Discontinuities in fundamental frequency: When do they really matter in synthetic speech?

Nespojitosti základní frekvence: kdy mají v syntetické řeči vliv?

Tomáš Bořil and Radek Skarnitzl

Charles University, Faculty of Arts – Institute of Phonetics, náměstí Jana Palacha 2, 116 38 Praha 1

Attempts at improving the naturalness of synthetic speech have typically led to penalizing unsuitable candidates or large differences in acoustic parameters around the concatenation point. This paper reports a perceptual experiment which aimed at the opposite: relaxing the criteria for concatenation cost in the domain of fundamental frequency (f_0), specifically when concatenating diphones pertaining to voiced consonants. A listening test which involved several types of artificial f_0 discontinuities was administered to 21 respondents. The results suggest that f_0 discontinuities only matter in sonorant consonants (nasals and approximants) and only when they exceed 1 semitone. Most importantly, the direction of f_0 change should be taken into account, and not only the values around the concatenation point.

1. Introduction

Concatenative speech synthesis systems based on dynamic unit selection continue to dominate real-life applications, although research endeavours have, to a large extent, moved away from this relatively costly approach to generating artificial speech. It is the still superior naturalness of concatenative speech synthesis which lies behind this continued preference [1]. However, the output of concatenative synthesis may suffer from the sporadic occurrence of audible discontinuities. These artefacts, which may have an intrusive effect on the listener, may have several causes, as summarized by [2]. First, the database from which units (typically diphones) are selected for synthesis may feature some errors, either random or systematic (see [3] and also [4] for a proposal to eliminate some of the latter ones from the Czech synthesis system ARTIC [5]); this is the case especially in languages with a more or less straightforward relationship between spelling and pronunciation like Czech. Second, the *target cost* and *concatenation cost*, two functions governing the selection of units from the database, may not correlate perfectly with human perception and may thus fail to capture some audible discontinuities. Finally, because selection algorithms typically prefer a low global cost over a low local cost, the globally "cheapest" set of selections may feature a local artefact at a specific concatenation point.

A number of experiments have addressed the question of artefacts in concatenative synthesis. The most intrusive effect on the listener seems to be exerted by "jumps" in the fundamental frequency (f_0) of the voice [6], [7] and by discontinuities in the spectral domain [8], [9]. Many past attempts at improving the specification of the *target* and *concatenation cost* have focused on stipulating penalties concerning, for instance, the permissible difference in the acoustic parameters of neighbouring diphones or the context in which the source and target diphones could appear.

Naturally, the more rules there are and the more potential diphone candidates are penalized, the fewer units remain for selection. That is why, in our most recent attempts at improving the ARTIC synthesis system, we have adopted an opposite perspective: we are applying phonetic experimentation to investigate in which specific contexts a given acoustic difference does need to be taken into account in calculating the concatenation cost, and when a difference of, stated objectively, the same or even greater magnitude may be ignored because the acoustic discontinuity is not perceptually detectable. This study addresses fundamental frequency which, according to our informal observations, continues to be one of the most frequent sources of intrusive artefacts in the ARTIC synthesis system. In the current implementation of ARTIC [5], the transition of f_0 between neighbouring diphones is part of the *concatenation cost* calculation in all voiced segments. The aim of this study is to verify whether this is necessary, or whether acoustic (objective) discontinuities may be ignored in some contexts because they are not perceptible.

2. Fundamental frequency vs. perceived pitch

First of all, it must be emphasized that the f_0 contour (an output of a f_0 extractor) does not correspond to the pitch contour (the subjective percept of pitch movements); in other words, listeners do not perceive pitch objectively. There are several components of the discrepancy between an f_0 contour and its corresponding pitch contour. Researchers often talk about *pitch contour stylization*, which refers to such an approximation of the extracted f_0 contour so that it is perceptually indistinguishable (or at least so that it perceptually resembles) from the original [10], [11].

Přijato 14. října 2018, akceptováno 21. ledna 2019.

