Last Syllable Unit Penalization in Unit Selection TTS

Markéta Jůzová[1], Daniel Tihelka[1(✉)], and Radek Skarnitzl[2]

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

[2] Faculty of Arts, Institute of Phonetics, Charles University, Prague, Czech Republic
radek.skarnitzl@ff.cuni.cz

Abstract. While unit selection speech synthesis tries to avoid speech modifications, it strongly depends on the placement of units into the correct position. Usually, the position is tightly coupled with a distance from the beginning/end of some prosodic or rhythmic units like phrases or words. The present paper shows, however, that it is not necessary to follow position requirements, when the phonetic knowledge of the perception of prosodic patterns (mostly durational in our case) is considered. In particular, we focus on the effects of using word-final units in word-internal positions in synthesized speech, which are often perceived negatively by listeners, due to disruptions in local timing.

Keywords: Speech synthesis · Unit selection · Target cost · Word final lengthening

1 Introduction

Interactions between the segmental and prosodic levels of speech may be exemplified in several ways, but it is perhaps the temporal domain in which this interaction can most readily be observed. Temporal segmentation and the rhythm of speech have been shown to affect listeners to a great extent. Arhythmical or temporally unpredictable speech leads to longer reaction times in monitoring experiments [3,19] and, therefore, to increased cognitive processing of the incoming speech signal. In addition, speech with unnatural durational patterns has been shown to decrease intelligibility [18], or to induce negative perceptions regarding the speakers, making the speakers sound, for instance, more nervous and anxious [30] or less competent.

In the field of text-to-speech (TTS) synthesis, the durational patterns (as well as other prosody patterns) can either be modelled explicitly, i.e. by

This research was supported by the Czech Science Foundation (GA CR), project No. GA16-04420S, and by the grant of the University of West Bohemia, project No. SGS-2016-039.

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 317–325, 2017.
DOI: 10.1007/978-3-319-64206-2_36

means of an estimator, assigning a particular duration value to each speech unit to be synthesized [8,12,20], or by a symbolic description, defining only deep-level description (discriminative features defining what the prosody should/should not be) and expecting that the appropriate surface-level duration patterns (the particular unit durations) will emerge as a result of this description. While the former approach is used mainly in generative approaches like HMM [11], DNN [33] or single instance speech synthesis [16,27], the latter is mostly employed in unit selection speech synthesis, where it provides more robust selection criteria when compared to the following of generated prosody contours [24]. The insidious part is, however, the definition of the discriminative features. The usual way is to describe the position of speech units within a prosodic pattern (whether a phrase, word or syllable) in the source speech recordings and to select each particular unit into the most similar position in the synthesized phrase. The side effect is the increased pressure on the selection criteria, trying to balance the trade-off for all the features used, which is in details described in Sect. 3. First, however, let us look at the specific cases where it is important to consider temporal/durational properties of speech units and why this is necessary.

2 The Special Status of Final Syllables

Concatenative speech synthesis may be regarded as one of the key stimulating factors in the research of the effects of the prosodic domain on the duration of speech segments. Klatt [13] showed how segmental duration interacts with linguistic characteristics at various levels and devised a series of rules which predicted the durations of speech sounds in American English through the multiplication of a base value by coefficients related to various segmental and prosodic attributes (see [5,22] for similar studies).

The above-mentioned and other studies document several ways in which segmental duration is affected by the prosodic structure of utterances. In many languages, accented syllables are realized as longer than unaccented ones, or syllables tend to shorten in longer words [9]. Probably the most salient reflection of prosodic structure in the segmental strand is called phrase-final lengthening, which appears to be related to the general declination observed on a number of levels: ends of prosodic phrases are thus marked by lower speech rate and a drop in fundamental frequency [10,14], as well as larger and longer articulatory gestures [4] and laxer phonation [17]. Phrase-final lengthening has been documented in many languages [9], including Czech [7,29], and is considered to be universal in speech. A number of studies (see the Refs. in [9]) have agreed that the rate lowering concerns the rhyme of the final syllable (i.e., the vowel nucleus plus any consonants in the coda following the vowel).

Lengthening is thus a well-documented phenomenon at the level of prosodic phrases. In this study, we are interested in temporal adjustments at the level of individual words, which have been researched considerably less (although mentioned in literature). When comparing the duration of the schwa vowel framed in sentences /*Herpopp*a *posed a problem*/ and /*Her pop* o*pposed* the marriage/,

Beckman and Edwards [2] found the schwa in the first sentence significantly longer. Similar results were obtained by [6], who compared pairs like */lettuce–let us/* or */inquires–in choirs/*, though the effect was relatively small. More recently, word-final lengthening is discussed in [31], who, however, did not reach a definitive conclusion, and in [32], who report significantly longer durations in word-final positions. All of these studies consider English, though.

