Babeș-Bolyai University
Faculty of Mathematics and Computer Sciences

AUGMENTED LEXICAL FEATURES AND LIMITED DATA SPEAKER ADAPTATION APPLICATIONS FOR ROMANIAN NEURAL SPEECH SYNTHESIS

PhD Thesis Summary

PhD Student:
Beáta LÁZÁR-LŐRINCZ

Scientific Supervisor:
Prof. dr. Bazil PÂRV

2022

Contents

Thesis table of contents

List of publications

Conference proceedings

1. Beáta Lőrincz, Maria Nuțu, Adriana Stan, Romanian Part of Speech Tagging using LSTM Networks, In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, pp. 223-228, 2019.

   Rank C, 2 points.

2. Maria Nuțu, Beáta Lőrincz, Adriana Stan, Deep learning for automatic diacritics restoration in Romanian, In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, pp. 235-240, 2019.

   Rank C, 2 points.

3. Beáta Lőrincz, Maria Nutu, Adriana Stan, Mircea Giurgiu, An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data, In 2020 IEEE 10th International Conference on Intelligent Systems (IS), IEEE, pp. 437-442, 2020.

   Rank C, 1 points.

4. Beáta Lőrincz, Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks, In Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, Procedia Computer Science, 176, pp. 108-117, 2020.

   Rank B, 4 points.

5. Beáta Lőrincz, Adriana Stan, Mircea Giurgiu, Speaker verification-derived loss and data augmentation for DNN-based multispeaker speech synthesis, In 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, pp. 26-30, 2021.

   Rank B, 4 points.

6. Beáta Lőrincz, Adriana Stan, Mircea Giurgiu, An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis, In Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021, Procedia Computer Science, 192, pp. 756-765, 2021.

   Rank B, 4 points.

7. Stefan Daniel Dumitrescu, Petru Rebeja, Beáta Lőrincz, Mihaela Gaman, Andrei-Marius Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George-Andrei Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, Viorica Patraucean LiRo: Benchmark and leaderboard for Romanian language tasks, In

Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 pre-proceedings (NeurIPS Datasets and Benchmarks 2021), 2021.

Rank A*, 0.6 points.

8. Beáta Lőrincz, Bogdan Iudean, Andreea Vescan, Experience report on teaching testing through gamification, In Proceedings of the 3rd International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, pp. 15-22, 2021.

   Rank A, 8 points

9. Beáta Lőrincz, Contributions to neural speech synthesis using limited data enhanced with lexical features, In Proceedings of the 1st ISCA Symposium on Security and Privacy in Speech Communication, pp. 83-85, 2021.

   Rank D, 1 points

10. Adriana Stan, Beáta Lőrincz, Maria Nutu, Mircea Giurgiu, The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data, In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE, pp. 85-90, 2021.

    Rank D, 0 points

Book chapters

1. Adriana Stan, Beáta Lőrincz, Generating the Voice of the Interactive Virtual Assistant, In Virtual Assistant, IntechOpen, 2021.

   Rank D, 1 points

Publication score: 27.6 points.

1. Introduction

Speech synthesis systems are commonly used in our everyday lives, incorporated into virtual assistant applications, navigation systems or any software application, that allows its textual content to be converted into synthesised speech. With the rise of the deep learning technologies, neural network based solutions are frequently engaged for such applications, as the resulting synthesised speech achieves naturalness that is close to that of human speech.

However, there are still a number of issues that present challenges for neural speech synthesis. In order to achieve natural sounding speech output, large amounts of high quality audio data are required for training. This holds true both for synthesis systems that include a single speaker or multiple speaker identities. The naturalness is desired, as synthesis systems that generate intelligible speech, but are less natural sounding, are not feasible for applications generating longer speech segments. Achieving high speaker similarity, between the synthesised output and the human voice recorded for the training data, is especially beneficial for synthesis systems that aim to recreate the voices for those who have lost the capability to speak.

Besides the naturalness and speaker similarity, the expressivity of the synthesised speech contributes to its efficiency and applicability, as it conveys additional information about the intonation and expressions associated with the textual content. Expressive speech synthesis is particularly suitable for the narration of audio books. Modelling of the prosodic features, that result in an expressive synthesised speech output, currently poses a challenge for neural speech synthesis systems.

The research and industrial communities have published a large number of neural architectures, that propose models for speech synthesis. The majority of these models focus on mainstream languages, such as English, but the under-resourced languages, such as Romanian, still suffer from the lack of resources and tools that are capable of generating high quality synthesised speech.

