INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Accentor: An Explicit Lexical Stress Model for TTS Systems

*Diana Geneva*[4,*], *Georgi Shopov*[1], *Kostadin Garov*[5,*], *Maria Todorova*[2], *Stefan Gerdjikov*[1,3], *Stoyan Mihov*[1]

[1]IICT – Bulgarian Academy of Sciences, [2]IBL – Bulgarian Academy of Sciences, [3]FMI – Sofia University, Bulgaria, [4]Chaos, [5]INSAIT

diana.geneva@chaos.com, gshopov@lml.bas.bg, kostadin.garov@insait.ai, maria@dcl.bas.bg, stefangerdzhikov@fmi.uni-sofia.bg, stoyan@lml.bas.bg

## Abstract

The accurate placement of word stress is a critical component of the correct pronunciation of words. Contemporary publicly available text-to-speech (TTS) datasets have a relatively narrow coverage of unique words, which causes modern neural TTS systems to synthesize speech that often suffers from lexical stress errors. In this work, we propose an efficient approach for explicitly modeling lexical stress knowledge with a dedicated Accentor neural network. The Accentor is trained separately on a large lexically diverse stress-annotated text corpus that is automatically compiled using an automatic speech recognition system. We demonstrate that the Accentor can be combined with a TTS acoustic model to reliably control the word stress encoded in the generated acoustic features. Experiments show that our approach increases the stress prediction accuracy by a factor of 12 in comparison to other modern TTS systems and improves the naturalness and comprehensibility of the synthesized speech.

**Index Terms**: lexical stress modeling, lexical stress corpus, controllable acoustic model, text-to-speech

## 1. Introduction

Knowledge of the proper pronunciation of words and phrases is very essential for any state-of-the-art TTS system. It is crucial especially for languages with orthographic ambiguities caused by complex letter-to-sound relations and homographic word forms. A critical component of the correct pronunciation of words is the accurate placement of lexical stress.

### 1.1. Unpredictability and importance of lexical stress

Lexical stress is a prominent feature of many spoken languages. The incorrect placement of stress can render speech almost incomprehensible in languages such as English, Russian, Bulgarian, etc., where word stress is phonemic, i.e. many word forms are distinguished from one another only by stress position. Furthermore, standard written languages do not typically mark word stress. [1]

In the above-mentioned languages, variations in the lexical stress position often cause ambiguities with respect to the meaning of the word form. For example, *cònstruct* and *constrùct* in English are homographic word forms of different lexemes sharing the same root. Another type of ambiguity is caused by homographic word forms with different roots (e.g *dèsert*, *desèrt* in English; *dorògа* (road), *dorogà* (precious) in Russian; *čèta* (armed group), *četà* (read) in Bulgarian), which in some cases share identical morphosyntactic values (e.g. *zàmok* (castle), *zamòk* (lock) in Russian; *v`ălna* (wool), *vălnà* (wave) in Bulgarian). A very common example of ambiguity in Bulgarian and Russian, which is resolved by the stress position, is the homography of word forms of the same lexeme (e.g. *tèla* (body), *telà* (bodies) in Russian; *govòri* (speak), *govorì* (speak, imperative) in Bulgarian). The use of functional words, which are generally unstressed, causes another type of lexical stress irregularity because in certrain contexts prepositions, conjunctions, articles and auxiliary verbs require the placement of lexical stress.

Table 1: *Frequencies of functional words and words with ambiguous stress pattern in the Bulgarian language.*

| Frequency | Ambiguous | Functional |
|---|---|---|
| Word-level | 4.43% | 31.77% |
| Utterance-level | 34.37% | 87.48% |

It is easily noticeable that these situations cannot be resolved at the character level within a word; only detailed morphosyntactic analysis and additional contextual and/or discourse information could resolve these ambiguities and offer the correct stress predictions. What's more, Russian and Bulgarian are highly inflected languages and the aforementioned ambiguities occur very often in an unrestricted text. Table 1 demonstrates that for Bulgarian one in every 23 words has an ambiguous stress pattern and every third utterance contains at least one lexical stress ambiguity. [2] Even in situations where misunderstandings cannot occur, an improper placement of lexical stress increases the cognitive load on the listener. Thus, the intelligibility, comprehensibility and usability of a TTS system is heavily reliant on its ability to correctly predict lexical stress.

