INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

T5G2P: Using Text-to-Text Transfer Transformer for Grapheme-to-Phoneme Conversion

Markéta Řezáčková, Jan Švec, Daniel Tihelka

New Technologies for the Information Society and Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

juzova@ntis.zcu.cz, honzas@ntis.zcu.cz, dtihelka@ntis.zcu.cz

Abstract

Despite the increasing popularity of end-to-end text-to-speech (TTS) systems, the correct grapheme-to-phoneme (G2P) module is still a crucial part of those relying on a phonetic input. In this paper, we, therefore, introduce a T5G2P model, a Text-to-Text Transfer Transformer (T5) neural network model which is able to convert an input text sentence into a phoneme sequence with a high accuracy. The evaluation of our trained T5 model is carried out on English and Czech, since there are different specific properties of G2P, including homograph disambiguation, cross-word assimilation and irregular pronunciation of loanwords. The paper also contains an analysis of a homographs issue in English and offers another approach to Czech phonetic transcription using the detection of pronunciation exceptions.

Index Terms: grapheme-to-phoneme, phonetic transcription, T5, transformers, TTS system

1. Introduction

In text-to-speech (TTS) systems, the grapheme-to-phoneme (G2P) module is a crucial part affecting the overall intelligibility and correctness of the synthesized text, as well as the process of voice building itself where the correctness of phonetic transcription affects the overall quality of the synthetic voice [1]. Although there is a common shift towards the end-to-end TTS systems, taking a raw written text as an input [2, 3, 4], even these must carry out some kind of phonetic transcription internally, hidden in the neural network structure. However, the explicit handling of G2P transcription and the possibility to analyse and evaluate its output simply by text comparison, as well as the ability of fast-fixing of inappropriate pronunciations are among the reasons why the G2P research remains highly topical.

Traditional approaches to G2P conversion usually depend on the sets of phonetic rules or pronunciation dictionaries, but mostly on a combination of both. The problem with the rule-based approach is that even in the very regular languages, from the pronunciation point of view, there are some exceptions in phonetic transcription, especially for words originating from foreign languages (loanwords). On the other hand, the dictionary-based approaches suffer from the incompleteness of the dictionary used, which, despite frequent updating, can never contain all the words of a given language – not mentioning huge dictionaries for inflectional languages (containing all possible word forms), or additional automatic morphological analysis of words required in the case of lemma pronunciation dictionaries (possibly introducing other errors into the system). The dictionary+rules combination can usually ensure a sufficiently accurate conversion of the input text into its phonetic form, both in the case of primary use of a dictionary supplemented by post-processing rules (e.g. for distinguishing pronunciation for "`the`" in English based on the following phoneme), or vice versa the use of rules for languages with regular pronunciation (e.g. Slavic languages) and dictionaries look-up only for the conversion of some exceptions to the rules.

A big challenge for languages with more or less irregular pronunciation (with e.g. English spelling being notoriously irregular) is an automatic G2P conversion of hitherto unknown words. While automatic converters often cannot predict correct pronunciation, that is rarely a problem for native speakers of a given language thanks to the human ability to generalize things already learned. This thought was at the birth of G2P approaches using different types of neural networks [5, 6], including the state-of-the-art approaches based on transformer-based networks with an attention mechanism [7, 8]. As showed in the studies mentioned before, the neural networks model can generalize well and predict the correct pronunciation of unseen words. Such feature is usable for the use in dialog systems where the synthesised entities are not known in advance [9].

A large number of approaches work only at the word level, trained from large dictionaries [5, 6, 7, 8]. Such an approach, however, still requires some level of post-processing of the sentence/phrase phonetic transcription, including the consideration of relationships between the words, homographs disambiguation etc. For this reason, it seems better to train G2P models directly on whole sentences or phrases, and let the model itself to learn the inter-sentence word relations and cross-words specifics of the particular language [10, 11]. The training the model for the phonetic transcription of input text on whole sentences actually corresponds to the task of machine translation [12, 13, 7, 14]: the input sequence of words in one language (orthographic words in our case) is converted to the output sequence of words in an another language (here, the sequence of phonetic word forms).

