First Steps Towards Hybrid Speech Synthesis in Czech TTS System ARTIC

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

[1] New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
{dtihelka,zhanzlic}@ntis.zcu.cz

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

Abstract. The hybrid speech synthesis, combining an HMM-based parameter trajectories generator and unit selection, was reported to achieve high speech output quality, in some cases even outperforming the "classic" unit selection method, while having reasonable cost of hardware requirements increase, especially when compared to modern DNN-based (e.g. WaveNet) speech synthesis methods.

The present paper introduces one of this hybrid approaches, facing up the mismatch between rather smooth flow of parameters when generated by a model and between their varying evolution when obtained from speech. We also describe several modifications of target cost computation, influencing the selection of units being close to the required parameters, while our aim is to obtain a notion of the mutual interactions within the modified selection process. The overall conclusion is covered by listening tests, showing comparable quality of the trial hybrid synthesis described to unit selection method tuned through the years.

Keywords: Statistical-parametric synthesis · HMM speech synthesis
Unit selection · Hybrid speech synthesis · Target cost

1 Introduction

In the past few years, there have been multiple concurrent approaches to speech synthesis at the centre of interest, ranging from traditional unit selection [3], through statistical-parametric synthesis (SPS, [32]), to the use of deep neural networks [17]. Each of the approaches, however, has its advantages and drawbacks – unit selection suffers from artefacts, "raw" SPS synthesis from parameters flattening and vocoder imperfection, and the DNN requires powerful hardware to run on. Therefore, there is research interest in *hybrid speech synthesis*, trying to combine the advantages of HMM or DNN acoustic parameters generation, driving then the unit selection module using either natural signals to generate

© Springer Nature Switzerland AG 2018
A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 676–686, 2018.
https://doi.org/10.1007/978-3-319-99579-3_69

the speech [16,19,31], or combining real signals with signals built by SPS in cases where no suitable speech segments are found for the generated parameter trajectories [18,20,29].

The present paper describes our attempt to employ the hybrid speech synthesis approach within the Czech TTS system ARTIC [25]. As the system contains both unit selection and SPS modules, it was a natural choice to join them together, as it has even been reported that a hybrid approach can achieve higher naturalness of speech it generates [1,16,18,20,31], except [31] on smaller speech corpora than we use, though. Since we have quite large speech corpora recorded by a professional (or semi-professional) speakers [25], we are not going to mix real and artificial speech signals in this paper due to our concerns about their different quality. Instead, we will drive the unit selection exclusively by the SPS-generated trajectories, but only real speech units will be concatenated.

2 Hybrid Speech Synthesis

The first experiments with hybrid synthesis started more than a decade ago with frame-based units [13], and were extended in various ways e.g. in [1,7,14,18,20,29].