Text-to-speech is one of the fundamental inclusive technologies for the visually impaired people – it empowers them to pursue many education and career paths by enabling them to consume and create written content. Therefore, the accuracy of the lexical stress prediction is crucial for the blind TTS users, since it strongly affects the intelligibility and comprehensibility of the synthesized speech.

### 1.2. Lexical stress in end-to-end TTS acoustic models

Recently, monolithic neural TTS models have been proposed that are trained end-to-end, like the Tacotron 2 [1] and FastSpeech 2 [2] acoustic models, which jointly learn the traditional front-end steps (linguistic analysis) and back-end processing

*Work done while at IICT – Bulgarian Academy of Sciences.

[1]Throughout this article Cyrillic is transliterated using the scientific transliteration scheme. An additional grave accent above vowels is used to indicate stress position.

[2]The statistics are derived from the 70M words aligned data described in Section 3.



10.21437/Interspeech.2023-433

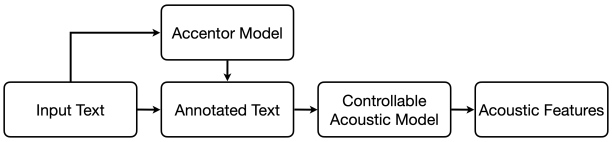


Figure 1: *The composition of the Accentor and the controllable acoustic model during inference.*

(acoustic features generation) simultaneously. End-to-end models must learn to generalize from character input to acoustic output from sets of parallel text and speech audio data, and in this way to implicitly learn pronunciation and lexical stress knowledge. Popular publicly available datasets such as the single-speaker LJ Speech [3] and multi-speaker VCTK [4] and LibriTTS [5] contain respectively 24, 44 and 585 hours of speech. Despite their lengths in terms of total hours of speech, these datasets have relatively narrow coverage of unique words. In [6] it is shown that the diversity and balance of lexical coverage of those datasets is significantly lower than that of a standard lexicon. The experiments in [6] suggest that limited and unbalanced lexical coverage in end-to-end training data may hinder the accuracy of learned pronunciation and lexical stress knowledge. The result of these experiments is supported by the error analysis of the Tacotron 2 model in [1] where the authors note that out of 100 utterances 23 contained prosody errors including incorrect lexical stress placement.

### 1.3. Improving lexical stress accuracy in TTS

Training modern end-to-end acoustic models on lexically diverse datasets consisting of thousands of hours of speech and transcriptions of tens of millions of words is extremely computationally expensive because of the high resolution of the output acoustic features (typically representing 10 ms frames). Moreover, our experiments show that end-to-end acoustic models tend to produce unclear and blurred stress characteristics in their output features (see Section 6 and the provided audio samples).

Therefore, we make the task computationally more tractable and the stress pronunciation clearer by separating the prediction of lexical stress (which needs a large lexically diverse annotated text corpus) from the prediction of the acoustic features (which can be achieved with a TTS dataset of standard size). Thus, our approach consists of the following subtasks:

- compilation of a large lexically diverse text corpus automatically annotated with lexical stress marks,
- training of an Accentor model on the compiled corpus that aims to resolve ambiguities and correctly place lexical stresses in text; this model is used during inference to annotate the input text with stress marks before it is passed to the acoustic model (see Figure 1),
- training of a controllable acoustic model on regular sized TTS dataset with transcriptions augmented with lexical stress information; the aim is to make the stress encoded in the output acoustic features controllable by the stress marks in the input text sequence.

In what follows, we demonstrate the effectiveness of our approach for Bulgarian TTS. The developed TTS system was commissioned by the Union of the Blind in Bulgaria and is currently widely used by the visually impaired people in the country. Nevertheless, it should be noted that the proposed method is not dependent on any specific characteristic or peculiarity of the Bulgarian language.

## 2. Related work

Our approach relies on ideas and techniques from two different but related tasks on lexical stress classification.

*Lexical stress detection* consists of the identification of lexical stress in a speech recording given its corresponding transcription. A common approach for this task is to use an acoustic model to align the audio with its transcription, use the alignment to extract acoustic features (F0, energy and duration) for each syllable and then classify the syllables as stressed or unstressed based on their acoustic features and textual representations [7]. In order to avoid the manual annotation of the training text, Ramanathi et al. [8] propose to train an automatic speech recognition (ASR) system with a pronunciation lexicon containing all canonical stress markings. For prediction of the stress labels they perform decoding of the utterance permitting all possible placements of lexical stress in the transcription and choose the acoustically most probable one.