As one of the main challenges of neural speech synthesis is the necessity of large amounts of high quality training data, in this work, the issue of limited training data is addressed. This scenario is examined, both in a single and multi-speaker speech synthesis setup. The methods presented aim to achieve naturalness and high speaker similarity for models trained on the Romanian language.

The control over prosodic features of the output speech is analysed, in order to improve the expressive effect of the output speech. For this reason, a new expressive Romanian speech dataset is introduced. Linguistic features are examined in this process, the input text is enhanced with linguistic features to examine the effect of this additional information over the synthesised speech output.

The main objective of the thesis encompasses the improvement of naturalness and speaker similarity of the synthesised speech output, using augmented lexical features for the input text and limited amount of audio data for neural speech synthesis systems, focused on the Romanian language.

1.1. Organization of the thesis

The organization of the thesis is presented below.

Chapter 1 describes the main motivation and objectives of the thesis.

Chapter 2 presents concepts related to speech generation, and analyses the two major components of speech synthesis: text processing and acoustic modeling. We detail the components and related concepts, the historical background and current state-of-the-art methods for speech synthesis. The chapter is concluded with listing evaluation methods for speech synthesis models.

Chapter 3 describes text processing tools that contribute to the front-end components of text-to-speech systems. We examine part-of-speech tags, phonemic transcription, syllabification and lexical stress assignment, that can be used to enhance the input text of speech synthesis systems. In addition, we mention a sentiment analysis, diacritic restoration experiment and contributions to a Romanian benchmark for language evaluation tasks.

Chapter 4 details the main contributions of this thesis, as it presents the experiments conducted with various speech synthesis architectures. The first section of the chapter introduces the postfitering method used to enhance the output of a feed-forward neural network by engaging speaker adaptation. The next section presents two multi-speaker experiments evaluated on current state-of-the-art architectures and proposals to enhance and evaluate the output of these systems trained on the Romanian language. This chapter also discusses the introduction of a Romanian expressive speech corpus and experiments with synthesis systems aimed at providing an interactive control over the prosodic features of the output speech. Speaker adaptation techniques are discussed in various sections of the chapter.

The thesis concludes with **Chapter 5** summarising the main conclusions and lists future development proposals.

1.2. Thesis contributions

The contributions of the thesis are summarised below with the corresponding chapter number.

In Chapter 2 a basic set of terms related to digital speech processing were elaborated. The text processing component, that converts the input text into textual features used for speech generation, and the acoustic modelling component were detailed, including a historical overview of the speech synthesis systems. This overview provided the speech processing fundamentals and background used as starting point for our experiments.

Enhancing the input text with lexical features can enhance the quality of the output speech. Based on this observation, in Chapter 3 we presented experiments that predict the part of speech of the input sequence, the phonemic transcription, syllabification and lexical stress. Providing this information to speech processing applications can contribute to its expressivity, and especially in the case of phonetically rich languages, the phonemic transcription can contribute to a better pronunciation of the input text. Additional experiments presented include diacritic restoration that can reinstate the diacritic symbols of the input text, sentiment analysis that can be used to detect the affective state of the input text. The contributions to the benchmark for language tasks presented encourages research for Romanian natural language processing practitioners.

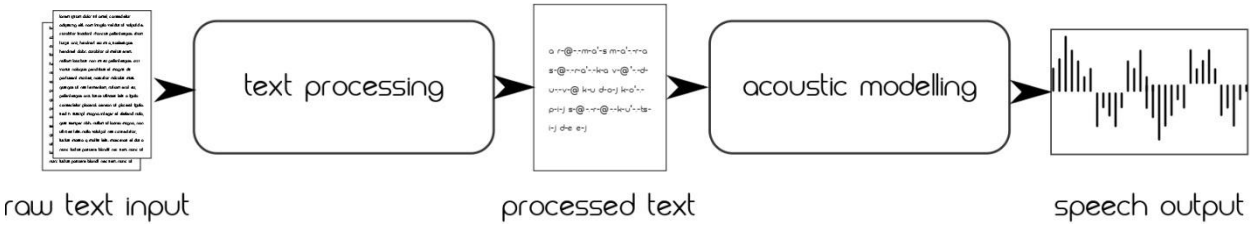
Text-to-speech synthesis systems have achieved naturalness that is close to that of human speech, however these rely on large amounts of high quality speech data. In Chapter 4 we present experiments that strive to enhance the quality of the output speech using limited amount of data. The methods of speaker adaptation, the engagement of a speaker verification network, data augmentation, careful speaker data selection were presented that aim to produce high quality synthesised speech for the Romanian language. To contribute to the naturalness of the output speech we also presented experiments that target the prosody control of the synthesised speech.