Until recently, no data on word-final lengthening have been available for Czech. A new study on the acoustic characteristics of Czech lexical stress [23] showed, however, that duration is the only parameter which systematically varies with lexical stress – note that Czech is a language with stress fixed on the first syllable of the base rhythm unit, also called the *prosodic word* [21]. Figure 1 shows part of the duration data taken from multi-syllabic words in spontaneous speech; only words which did not appear as phrase-final were analysed in the study. The results differ from what is typical in most languages: the vowel in the stressed syllable tends to be shortest and that in the last syllable is always longest (though not always statistically significant).

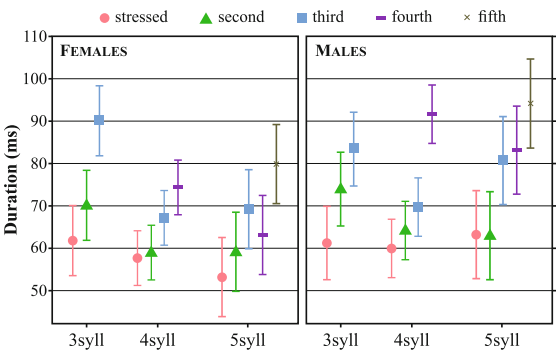

Fig. 1. Vowel duration in individual syllables of 3-, 4-, and 5-syllabic words; based on [23].

The account presented in this section indicates that phrase-final lengthening is essentially a ubiquitous and very salient phenomenon, but even word-final lengthening is non-negligible. The present study examines speech sounds in the last syllable of phonological words from the perspective of concatenative unit-selection synthesis of Czech. The results presented by [23] suggest that the word-final syllable may enjoy a special status in the prosodic hierarchy of Czech, and this may need to be reflected in the speech synthesis algorithm. Specifically, it is possible that units which appeared in the word-final position in the source recordings should not be selected for synthesis within words. This study therefore tests the hypothesis that the presence of word-final (source) units in word-internal positions in synthesized speech will be perceived negatively by listeners, since the lengthening will disrupt the local timing.

3 Handling of Unit Position in TTS Systems

The previous section indicates the great importance of the last syllable. We decided to check this in our TTS system ARTIC based on unit selection speech synthesis method. The highest influence of the unit position has the design of *target cost* (TC).

The target cost, as usually handled, is set to strive for putting the units into the same suprasegmental surroundings as they originally had in the source recordings, which is achieved when target features match. Thus, to make sure that the last-syllable position is handled properly, as described in Sect. 2, we could simply define an additional feature with *onset, nucleus, coda* symbolic values assigned to units recorded in the last syllable, and with *not-last* value for units from other syllables. Based on the match or mismatch of such feature values, an additional penalty would be added to the total TC, encouraging (in theory) the placement of units into the required position. However, in case of value mismatch, there is the same penalty no matter the value mismatch – i.e. the same penalty for e.g. *unknown*↔*nucleus* as for *coda*↔*onset* (unless a feature exchange matrix is defined somehow). An alternative way of defining the last-syllable position as a binary feature, i.e. a unit either *belongs* to the last syllable or *not*, is not going to improve the situation very much – there is better change to avoid mismatch of last/non-last syllable units, but units within the last syllable may still be interchanged between coda and onset parts (while both match the feature), which is not better either.

In our current TTS version, therefore, the computation of position within a prosodic word (hereafter referred to as *p-word*) is based on the assumption that there is no need to put a unit into the identical position in which the unit was originally recorded [26,28]. This is achieved through the definition of a *suitability* to the required position related to the beginning/middle/end of each p-word, represented by 3 values, each corresponding to one von Hann window spanned through the p-word, as illustrated in Fig. 2. Such a position feature can avoid a completely wrong placement (i.e. unit originating at the beginning of a p-word to be placed to its end), while it still allows some kind of flexibility in units interchanging, when the unit originates in a position which is "close" to the one in which it is to be placed. From the point of view of the selection

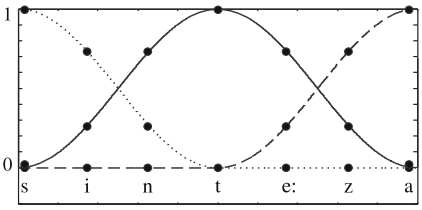

Fig. 2. The illustration of the correspondence of windowing functions to a prosodic word "synthesis". Individual windows are distinguished by line style, points correspond to the values describing the candidates.

algorithm, the set of candidates suitable for the given p-word position (as defined by the three values) is wider than if the position were defined as, for example, distance from the beginning and from the end of the p-word. And the wider set to select from gives the algorithm a better chance to choose a more appropriate unit, when the other target and concatenation features are taken into account as well.