*Lexical stress prediction* is the task of assigning lexical stress in a text using only its spelling. The common approaches using dictionaries and hand-crafted rules [9, 10], statistical methods [11, 12, 13, 14], and deep neural networks [15, 16] have predominantly been focused on the prediction of the stress patterns of separate words without taking their contexts into consideration. As we have discused in Subsection 1.1, those approaches cannot be used to solve the homographic ambiguities that are present in many languages. An attempt to resolve those ambiguities has been made in [17] using limited context (the surrounding four tokens) and part-of-speech tags.

## 3. Lexical stress corpus

Text corpora of adequate lexical diversity and size annotated with stress information are a very scarce resource for any language. To our knowledge there is no such corpus for Bulgarian that is appropriate for the effective training of deep neural models. Thus, we automatically compile a large text corpus annotated with lexical stress marks by employing the method described in [8] for lexical stress detection.

### 3.1. Data sources

The website of the Bulgarian Parliament [3] provides video files for all plenary sessions from 2010 up to now along with their manual transcriptions. We downloaded the available data from year 2010 to year 2018. Additionally, we used professional audiobook recordings provided to us by the Union of the Blind in Bulgaria as part of a joint project. The audiobook texts were separately retrieved from the Chitanka [4] online library and only the audiobooks that could be successfully paired with their corresponding texts were used. The acquired data consists of 11.7K hours of speech and texts with a total of 87M words (tokens) and 435K unique words (types).

### 3.2. Automatic alignment

The collected raw data contains many discrepancies between the audio and its corresponding text, which necessitates preliminary processing and alignment. Some of the specifics of the dataset are as follows: unintended repetitions, grammatical errors or lexical mistakes that are present in the audio are corrected in the transcriptions; on some occasions the texts are modified by changing the word order and inserting or deleting

[3] https://www.parliament.bg
[4] https://chitanka.info

Table 2: *Lexical stress corpus statistics.*

| Source | Utterances | Audio | Tokens | Types |
|---|---|---|---|---|
| Parliament | 1.9M | 2695h | 22M | 121K |
| Audiobooks | 5M | 5275h | 48M | 254K |
| Total | 6.9M | 7970h | 70M | 278K |
| + Dictionary | **7.5M** | **7970h** | **71M** | **829K** |

words to increase clarity; the texts contain additional annotations and notes that are not read by the speaker; the texts are not verbalized – digits are used to express number, date, time and currency information, common abbreviations occur as well.

To overcome the mentioned discrepancies we use a hybrid DNN-HMM ASR system and the methodology developed in [18, 19] to decode the audio and simultaneously verbalize and align the transcriptions with the recognized text. The lexicon used during decoding consists of 287K word types and covers 99% of the word tokens in the dataset. To deal with non-lexical units such as numbers, dates, times, metric units and abbreviations, the alignment of the recognized texts with the transcription texts is done using a phoneme-level Levenshtein distance that takes into account all possible verbalizations of the tokens in the transcribed text [19]. From the computed alignments we extract entire sentences or parts of sentences where the audio completely matches with its transcription. The first three rows of Table 2 summarize the statistics of the aligned data.

### 3.3. Automatic annotation

In order to annotate the aligned data with stress information, we need a specialized ASR acoustic model which can differentiate between stressed and unstressed vowels. To this end, we use a phonetic system with separate phonemes for stressed and unstressed vowels. Also, we add to the lexicon all of the linguistically correct (canonical) phonetizations and stress patterns for each word form. We use the Kaldi ASR Toolkit [20] and the LibriSpeech [21] recipe to train a time delay deep neural network [22] acoustic model on the BG-PARLAMA corpus [18] extended with some of the aligned audiobook recordings. The resulting training dataset consists of 749 hours of speech.

With the trained ASR acoustic model we decode the aligned dataset to classify the vowels in each utterance as stressed or unstressed. For each utterance we compile a separate decoding graph that encodes all possible stress patterns of the words in its corresponding sentence. Figure 2 depicts the decoding graph for the sentence *"ne poveli da pravi beli"* (optional silence between words is intentionally omitted.). All of the words in this sentence have two canonical variants (e.g. *povèli* and *poveli*) which differ only in the position of the lexical stress. Thus, the responsibility of the model is to choose the most likely path or in other words the most likely placement of lexical stress in the sentence based on the pronunciation in the audio.