2. Traditional G2P

In the present paper we will focus on two languages – English and Czech. The reason of their choice was the different nature (irregular vs. regular) of how the pronunciations of words from their orthographic forms are created, and thus the different way of dealing with the problem of phonetic transcription in TTS systems.

Taking the *ARTIC* TTS system as an example [15], the G2P module for English is based on a large pronunciation dictionary containing about 300,000 words. That dictionary is accompanied by more than 1,000 automatically derived rules [16] used as a fallback to handle out-of-vocabulary words. On the contrary, for Czech, an inflected language with rather regular phonetic transcription, there is a set of approximately 100

http://dx.doi.org/10.21437/Interspeech.2021-546

context-dependent rules designed manually by phonetic experts [17] accompanied by a dictionary with the correct transcription of more than 170,000 irregular word-forms (words in their inflected forms) in which the input is searched prior the use of rules.

In an effort to avoid the manual maintenance of the relatively large dictionaries for all the different languages supported by [15], we have started to search for an universal and flexible approach to G2P. The first attempt, based on sequence-to-sequence DNN-based approach inspired by machine translation [10] was able to reach similar or even higher accuracy than the traditional baseline approach for the words of the most common lengths. However, the accuracy started to drop in case of words longer than 10 characters – sometimes the model was not able to "remember" the whole word and the failures started to appear towards the end of such words, as shown here[1] :

- suspiciously (EN) `s@spIS@sl**ll**`
  (correct: `s@spIS@sl**I**`)
- autobiography (EN) `O:t@Ub**I**gr**Af**fI`
  (correct: `O:t@Ub**aI**Qgr**@**fI`)
- přibývajícími (CZ) `pQibi:vaji:ci:m**i:**`
  (correct: `pQibi:vaji:ci:m**i**`)
- organizátorek (CZ) `!organiza:t**a:t**ek`
  correct:`!organiza:t**or**ek`

Despite increasing encoder/decoder capacity, the model did not yield better results and the accuracy on longer words was still significantly lower, as thoroughly shown in histogram in [11] and Figure 1. To mitigate these failures, we draw our attention to Transformer-based model.

3. T5-based G2P

To examine the ability of transformer-based DNN to deal with the problematic cases mentioned before, we used the pre-trained Text-to-Text Transfer Transformer (T5) model [19]. In general, the T5 model is trained as the full encoder-decoder transformer in a semi-supervised manner from a large text corpus. The input sentence is perturbed and the goal of the model is to generate the output which corrects the perturbed input into an original one. More specifically for T5, the random sub-sequences of tokens in the input are masked with a single token and the model predicts the tokens hidden behind the masked token. The general advantage of the T5 model is the ability to perform many text-to-text tasks like text summarization, topic detection or sentiment analysis.

3.1. T5-model pre-training

For experiments with English data, we used the Google's T5-base English model[2] trained from Common Crawl data[3]. We replicated the same pre-processing procedure to obtain the Czech data and we pre-trained our own T5 models for these languages. The pre-processing steps correspond with the steps presented in [19] for building the Colossal Clean Crawled Corpus (C4) on which the Google's T5 model was pre-trained. Such rules are generally applied while processing web text [20]:

- Only lines ending in a terminal punctuation are retained. Short pages and lines are discarded.
- Pages with dirty and obscene words are removed.
- Lines with the word "JavaScript" and the curly braces { } are removed (remains of incorrect crawling of the webpage).
- The pages in the corpus were de-duplicated. The resulting corpus contains each three-sentence span just once.

For the Czech language, we have collected the CommonCrawl corpus at the end of August 2020. It contains 47.9GB of clean text, 20.5M unique URLs and 6.7B running words. The Czech T5 training procedure followed the original procedure described in [19].

For both the English and Czech experiments, we used the `t5-base` architecture consisting of 220M parameters, 2×12 transformer block in the encoder and the decoder. The dimensionality of hidden layers and embeddings was 768. The attention mechanism uses 12 attention heads with inner dimensionality 64.