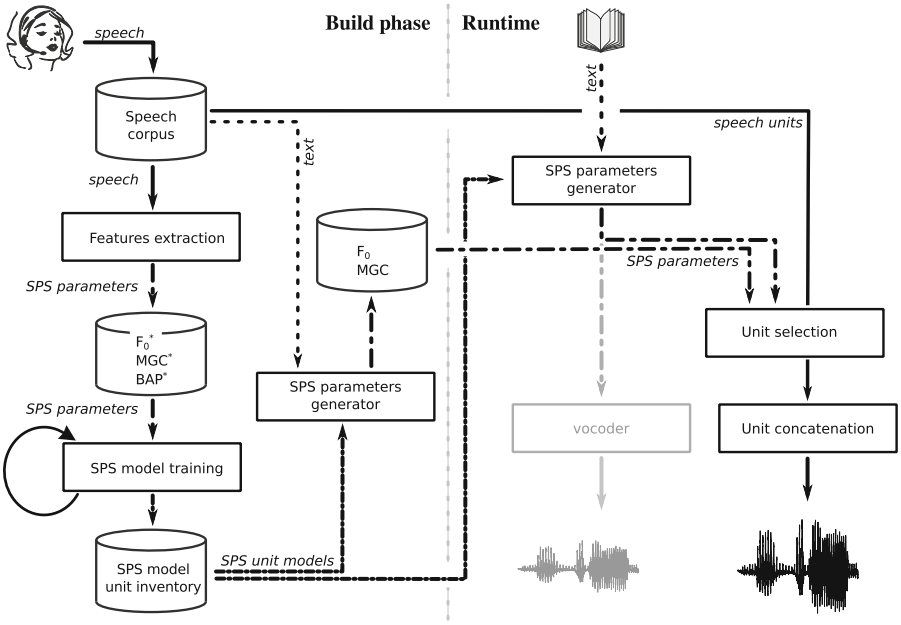

Fig. 1. Scheme of hybrid speech synthesis system. The gray parts of *runtime* are not used in this paper.

The most common approach is to replace the *target cost* [22] component of unit selection with a measure of similarity (or closeness) between *target features*[1] generated for the input text by an HMM model (or a DNN in last time) and the same feature extracted from the unit selection speech corpus from where the units are taken. We have also occasionally tried to employ hybrid speech synthesis in recent years, based on some of the papers cited, but until recently our experience was that the more unit selection path was approached, the better the output quality was.

The key issue was that we compared the generated target parameters to the parameters extracted from natural unit candidates, since there is a significant mismatch between the target (relatively smooth) and the candidate (varying), as was illustrated for F_0 in [24] and a parameter coefficient in [29]. However, one of the fundamental operations of statistical modeling is the averaging of model parameters during the training and the generation of novel values during synthesis – we can liken this to the interpolation and extrapolation of the values found in the training data. Therefore, in the present paper, we *re-generate* the whole speech corpus with the same model as used to predict the target parameters, and each unit candidate in the acoustic units inventory is tied to the parameters generated by the model for frames[2] belonging to the candidate – see Fig. 1. By this re-assignment, the behaviour of the parameters used to drive the selection is unified (the parameters for candidates behave as smoothly as the parameters for the target), and thus the "closeness" of the target parameters, as represented by a distance measure (see Sect. 2.2), does make much more sense. Also, it still does not violate one of the unit selection assumptions – to have the target cost = 0, selecting the natural unit sequence, when the whole phrase from the corpus is required to be synthesized.

2.1 The Generation of Target Parameters

In the SPS method, the speech signal is represented as a sequence of parameters extracted at fixed 5 ms frame rate. In out case, these are 40 mel-generalized cepstral coefficients (MGC) extracted from STRAIGHT spectra [11], logarithm of fundamental frequency extracted from the glottal signal [12] by the PRAAT [2], and 21 band-aperiodicity coefficients (BAP) derived from STRAIGHT aperiodicity spectra. Thus, each phrase from the source speech corpus is represented by the p-dimensional vector $F_0^\star = [f(1), f(2), \ldots, f(p)]$, the $p \times 40$ matrix $\mathrm{MGC}^\star = \left[\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(p)\right]$ and the $p \times 21$ matrix $\mathrm{BAP}^\star$ (naturally, p here depends on the length of a phrase). The corpus represents the common speech base also used as the source of speech units to be concatenated, either by baseline

[1] The features are called *target* from the point of view of unit selection, since we want to select units having the feature values as close to these as possible. From the SPS point of view, these are destination, since the output speech can be generated from them.

[2] Let us note here that even the generated length of a candidate (a number of frames assigned to it) may differ from the real length of that candidate.

unit selection or the hybrid version described here. Then, the $\star$ parameters are used to train 5-stated hidden semi-Markov models, which involves a few repetitions of 3 main stages – initialization and training phone models (disregarding the context), training of full-context models and model clustering [4,5].

Once the models are trained, for the given sequence of units required to be synthesized, the streams of F_0, MGC and BAP parameters are *generated* by using a parameter generation algorithm considering global variance [30], and are passed to the vocoder which uses them to build the output speech.