The first step in bringing f_0 and pitch closer to each other consists in expressing differences in a psychoacoustic unit rather than in the physical unit Hertz; it was found that semitones (ST) best correspond to the perceptual impression of pitch [12]. The next important component that has to be accounted for concerns the so-called microprosodic variations [13], [14], where f_0 is affected by the voicing status of the surrounding consonants; these small perturbations are not perceptible and have to be eliminated. We can state in general that f_0 changes of short durations and small magnitudes are not perceptible [10], [11]; that is why the f_0 contours should always be smoothed (i.e., lowpass-filtered).

Another important aspect of pitch perception is the alignment of perceived pitch to the segmental chain. As summarized by [11] or [15], evidence suggests that we perceive pitch mostly in syllabic nuclei (i.e., typically vowels, sometimes sonorant consonants), most likely in their central portion. Most frequently, every syllable is perceived as having one tone; it is only in final syllables of prosodic phrases, which carry the nuclear tone and where syllabic nuclei are sufficiently lengthened, where we perceive melodic changes [11].

If we consider these findings from the opposite perspective, it is clear that f_0 changes in consonants should not contribute to the perceived pitch contour. That does not automatically mean, however, that larger f_0 jumps occurring within consonants may not be audible. The main research question of the current study therefore is whether discontinuities in fundamental frequency, when concatenating diphones pertaining to a consonant, will have an intrusive effect on listeners. More specifically, we want to examine whether there is a threshold beyond which the f_0 jump is already perceptible, whether a larger context of the f_0 contour may play a role in the perceptual judgements, and whether this effect applies to all consonant classes. Since this is an exploratory study, we only formulate a general hypothesis: it is predicted that listeners will not be equally sensitive to all types of f_0 discontinuities.

3. Method

To investigate the effect of f_0 discontinuities, it was essential to use very short sound stimuli and manipulate them in a strictly controlled manner. As source material, we used recordings of [aCa] disyllables, where the voiced intervocalic consonant (C) included two plosives [b, d], two fricatives [z, ʒ], two nasals [m, n], two liquids [l, r], and also [v], a voiced fricative which, however, retains some properties of sonorant sounds [13]. These source disyllables were recorded by 4 female and 4 male native speakers of Czech; an EGG signal using the VoceVista system [14] was recorded alongside the audio to ensure completely reliable f_0 values. Attention was paid during the recording that the intervocalic consonant was pronounced with full voicing (obstruents frequently lose some voicing in intervocalic positions [15]).

The subsequent manipulations of f_0 were performed by means of PSOLA [16] in Praat [20] on these source disyllabic recordings, using a Praat script. The time points used for the manipulations, stipulated based on [21], are shown in Fig. 1.

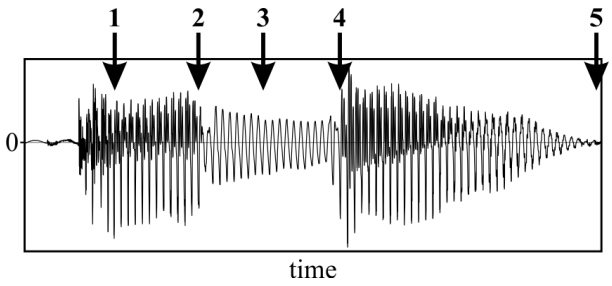

Figure 1: Time points of disyllables indicated in the waveform of [aba]. 1: onset of the periodic part of vowel 1; 2: consonant onset; 3: midpoint of consonant (or of its closure phase); 4: consonant offset; 5: offset of vowel 2

We simulated f_0 jumps of 1 and 5 semitones (ST) around time point 3; these intervals were selected because they seem to correspond to the range of f_0 discontinuities encountered during an analysis of the ARTIC synthesis outputs. There were two types of experimental manipulations, as illustrated in the left panel of Figure 2. The two types differ in how f_0 changes between time points 2 and 4, i.e., during the target consonant (or, in case of plosive sounds, during their closure phase). In the first type, f_0 remained stationary before and after the jump itself; this type, which is based on the Heaviside step function, will be henceforth referred to as type H (see the f_0 contours in H1 and H5 in Fig. 2). In the second type, f_0 was manipulated so that it changes during the consonant beyond the 1- or 5-ST jump itself; importantly, the change is in the opposite direction with respect to the target jump, resembling a sawtooth. Specifically, f_0 remained stationary in the vowel, then dropped by 0.5 or 2.5 ST respectively during the first half of the consonant (or, in the case of plosives, of the closure phase, between time points 2 and 3), jumped up abruptly by 1 or 5 ST respectively (this is the target f_0 jump), and dropped again by 0.5 or 2.5 ST during the second half of the consonant, between time points 3 and 4. This sawtooth-like change is henceforward referred to as type S. The target f_0 jump always occurred within 2 milliseconds, which is comparable to jumps occurring in synthetic speech. In total, this yielded four types of modified stimuli.

