Correction of Formal Prosodic Structures in Czech Corpora Using Legendre Polynomials

Martin Matura(✉) and Markéta Jůzová

Department of Cybernetics and New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
{mate221,juzova}@kky.zcu.cz

Abstract. Naturalness is a very important aspect of speech synthesis that is necessary for a pleasant and undemanding listening and understanding of synthesized speech. However, in a unit selection, unexpected changes in F_0 caused by units transitions can lead to an inconsistent prosody. This paper proposes a two-phased classification-based method that improves the overall prosody by correcting a formal prosodic description of speech corpora. For speech data representation, the authors decided to use *Legendre polynomials*.

Keywords: Anomaly detection · One-class SVM · Multiclass SVC
Formal prosodic grammar · Prosodemes
Unit selection speech synthesis

1 Introduction

In human speech, the fundamental frequency values varies within a sentence. The F_0 contour, in general, is closely related to the position of stressed syllables and also to the phrasing of the sentence. The F_0 movements (increases/decreases), especially at the phrase-final position, have a communication function in the particular language – the mismatch in these movements can cause the misunderstanding of the sentence's meaning [15,24]. Therefore, it is evident that the correct prosodic description of speech corpora is one of the crucial issues in text-to-speech synthesis.

In general, in the unit selection method, the *join* and *target* costs are computed to ensure that the optimal sequence of units is selected. These costs control the smoothness of the concatenated neighbouring units, as well as the unit's suitability for the required position in the synthesized sentence. In our TTS *ARTIC* [11,20], besides concatenation smoothness, the symbolic prosody features, called *prosodemes* (Sect. 3, [17,18]), are used in the target cost to ensure the synthesized speech keeps the required communication function (i.e. listeners are able to distinguish declarative sentences from questions) [10]. However, due to some inaccuracies in the formal prosodic description of speech data, speech

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 387–397, 2018.
https://doi.org/10.1007/978-3-319-99579-3_41

units are sometimes used in a different context than they were pronounced by a speaker and than they belong to. This may be manifested in the synthetic speech e.g. by unnatural dynamic melody or by inappropriate stress perception.

The presented paper focuses on the symbolic prosodic labels in our speech corpora and, using powerful *Legendre polynomials* (Sect. 2), offers the two-phase algorithm for their correction. The initial experiments were carried out in [14] and showed that the description of an F_0 contour based on the Legendre polynomials is sufficient for classification-based approaches.

2 Legendre Polynomials

To describe the F_0 contours, the authors used *Legendre polynomials* [9] – contrary, e.g. to usage of Gaussian mixture models by the author of [7], or HMM models used in [5,6] for the correction of wrongly labelled formal prosodic structures in speech corpora. These polynomials are frequently encountered in physics and other technical fields.

Legendre polynomials are defined by Eq. 1,

$$L_n(x) = 2^n \sum_{k=0}^{n} x^k \binom{n}{k} \binom{\frac{n+k-1}{2}}{n}, \tag{1}$$

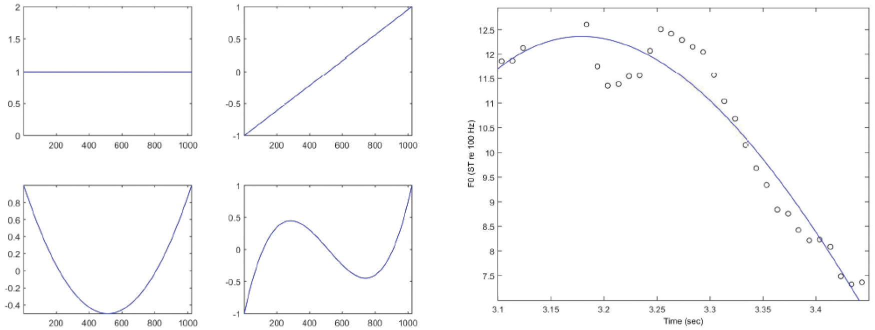
and they form an orthogonal basis (i.e., non-correlated) suitable for modelling of F_0 contours [4,23]. An F_0 contour is described by coefficients as a linear combination of these polynomials. Because of the orthogonality, the coefficients can be estimated using cross-correlation at a time lag of 0 (i.e., a mutual energy of F_0 contour and Legendre polynomial).