In the case of hybrid synthesis, though, instead of using a vocoder, the stream is passed to the unit selection module which then selects "close enough" real speech chunks (candidates) to be concatenated, see Sect. 2.2. Note here that while all the streams must be used by the vocoder, only the F_0 and cepstral coefficients were used to drive the unit selection, the aperiodicity was omitted since it is not supposed to bring any significant information to select according to. Although it could be used to generate speech of a unit as a replacement of a raw unit signal (a.k.a. "multiform" synthesis) when no candidate with match "close enough" is found [18,20,21,29], this has been dismissed here.

2.2 Unit Selection

In the most common unit selection scheme, a number of metrics are used to define target and concatenation costs. While the former measures how feature values of a speech unit from acoustics unit inventory match the prescribed (target) values of the features (what is aimed at to be expressed), the latter attempts to evaluate how smooth units will be perceived when joined together (how it will sound).

Target Cost. Having the trajectories of parameters $F_0 = [f(1), f(2), \ldots, f(r)]$ and MGC $= \big[\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(r)\big]$ generated by the SPS module for the given phrase to synthesize, we have used the scripting interface of ARTIC TTS [25] to modify the unit selection in such a way that the use of *symbolic* features [15,26], denoted by [23] as *independent feature formulation* – IFF, was replaced by the measure of mismatch between parameters from the parameter streams generated by the HMM model.

Contrary to SPS, where the parameters are passed through a vocoder as they are, the unit selection works with unit candidates (in [19] called "tiles"). Thus, we define t_j as a target unit within a synthesized phrase $\mathbf{T} = t_1, t_2, \ldots$. Each target unit is tied to $N(t_j)$-dimensional vector $F_0(t_j) = [f(1), f(2), \ldots, f\big(N(t_j)\big)]$ and $N(t_j) \times 40$ matrix MGC$(t_j) = \big[\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}\big(N(t_j)\big)\big]$ corresponding to the generated parameters for that unit in the phrase being currently synthesized; $r = \sum_{\forall j} N(t_j)$.

Similarly, we define a unit candidate c_i as a constituent of a phrase in the speech corpus the unit candidates are selected from. These unit candidates were tied to $F_0(c_i) = [f(1), f(2), \ldots, f\big(N(c_i)\big)]$ and MGC$(c_i) =$

$\left[\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}\left(N(c_i)\right)\right]$ in the same way as target units, except during the speech corpus re-assignment.

The target cost in the hybrid synthesis experiment described here was computed using various schemes (see Eqs. (2), (3), (4) and (5)), as we first want to obtain a notion of the mutual interaction between target and concatenation costs and the relevance of the individual features in the target cost. All target cost definitions, however, used both F_0 and MGC features anyway. Contrary to [19], these parameters were not normalized in this experiment. Instead, the ad-hoc defined weights were assigned to them in order to balance the importance of the individual features, as described in Sect. 3.

Regarding the duration of units, it is expressed by the number of parameter vectors $N(t_j), N(c_i)$. In [19], the authors always choose the candidates with the same duration as the target has. However, our initial experiment with such setting showed worse speech quality. Therefore, we allowed $N(t_j) \neq N(c_i), \forall i, j$, with a small penalty added to the target cost when this non-equality occurred. The number of parameters the cost was computed from was set to $N(t_j, c_i) = \min\left(N(t_j), N(c_i)\right)$, with the indexes aligned to the center of the parameter vectors, i.e. $N(t_j, c_i)/2 = \{N(t_j)/2, N(c_i)/2\}$.

Concatenation Cost. The handling of concatenation cost CC was the same as in the "raw" unit selection [15], i.e. absolute difference of "static" F_0 (as described in [27]) and energy, and the Euclidean distance of 12 MFCC coefficients.