As shown in the right panel of Fig. 2, a "default" version was created as a control to each of the experimental manipulations, which involved either flat f_0 or a "natural" jump (i.e., one which may occur in ordinary speech), around time point 2 (i.e., at the onset of the intervocalic consonant). The objective was to generate pairs of disyllables which – if the performed manipulation were not per-

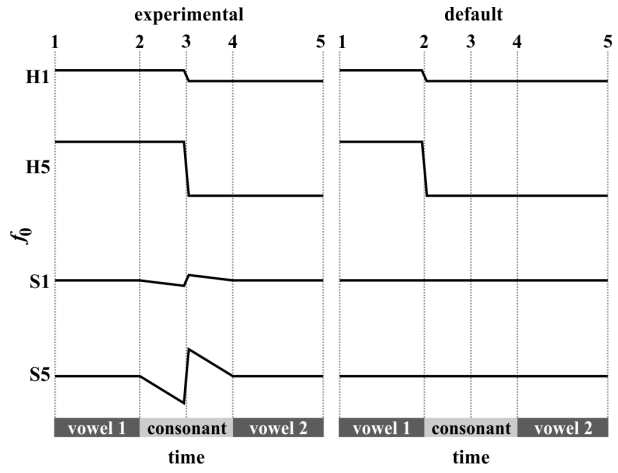

Figure 2: Four types of experimental manipulations on the left – Heaviside (H) and sawtooth (S) jumps by 1 and 5 ST – with their corresponding default versions on the right (see text)

ceptible – would have identical perceptual effect (i.e., their perceived intonation would be the same).

In total, the study is based on 8 speakers, 4 female and 4 male. Since we did not want listeners to have to perceptually "switch" between male and female voices, as the acoustic differences between the stimuli (caused by the manipulations) are very small, we created two tests, and the listeners were randomly divided into two groups, listening only to male or only to female stimuli. The manipulated and default variants were used to create a listening test. Each test item consisted of a pair of stimuli, one default and one manipulated. In total, the listening test contained 144 items (4 speakers × 9 consonants × 8 variants). No test items were repeated.

The listening test was administered to 21 respondents via ARTIC-Tests 3.0, a web-based environment created by the West Bohemian University in Pilsen (11 respondents evaluated the female stimuli, 10 evaluated the male stimuli); all were students at Charles University, Faculty of Arts. The respondents' task was to listen to random-order sorted items consisting of two sounds (one always being the manipulated, the other the default version, in random order) and to decide whether one of them sounded intrusive or whether both sounded the same. They indicated their choice by clicking one of three radio buttons: the first sound is worse, the second sound is worse, they are both of equal quality. They were allowed to repeat each sound at will. The listeners were instructed to use closed headphones. Since the sound stimuli were very short, the entire listening test, with the 144 items, did not last longer than 15 minutes.

Statistical analyses were carried out using R [22], and graphical outputs were created using the R package *ggplot2* [23].

4. Results

The listeners' responses were associated with values as follows: 1 = the manipulated stimulus sounds worse; 0 = both sounds are of equal quality; and –1 = the default stimulus sounds worse. Figure 3 shows these results split into groups by combining the consonant in the disyllable, manipulation type (H and S), and size of the manipulation interval (1 and 5 ST). For each group we calculated the mean value and estimated confidence intervals using the bootstrap method with a significance level of 0.05 (Bonferroni-corrected for multiple testing). This means that a null hypothesis of no noticeable hearing difference between the manipulated and default version of a stimulus cannot be rejected if the confidence interval includes the value of 0.