Keywords: text-to-speech synthesis, neural speech synthesis, text processing, deep neural networks, Romanian

2. Background

In this chapter we discuss the components and methods used for speech synthesis, also referred to as text-to-speech (TTS). We detail the steps of the process starting from text processing through acoustic modeling in order to reach the speech output. The methods used for this flow are examined starting from rule based approaches up to the current deep neural network (DNN) based, state-of-the-art models.

Text-to-speech synthesis refers to the process of generating the speech signal from text. Starting from the raw text through several intermediate steps, the acoustic signal or speech waveform is generated. An overview of speech synthesis is illustrated on the following figure:



At a first glance, this process seems like a straight forward mapping from the characters of the input text to their acoustic realisation. However, there are a number of technical issues that make this an extremely complex problem, such as speaker and speech variability or the modelling of prosodical features.

Before diving into the historical background of speech synthesis, we define a set of terms related to the speech generation, its representations and the challenges imposed by this process. The description of speech fundamentals is then followed by sections describing the two major components of TTS systems: text processing and acoustic modeling. The approaches used for these in the past few decades are detailed. The chapter is concluded with the description of methods used for evaluation of TTS systems.

The outline of this chapter relies on the book chapter published in [1].

3. Text processing contributions

The first component of text-to-speech includes the text processing steps, which transform the input text into a representation, that will be used for acoustic modelling. As the output of the synthesis system is the acoustic realisation of the input text, the textual features effect the speech output. Augmenting the input text with lexical information or content related features can enhance the output speech in terms of expressivity and naturalness.
The complexity of the text processing steps is highly dependent on the language used; for this reason, at the selection of these steps, the language characteristics need to be considered.

In this chapter, we present text processing tasks that can be used as augmenting features for TTS systems. The first task described aims to predict the part-of-speech (POS) tag of words. Providing the POS tags for the words of the input text is helpful for the disambiguation of words that are spelled the same way, but are pronounced differently. The second section presents concurrent solutions for the tasks of phonemic transcription, lexical stress assignment and syllabification. The phonemic transcription is advantageous for phonetically rich languages, while the lexical stress and syllable boundaries carry information about the intonation and rhythm of the speech. Three additional tasks presented are diacritics restoration, sentiment analysis and debiasing of word embeddings. For languages that contain diacritic symbols, such as Romanian, it is important to include these in the input text, to preserve the original meaning of the text. Sentiment analysis provides information about the affective state of the text, and can be used to label the input text with sentiments. Word embeddings can be used as an alternative for the representation of the textual information.

We analyse these tasks due to their importance in speech processing tasks. At the end of the chapter, we present a benchmark for evaluation of text processing tasks, focused on the Romanian language.

3.1. Part-of-speech tagging

Part-of-speech (POS) tagging is one of the key tasks of natural language processing (NLP). It refers to the identification of the part-of-speech of a given word and optionally, additional grammatical properties inherent to a particular POS. Along with other components of NLP, such as lemmatization, stemming, syllabification, word or sentence boundary detection and many others, it aims to help computers understand and process natural language.

The difficulty of the task lays in the fact that the same orthographic form of a word can have a different meaning, depending on the context (i.e. homographs). Another problem is that the declination and inflections of the words are not regular, especially in morphologically rich languages, such as the case of Romanian.

As a result, to define the correct POS of a word aside from its spelling, we also need to take into account the semantic links between words.

For the task of POS tag prediction we examine neural network models composed of recurrent layers and compare these with sequence-to-sequence models containing recurrent encoders and decoders. The highest accuracy achieved for the root POS tag is 99.18% and 98.25% for the extended tags that contain additional grammatical information next to the root POS.

3.2. Phonemic transcription, syllabification and lexical stress assignment

Other text processing tools frequently used for TTS are the tasks of phonemic transcription, syllabification and lexical stress assignment.

Phonemic transcription or grapheme-to-phoneme conversion refers to the process of representing the written form of a word into an acoustically-derived form, suited for a correct reading or pronunciation of the respective word. Lexical stress assignment refers to determining the most prominent syllable(s) within a word. Syllabification is the process of splitting a word into smaller units, generally pertaining to a single pronunciation block of phonemes.