### 3.4. Extension with dictionary

As shown in Table 2, the aligned data contains 278K word types. A standard Bulgarian dictionary consists of over 1M word types, more than 90% from which have an unambiguous stress pattern. Thus, in order to extend the coverage of the corpus, we add 551K word types with unambiguous stress pattern, which are not present in the aligned data. Each word type is added as a separate utterance. The last row of Table 2 shows the statistics of the extended corpus and demonstrates its superior lexical diversity compared to a standard Bulgarian TTS corpus which contains only around 32K word types (see Section 5).

## 4. Accentor model

The Accentor model performs the task of stress placement in a text. As we have already discussed, there are many cases in which the correct placement of lexical stress is highly dependent on the use of contextual information. Therefore, we opt for a model that has the opportunity to attend to the whole input sequence when classifying a certain vowel as stressed or unstressed. This is achieved via the utilization of FFT blocks [23] – a feed-forward structure based on multi-head self-attention and 1D convolutions, which have been successfully applied in the recently proposed FastSpeech 2 [2] acoustic model to obtain an intermediate representation of the input text from which per symbol features such as duration, pitch and energy can be efficiently predicted. Those acoustic features are regarded as important measurable characteristics for lexical stress detection – vowels in stressed syllables tend to be longer, louder, and higher in fundamental frequency than vowels in unstressed syllables.

The architecture of the Accentor model consists of an embedding layer, 6 FFT blocks and a predictor. The size of the input vocabulary is 41 – it contains the 30 letters of the Bulgarian alphabet, 10 punctuation symbols and a word boundary symbol. The dimension of the embeddings and the hidden size of the self-attention in the FFT blocks are set to 256. The number of attention heads is 4, each of dimension 128. The kernel sizes of the 1D convolutions in the 2-layer convolutional network are set to 9 and 1, with input/output size of 256/1024 for the first layer and 1024/256 for the second layer. The output of the last FFT block goes through the predictor, which consists of two 1D convolutions with kernel size 3 and 256 output channels. The result from the predictor is projected to a scalar – the logit of the probability of the corresponding input vowel being stressed. As a loss function we use the binary cross entropy between the predicted logits of the vowels and the binary characteristic vector of the stressed vowels. To reduce the size of the resulting model and increase its inference speed we replace all 1D convolutions with depth-wise separable convolutions which are much more memory and computationally efficient (see LightSpeech [24]).

The resulting Accentor model has 6.5M parameters and is trained on 90% of the lexical stress corpus for 4.3M iterations (6 days on one V100 GPU) using a batch size of 512. The remaining 10% of the corpus are equally split for validation and testing. The final model is obtained by averaging the 10 best performing models on the validation set. Over the test set, the model achieves 0.119% word-level stress error rate (SER). Note that, since the corpus is automatically compiled and could contain inaccuracies, this performance might not carry over to a real world scenario. In Section 6, we conduct further experiments to give a more accurate estimate of the actual SER of the Accentor.

## 5. Controllable acoustic model

The aim of the acoustic model is to process text annotated with stress information and produce acoustic features that respect the stress patterns present in the input. Thereby, the model enables the control of stress placement in the synthesized speech. In a related work [25, 26], it has been demonstrated that control over the pronunciation in a TTS acoustic model can be achieved by training on texts that contain a mixture of letters and phonemes.

We use the StreamSpeech [27] acoustic model, which im-

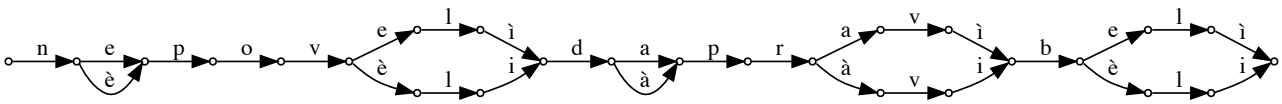


Figure 2: *Decoding graph that encodes all canonical stress patterns of the sentence* "ne poveli da pravi beli".