It is immediately apparent that the listeners perceived no clear difference in the quality of the sound when Heavi-

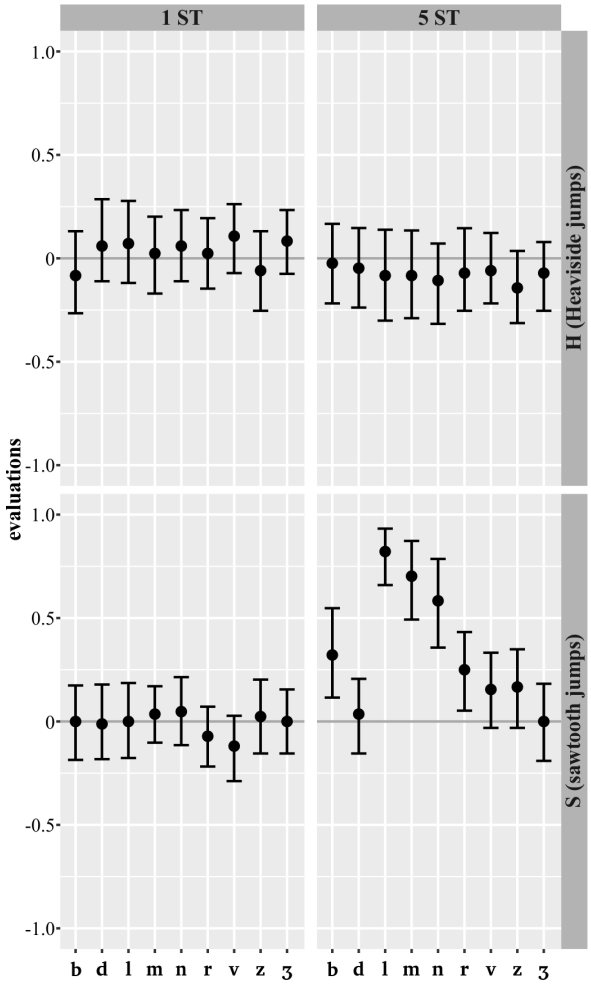

Figure 3: Responses for individual consonants to the Heaviside (H, top) and sawtooth (S, bottom) discontinuities of 1 semitone (left) and 5 semitones (right); see section 3 for more details. The evaluation of 1.0 corresponds to the manipulated stimulus sounding worse, 0 to no difference in evaluation (i.e., chance level), and the evaluation of −1.0 to the default stimulus sounding worse