3 Experiments

Let us emphasize that for each sequence of unit candidates $\mathbf{C}_p = c_1, c_2, \ldots, c_{N(p)}$ from the p-th phrase in the speech corpus, the sequence of units $\mathbf{T}_p = t_1, t_2, \ldots, t_{N(p)}$ was generated and $\mathrm{F}_0(c_j) \stackrel{def}{=} \mathrm{F}_0(t_j), \mathrm{MGC}(c_j) \stackrel{def}{=} \mathrm{MGC}(t_j), \forall j = 1, 2, \ldots, N(p)$ was assigned. In this way, it is ensured that the continuous (i.e. natural) sequence of unit candidates is chosen from a phrase from the corpus when that phrase is to be synthesized.

The generic target cost definition is the weighted sum of F_0 and MGC subcosts with the penalty value $\wp > 0$ set in case of $N(t_j) \neq N(c_i)$:

$$TC_x(t_j, c_i) = w_x^{\mathrm{TC}} \cdot \frac{w_x^{\mathrm{F}_0} \cdot TC_x^{\mathrm{F}_0}(t_j, c_i) + w_x^{\mathrm{MGC}} \cdot TC_x^{\mathrm{MGC}}(t_j, c_i) + 5 \cdot \wp(t_j, c_i)}{w_x^{\mathrm{F}_0} + w_x^{\mathrm{MGC}} + 5} \tag{1}$$

where c_i is here the i-th candidate for j-th unit $u_j = \{c_1, c_2, \ldots\}$ in the synthesized sequence $\{t_1, u_1\}, \{t_2, u_2\}, \ldots, \{t_J, u_J\}$. The x here denotes the number of experiment.

The very first experiment was designed to simply replace the "symbolic"-features-driven target cost (the *baseline*) with the target cost following the behaviour of parameters prescribed by the SPS generator.

$$
\begin{aligned}
w_1^{\mathrm{TC}} &= 1.0 \\
w_1^{\mathrm{F}_0} &= 1.0 \\
TC_1^{\mathrm{F}_0}(t_j, c_i) &= \sum_{\{n^t, n^c\}}^{N(t_j,c_i)} \left| \mathrm{F}_0(t_j, n^t) - \mathrm{F}_0(c_i, n^c) \right| \\
w_1^{\mathrm{MGC}} &= 1.0 \\
TC_1^{\mathrm{MGC}}(t_j, c_i) &= \sum_{\{n^t, n^c\}}^{N(t_j,c_i)} \sqrt{\sum_{\substack{\mathbf{f}^t=\mathrm{MGC}(t_j,n^t) \\ \mathbf{f}^c=\mathrm{MGC}(c_i,n^c)} (f^t - c^t)^2}}
\end{aligned}
\tag{2}
$$

Thus, the F_0 sub-cost was the sum of log F_0 differences, the MGC-sub cost was the sum of Euclidean distances of cepstral vectors, both through the $\{n^t, n^c\}$ couples. Similar to [15], we did distinguish between voiced and unvoiced speech units (or their parts), since the SPS module fills $\mathrm{F}_0(t_j, n) = -\inf$ when n corresponds to an unvoiced frame. But contrary to the baseline unit selection, where voicedness is checked only at units beginning and end, we set value 10 for each $\{n^t, n^c\}$ with voice/invoice mismatch between $\mathrm{F}_0(t_j, n^t)$ and $\mathrm{F}_0(c_i, n^c)$.

Through the rough look at the CC and TC values in the first experiment it has been found that the TC value was about 10× higher than the CC value. Therefore, for the second experiment, the weight w_2^{TC} was adjusted to make the gap between costs lower:

$$
w_2^{\mathrm{TC}} = 0.1 \tag{3}
$$

In the following experiment $x = 3$, we tried to put greater emphasis on the F_0 contour (relative to the MGC parameters). The same rough look at data as in the previous experiment suggested that the TC^{F_0} cost values were found about 10× lower than the values of TC^{MGC}. Therefore, the weight of F_0 match was increased:

$$
w_3^{\mathrm{TC}} = 0.1 \qquad w_3^{\mathrm{F}_0} = 50.0 \tag{4}
$$

The overall TC value, however, was kept at the range similar to the CC, as it was in the previous experiment.