Table 3: *Word-level and utterance-level SER comparison of our approach with a standard FastPitch model.*

| Stress error | FastPitch | Accentor |
|---|---|---|
| Word-level | 4.43% | **0.37%** |
| Utterance-level | 38.55% | **4.45%** |

Table 4: *Comparison of our approach with a standard FastPitch model with respect to SER on different word classes.*

| Word class | FastPitch | Accentor |
|---|---|---|
| Unambiguous | 5.51% | **0.31%** |
| Ambiguous | 18.08% | **2.45%** |
| Functional | 0.45% | **0.23%** |

proves the efficiency of the FastSpeech 2 [2] architecture by employing depth-wise separable convolutions in the encoder and replacing the attention-based decoder with a lightweight convolutional decoder. The input vocabulary of the model consists of 47 different symbols – the 41 used by the Accentor and 6 new symbols representing stressed vowels.

The resulting acoustic model has 7.7M parameters and is trained for 370K iterations (30 hours on one V100 GPU) using a batch size of 64. We use a regular sized TTS dataset consisting of 23 hours of speech (190K word tokens and 32K word types) recorded by a professional female speaker with the corresponding transcriptions annotated with the methodology described in Section 3. Manual examination of the speech synthesized using the resulting model revealed that the stress information encoded in the acoustic features strictly and accurately follows the lexical stress labels specified in the input text. Samples demonstrating the accomplished control are available online [5].

## 6. Experiments

To evaluate the accuracy of the presented approach we conduct experiments on the Bulgarian Brown [28] corpus and the BulTreeBank [29] corpus, since they are lexically and grammatically representative of the Bulgarian language. From each of the corpora we sample uniformly 1000 sentences – the excerpt from the Bulgarian Brown corpus consists of 14637 words and the excerpt from the BulTreeBank consists of 11559 words. The frequencies of functional words and words with ambiguous stress pattern in the excerpts coincide with those reported in Table 1.

The extracted sentences are annotated with the Accentor model (see Section 4) and then synthesized with the controllable acoustic model (see Section 5) and an LPCNet [30] vocoder. To determine the SER of the resulting TTS system we manually check the correctness of the stress placement in the synthesized speech [6]. As a baseline we use a vanilla FastPitch [31] model with 44.7M parameters (in place of the Accentor and the controllable acoustic model) trained on the same 23 hours of speech without using the stress information in the transcriptions. We use FastPitch instead of FastSpeech 2 because of the availability of an official implementation and because the two architectures are almost identical and perform similarly.

The overall results are presented in Table 3. Errors are considered both at the word and at the utterance level. As shown, the proposed approach (referenced as Accentor) makes on average one stress error in 270 words or 22 utterances. In comparison, the FastPitch model makes one stress error in 23 words or 2.6 utterances. In Table 4, we view separately the SER on several different word classes. We observe that our approach achieves 18 times lower SER on words with unambiguous stress pattern, 7.4 times lower SER on words with ambiguous stress pattern, and 49% lower SER on functional words. Samples from the conducted evaluation are available online [5].

A more careful analysis reveals that 14.85% of the words in the evaluated sentences are unknown to the FastPitch model, while only 0.95% are unknown to the Accentor model. What's more, 69% of the words with unambiguous stress pattern that have been mistaken by the FastPitch model have not occurred in its training dataset. Those observations confirm the importance of the lexical coverage of the datasets on which modern TTS systems are trained. Also, the significant reduction in the stress errors in words with ambiguous stress pattern demonstrates the ability of the proposed Accentor architecture to model and resolve contextual homographic ambiguities. In addition, we noticed that the speech synthesized using the standard FastPitch model often suffers from unclear or blurred lexical stress characteristics and inaccurate phoneme pronunciations, which make the speech sound unnatural and reduces its comprehensibility. Those defects are not present in the speech synthesized with the Accentor and the controllable acoustic model.

The manually measured 0.37% word-level SER of our approach is significantly higher than the 0.119% measured on the Accentor test set (see Section 4). Our analysis reveals that most of the observed stress errors in the manual evaluation are caused by non-standard words such as foreign names and loanwords, which are under-represented in the Accentor training set.

## 7. Conclusion

The main contribution of the presented work is a new approach for augmenting a TTS system with a dedicated Accentor neural network that explicitly models lexical stress knowledge. We described a methodology for the automatic stress annotation of a large text corpus, required for the separate training of the Accentor model. Furthermore, we showed that the composition of the Accentor with a controllable acoustic model achieves 12 times lower word-level and 8.7 times lower utterance-level SER compared to modern monolithic neural TTS systems. The proposed approach was implemented in the Bulgarian TTS engine NeuralSpeechLab, which is widely used by the visually impaired people in Bulgaria.