3.2. T5-model fine-tuning

For training the T5-based model for grapheme-to-phoneme task (T5G2P) we used the Tensorflow implementation of HuggingFace Transformers library [21] together with the T5s[4] library, which simplifies mainly the training and prediction process by accepting a simple tab-separated input-output pairs. The T5s library also uses the variable sequence length and variable batch size to fully utilize the underlying GPU used for training.

In the experiments, we used the ADAM optimization with learning rate decay proportional to inverse square root of the number of learning epochs.

For both languages, we have a large amount of proprietary sentences with their phonetic transcriptions at our disposal; specifically, we have 128,532 English and 442,029 Czech unique sentences. For both languages, the data were randomly split into *train*, *valid* and *test* (80%, 10% and 10%). For the individual language-related models, whole sentences were used as an input and the model was trained (fine-tuned) to generate the corresponding phonetic transcription of them. This training lasted 50 epochs, with 2000 steps per each.

4. Experiments and Results

The fine-tuned T5G2P model described in Section 3.2 was used to predict the transcription of the *test* data (the selected 10% of the fine-tuning dataset). The output, i.e. the predicted sequence of phones, was compared to the reference using the following measures: *sentence accuracy*, *word accuracy* and *phoneme accuracy*. The word accuracy clearly reflects the correctness of a transcription similarly to word-error-rate in ASR evaluation, as we can suppose invalid phonetic form as an invalid or unintelligible word. On the other hand, the phoneme accuracy does not tell much about intelligibility (there may be few misses in many words or many misses in few words), but it is used in other studies.As for the sentence accuracy, it gives us a higher-level overview of the failures and can be related to the presence of uncommon words in the sentences.

The overall results for English and Czech are shown in Table 1. The table compares the T5-based approach to the both baseline approaches, first based on pronunciation dictionaries and a sets of rules (see Section 2), and second the encoder-decoder DNN-based G2P converter presented in [10, 11]. For the Czech, there is also an extra row showing the accuracy used

[1] All the transcriptions are in SAMPA alphabet [18].

[2] `https://github.com/google-research/text-to-text-transfer-transformer`

[3] `https://commoncrawl.org/`

[4] `https://github.com/honzas83/t5s`

Table 1: *Results of the tested English and Czech G2P outputs compared to the reference transcription.*

	approach	A_sent	A_word	A_phoneme
English	dict + rules	54.49 %	90.93 %	97.20 %
	encoder-decoder	82.75 %	95.72 %	97.18 %
	T5	91.84 %	99.04 %	99.68 %
Czech	only rules	56.74 %	90.97 %	99.36 %
	dict + rules	98.86 %	99.99 %	99.99 %
	encoder-decoder	88.64 %	98.69 %	99.51 %
	T5	98.77 %	99.89 %	99.97 %

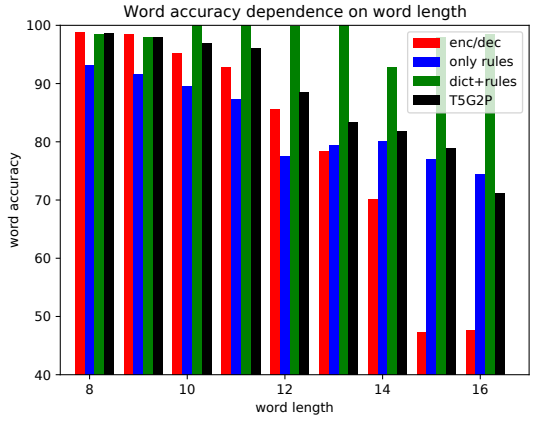

Figure 1: *Word accuracy comparison for longer words in Czech.*

when the rules are used only, i.e. without the dictionary of exceptions. Let us note that the numbers differ slightly from those presented in the papers mentioned (here the accuracy of baselines is slightly higher), as there were some error corrections made and the dictionaries were extended over the past two years.

The results clearly show the advantages of the proposed approach. In general, it outperforms the basic encoder-decoder model for both languages in all three evaluation metrics. The detailed analysis of the outputs showed that the new proposed approach makes significantly fewer mistakes in long words, which was the main problem of the previous encoder-decoder model [11], as demonstrated in Figure 1. (Also, all the examples from section Section 2 were transcribed correctly by our T5G2P model.)