These tasks are highly dependent on the language used and are vital in many speech and language processing flows. Knowing how the words are pronounced is essential for producing high quality speech synthesis and speech recognition systems [2]. The exact pronunciation of a word in a language depends mostly on its constituent phones. However, the syllable structure and lexical stress play an important role in assigning the correct meaning and emphasis to it [3].

We have analysed these tasks, and proposed a solution for jointly solving them. The need for the concurrent solution comes from an important aspect of speech synthesis, which is real time processing. Having separate modules for each of the tasks translates into an overhead for the processing flow. As a countermeasure, we proposed a concurrent solution for the phonemic transcription, lexical stress assignment and syllabification. The goal is to evaluate the feasibility of a single model which jointly predicts all three linguistic information. The languages used in the evaluation are English and Romanian.

In terms of algorithmic approach, recent studies showed that deep neural network based sequence-to-sequence models are highly efficient in solving the tasks of phonemic transcription, lexical stress assignment and syllabification in various languages [4].

Starting from this observation, the current study employs recurrent, convolution and attention-based neural network architectures as the main training strategies, and evaluates several combinations and variations of a sequence-to-sequence learning scenario. Two case studies are presented, the first one is based on recurrent and convolutional architectures

combined with attention models, while the second relies on the fully attentional Transformer architecture. The best results for the first case study were obtained with a combination of convolution and attention layers, where the accuracy of the joint prediction for the three tasks was of 58.96% for English and 86.64% for Romanian. The second set of experiments obtains enhanced results, with a word error rate (WER) of 5.6% for the Romanian dataset (on a newly introduced large Romanian dataset called RoLEX) and a WER of 24.11% for the English dataset.

3.3. Sentiment analysis

Labeling the input text with its affective state can be used as an additional feature for TTS systems to condition the output speech based on the sentiment of the input text. Sentiment analysis is a fundamental task of NLP that determines the affective state of the text, most frequently using positive and negative as labels. This task has been addressed with various rule based, statistical or machine learning methods.

The applications of sentiment analysis are various; next to its application in expressive TTS, it can be used for social media monitoring, customer feedback analysis or any other text analysis tasks, where the sentiment of the text carries a valuable information for the application. In our research, we also applied sentiment analysis to evaluate the students feedback related to educational activities.

3.4. Diacritics restoration

The task of automatic diacritic restoration (ADR) refers to reinstating the diacritic symbols of a text being stripped of these symbols. This restoration process is useful and necessary as many of the electronic content available online does not contain diacritic symbols.

In the case of numerous languages the usage of the diacritics can change the meaning and the pronunciation of words. For this reason the input text of TTS systems need to contain the diacritic symbols in order to facilitate the correct learning of word pronunciations in the speech output.

In our experiments neural network architectures composed of recurrent and convolutional layers were presented for the task of diacritics restoration. The best accuracy results achieved on word level was 97%.

3.5 Gender bias in word embeddings

Word embeddings are frequently used in any of the mainstream NLP applications. Words are represented by real-valued vectors, so that words with similar meaning are closer in the vector space. However, for languages with grammatical gender, the word embeddings can

encode a gender bias inherited from the training corpora. To eliminate the unwanted consequences, efforts have been put towards reducing and analysing of the existing gender bias in language models [5]. The nouns of the Romanian language carry a grammatical gender being categorized into masculine, feminine or neutral nouns. Many studies focusing on mitigating gender bias target the English language, but the methods proposed cannot be easily transferred to other languages with grammatical gender. We follow the proposal of [5] and reproduce the results presented for Spanish and French on the Romanian language.

3.6. Benchmarks for language evaluation tasks

Benchmarks for natural language understanding or other domains encourage the domain's research, provide resources for training and analysing tasks, and promote the standardization of the selected tasks. The LiRo platform was proposed to facilitate the NLP research for the Romanian language. 10 text processing tasks are available in the benchmark, and 3 new datasets were created to extend the available resources for the Romanian language.

The results and experiments presented in this chapter are based on the original papers published in [6-10].

4. Speech synthesis contributions

Recent research in speech synthesis focuses on the naturalness and expressivity of the speech output, in order to increase the applicability of TTS systems. It is reasonable to admit that, the more natural a synthetic voice sounds like, the easier and more pleasant is to use it in any environment.

