logs collected during the synthesis of the 150,000 phrases, following the methodology described in [17]. The difference criteria $\delta(a, b)$ were defined as:

1. the number of different candidates in the selected sequence;
2. the number of expected/detected voicing mismatches (as defined by Eq. 1)

and both criteria were originally expected to be evaluated for combinations

1. baseline × no voicing cost (*no-VC*)
2. baseline × voicing cost moved to F_0 concatenation sub-cost (F_0-*VC*)
3. *no-VC* × F_0-*VC*

After the analysis of the results, however, the unit sequences selected by *no-VC* and F_0-*VC* were found very similar, and thus their independent comparison to baseline was omitted.

For each of the criteria and voice, 10 unique phrases with the highest criteria value were selected for further evaluation by means of informal listening tests. The test itself was the simple `ABX` preference format, with `A` and `B` stimuli shuffled at random through the whole test (but not through the listeners). 6 listeners participated in the listening test, all of them being experts in speech technologies. While 6 may seem to be little, all have experience in phonetics and due to the specific test configuration there is also no reason to expect significantly different results with larger number of listeners.

In Table 1, the overall results are collected. For both voices, the `X` variant (i.e. no preference) was chosen the most frequently. From the evaluation point of view, the most interesting are the cases where the baseline system was evaluated as better. Further analysis showed that there is another cause of the quality deterioration, not related to voicing mismatch (e.g. slightly more fluctuations in F_0).

Table 1. The results of `ABX` listening test. The numbers represent the count of preferences given to the corresponding system, the *total* is the sum through evaluation of sentences with the highest differences in candidate sequence and voicing mismatches.

	Candidates diff. no.			Voice mismatch no.			Candidates diff. no.		
	baseline	none	*no-VC*	baseline	none	*no-VC*	*no-VC*	none	F_0-*VC*
Jan	13	30	17	11	25	24	14	41	5
Kateřina	14	36	10	14	33	13	9	32	19
Total	all collected								
	52	124	64						

To test the statistical significance of the result, we have carried out a sign test with the null and alternative hypothesis:

H0: *The outputs of the both systems are perceived as equally good*
H1: *The output of one system sounds better*

The null hypothesis testing the same quality was chosen intentionally, as we need to check whether or not omitting the mismatch check will have a negative impact on the quality.

The sign test proved (at significance level $\alpha = 0.05$) that the version not considering the voicing mismatch is better for voice Jan (p-value = 0.0042) and that both version are of the same quality for voice Kateřina (p-value = 0.1053). Taking all the results together, the test proved that not considering voicing mismatch does not decrease the quality of the output, i.e. both systems are perceived as equally good (p-value = 0.3502).

5 Conclusion

We have identified the reason of the suspicious behaviour reports and narrowed it by removing the counterproductive voicing mismatch evaluation from unit selection cost computations. Using the listening tests designed to check a "worst-case" scenario, and knowing that the new system behaviour does not affect the quality of synthetic speech in any negative way, we can use *no-VC* in out TTS system now.

Despite the fact that the DNN-based speech synthesis is naturally moving to the centre of speech research, the relative ability to identify problems and tune and fix the behaviour of TTS system in relatively straightforward way remains one of the strengths of the unit selection approach.

Let us also emphasize that the observations we point out are cross-language (we have found similar issues in English and Russian voices), so the results can not only be extended to other speech synthesizers but also to other fields where phone voicing needs to be considered; at least as a caution that there may be such an uncerntainty.

Acknowledgments. This research was supported by the Technology Agency of the Czech Republic (project No. TH02010307), and by the grant of the University of West Bohemia, (project No. SGS-2019-027).

References

1. Železný, M., Krňoul, Z., Císař, P., Matoušek, J.: Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. Sig. Process. **12**, 3657–3673 (2006)
2. Hanzlíček, Z., Vít, J., Tihelka, D.: WaveNet-based speech synthesis applied to Czech: a comparison with the traditional synthesis methods. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2018. LNCS (LNAI), vol. 11107, pp. 445–452. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00794-2_48
3. Hanzlíček, Z., Vít, J., Tihelka, D.: LSTM-based speech segmentation for TTS synthesis. In: Ekštein, K. (ed.) TSD 2019. LNCS (LNAI), vol. 11697, pp. 361–372. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27947-9_31

4. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: ICASSP 1996, Proceedings of International Conference on Acoustics, Speech, and Signal Processing, IEEE, Atlanta, Georgia, vol. 1, pp. 373–376 (1996)
5. Kala, J., Matoušek, J.: Very fast unit selection using Viterbi search with zero-concatenation-cost chains. In: ICASSP 2014, Proceedings of International Conference on Acoustics, Speech, and Signal Processing, IEEE, Florence, Italy, pp. 2569–2573 (2014)
6. Kalchbrenner, N., et al.: Efficient neural audio synthesis. arXiv preprint arXiv:1802.08435 (2018)
7. Legát, M., Matoušek, J., Tihelka, D.: A robust multi-phase pitch-mark detection algorithm. In: Interspeech, vol. 2007, pp. 1641–1644 (2007)
8. Lorenzo-Trueba, J., et al.: Towards achieving robust universal neural vocoding, pp. 181–185 (2019)
9. Machač, P., Skarnitzl, R.: Principles of Phonetic Segmentation. Epocha, Prague (2013)
10. Matoušek, J., Legát, M.: Is unit selection aware of audible artifacts? In: SSW 2013, Proceedings of the 8th Speech Synthesis Workshop, ISCA, Barcelona, Spain, pp. 267–271 (2013)
11. Matoušek, J., Romportl, J.: Automatic pitch-synchronous phonetic segmentation. In: INTERSPEECH 2008, Proceedings of 9th Annual Conference of International Speech Communication Association, ISCA, Brisbane, Australia, pp. 1626–1629 (2008)
12. Matoušek, J., Tihelka, D.: Using extreme gradient boosting to detect glottal closure instants in speech signal. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, Great Britain, pp. 6515–6519 (2019)
13. van den Oord, A., et al.: WaveNet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
14. Romportl, J.: Structural data-driven prosody model for TTS synthesis. In: Proceedings of the Speech Prosody 2006 Conference, pp. 549–552. TUDpress, Dresden (2006)
15. Romportl, J., Matoušek, J.: Formal prosodic structures and their application in NLP. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 371–378. Springer, Heidelberg (2005). https://doi.org/10.1007/11551874_48
16. Taylor, P.: Text-to-Speech Synthesis, 1st edn. Cambridge University Press, New York (2009)
17. Tihelka, D., Grůber, M., Hanzlíček, Z.: Robust methodology for TTS enhancement evaluation. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 442–449. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_56
18. Tihelka, D., Hanzlíček, Z., Jůzová, M., Matoušek, J.: First steps towards hybrid speech synthesis in Czech TTS system ARTIC. In: Karpov, A., Jokisch, O., Potapova, R. (eds.) SPECOM 2018. LNCS (LNAI), vol. 11096, pp. 676–686. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99579-3_69
19. Tihelka, D., Hanzlíček, Z., Jůzová, M., Vít, J., Matoušek, J., Grůber, M.: Current state of text-to-speech system ARTIC: a decade of research on the field of speech technologies. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2018. LNCS (LNAI), vol. 11107, pp. 369–378. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00794-2_40

20. Tihelka, D., Matoušek, J., Hanzlíček, Z.: Modelling F0 dynamics in unit selection based speech synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS (LNAI), vol. 8655, pp. 457–464. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10816-2_55
21. Tihelka, D., Matoušek, J., Kala, J.: Quality deterioration factors in unit selection speech synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 508–515. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74628-7_66
22. Tihelka, D., Romportl, J.: Exploring automatic similarity measures for unit selection tuning. In: INTERSPEECH 2009, Proceedings of 10th Annual Conference of International Speech Communication Association, ISCA, Brighton, Great Britain, pp. 736–739 (2009)
23. Vít, J., Hanzlíček, Z., Matoušek, J.: Czech speech synthesis with generative neural vocoder. In: Ekštein, K. (ed.) TSD 2019. LNCS (LNAI), vol. 11697, pp. 307–315. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27947-9_26
24. Wells, J.C.: SAMPA computer readable phonetic alphabet. In: Gibbon, D., Moore, R., Winski, R. (eds.) Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin and New York (1997)
25. Wu, Z., Watts, O., King, S.: Merlin: an open source neural network speech synthesis system. In: 9th ISCA Speech Synthesis Workshop (2016), pp. 218–223, September 2016