For English, the T5 model is able to beat both of the baselines – the explanation for that lies in a large number of homographs in English which the T5G2P is able to deal with (contrary to the dictionary+rules approach) as showed later in Section 4.2, and also in more accurate transcription of longer words (contrary to the encoder-decoder DNN model). For Czech, the accuracy on word and phone level is very close to the baseline dictionary+rules combination. Let us note, however, that the baseline has a significant advantage since the rules are well tuned and the dictionary has been extended for many years since [22].

As mentioned in the Section 1, most G2P studies are developed and evaluated on dictionaries, e.g. NetTalk or CMUdict[5] – in recent years, the published results range between 20% and 30% word error rate, which means the accuracy 70-80% (e.g. [8]). Our *word accuracy* value on English set is much higher, however, we intentionally work on a sentence level to capture cross-word assimilation [23, 24]. Therefore, although our test data contained only unseen sentences, many words occurred in both training and testing data. More precisely, the English training data contained almost 130 thousand of unique words, the testing data contained less then 17 thousand and only 5% of them were unseen. To compare our results on word level to others, we extracted the unseen words only, for which the error rate was 33.8%. This may seem to be relatively high value, but let's keep in mind that the unseen words are also rather rare in our case (unseen in the whole text set).

The same evaluation for the Czech language, with quite regular pronunciation, showed the error rate 2.3% for more than 10,000 unseen words from all 78 thousand words in testing data (the training data contained more than 350 thousand unique words).

4.1. Loanwords detection in Czech

As written in Section 2, the Czech G2P transcription can be very reliably handled by the relatively small set of rules, except for the loanwords which are not that rare in Czech [25]. Thus, all the errors for *dict + rules* system in Table 1 are caused by such words, not included in the dictionary. Therefore, we explored an idea to use a trained T5 model to automatically detect the words to be added to the exceptions dictionary, or to convert such words to the form which can then be correctly transcribed by the rules (all loanwords can be transcribed to such a form). Although this solution would not be as elegant and universal as the T5G2P transcriber, it would avoid some of the errors caused by T5 on cases handled correctly by the rules.

To explore this possibility, we have trained another T5 model, the task of which is to detect exceptions from the rule-based transcriptions. The same split of fine-tuning Czech data as described in Section 3.2 has been used, but the phonetic output was modified in such a way that the words transcribed by the rules were replaced by "0" and the words from the exceptions dictionary were marked by "1". To decide about the exception, a dictionary with more than 170,000 loanwords from the baseline system, described in Section 2, has been used.

Similarly to the T5-model tuning described in Section 3.2, the training lasts 50 epochs, but with 1000 steps per epochs only. The ability of the model to predict these exceptions from regular pronunciation was evaluated on the set of *test* data, which contained about 509 thousand words, 3.3% of which were the exceptions from the dictionary. Having approx. 44 thousand test sentences, a loanword appears in every second/third sentence.

The results of loanwords detection are presented in Table 2 – as the task is actually a classification into two unbalanced classes, in addition to the *accuracy* we also show *precision*, *recall* and *F1-score*. High P, R and F1 values indicate that the T5-model has learned to recognize exception-marked words very well. Regarding the errors, there were only 47 false negatives (missed exceptions) and 84 false positives (extra detected exceptions) in all the testing sentences.

Naturally, we are aware that the dictionary does not contain all the possible loanwords and/or their word forms. Thus, some of the words might be marked by "0" even when they should be marked by "1". To find whether the trained T5 model was able to detect such words, we have analysed the false positive cases,

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Table 2: *Evaluation of the T5 model for loanwords detection on testing data*

Accuracy	Precision	Recall	F1-score
99.97 %	99.51 %	99.72 %	99.62 %

Table 3: *Results of the tested Czech G2P outputs using loanwords detection.*

approach	A_sent	A_word	A_phoneme
LW detection + rules	98.31 %	99.71 %	99.94 %
LW detection + T5	98.93 %	99.90 %	99.98 %

since these are expected to be cases recognised by T5 model but not included in the dictionary. As an example, the T5 model marked the following words which are not directly contained in the dictionary, but derived from items being there:

- `neoklasicismu` (in dictionary: `klasicismus`)
- `meziresortními` (in dictionary: `meziresortní`)

Having a model trained to detect words with irregular pronunciation, we tested another approach for phonetic transcription of Czech. Contrary to T5 in Table 1, the main idea was to detect and highlight loanwords which can be then added to the dictionary by a user. The whole phrase has, therefore, been transcribed to phones either by the rules (without dictionary) or by the T5G2P model described in Section 3, and the words detected by the T5 detector were excluded from the comparison, simulating the case of their manual check/transcription/correction by a user. On average, there were about four such words excluded in every 10 sentences, which may seem to be a laborious work to check them all, but the majority of such words are already in the dictionary.

The results of the modified G2P transcription are shown in Table 3. The combination of the loanwords (LW) detection and the set of rules led to a significant improvement over the rules-only approach in Table 1. The results came, naturally, very close to the values for dict+rules approach in Table 1, since all the loanwords are now treated as transcribed correctly. The table also shows a slight improvement for the T5 model if the LW T5-based detector is used, since the mistakes of the T5G2P in these words are not affecting the results.

4.2. T5 and homographs in English

There are several homographs in English among the words used in ordinary communication, e.g. *live*, *read*, *record*. The pronunciation of these can be inferred from the context or additional meta-information, e.g. some part-of-speech (POS) analysis [26, 27]. Although the presented T5G2P model (Section 3) does not explicitly use any of such information, it is able to distinguish between the two pronunciation variants quite reliably, as shown in Table 4.

To demonstrate it, we selected the above-mentioned homographs and evaluated the word accuracy measure for all of their transcriptions in the testing data. It can be seen from Table 4 that the ability of the model to decide between the pronunciation variants of the homographs is very high.

Looking at the failures in "read" transcription, they occurred in the following phrases:
I am furious that they read *her letter, it is like being in prison.*

Table 4: *Results of phonetic prediction for representative English homographs*

homograph	variant	errors	total	Accuracy
live	`[laIv]`	0	19	100 %
	`[lIv]`	0	48	100 %
read	`[rEd]`	2	20	90 %
	`[ri:d]`	0	22	100 %
record	`['rEkOd]`	0	19	100 %
	`[ri'kOd]`	0	6	100 %

– reference: [rEd], predicted: [ri:d]

You read *the worst kinds of things about them.*
– reference: [rEd], predicted: [ri:d]

The analysis of the fine-tuning texts used for the training showed that there is $6\times$ *you read*[rEd] and $11\times$ *you read*[ri:d], $1\times$ *they read*[rEd] and $1\times$ *they read*[ri:d]. Thus, we tend to think that the model learned the more frequent context, since it is unable to get a meaning or understanding of the phrase. Further investigation is, of course, needed.

5. Conclusions

In the present paper, we focused on the use of T5G2P, the Text-to-Text Transfer Transformer (T5) model in the task of grapheme-to-phoneme (G2P) transcription, which is a natural choice due to the similarity of G2P and machine translation tasks. The T5 model has been trained on Common Crawl data and then fine-tuned on a large set of texts with phonetic transcription. For English, it significantly outperforms both the baseline traditional dictionary+rules and more modern encoder-decoder approaches, and it is also capable to deal with homographs surprisingly well. For Czech, it outperforms the encoder-decoder approach and is very close to the very well tuned dictionary+rules G2P module. Let us emphasize that all the G2P transcriptions work on the sentence level, and thus the cross-word assimilation is taken into account in the results.

Moreover, the same data were used to fine-tune T5 model detecting words with irregular pronunciation in Czech. It has also been shown that a "hybrid" approach, where the irregular words are first marked for a manual inspection, and the rest of the text is transcribed "as usual", can iteratively improve the precision of the G2P transcription with high reliability and thus without too much labour put into the extension of exceptions dictionary used to handle loanwords with irregular pronunciation.

For the future work, we also aim to verify the model on other languages, such as Russian, German or Spanish.