It can be expected that the cost value depends on the number of parameters $N(t_j, c_i)$ describing a unit, i.e. longer units may achieve higher cost values. To minimize this effect, the costs were normalized by unit lengths as:

$$
\begin{aligned}
TC_4^{\mathrm{F}_0}(t_j, c_i) &= \frac{TC_1^{\mathrm{F}_0}(t_j, c_i)}{N(t_j, c_i)} \\
TC_4^{\mathrm{MGC}}(t_j, c_i) &= \frac{TC_1^{\mathrm{MGC}}(t_j, c_i)}{N(t_j, c_i)} \\
w_4^{\mathrm{F}_0} &= 50.0
\end{aligned}
\tag{5}
$$

Still, we emphasize the F_0 feature, but we do not explicitly lower the whole target cost value since it has been lowered implicitly by the length normalization (thus, TC values already are in range comparable to CC values).

Let us also note that we have tried to use normalized F_0 and MGC values in the costs $TC_{1,\ldots,4}(t_j, c_i)$ with the weights adjusted appropriately. The values were z-score normalized per-phrase, both when re-synthesizing the speech corpus and when synthesizing a text. However, the quality of speech was noticeably lower than the quality of speech evaluated in this paper. It remains to be answered if per-corpus normalization would improve this situation.

4 Results

First, let us look at the behaviour of the individual parameter values and their mutual relations. In Sect. 3, the design of the individual costs/weights was based on the rough values analysis. To have a more precise view of the relation among the cost values, we have additionally analysed the 75922 values from 1,895 synthesized phrases, with the averages summarized in Table 1 and random selected subset plot in Fig. 2. Let us note that the cost values differ even when computed by the same equation (e.g. $TC_x^{F_0}$ for $x = 1, 2, 3$), since the values are taken from the sequences of *selected* units, which naturally differ through the experiments as the overall costs computation changes.

To evaluate the quality of the hybrid synthesis methods, we have carried out informal listening tests. Due to the larger number of versions, we have decided to use MUSHRA-like tests [8], without anchor and reference prompts, though.

Table 1. The mean of 75922 individual costs collected through 1,895 synthesized phrases. The *ratio* was computed as $w_x^{TC} \cdot \frac{TC_x}{CC}$ and $w_x^{F_0} \cdot \frac{TC_x^{F_0}}{TC_x^{MGC}}$.

Voice	x	CC	TC_x	Ratio	$TC_x^{F_0}$	TC_x^{MGC}	Ratio
Jan	1	0.197	2.160	10.98	1.519	8.379	0.18
	2	0.087	0.255	2.93	2.130	9.400	0.23
	3	0.080	0.110	1.39	0.938	11.430	4.10
	4	0.074	0.061	0.82	0.377	0.377	4.89
Stanislav	1	0.193	2.302	11.90	1.405	9.388	0.15
	2	0.085	0.269	3.18	1.883	10.540	0.18
	3	0.065	0.114	1.77	0.963	12.602	3.82
	4	0.065	0.050	0.77	0.028	0.393	3.60
Iva	1	0.209	2.226	10.63	1.029	9.163	0.11
	2	0.089	0.265	2.97	1.520	10.470	0.15
	3	0.072	0.078	1.09	0.555	12.839	2.16
	4	0.069	0.046	0.68	0.022	0.379	2.93
Radka	1	0.217	2.550	11.74	1.881	10.480	0.18
	2	0.096	0.3	0.10	2.642	11.707	0.23
	3	0.089	0.127	1.42	1.058	14.137	3.74
	4	0.080	0.065	0.81	0.039	0.046	4.80