The first four polynomials $L_0(x)$, $L_1(x)$, $L_2(x)$ and $L_3(x)$ (see Fig. 1a) match linguistic interpretation as $L_0(x)$ responds to mean value of the pitch, $L_1(x)$ to rise or fall depending on the positive or negative sign of the coefficient (the slope is determined by its absolute value), $L_2(x)$ to peak or valley and $L_3(x)$ to the wave shape of F_0 contour.

For the purposes of the presented experiment, the authors used mPraat toolbox for Matlab [1] and for each F_0 contour, the frequency values has been transferred to semitone scale, interpolated the contour in 1,000 equidistant points and estimated the first four Legendre coefficients (for example, see Fig. 1b, coefficients are 10.7407 (*mean value*), −2.6880 (*falling slope*), −1.5522 (*valley shape*), 0.1685 (*only a slight wave curvature*)).

3 Symbolic Prosody Features in Speech Corpora

The authors of [17,18] introduced a new formal prosodic model to be used in text-to-speech systems to control the appropriate usage of intonation schemes within the synthesized sentence, the original idea was based on the Czech classical phonetic view described in [15]. This grammar parses the given text sentence



(a) The first four polynomials - mean value, slope, valley shape and curvature.

(b) Interpolation of an F_0 contour to estimate Legendre coefficients.

Fig. 1. Setup of the experiment.

in a derivation tree and each prosodic word (PW, i.e. a group of words with only one words stress) is assigned with an abstract prosodic unit, a *prosodeme*, marked as P_X. The former grammar was focused mainly on the differentiation of phrase-final and other PW in the sentence since phrase-final words are, in general, characterized by a distinct increase/decrease of an F_0, they have a certain communication function. However, as showed in [8], the phrase-initial PWs also distinguish from the following words, especially by the increase of the F_0 [24]. Recently, based on these observations, the grammar was extended to describe the phrase-initial PWs by a new prosodeme type ($P_{0.1}$, see below).

In our TTS *ARTIC* [11,20], we distinguish the following prosodeme types assigned to each PW (see also Fig. 2):

- P_1 – prosodeme terminating satisfactorily (the last PWs of declar. sentences)
- P_2 – prosodeme terminating unsatisfactorily (the last PWs of questions)
- P_3 – prosodeme non-terminating (the last PWs in intra-sentence phrases)
- P_0 – *null* prosodeme (assigned otherwise)
- $P_{0.1}$ – special type of *null* prosodeme (assigned to the first PW in phrases)

The prosodeme types are used in speech synthesis to ensure the required communication function on the phrase level of synthesized sentences [10,22] – the usage of a correct prosodeme type is controlled by the *target cost* computation in the unit selection method. Unfortunately, despite the professional speakers were recording the speech corpora for the purposes of TTS, the prosodic description of recorded sentences (based on the formal prosody grammar applied on texts of segmented sentences) sometimes do not correspond to the real F_0 contours. The problems mainly appear within the *null* prosodeme where a *"neutral"* speech is expected, but the speaker could pronounce a word in an unexpected way regarding prosody. This inaccurate description (and thus the wrong usage of some speech units in the synthesis itself) may lead to an unnatural excessive increase or decrease of the F_0 contour in a non-phrase-final prosodic word with

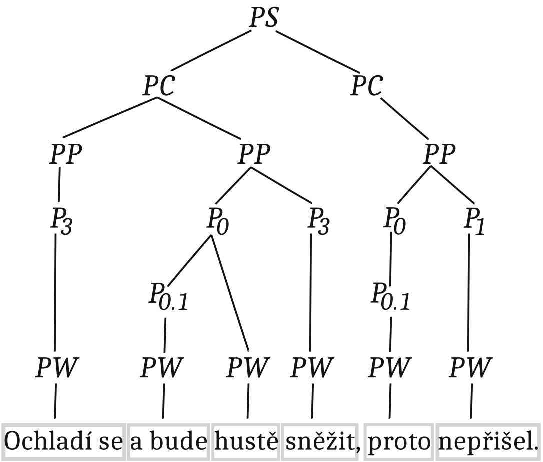


Fig. 2. The illustration of the tree built using the extended prosodic grammar [8,18] for the Czech sentence "It will get colder and it will snow heavily, so he did not come."

the *null* prosodeme which could be manifested by an inappropriate stress or an unnatural melody or, eventually, it may result in a misunderstanding due to not keeping the required communication function.