6. Acknowledgements

This research was supported by the Czech Science Foundation (GA CR), project No. GA19-19324S, and by the grant of the University of West Bohemia, project No. SGS-2019-027.

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructure.

7. References

[1] J. Matoušek and D. Tihelka, "Annotation errors detection in TTS corpora," in *Proceedings of INTERSPEECH 2013*, Lyon, France, 2013, pp. 1511–1515. [Online]. Available: http://www.kky.zcu.cz/en/publications/MatousekJ_2013_AnnotationErrors

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[3] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *ICLR*, 2017.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2021.

[5] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4225–4229, 2015.

[6] K. Yao and G. Zweig, "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion," *CoRR*, vol. abs/1506.00196, 2015.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: arXiv:1706.03762

[8] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, "Transformer Based Grapheme-to-Phoneme Conversion," in *Proc. Interspeech 2019*, 2019, pp. 2095–2099.

[9] J. Švec and L. Šmídl, "Prototype of czech spoken dialog system with mixed initiative for railway information service," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer Berlin Heidelberg, 2010, vol. 6231, pp. 568–575.

[10] M. Jůzová, D. Tihelka, and J. Vít, "Unified language-independent DNN-based G2P converter," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2085–2089.

[11] M. Jůzová and J. Vít, "Using auto-encoder biLSTM neural network for Czech grapheme-to-phoneme conversion," in *Text, Speech, and Dialogue - 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11-13, 2019, Proceedings*, ser. Lecture Notes in Computer Science, K. Ekstein, Ed., vol. 11697. Springer, 2019, pp. 91–102.

[12] K. Cho, B. van Merrienboer, C. Gülcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation." in *EMNLP*, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1724–1734.

[13] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[14] X. Liu, K. Duh, L. Liu, and J. Gao, "Very deep transformers for neural machine translation," 2020. [Online]. Available: https://arxiv.org/abs/2008.07772

[15] D. Tihelka, Z. Hanzlíček, M. Jůzová, J. Vít, J. Matoušek, and M. Grůber, "Current state of text-to-speech system ARTIC: A decade of research on the field of speech technologies," in *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2018, vol. 11107, pp. 369–378.

[16] J. Zelinka and L. Müller, "Automatic general letter-to-sound rules generation for german text-to-speech system," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopeček, and K. Pala, Eds., vol. 3206. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 537–543.

[17] J. Matoušek, D. Tihelka, and J. Psutka, "New Slovak unit-selection speech synthesis in ARTIC TTS system," in *Proceedings of the World Congress on Engineering and Computer Science 2011*, San Francisco, USA, 2011, pp. 485–490.

[18] J. C. Wells, "SAMPA computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore, and R. Winski, Eds. Berlin and New York: Mouton de Gruyter, 1997.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2020. [Online]. Available: arXiv:1910.10683

[20] J. Švec, J. Lehečka, P. Ircing, L. Skorkovská, A. Pražák, J. Vavruška, P. Stanislav, and J. Hoidekr, "General framework for mining, processing and storing large amounts of electronic texts for language modeling purposes," *Language Resources and Evaluation*, vol. 48, no. 2, pp. 227–248, 2014.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[22] J. Matoušek, "Building a new Czech text-to-speech system using triphone-based speech units," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopeček, and K. Pala, Eds. Berlin–Heidelberg, Germany: Springer, 2000, vol. 1902, pp. 223–228.

[23] P. Roach, *English Phonetics and Phonology: A Practical Course*. Cambridge University Press, 1983.

[24] P. Machač and R. Skarnitzl, *Principles of Phonetic Segmentation*. Epocha, 2013.

[25] J. Lehečka and J. Švec, "Improving speech recognition by detecting foreign inclusions and generating pronunciations," in *Text, Speech, and Dialogue*, I. Habernal and V. Matoušek, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 295–302.

[26] S. H. Burton, "The parts of speech: A chapter for reference," in *Mastering English Grammar*. London: Palgrave Macmillan UK, 1984, pp. 115–141.

[27] S. Bird, E. Klein, and E. Loper, "Categorizing and tagging words," in *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009, pp. 179–220.