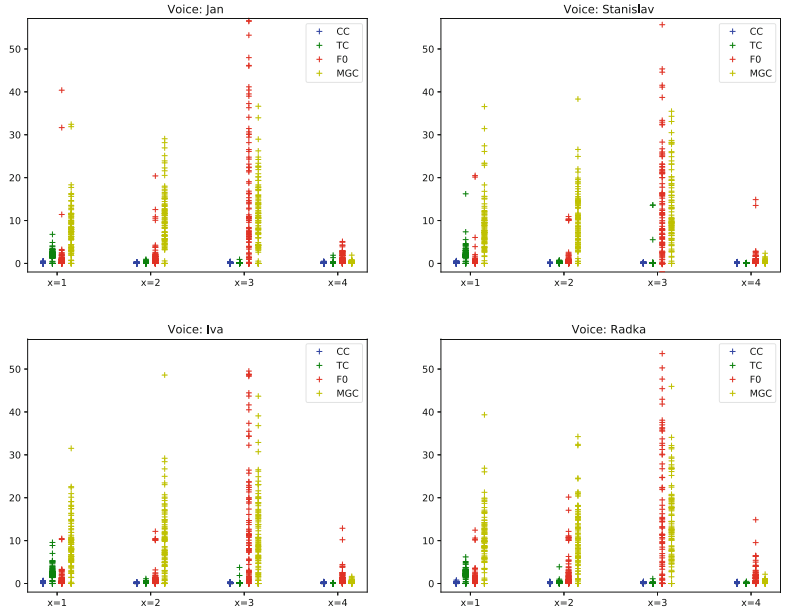

Fig. 2. Plot of 100 random selected values from the total set of 75922 costs. All the values include their wights applied.

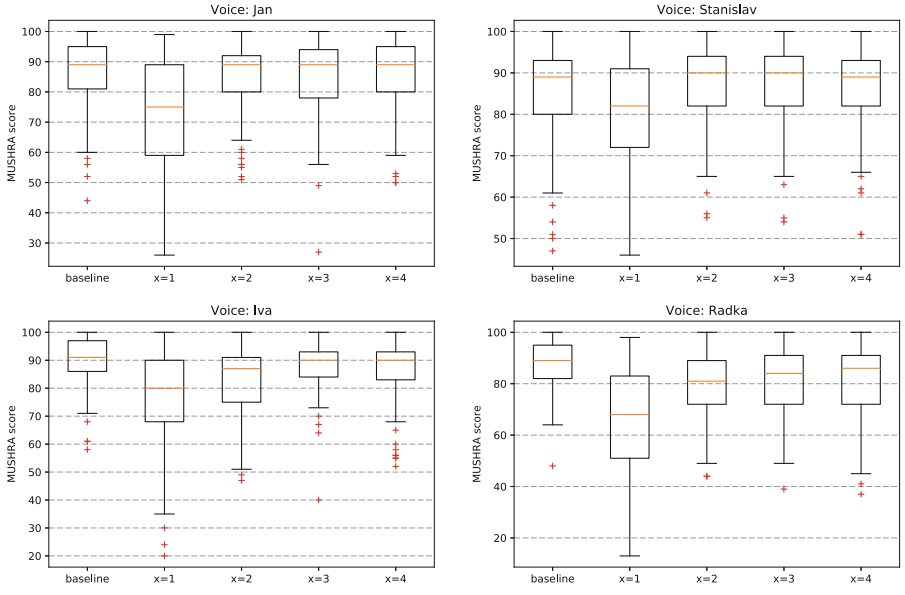

Fig. 3. Detailed plot of MUSHRA test results for all 4 professional voices.