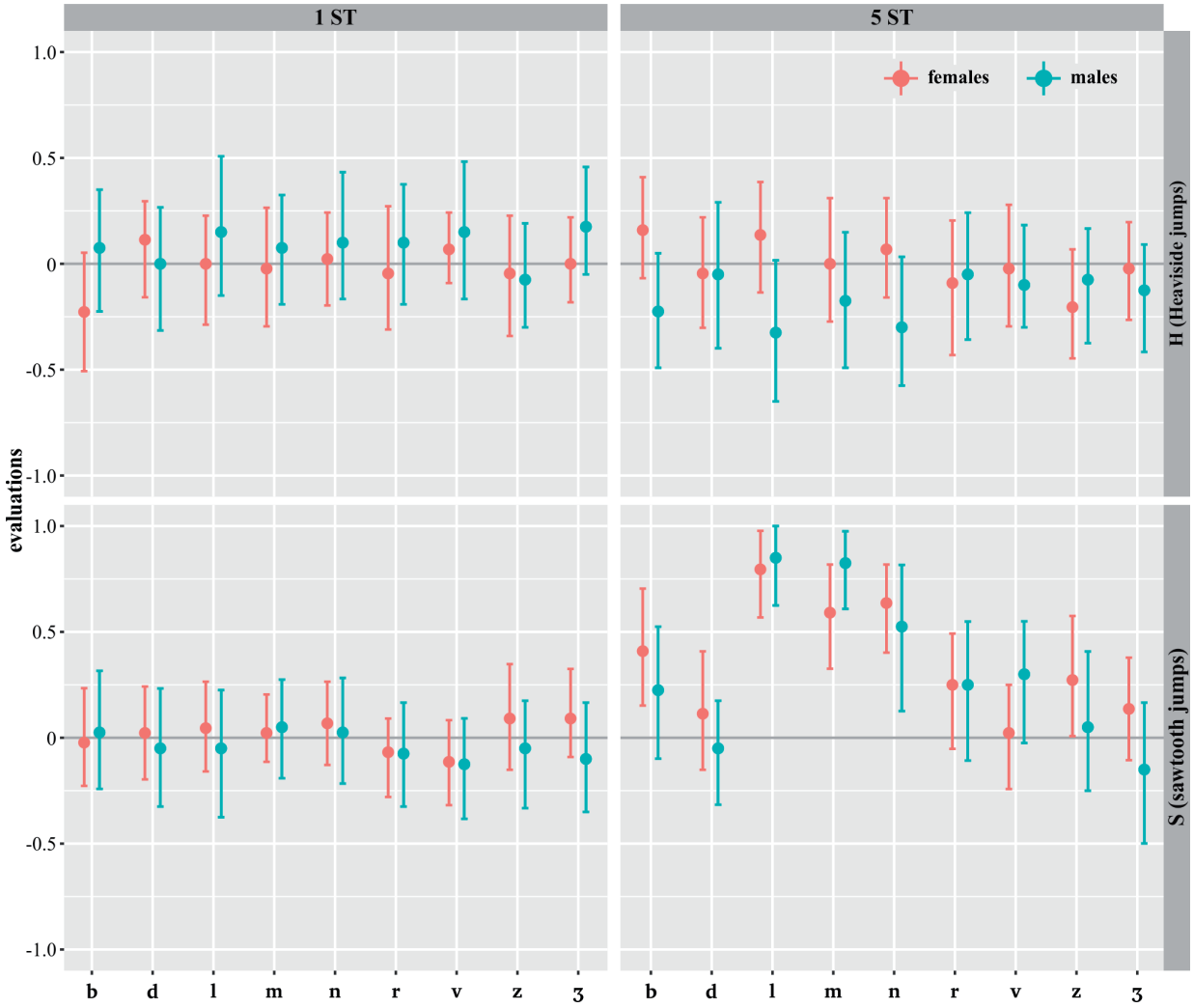

Figure 4: Responses for individual consonants to the Heaviside (H, top) and sawtooth (S, bottom) discontinuities of 1 semitone (left) and 5 semitones (right), separately for male and female speakers; see section 3 for more details. The evaluation of 1.0 corresponds to the manipulated stimulus sounding worse, 0 to no difference in evaluation, and the evaluation of −1.0 to the default stimulus sounding worse

side jumps (marked H) were concerned: confidence intervals for all consonants intersect the value of 0, irrespective of the f_0 manipulation interval. While the same applies for all the consonants in which a sawtooth (S) discontinuity of 1 semitone was introduced (see the bottom left panel of Figure 3), some sawtooth f_0 discontinuities in the order of 5 semitones clearly do matter. It can be seen that it is especially the sonorant consonants (i.e., [l, m, n, r]) and also the plosive [b] where the listeners could hear a difference in the quality of the sound. Specifically, the manipulated stimuli were perceived as significantly inferior in comparison with the default versions.

In Figure 4 the results of the listening test are shown separately for the female and male speakers. Each female-speaker group consists of 44 values (4 speakers × 11 respondents) and each male-speaker group consists of 40 values (4 speakers × 10 respondents). It was not the purpose of this study to examine the effect of speaker sex; for that our data would not be sufficient. The figure merely shows that there may be some small differences in the results, which may be caused by the specificity of the individual voices.

Naturally, these separated results are comparable to the pooled data presented in Figure 3. First, the Heaviside discontinuities in f_0, of either 1 ST or 5 ST, do not seem to be perceptually salient in any of the examined consonants, as indicated in the top panel of Figure 4. Again, the same applies for the sawtooth discontinuities of 1 ST. In addition to these similarities, however, there are some differences in the bottom right quadrant of the figure which are worth pointing out.

Most importantly, it can be seen that the manipulated stimuli of the sonorants [l, m, n] were evaluated as significantly worse in quality than their corresponding default stimuli. While the pooled evaluation for the trill [r] did reach statistical significance, the evaluation is not signifi-

cant when the stimuli from male and female speakers are considered separately. The figure also suggests that the significant effect in the assessment of the plosive [b] was pulled by the responses to the female speakers' stimuli.