[5] https://lml.bas.bg/~gshopov/accentor.html

[6] The evaluation is performed auditorily by a group of 5 native Bulgarian speakers, supervised by a professional linguist.

# 8. References


[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] 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," in *International Conference on Learning Representations*, 2021.

[3] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[4] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," https://datashare.ed.ac.uk/handle/10283/3443, 2019.

[5] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *INTERSPEECH*, 2019.

[6] J. Taylor and K. Richmond, "Analysis of pronunciation learning in end-to-end speech synthesis," in *INTERSPEECH*, 2019.

[7] B. Lin, L. Wang, X. Feng, and J. Zhang, "Joint detection of sentence stress and phrase boundary for prosody," in *INTERSPEECH*, 2020, pp. 4392–4396.

[8] M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "ASR inspired syllable stress detection for pronunciation evaluation without using a supervised classifier and syllable level features," in *INTERSPEECH*, 2019, pp. 924–928.

[9] M. Andreeva, I. Marinov, and S. Mihov, "SpeechLab 2.0 – a high-quality text-to-speech system for Bulgarian," in *Proceedings of the RANLP*, 2005, pp. 52–58.

[10] O. Yakovenko, I. Bondarenko, M. Borovikova, and D. Vodolazsky, "Algorithms for automatic accentuation and transcription of Russian texts in speech recognition systems," in *International Conference on Speech and Computer*. Springer, 2018, pp. 768–777.

[11] R. Sproat and K. Hall, "Applications of maximum entropy rankers to problems in spoken language processing," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] C. Ungurean, D. Burileanu, and A. Dervis, "A statistical approach to lexical stress assignment for TTS synthesis," *International Journal of Speech Technology*, vol. 12, no. 2, pp. 63–73, 2009.

[13] T. Anbinderis, "Automatic stressing of Lithuanian text using decision trees," *Information Technology and Control*, vol. 39, no. 1, 2010.

[14] M. Gams *et al.*, "Automatic lexical stress assignment of unknown words for highly inflected Slovenian language," in *International Conference on Text, Speech and Dialogue*. Springer, 2002, pp. 165–172.

[15] D. van Esch, M. Chua, and K. Rao, "Predicting pronunciations with syllabification and stress with recurrent neural networks," in *INTERSPEECH*, 2016, pp. 2841–2845.

[16] B. Lőrincz, "Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks," *Procedia Computer Science*, vol. 176, pp. 108–117, 2020.

[17] K. Gorman, G. Mazovetskiy, and V. Nikolaev, "Improving homograph disambiguation with supervised machine learning," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[18] D. Geneva, G. Shopov, and S. Mihov, "Building an ASR corpus based on Bulgarian parliament speeches," in *International Conference on Statistical Language and Speech Processing*. Springer, 2019, pp. 188–197.

[19] D. Geneva and G. Shopov, "Towards accurate text verbalization for ASR based on audio alignment," in *Proceedings of the Student Research Workshop Associated with RANLP 2019*. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 39–47.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[22] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[23] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] R. Luo, X. Tan, R. Wang, T. Qin, J. Li, S. Zhao, E. Chen, and T.-Y. Liu, "LightSpeech: Lightweight and fast text to speech with neural architecture search," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5699–5703.

[25] K. Kastner, J. F. Santos, Y. Bengio, and A. Courville, "Representation mixing for TTS synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5906–5910.

[26] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *International Conference on Learning Representations*, 2018.

[27] G. Shopov, S. Gerdjikov, and S. Mihov, "StreamSpeech: Low-latency neural architecture for high-quality on-device speech synthesis," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[28] S. Koeva, S. Leseva, I. Stoyanova, E. Tarpomanova, and M. Todorova, "Bulgarian tagged corpora," in *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages*, 2006, pp. 78–86.

[29] K. Simov, P. Osenova, A. Simov, and M. Kouylekov, "Design and implementation of the Bulgarian HPSG-based treebank," *Journal of Research on Language and Computation. Special Issue*, pp. 495–522, 2005.

[30] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[31] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.