Table 2. Results of MUSHRA listening test for all target computation schemes $x = 1, \ldots, 4$ and the baseline, presented for the speakers independently as well as collected to the overall results.

x	Jan		Stanislav		Iva		Radka		All voices	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1	72.39	75.00	80.30	82.00	75.51	80.00	65.68	68.00	73.47	76.50
2	84.83	89.00	87.26	90.00	83.05	87.00	79.20	81.00	83.51	87.00
3	85.20	89.00	87.19	90.00	87.52	90.00	79.84	84.00	84.94	88.00
4	85.58	89.00	86.13	89.00	86.66	90.00	80.75	86.00	84.78	88.00
Baseline	86.24	89.00	85.04	89.00	89.55	91.00	87.12	89.00	86.99	90.00

In the test, the 7 speech technologies experts were instructed to evaluate 15 shorter prompts for 4 of our professional unit selection voices [25], each prompt containing random-ordered 5 versions of a single prompt; the evaluation could use scale from 0 to 100 points (100 should be assigned to natural a sounding prompt, 0 to wrack). One of prompts was generated by the baseline unit selection [15] and four by the hybrid synthesizer with target computed as described in Sect. 2.2. The tests were reported rather demanding, with some of the versions sounding fairly similar, as illustrated by Fig. 3 and Table 2.

5 Conclusions

It can be seen from the very first experiments that the hybrid synthesis is able to achieve comparable speech quality as the raw unit selection, using symbolic features (IFF) in the target cost. And for the very first time it is evaluated on Czech language. Although it has been reported that the hybrid synthesis should be able to outperform the unit selection, it was not clearly shown in this paper, when large speech corpora has been used. On the other hand, we must emphasize that there still is room for improvement and thus the method can show its expected potential.

Regarding the future work, we aim at further experiments with target cost computation, for example to use z-score normalized coefficients, or a computation more aligned with [19] or the other approaches reported quality improvements, e.g. [16]. Also, stressing both methods with reduced speech unit inventory [6], or focusing on known unit selection failures [9,10] could provide valuable insights. In addition to target cost, the authors in [19,31] also adjusted the computation of concatenation cost, using cross-correlation to find the optimal join point for each unit join. This, however, can help unit selection in general, if proved being beneficial.

Naturally, the hardware requirements of this method are higher than what is required for "classic" unit selection [28]. It is due to both the SPS parameters generation for target and more complex target cost computation. However, being able to outperform unit selection quality, it is the price one is ready to pay, especially when compared with DNN-based approaches.

Acknowledgements. This research was supported by the Czech Science Foundation, project No. GA16-04420S.

References

1. Black, A.W., et al.: CMU blizzard 2007: a hybrid acoustic unit selection system from statistically predicted parameters. In: Blizzard Challenge (2007)
2. Boersma, P., van Heuven, V.: Praat, a system for doing phonetics by computer. Glot Int. **5**(9/10), 341–347 (2001)
3. Clark, R., Richmond, K., King, S.: Multisyn: open-domain unit selection for the festival speech synthesis system. Speech Commun. **49**(4), 317–330 (2007)
4. Hanzlíček, Z.: Czech HMM-based speech synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 291–298. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15760-8_37
5. Hanzlíček, Z.: Optimal number of states in HMM-based speech synthesis. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 353–361. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_40
6. Hanzlíček, Z., Matoušek, J., Tihelka, D.: Experiments on reducing footprint of unit selection TTS system. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 249–256. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_32
7. Hirai, T., Yamagishi, J., Tenpaku, S.: Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis. In: Proceedings of SSW6, pp. 81–84. ISCA, Bonn (2007)
8. ITU Recommendation BS.1534-2: Method for the subjective assessment of intermediate quality level of coding systems. Technical report, International Telecommunication Union (2014)
9. Jůzová, M., Tihelka, D., Skarnitzl, R.: Last syllable unit penalization in unit selection TTS. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 317–325. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_36
10. Jůzová, M., Tihelka, D., Volín, J.: F0 post-stress rise trends consideration in unit selection TTS. In: TSD 2018. LNCS. Springer (2018, to appear)
11. Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. **27**, 187–207 (1999)
12. Legát, M., Matoušek, J., Tihelka, D.: On the detection of pitch marks using a robust multi-phase algorithm. Speech Commun. **53**, 552–566 (2011)
13. Ling, Z.H., Wang, R.H.: HMM-based unit selection using frame sized speech segments. In: Proceedings of Interspeech 2006 - ICSLP, pp. 2034–2037. ISCA, Pittsburgh (2006)
14. Ling, Z.H., Wang, R.H.: HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion. In: Proceedings of ICASSP, pp. 1245–1248. Honolulu, Hawaii (2007)
15. Matoušek, J., Legát, M.: Is unit selection aware of audible artifacts? In: Proceedings of SSW8, pp. 267–271. ISCA, Barcelona (2013)
16. Merritt, T., Clark, R.A.J., Wu, Z., Yamagishi, J., King, S.: Deep neural network-guided unit selection synthesis. In: Proceedings of ICASSP, pp. 5145–5149. IEEE, Shanghai (2016)