5. Discussion and conclusions

The objective of this exploratory study was to investigate in greater detail the perceptual aspects of concatenating diphones, where the concatenation involves various kinds of discontinuities in the fundamental frequency of the voice (f_0). Although the listening test itself was not excessively long and the web-based environment allowed the respondents to interrupt the experiment and resume it later, informal post-hoc queries from some of the respondents indicated that the listening was tedious. More specifically, what may have been slightly frustrating for the listeners was the inevitable tendency that, in line with our predictions, many stimuli pairs would sound the same in terms of their quality. We therefore believe that the fact that positive results were obtained – i.e., that the listeners diligently compared the stimuli throughout the 144 items – is worth emphasizing.

The results of the presented experiment are positive in several aspects. First, they confirm previous findings related to the perception of pitch (see [11] or [15]), but make them more detailed. The most important implications are related to our ultimate aim, which was to simplify the selection of diphones for concatenative speech synthesis using dynamic unit selection. Our results show that acoustic discontinuities at the point of concatenation within a consonant which are smaller than 1 semitone do not seem to be perceptually relevant. Based on this finding, f_0 jumps smaller than (at least) 1 ST can be ignored when concatenating diphones pertaining to any voiced consonant.

More interesting are our findings regarding the nature of the introduced discontinuity. The upper right panels of Figures 3 and 4 suggest that even discontinuities of 5 semitones do not lead to an intrusive perceptual effect, if they are "smooth" in the sense that the f_0 contour in the vicinity of the jump does not involve movement contrary to the jump (these changes were labelled H, as they resemble the Heaviside function). It is only 5-ST jumps which involve a more salient change of direction of the f_0 contour – these were labelled S for sawtooth – that have resulted in a significant perceptual effect. The conclusion that can be drawn from this result is that it would be highly beneficial to calculate f_0 not only in the frames closest to the concatenation point, but to also incorporate the direction of f_0.

To provide a more specific example, extrapolating on our results, we may hypothesize that the discontinuity in the f_0 track marked as A in Figure 5 will not be perceptually salient, while that marked as B – which involves exactly the same jump around the concatenation point in terms of its magnitude but one that is sawtooth-like – most likely will.

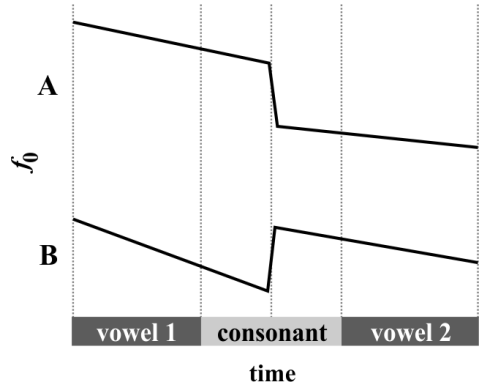

Figure 5: Schematic illustration of two "identical" f_0 discontinuities around the concatenation point (see text)

Finally, let us turn to the finding which concerns the individual consonants (or more precisely, consonant types). As mentioned in the Introduction, the current implementation of the ARTIC system [5] considers f_0 in all voiced segments to determine the *concatenation cost*. The results of this study prove that this not necessary. Perceptually salient artefacts have been conclusively obtained only for the sonorant sounds, specifically the nasals (in Czech, this would be [m n ɲ ŋ]) and the lateral approximant [l]; it may be assumed that the same would apply to the palatal glide [j]. The manipulated stimuli of the trill [r] – also classified as a sonorant sound – were also evaluated as worse in the pooled data. On the other hand, the significantly worse evaluation of manipulated [b] items is not straightforward.

To conclude, this experiment aimed at simplifying unit selection when it comes to incorporating f_0 in the *concatenation cost* when concatenating diphones pertaining to voiced consonants. The results show that the direction of f_0 change needs to be taken into account, that only sonorant sounds should be considered, but only when the discontinuity exceeds 1 semitone. It may be worthwhile to conduct a mode detailed experiment which would determine with greater precision where between 1 and 5 ST the boundary of perceptual intrusiveness lies.

While this study was motivated by audible artefacts in the Czech speech synthesis ARTIC [5], it is to be expected that our results may be applicable in any speech synthesis algorithm which makes use of the f_0 criterion in the computation of the *concatenation cost*. Although savings in terms of computation time or in terms of the number of diphones which would have been previously eliminated from selection and retained after the inclusion of the proposed relaxed criteria have not been examined, we assume that especially the latter aspect – having more diphones available for concatenation – is an important result of this experiment.