In the presented paper, the experiments are carried out on two large speech corpora – AJ and MR [12,20]. The male synthetic voice, built from AJ corpus, is widely used in commercial products for its high naturalness. On the other hand, the female synthetic voice, built from MR corpus, is not very consistent in prosody (her prosody is very dynamic) – given the original prosodic description baseline, synthesized sentences quite often contain an unnatural intonation pattern (especially in the *null* prosodeme). The complete statistics of the corpora are listed in Table 1.

Table 1. Number of prosodic words labelled by a specific prosodeme type.

Corpus	No. of sentences	No. of PWs	P_0	P_1	P_2	P_3	$P_{0.1}$
AJ	12,277	84,733	35,781	9,850	922	12,141	26,039
MR	12,308	83,486	41,728	11,017	905	7,953	21,883

4 Correction Process

The basic idea behind the correction process is simple. With inconsistent prosody, the speech created by the unit selection does not sound naturally and it is unpleasant to listen due to the speech artefacts. If we were able to correct wrongly marked prosodic words, we might achieve more fluent and consistent prosody, which would lead to a better synthesis. The correction process has two phases and a choice of a suitable data description is a principal issue. Despite prosodemes (Sect. 3) being the only symbolic prosody features, each prosodeme type corresponds to the specific changes in the F_0 contour – these could be modelled by the *Legendre polynomials* (Sect. 2) whose first four coefficients are used as the only representation of our data in the presented experiment.

In the first phase, anomalies among the *null* prosodemes are detected (Sect. 4.1). In the second phase, the detected outliers are classified by a multiclass classifier that gives them new labels (Sect. 4.2). Both phases are described below in detail. After the correction, the evaluation by listening tests was performed (see Sect. 5).

4.1 Phase One: Anomaly Detection

Anomaly (or novelty) detection [2,13] is a well-known approach which is used to find items that do not have the same or similar properties as other items in a dataset. Our previous study [14] showed that, among other classification methods, the One-class Support Vector Machine (OCSVM) is the most suitable for this experiment. We are using the implementation of OCSVM from *scikit-learn* [16] which is based on libsvm [3] with radial basis function as a kernel and $\gamma = 0.1$. The parameter ν, which influences an upper bound on the fraction of training errors, was set to 10% – this value is the authors' estimation of possible wrongly labelled PWs in corpora. Since we are looking for anomalies only in our closed dataset, we can afford to train the OCSVM model on the whole dataset to get the best decision function possible.

We trained two OCSVM models. The first one was trained by using 35,781 P_0 prosodemes from AJ corpus and the second one by using 41,728 P_0 prosodemes from MR corpus. After training the models, we tested how these models react to the different types of prosodemes and also to the training data. We detected anomalies in each group of prosodemes using the OCSVM model to obtain the number of outliers for each group. Since the model was trained with P_0 prosodemes, where we supposed 10% of anomalies, we expected the number of outliers to be about 10% for P_0 and significantly higher for the other groups. The results shown in Table 2 confirm our assumption – most of the P_1 prosodemes were correctly detected as anomalous by OCSVM model trained on P_0. All the results are described in [14].

Table 2. Number of anomalies detected by OCSVM.

Corpus	P_0	P_1
AJ	3,578 (10.0%)	8,508 (86.4%)
MR	4,174 (10.0%)	10,317 (93.6%)

4.2 Phase Two: Outliers Classification

By detecting the anomalies in P_0 prosodemes, we obtained a group of outliers whose $F0$ does not have *"neutral"* contour. These outliers can be either strongly penalized to exclude them from speech synthesis process as described in Sect. 5.1 (see [14]), or classified to another prosodeme class – as mentioned in Sect. 3, apart from P_0, we distinguish another 4 different prosodeme types: P_1, P_2, P_3 and $P_{0.1}$. To perform the multi-class classification of the P_0 outliers, we had to train an appropriate model for each corpus.