On the other hand, such a position measure (and either of similar) is not able to handle a "last-syllable" occurrence feature reliably. We observed unnatural unit lengthenings close to the end of a p-word which occurred when a unit from the last syllable was placed to the penultimate syllable, since its position was suitable to the position required, as measured by the current approach.

4 New Unit Position Reflecting Final Syllable Status

Based on the findings from Sect. 2, we thus re-defined the unit position feature in a way that instead of measuring appropriateness for the given position, we strongly penalize placements into inappropriate positions, while expecting any other placements to be equally suitable, i.e. not penalizing them in any way. Let us emphasize that this completely replaces the original position function from Sect. 3, instead of just being added as an extra measure. The reason is that adding a feature would increase the pressure on the selection algorithm, lowering the set of candidates matching the feature and thus increasing the chance that a unit not matching the required position will be used after all. And moreover, ensuring a position placement is not that significant, as suggested in Sect. 2.

Let us also note that there was no exact syllabification used to identify the last p-word syllable (neither is it possible to detect syllables exactly [15]). Instead, we simply found the last vowel (or syllabic consonant) V in each p-word, and the diphones *[*-V]*, *[V-*]* as well as the remaining diphones until the end of p-word were considered to be the syllable constituents (cf. references in Sect. 2 which show that the rhyme of the final syllable is most affected by final lengthening).

Having the new positional feature, we synthesized more than a million phrases by the original (marked as TTS_{base} hereafter) and by the modified version of unit selection method (TTS_{syll}) embedded into our TTS system. The resulting sequences of units provided by TTS_{base} for each phrase were further analysed – both sequence of the phrase units and its corresponding selected units were passed through TTS_{syll} module which returned the high TC value for units with inappropriate syllable position. This number of position misplacements was used as the base score for the selection of phrases to be evaluated by listening tests.

4.1 Listening Tests Overview

To verify our presumption, we carried out a large 3-scale preference listening test, where two variants of the same phrase, synthesized by TTS_{base} and TTS_{syll}, were compared. To select the phrases for the evaluation, the set of all the synthesized

phrases was first limited to those having 40 characters at most, since smaller differences in long phrases are rather hard for listeners to recall [1]. Then, only the phrases with two or more misses, as described in Sect. 4, were kept. The final set of 25 phrase pairs was selected randomly, while 4 different professional synthetic voices, two male and two female, were used – in total 100 phrases were thus compared.

22 listeners participated in the test, 7 of them being speech synthesis experts, 6 being phoneticians and 9 naive listeners. We intentionally did not inform any of them about the experiment's details, since the aim was to test the overall quality of the new TTS system. Furthermore, during the listening test, the order of the samples was randomized across the listening prompts, so the listeners did not know which one was synthesized by TTS_{base} and by TTS_{syll}. The listeners were able to listen to each stimulus repeatedly, they were instructed to use earphones, and had to choose one of the following choices: *sample A sounds better/samples are of the same quality/sample B sounds better*. The A/B assignments were then normalized to $A = 1$ where TTS_{syll} variant was preferred, $A = -1$ where TTS_{base} was preferred, and $A = 0$ otherwise. The final score s of the listening test T was then computed using the Eq. 1

$$s = \frac{\sum_{A \in T} A}{\sum_{A \in T} 1} \tag{1}$$

Thus, the positive value of s indicates the improvement of the overall quality when using the new last-syllable feature.

Let us note that for the purpose of this testing, we did not use the procedure designed in [25]. The main reason is that we do not have to rely on numbers of changed units when comparing the output sequences from the two system versions; instead we can simply detect wrong cases in TTS_{base}, as described in Sect. 4. Nevertheless, we plan to use the procedure before the final deployment of TTS_{syll} to the production-ready version of our TTS system.

5 Results

The results of the listening tests are shown in the Table 1. All the score values s are positive, indicating a considerable improvement of the quality of synthesized samples.