In this chapter we will address some of the main challenges of TTS systems. The first section is dedicated to analysing methods for improving the speech output when limited data is available. This case is encountered frequently in practice, especially for under-resourced languages, such as Romanian. In general, our goal is to synthesise voices with as little data as possible. The second section presents two sets of experiments for multi-speaker speech synthesis, aimed at improving the output speech, while also focusing on the availability and characteristics of the input text and audio data. The input text is augmented with lexical information based on the text processing steps presented in the previous chapter. Various amounts of audio data is used for the experiments, and the characteristics of the audio recordings are analysed with respect to the quality of the resulting speech output. Methods for speaker adaptation with different amounts of training data are also explored. The last section discusses the speech expressivity that contributes a great deal to the naturalness of the speech output. Characteristics such as prosody, rhythm and pitch that make the speech more or less human sounding are examined.

4.1. Speech synthesis with limited data

In neural speech synthesis, obtaining a high quality synthesised speech relies on the availability of large amounts of high quality speech data. Machine learning methods in general, including neural networks, rely on large amounts of training data. When this data is not readily available, other methods are sought to enhance the quality of the speech output.

In this chapter we address the issue of limited data by using the method of postfiltering achieved through speaker adaptation. We apply the speaker adaptation to enhance the quality of the output speech. Results show that even with as little as 10 minutes of speech recordings, the quality of the synthetic speech output is improved by the postfiltering process.

4.2. Multi-speaker synthesis with various amounts of data

Speech synthesis systems are capable of synthesising speech for a single or multiple speakers, whose identities are learnt during the training process. Multi-speaker TTS systems have the advantage of incorporating multiple vocal identities into a single network, and allow the user to select any of these identities without the need to change the architecture.

This scenario also enables the use of fewer samples per speaker, however in the resulting acoustic model not all speakers exhibit the same synthetic quality. In this study we propose methods for enhancing the speaker identity learning of multi-speaker speech synthesis systems.

The first two directions explored are: (1) forcing the network to learn a better speaker identity representation by adding an additional loss term based on speaker verification metrics; and (2) augmenting the input data pertaining to each speaker using waveform manipulation methods. We showed that both methods are efficient, the additional loss term aids the speaker similarity, while the data augmentation improves the intelligibility of the multi-speaker TTS system.

The third approach exploits a Romanian parallel corpus that contains speech recorded under different conditions, and evaluates if particular voice characteristics or recording conditions influence the resulting quality of the respective voice identity in a multi-speaker TTS scenario. It also analyses if different text representations fed to the neural TTS system boost or deteriorate the synthesised speech of the different speakers. The results showed that the recording conditions influence the most the quality of the speech output.

4.3. Expressive speech synthesis

Expressive speech synthesis has been a focus of the research community as the naturalness of the synthesis systems are comparable to that of human speech [11], and controlling the prosody, expressivity of such systems allows a more varied range of applications.

In case of neural network based end-to-end models, the representation of speech prosody is commonly learnt as an internal representation within the model. As the prosody is not explicitly learnt by the model, the expressivity of the output speech also depends on the speech corpora used as training data for neural models. However, speech corpora resources for the non-mainstream languages such as Romanian are limited, and the lack of resources hinders the development process of expressive speech synthesis. For this reason, we introduced a new Romanian speech corpus called MARA, the first Romanian expressive speech corpus release. Using the newly presented speech corpus the usage of synthesised speech as training data was examined.

The results and experiments presented in this chapter are based on the original papers published in [12-15].

5. Conclusions and Future Work

In our work we have focused on neural speech synthesis, and aspects of natural language processing tasks that can be used to enhance synthesis systems. Our experiments relied mostly on Romanian datasets, as our goal was to contribute to the tools and systems trained on under-resourced languages, such as Romanian. To be comparable to the scientific literature, a selected number of experiments were presented using English datasets as well.

Before presenting the text and speech processing related experiments, we provided an overview in Chapter 2 about speech processing fundamentals, text processing elements relevant for speech processing, respectively text-to-speech systems developed starting with the rule-based synthesis from before the 90's up to the current state-of-the-art neural network based models.

Starting from this overview, in Chapter 3 we presented a set of text processing related experiments that can be incorporated into speech applications to enhance the output speech. The first experiment predicted the part-of-speech tag of words that can be used as augmenting feature for the input text. The next experiment presented models for concurrently predicting the phonemic transcription, syllabification and lexical stress of the input text. Similarly to the part-of-speech tags, this information can also be used to enrich the textual information provided to the speech synthesis systems. An additional set of experiments is described including diacritics restoration and the contribution of the author to a Romanian benchmark for language evaluation tasks.