We collected all available prosodeme data from one corpus to cover all the prosodeme types and then we trained a Support Vector Classifier (SVC) from scikit-learn as our multi-class model. SVC uses *one-vs-all* decision function and since our data are not evenly distributed among all types of prosodemes, we set the parameter for class weight to *"balanced"*, which means the weight of each class is adjusted inversely proportional to the class frequencies in input data. As in the previous case, we were working on the closed dataset and therefore we could used all data to train the classification model.

The classification and relabelling of P_0 outliers was done again for both corpora. We classified 3,578 outliers in AJ corpus and 4,174 outliers in MR corpus; the classification results are listed in Table 3.

Table 3. Classification of P_0 outliers.

Corpus	P_0 outliers	P_0	P_1	P_2	P_3	$P_{0.1}$
AJ	3,578	1,559 (43.6%)	189 (5.3%)	328 (9.2%)	328 (9.2%)	1,174 (32.8%)
MR	4,174	988 (23.7%)	385 (9.2%)	145 (3.5%)	817 (19.6%)	1,839 (44.1%)

It is obvious, that most of the P_0 outliers (76.3%) from MR corpus were labelled as a different type of prosodeme. However, 23.7% of them were given the P_0 label again. These outliers were picked by the OCSVM model as anomalies, because their properties were somehow different from the other P_0 data. Nevertheless, the properties of these outliers are still more similar to the P_0 prosodeme than to another prosodeme type, hence the SVC labelled them as P_0. The situation for AJ corpus is analogous with the difference that even more outliers were labelled back to P_0. This is probably caused by a different prosody consistency of each corpus. The intonation of AJ speaker was more consistent and precise compared to the MR speaker and therefore, classifier marked them back to type P_0 more often than in the case of MR corpus. The evaluation of the prosodeme corrections will be further described in Sect. 5.2.

5 Evaluation

To evaluate the process proposed in Sect. 4, we carried out two listening tests (see Sects. 5.1 and 5.2) in our new listening test framework. Both tests had the same structure, both were 3-scale preference listening test. The listeners were comparing sentences generated by our baseline TTS system *ARTIC* (with original corpora, *TTS-base*) and those generated by a modified system *TTS-new* build on the fixed corpora (based on the classification described in Sects. 4.1, 4.2 respectively). They were instructed to use earphones and to compare the overall quality of samples A and B in each pair by selecting one from these options:

- *Sentence A sounds better.*
- *I cannot decide.*
- *Sentence B sounds better.*

The answers where normalized for each listener and pair of samples in the listening test to $p = 1$ where the *TTS-new* output was preferred, to $p = -1$ where the *TTS-base* output was preferred and $p = 0$ otherwise. These values were used for the final computation of the listening test score s, defined by Eq. 2,

$$s = \frac{\sum_{p \in T} p}{|T|}\ , \tag{2}$$

where T is a set of all answers from all listeners. The positive value of s indicates the improvement of the overall quality when using *TTS-new*.

5.1 Evaluation of the Phase One

First, the authors evaluate the *phase one*, the anomaly detection using OCSVM in Sect. 4.1, directly in the unit selection speech synthesis itself [14]. In this evaluation, the modified *TTS-new* represents a system which highly penalizes units originated from anomalous PWs (those detected by OCSVM) during the Viterbi search [21]. This “ban” should ensure that these “strange” (anomalous) units are not used in the synthesis and it may, hopefully, increase the naturalness of speech synthesis. On the other hand, about 10% of all P_0 units are dropped by this approach – this should, however, not be a big problem since the corpora are quite large and they were carefully designed [12] to cover all the different units sufficiently. In any case, this approach results in a different sequence of units compared to that generated by *TTS-base*.

To select the sentences for the listening test, we synthesized 6,000 sentences by *TTS-base* and *TTS-new* and we randomly selected 20 sentences for each voice so that they fulfilled the criterion of having 8 or more anomaly units (similarly to [19], but the selection criterion was the number of anomalous units occurrences in *TTS-base* sentences in this experiment). Thus, the whole listening test contained 40 pairs of synthesized sentences, each pair included two variants of the same sentence – one generated by *TTS-base* and one generated by the modified system *TTS-new*.