Since the score value for *male speaker 1* obtained from phonetics experts' answers seems to be low and not so conclusive, we decided to prove the statistical significance of this result. We have carried out the sign test with the null hypothesis H0:*the outputs of the both systems are of the same quality*, and alternative hypothesis H1:*the output of one system sounds better*. The computed p-value $= 0.0193$, so we can reject the null hypothesis H0 at $\alpha = 0.05$ significance level, concluding that the quality of TTS_{syll} system is really higher.

However, there is a considerable number of "*same quality*" evaluations. It suggests that not all candidates from the last syllable cause a speech artefact

Table 1. The results obtained from listening tests. The table contains the number of listener answers and the score values s computed by Eq. 1.

	Male spkr 1	Male spkr 2	Female spkr 1	Female spkr 2	All speakers
	Numbers of answers in percents				
TTS_{syl} better	58.7%	56.4%	65.1%	53.6%	58.5%
Same quality	20.0%	27.1%	21.8%	29.3%	24.5%
TTS_{base} better	21.3%	16.5%	13.1%	17.1%	17.0%
	Score value s				
All listeners	**0.375**	**0.398**	**0.520**	**0.365**	**0.415**
TTS experts	0.531	0.480	0.531	0.457	0.500
Phonetics experts	0.187	0.313	0.520	0.387	0.352
Naive listeners	0.378	0.391	0.511	0.280	0.390

when placed to non-last syllable position – note that, as described in Sect. 4, we *know* that there is a unit from the last syllable placed to non-last position in the TTS_{base}.

On the other hand, the listeners sometimes preferred the TTS_{base} variant. Therefore, we inspected the problematic prompts, trying to find the cause of the preference. The main problem was a disturbing artefact of another kind (not directly related to syllable position) in TTS_{syl} variant, while none of the syllable position misses in TTS_{base} was perceived negatively, nor was there another artefact. In a few cases, moreover, even TTS_{syl} still contained unnatural inter-phrase lengthenings. These originated from the failures of p-word tokenization in source recordings, and thus the system used the last-syllable unit into inappropriate position without being able to realize it.

6 Conclusion

The results of the listening tests clearly confirm the importance of the correct handling of the last syllable of a p-word. Let us also note that the change of paradigm, when instead of "forcing" units into the expected position we try to avoid their use in positions where they are known to cause audible artefacts, follows the principle of units synonymy/homonymy established in [24,26]. We believe that this is the right direction towards the tuning of unit selection features, as it allows the use of units in a much wider range of placements (than the one in which the unit has been placed in the source recordings), and it also avoids the definition of a penalty function which would evaluate a distance of what the unit is (where it *is* placed) to what it is required to be (where we *try* to place it).

One of the possible further improvements will now be the focus of p-word tokenization which was found to be incorrect in some of the cases examined and proved to be the clear cause of TTS_{base} preference in these. Also, the findings

in Sect. 5 suggest that there may be some durational (or prosodic, in general) patterns, allowing the exchange of the last syllable and non-last syllable units under some conditions. Answering this phenomena would even more relax the pressure on the selection algorithm, thus widening the set of candidates to be used.