Acknowledgements

This research was supported by the Czech Science Foundation project No. 16-04420S.

References

[1] Dutoit, T.: Corpus-based speech synthesis, in Benesty, J., Sondhi, M., Huang Y. (Eds.), *Springer Handbook of Speech Processing*, p. 437–455. Springer, Dordrecht, 2008.

[2] Matoušek, J., Tihelka, D., Legát, M.: Is unit selection aware of audible artifacts? *Proc. 8th ISCA Speech Synthesis Workshop 2013*, p. 267–271, 2013.

[3] Matoušek, J., Tihelka, D.: Anomaly-based annotation error detection in speech-synthesis corpora, *Computer Speech & Language*, 46, p. 1–35, 2017.

[4] Skarnitzl, R.: Alofonická variabilita v češtině z pohledu řečové syntézy, *Akustické listy*, 24, p. 15–20, 2018.

[5] 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.), *Text, Speech, and Dialogue, TSD 2018*. Lecture Notes in Computer Science, vol. 11107. Springer, Cham, 2018.

[6] Legát, M., Matoušek, J.: Pitch contours as predictors of audible concatenation artifacts, *Proc. World Congress on Engineering and Computer Science 2011*, p. 525–529, 2011.

[7] Dutoit, T.: Corpus-based speech synthesis in Benes ty, J., Sondhi, M. M. Huang, Y. (Eds.), *Springer Handbook of Speech Processing*, p. 437–455. Springer, Berlin, 2008.

[8] Klabber, E., Veldhuis, R.: Reducing audible spectral discontinuities, *IEEE Transactions on Speech and Audio Processing*, **9**(1), p. 39–51, 2001.

[9] Bořil, T., Šturm, P., Skarnitzl, R., Volín, J.: Effect of formant and F0 discontinuity on perceived vowel duration: Impacts for concatenative speech synthesis, *Proc. Interspeech 2017*, p. 2998–3002, 2017.

[10] Hart, J., Collier, R., Cohen, A.: *A perceptual study of intonation: An experimental-phonetic approach to speech melody*, Cambridge University Press, Cambridge, 1990.

[11] Hermes, D. J.: Stylization of pitch contours, in Sudhoff, S., Lenertová, D. Meyer, R., Pappert, S., Augurzky P, Mleinek, I., Richter, N., Schließer, J. (Eds.), *Methods in Empirical Prosody Research*, p. 29–62. De Gruyter, Berlin, 2006.

[12] Nolan, F.: Intonational equivalence: An experimental evaluation of pitch scales, in *Proc. 15th ICPhS*, Vol. l, p. 771–774, 2003.

[13] Lehiste, I., Peterson, G. E.: Some basic considerations in the analysis of intonation, *Journal of the Acoustical Society of America*, 33, p. 419–425, 1961.

[14] Hanson, H. M.: Effects of obstruent consonants on fundamental frequency at vowel onset in English, *Journal of the Acoustical Society of America*, 125, p. 425–441, 2009.

[15] Volín, J.: Extrakce základní hlasové frekvence a intonační gravitace v češtině, *Naše řeč*, 92, p. 227–239, 2009.

[16] Skarnitzl, R., Volín, J.: Czech voiced labiodental continuant discrimination from basic acoustic data, in *Proc. Interspeech 2005*, 2921–2924, 2005.

[17] Miller, D. G., Nair, G., Schutte, H., Horne, R.: *VoceVista* version 3.2, 2017.

[18] Skarnitzl, R.: *Znělostní kontrast nejen v češtině*, Nakladatelství Epocha, Praha, 2011.

[19] Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, **9**(5–6), p. 453–467, 1990.

[20] Boersma, P., Weenink, D.: *Praat: doing phonetics by computer*, version 6.0.25, retrieved 12 February 2017 from `http://www.praat.org/`.

[21] Machač, P., Skarnitzl, R.: *Principles of Phonetic Segmentation*, Praha, Epocha, 2009.

[22] R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved 1 September 2018 from `https://www.R-project.org/`.

[23] Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*, New York, Springer-Verlag, 2016.