The main contributions of this thesis were elaborated in Chapter 4. The first case study presented aims to enhance the speech output of neural text-to-speech systems by engaging a postfiltering network. As emphasized in this section, our goal was to train synthesis systems on limited amount of data and still obtain a reasonable quality for the output speech. The postfiltering network is based on the speaker adaptation technique. The next section presented two multi-speaker speech synthesis case studies. The first one relied on a convolutional architecture and engaged a speaker verification network to obtain additional loss terms, respectively data augmentation methods to increase the volume of the training data. The second set of experiments focused on objective evaluations of the output speech based on recording conditions and speaker characteristics that can potentially be used to assess the quality of the training data. The additional experiments presented list the contribution of the author to the creation of the first expressive Romanian dataset called Mara, and experiments that target the controllability of speech expressiveness.

Neural speech synthesis is not considered a thoroughly solved problem, as it raises several open issues that are of interest to the TTS research community. One of the issues is related to prosody, as the output speech, especially in the case of long paragraphs, is required to model the adequate prosody that is not achieved by most of the state-of-the-art synthesis systems. In order to mimic the spontaneous speech, synthesised speech should also model

the mental process of developing and generating the message that might include repetitions or pauses. With regards to speaker adaptation, neural speech synthesis offers solutions for fast adaptation, but these methods can be improved by reducing the amount of adaptation data or training time.

As future work, we plan to continue the work on synthesis systems that incorporate the modelling of prosody, and allow the synthesis conditioned on aspects related to prosody with a strong emphasis on the Romanian language. Contributions to fast speaker adaptation will be in the focus of our research as well, aiming the development of pipelines that allow neural speech synthesis systems to be adapted to new voice identities with reduced amount of data.

We also plan to release synthesis systems trained on the Romanian language as an open API, including models trained on convolutional, recurrent and attention based architectures. The creation of text and speech corpora will remain in our interest, especially for the Romanian language to facilitate and encourage the work of other researchers in the domain.

Summary Bibliography

[1] Adriana Stan and Beáta Lőrincz. Generating the voice of the interactive virtual assistant. In Virtual Assistant. IntechOpen, 2021.

[2] Daan van Esch, Mason Chua, and Kanishka Rao. Predicting Pronunciations with Syllabification and Stress with Recurrent Neural Networks. In INTERSPEECH, pages 2841–2845, 2016.

[3] Paul Taylor. Text-to-Speech Synthesis. Cambridge University Press, 2009.

[4] Benjamin Milde, Christoph Schmidt, and Joachim Kohler. Multitask Sequence-to-Sequence Models for Grapheme-to-Phoneme Conversion. In INTERSPEECH, pages 2536–2540, 2017.

[5] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. arXiv preprint arXiv:1909.02224, 2019.

[6] Beáta Lőrincz, Maria Nuțu, and Adriana Stan. Romanian part of speech tagging using LSTM networks. In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), pages 223–228, 2019.

[7] Beáta Lőrincz. Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks. Procedia Computer Science, 176:108–117, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

[8] Beáta Lőrincz, Bogdan Iudean, and Andreea Vescan. Experience report on teaching testing through gamification. In Proceedings of the 3rd International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, pages 15–22, 2021.

[9] Maria Nuțu, Beáta Lőrincz, and Adriana Stan. Deep learning for automatic diacritics restoration in romanian. In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), pages 235–240. IEEE, 2019.

[10] Stefan Daniel Dumitrescu, Petru Rebeja, Beáta Lőrincz, Mihaela Gaman, Andrei-Marius Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George-Andrei Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, Viorica Patraucean, LiRo: Benchmark and leaderboard for Romanian language tasks, In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 pre-proceedings (NeurIPS Datasets and Benchmarks 2021), 2021.

[11] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A.

Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. CoRR, abs/1712.05884, 2017.

[12] Beáta Lőrincz, Maria Nuțu, Adriana Stan, and Mircea Giurgiu. An evaluation of postfiltering for deep learning based speech synthesis with limited data. In 2020 IEEE 10th International Conference on Intelligent Systems (IS), pages 437–442. IEEE, 2020.

[13] Beáta Lőrincz, Adriana Stan, and Mircea Giurgiu. Speaker verification-derived lossand data augmentation for DNN-based multispeaker speech synthesis. In 2021 29th European Signal Processing Conference (EUSIPCO), pages 26–30, 2021.

[14] Beáta Lőrincz, Adriana Stan, and Mircea Giurgiu. An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis. Procedia Computer Science, 192:756–765, 2021. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

[15] Adriana Stan, Beáta Lőrincz, Maria Nuțu, and Mircea Giurgiu. The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data. In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pages 85–90, 2021.