The results of the listening test, gained from 16 listeners (5 of them being speech experts), are presented in Table 4. *TTS-new* was preferred for both voice corpora, the results are statistically significant (as proved in [14]). The positive score values s indicate that the penalizing of outlier speech units (those originated from PW outliers detected by OCSVM using *Legendre polynomials* coefficients) leads to more natural synthetic speech.

Table 4. Results of the first listening test.

Corpus	*TTS-base* better	Same quality	*TTS-new* better	score s
AJ	62 (19.4%)	76 (23.7%)	182 (56.9%)	**0.375**
MR	104 (32.5%)	79 (24.7%)	137 (42.8%)	**0.103**
Total	166 (25.9%)	155 (24.2%)	319 (49.9%)	**0.239**

5.2 Evaluation of the Phase Two

The results presented in the previous section indicate the improvement of the quality of speech synthesis when penalizing the units originated from P_0 words detected as outliers by OCSVM. However, the outliers were in the *phase two* relabelled by a multi-class SVM classifier (described in Sect. 4.2) and so they could be used in the synthesis with the new label assigned. In this case, the *TTS-new* uses the same penalization of a mismatch of prosodeme types in the target cost computation as in the baseline *TTS-base*, the only difference of the two systems are the data with prosodeme labels – *TTS-new* uses the relabelled speech corpora, *TTS-base* uses the original speech corpora presented in Sect. 3.

Again, when designing sentences for the second listening test, we followed the methodology described in [19] with the selection criterion based on the number of relabelled units occurrences in *TTS-new* sentences. By this procedure, we randomly selected 10 sentences for the each non-*null* prosodeme type for both voices (80 sentences in total) to find out how the relabelled units performed in new prosodic contexts.

This listening test was finished by 16 listeners, 6 of them being speech synthesis experts. The results listed in Table 5 show that the relabelled prosodemes did not cause any serious problem in the synthesized sentences, the outputs of *TTS-new* were sometimes even much better evaluated by the listeners contrary to the *TTS-base* outputs.

6 Conclusion and Future Work

In the presented paper, we examined the usage of the Legendre polynomials for correction of formal prosody grammar. The corpora we have been working with contained inconsistencies in the prosody description – some prosodic words were labelled as *"neutral"* (P_0) in the meaning of prosody even though their $F0$ did not have a neutral contour. Therefore, we proposed the two-phased correction method to correct these wrongly labelled prosodemes. To represent our data, we took only the first four coefficients of the Legendre polynomials and then we trained One-Class Support Vector Machine (OCSVM) detector and multi-class Support Vector Classifier (SVC).

In the first phase, outliers among the P_0 prosodemes were detected by the OCSVM and then, in the second phase, we classified them with the multi-class SVC so we get the new labels for the P_0 outliers. Afterwards, we conducted

Table 5. Results of the second listening test.

Corpus	prosodeme	*TTS-base* better	Same quality	*TTS-new* better	score s
AJ	$P_{0.1}$	18 (11.3%)	111 (69.4%)	31 (19.4%)	**0.081**
	P_1	25 (15.6%)	103 (64.4%)	32 (20.0%)	**0.044**
	P_2	24 (15.0%)	46 (28.8%)	90 (56.3%)	**0.413**
	P_3	38 (23.8%)	65 (40.6%)	57 (35.6%)	**0.119**
MR	$P_{0.1}$	17 (10.6%)	124 (77.5%)	19 (11.9%)	**0.013**
	P_1	44 (27.5%)	68 (42.5%)	48 (30.0%)	**0.025**
	P_2	26 (16.3%)	72 (45.0%)	62 (38.8%)	**0.225**
	P_3	49 (30.6%)	47 (29.4%)	64 (40.0%)	**0.094**
AJ corpus - total		105 (16.4%)	325 (50.8%)	210 (32.8%)	**0.164**
MR corpus - total		136 (21.5%)	311 (48.6%)	193 (30.2%)	**0.089**
total		241 (18.8%)	636 (49.7%)	403 (31.5%)	**0.127**

two listening tests to evaluate the benefit of this approach. By the first test, we verified that the synthetic speech sounds better if we are not using the anomalous P_0 prosodemes. In the second test, we found out that if we relabel the anomalies to a different prosodeme type, we can still use them and the quality of speech will not decrease. Hence we do not need to penalize the anomalies or throw them away, which would be a waste of data. Furthermore, in some cases the synthesized speech even gets better with these relabelled prosodemes.