References

1. Baddeley, A.: Human Memory: Theory and Practice. Psychology Press, East Sussex (1997). Revised edn
2. Beckman, M., Edwards, J.: Lengthenings and shortenings and the nature of prosodic constituency. In: Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech, pp. 152–178. Cambridge University Press, Cambridge (1990)
3. Buxton, H.: Temporal predictability in the perception of English speech. In: Cutler, A., Ladd, D.R. (eds.) Prosody: Models and Measurements, vol. 14, pp. 111–121. Springer, Heidelberg (1983)
4. Byrd, D., Saltzman, E.: The elastic phrase: modelling the dynamics of boundary-adjacent lengthening. J. Phonetics **31**, 149–180 (2003)
5. Crystal, T.H., House, A.S.: Segmental durations in connected-speech signals: current results. J. Acoust. Soc. Am. **83**, 1553–1573 (1988)
6. Cutler, A., Butterfield, S.: Syllabic lengthening as a word boundary cue. In: Proceedings of the 3rd Australian SST, pp. 324–328 (1990)
7. Dankovičová, J.: The domain of articulation rate variation in Czech. J. Phonetics **25**, 287–312 (1997)
8. Fernandez, R., Rendel, A., Ramabhadran, B., Hoory, R.: Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In: Proceedings of Interspeech, pp. 2268–2272. ISCA (2014)
9. Fletcher, J.: The prosody of speech: timing and rhythm. In: The Handbook of Phonetic Sciences, pp. 521–602. Blackwell Publishing Ltd. (2010)
10. Gussenhoven, C.: The Phonology of Tone and Intonation. Cambridge University Press, Cambridge (2004)
11. Hanzlíček, Z.: Czech HMM-based speech synthesis: experiments with model adaptation. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 107–114. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23538-2_14
12. Holm, B., Bailly, G.: Generating prosody by superposing multi-parametric overlapping contours. In: Proceedings of ICSLP, pp. 203–206 (2000)
13. Klatt, D.H.: Linguistic uses of segmental duration in English: acoustic and perceptual evidence. J. Acoust. Soc. Am. **59**, 1208–1221 (1976)
14. Ladd, D.R.: Intonational Phonology, 2nd edn. Cambridge University Press, Cambridge (2008)
15. Matoušek, J., Hanzlíček, Z., Tihelka, D.: Hybrid syllable/triphone speech synthesis. In: Proceedings of 9th Interspeech (Eurospeech), Lisbon, Portugal, pp. 2529–2532 (2005)
16. Matoušek, J., Romportl, J., Tihelka, D., Tychtl, Z.: Recent improvements on ARTIC: czech text-to-speech system. In: Proceedings of Interspeech, Jeju Island, Korea, pp. 1933–1936 (2004)
17. NíChasaide, A., Yanushevskaya, I., Gobl, C.: Prosody of voice: declination, sentence mode and interaction with prominence. In: Proceedings of 18th ICPhS (2015). Paper 476

18. Quené, H., van Delft, L.E.: Non-native durational patterns decrease speech intelligibility. Speech Commun. **52**(11–12), 911–918 (2010)
19. Quené, H., Port, R.: Effects of timing regularity and metrical expectancy on spoken-word perception. Phonetica **62**(1), 1–13 (2005)
20. Romportl, J., Kala, J.: Prosody modelling in Czech text-to-speech synthesis. In: Proceedings of the 6th ISCA SSW, Bonn, pp. 200–205 (2007)
21. Romportl, J., Matoušek, J., Tihelka, D.: Advanced prosody modelling. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS, vol. 3206, pp. 441–447. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30120-2_56
22. van Santen, J.P.H.: Assignment of segmental duration in text-to-speech synthesis. Comput. Speech Lang. **8**, 95–128 (1994)
23. Skarnitzl, R., Eriksson, A.: The acoustics of word stress in Czech as a function of speaking style. In: Proceedings of Interspeech (2017)
24. Tihelka, D.: Symbolic prosody driven unit selection for highly natural synthetic speech. In: Proceedings of 9th Interspeech (Eurospeech), pp. 2525–2528. ISCA, Bonn (2005)
25. Tihelka, D., Grůber, M., Hanzlíček, Z.: Robust methodology for TTS enhancement evaluation. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS, vol. 8082, pp. 442–449. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40585-3_56
26. Tihelka, D., Matoušek, J.: Unit selection and its relation to symbolic prosody: a new approach. In: Proceedings of 9th ICSLP, vol. 1, pp. 2042–2045. ISCA, Bonn (2006)
27. Tihelka, D., Méner, M.: Generalized non-uniform time scaling distribution method for natural-sounding speech rate change. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 147–154. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23538-2_19
28. Tihelka, D., Romportl, J.: Exploring automatic similarity measures for unit selection tuning. In: Proceedings of 10th Interspeech, pp. 736–739. ISCA, Brighton (2009)
29. Volín, J., Skarnitzl, R.: Temporal downtrends in Czech read speech. In: Proceedings of Interspeech, pp. 442–445 (2007)
30. Volín, J., Poesová, K., Skarnitzl, R.: The impact of rhythmic distortions in speech on personality assessment. Res. Lang. **12**, 209–216 (2014)
31. White, L., Turk, A.E.: English words on the procrustean bed: polysyllabic shortening reconsidered. J. Phonetics **38**(3), 459–471 (2010)
32. Windmann, A., Šimko, J., Wagner, P.: Polysyllabic shortening and word-final lengthening in English. In: Interspeech 2015, pp. 23–40 (2015)
33. Wu, Z., Watts, O., King, S.: Merlin: an open source neural network speech synthesis system. In: Proceedings of 9th ISCA SSW, pp. 218–223, September 2016