17. van den Oord, A., et al.: WaveNet: a generative model for raw audio. CoRR abs/1609.03499 (2016)
18. Pollet, V., Breen, A.: Synthesis by generation and concatenation of multiform segments. In: Proceedings of Interspeech 2008, pp. 1825–1828. ISCA, Brisbane (2008)
19. Qian, Y., Soong, F.K., Yan, Z.J.: A unified trajectory tiling approach to high quality speech rendering. IEEE Trans. Audio Speech Lang. Process. **21**(2), 280–290 (2013)
20. Silén, H., Helander, E., Nurminen, J., Koppinen, K., Gabbouj, M.: Using robust Viterbi algorithm and HMM-modeling in unit selection TTS to replace units of poor quality. In: Proceedings of Interspeech 2010, pp. 166–169. ISCA, Makuhari (2010)
21. Sorin, A., Shechtman, S., Pollet, V.: Refined inter-segment joining in multi-form speech synthesis. In: Proceedings of Interspeech 2014, Singapore, pp. 790–794 (2014)
22. Taylor, P.: The target cost formulation in unit selection speech synthesis. In: Proceedings of Interspeech 2006 - ICSLP, vol. 1, pp. 2038–2041. ISCA, Pittsburgh (2006)
23. Taylor, P.: Text-to-Speech Synthesis, 1st edn. Cambridge University Press, New York (2009)
24. Tihelka, D.: Symbolic prosody driven unit selection for highly natural synthetic speech. In: Proceedings of Interspeech 2005 - Eurospeech, pp. 2525–2528. ISCA, Lisbon (2005)
25. 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: TSD 2018. LNCS. Springer (2018, to appear)
26. Tihelka, D., Matoušek, J.: Unit selection and its relation to symbolic prosody: a new approach. In: Proceedings of Interspeech 2006 - ICSLP, vol. 1, pp. 2042–2045. ISCA, Pittsburgh (2006)
27. Tihelka, D., Matoušek, J., Hanzlíček, Z.: Modelling F_0 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
28. Tihelka, D., Stanislav, P.: ARTIC for assistive technologies: transformation to resource-limited hardware. In: Proceedings of WCECS 2011, pp. 581–584. IANG, San Francisco (2011)
29. Tiomkin, S., Malah, D., Shechtman, S., Kons, Z.: A hybrid text-to-speech system that combines concatenative and statistical synthesis units. IEEE Trans. Audio Speech Lang. Process. **19**(5), 1278–1288 (2011)
30. Toda, T., Tokuda, K.: Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In: Proceedings of Interspeech 2005, Lisbon, Portugal, pp. 2801–2804 (2005)
31. Yan, Z.J., Qian, Y., Soong, F.K.: Rich-context unit selection (RUS) approach to high quality TTS. In: Proceedings of ICASSP 2010, Dallas, Texas, USA, pp. 4798–4801 (2010)
32. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Commun. **51**(11), 1039–1064 (2009)