As a future work, we would like to test this method on our other corpora (Czech, English, Russian, etc.) and we also want to compare the quality of synthesized speech without all the anomalies and with the relabelled variants of them.

Acknowledgements. The work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506 and by the grant of the University of West Bohemia, project No. SGS-2016-039. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

1. Bořil, T., Skarnitzl, R.: Tools rPraat and mPraat. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2016. LNCS (LNAI), vol. 9924, pp. 367–374. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45510-5_42
2. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011). Software http://www.csie.ntu.edu.tw/~cjlin/libsvm

4. Grabe, E., Kochanski, G., Coleman, J.: Connecting intonation labels to mathematical descriptions of fundamental frequency. Lang. Speech **50**(Pt 3), 281–310 (2007)
5. Hanzlíček, Z.: Classification of prosodic phrases by using HMMs. In: Král, P., Matoušek, V. (eds.) TSD 2015. LNCS (LNAI), vol. 9302, pp. 497–505. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24033-6_56
6. Hanzlíček, Z.: Correction of prosodic phrases in large speech corpora. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2016. LNCS (LNAI), vol. 9924, pp. 408–417. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45510-5_47
7. Hanzlíček, Z., Grůber, M.: Initial experiments on automatic correction of prosodic annotation of large speech corpora. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS (LNAI), vol. 8655, pp. 481–488. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10816-2_58
8. Jůzová, M., Tihelka, D., Volín, J.: On the extension of the formal prosody model for TTS. In: TSD. Lecture Notes in Computer Science. Springer (2018)
9. Legendre, A.M.: Recherches sur l'attraction des sphéroïdes homogènes. In: Mémoires de mathématique et de physique, presentés à l'Académie royale des sciences, par divers sçavans & lûs dans ses assemblées, Paris, pp. 411–435 (1785)
10. Matoušek, J., Legát, M.: Is unit selection aware of audible artifacts? In: SSW 2013. Proceedings of the 8th Speech Synthesis Workshop, pp. 267–271. ISCA, Barcelona, Spain (2013)
11. Matoušek, J., Tihelka, D., Romportl, J.: Current state of Czech text-to-speech system ARTIC. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 439–446. Springer, Heidelberg (2006). https://doi.org/10.1007/11846406_55
12. Matoušek, J., Tihelka, D., Romportl, J.: Building of a speech corpus optimised for unit selection TTS synthesis. In: LREC 2008, pp. 1296–1299. ELRA, Marrakech, Morocco (2008)
13. Matoušek, J., Tihelka, D.: Anomaly-based annotation errors detection in tts corpora. In: INTERSPEECH, pp. 314–318. ISCA, Dresden, Germany (2015)
14. Matura, M., Jůzová, M.: Using anomaly detection for fine tuning of formal prosodic structures in speech synthesis. In: TSD. Lecture Notes in Computer Science, Springer (2018)
15. Palková, Z.: Rytmická, výstavba prozaického textu. Studia ČSAV; čís. 13/1974. Academia (1974)
16. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
17. Romportl, J.: Structural data-driven prosody model for TTS synthesis. In: Proceedings of the Speech Prosody 2006 Conference, pp. 549–552. TUDpress, Dresden (2006)
18. 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
19. 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
20. 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. Lecture Notes in Computer Science (2018)

21. Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi search for fast unit selection synthesis. In: INTERSPEECH, pp. 174–177. ISCA, Makuhari, Japan (2010)
22. Tihelka, D., Matoušek, J.: Unit selection and its relation to symbolic prosody: a new approach. In: INTERSPEECH, vol. 1, pp. 2042–2045. ISCA, Bonn (2006)
23. Volín, J., Tykalová, T., Bořil, T.: Stability of prosodic characteristics across age and gender groups. In: INTERSPEECH, pp. 3902–3906. ISCA, Stockholm, Sweden (2017)
24. Volín, J.: Extrakce základní hlasové frekvence a intonační gravitace v češtině. Naše řeč **92**(5), 227–239 (2009)