INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Unified Language-Independent DNN-Based G2P Converter

Markéta Jůzová, Daniel Tihelka, Jakub Vít

Department of Cybernetics & New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
juzova@kky.zcu.cz, dtihelka@ntis.zcu.cz, jvit@kky.zcu.cz

Abstract

We introduce a unified Grapheme-to-phoneme conversion framework based on the composition of deep neural networks. In contrary to the usual approaches building the G2P frameworks from the dictionary, we use whole phrases, which allows us to capture various language properties, e.g. cross-word assimilation, without the need for any special care or topology adjustments. The evaluation is carried out on three different languages – English, Czech and Russian. Each requires dealing with specific properties, stressing the proposed framework in various ways. The very first results show promising performance of the proposed framework, dealing with all the phenomena specific to the tested languages. Thus, we consider the framework to be language-independent for a wide range of languages.

Index Terms: grapheme-to-phoneme, phonetic transcription, recurrent deep neural network, speech synthesis

1. Introduction

Grapheme-to-phoneme conversion (G2P), sometimes referred as *phonetic transcription*, is usually considered as "just working". However, to build an error-prone G2P module is not as easy a task as it may seem.

In general, there are two basic approaches to the G2P, which are combined together in real cases. The first is the use of a large dictionary of word forms matched with their phonetic form. It is suitable for languages with irregular pronunciation, such as English. For words not included in the dictionary (which may either be new or special words, as well as typos in the input text), there is a need to have a set of backup-rules to deal with them. On the other hand, the G2P may rely on a set of rules, which is an approach suitable for languages with regular pronunciation, such as Czech or Russian. Still, there has to be a dictionary of words not following the conversion rules, which is, in Czech, the case of words of foreign origin. Although the dictionary can be used for regular-pronunciation languages as well, it is not an optimal choice for inflected languages, where each word may have lots of forms, making the dictionary quite big. Moreover, there is the problem of cross-word assimilation, which is hard to handle effectively by a dictionary itself. On the contrary, the use of rules for irregular languages should rather be considered as a back-up solution, since a smaller set of rules is likely to miss the pronunciation and large set of rules is slow to duplicate the dictionary.

The present paper aims at building a unified G2P module able to cover both cases, without prior explicit considerations of language properties for which the module is being built for. We demonstrate the abilities of the G2P on three different languages. First it is English, where, due to irregular pronunciation, usually only a dictionary is used to build a G2P system. However, the use of dictionary alone is not enough to build a robust G2P for English – there are several homographs (*read*, *record*, . . .), requiring meta-information (e.g. POS) to be taken into consideration when deriving their pronunciation.

Have you read[**rEd**] the book?	You don't like to read[**ri:d**][1].
Don't touch live[**laIv**] wires.	I can't live[**lIv**] alone.

And also, there is a cross-word assimilation, albeit rather weak, which should be taken into consideration [2].

I leave[**li*f***] today[***t*@deI:**].	[liv] → [lif]
Those[**D@U*s***] shoes[***S*u:z**].	[D@Uz] → [D@Us]

The second language is Czech, an inflected language with rather regular pronunciation, but having strong inter- and cross-word assimilation, regressive in most cases, which has to be taken into account [3, 4, 5, 6].

Most[**most**] k[**k**] věži[**vjeZi**].	
Bridge to a tower.	
Most[**mo*zd***] k[***g***] dolu[**dolu**].	[most] → [mozd]
Bridge to a mine.	[k] → [g]
Bez[**bez**] sdružení[***zd*ruZeJi:**].	[sdruZeJi:] → [zdruZeJi:]
Without association.	
Bez[**be*s***] vzpírání[***fs*pi:ra:Ji:**].	[bez] → [bes]
Without weight lifting.	[vzpi:ra:Ji:] → [fspi:ra:Ji:]

Although all of these can be handled by a set of rules, there are some words of foreign origin, now common in Czech, that violate them.

Jásot[**ja:sot**] politiků[**poli*t*iku:**].	[policiku:] → [politiku:]
Exultation of polititians.	Ground rule is:
	transcribe *{d,t,n}i* as [{J\,c,J}i]
Hvizd[**hvi*st***] politiků[**poli*t*iku:**].	[hvizd] → [hvist]
A whistle of polititians.	

The last language is Russian which has similar properties to Czech, but has also non-fixed stress positioning [7] which can alter the meaning of homographs, making the sentence either nonsensical or changing its meaning.

У него два дома[**dom''a**].	У него нет дома[**d''oma**].
He has two houses.	*He has no house.*
Мало места[**m' ''esta**].	Красивые места[**m'est''a**].
Little space.	*Beautiful places.*

In addition to large pronunciation dictionaries and hand-crafted (or automatically trained) transcription rules, there are

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

http://dx.doi.org/10.21437/Interspeech.2019-2335

other approaches to automatic G2P conversion. The popular alternative approaches include joint-sequence alignments using *n*-grams [8], or weighted finite state transducers are used to build a conversion paths [9]. There are also studies using different classifiers, conditional random fields [10] and HMM [11]. And in recent years, different types of neural networks have been applied to the G2P problem, e.g. [12, 13].

As explained in [13], the great advantage of using neural networks, for example LSTM [14], for G2P conversion lies in the avoidance of the need for explicit G2P alignment which is not usually straightforward – although there is a one-to-one mapping of some graphemes to phonemes in different languages, many graphemes are not pronounced or, on the other hand, a grapheme may be rendered into a sequence of two or more phonemes. The neural network-based model is, nevertheless, able to make contextually-dependent decisions.

The majority of all G2P studies use the pronunciation dictionary as an input and their authors train the proposed models to be able to predict the phonetic transcription word by word. However, this approach cannot handle cross-word assimilation influence, unless various tricks are employed. Contrary to that, we have, therefore, decided to use phrases, split further into short sequences of words, for training our model. This allows us to simply and universally (i.e. in a language-independent way) handle both inter- and cross-word influences, either regressive or progressive, within a sequence of phones.

That assumption drives us closer to the machine translation approaches where the need for using the word sequences in the training data is unquestionable – the correct translation from one language to another in a word-by-word manner is unfeasible. The neural machine translation approaches commonly use the encoder–decoder architecture of the NN model, e.g. [15, 16, 17], where the encoder extracts the representation from an input sentence and the decoder generates a correct translation from this representation.

In this paper, we thus present our first G2P conversion experiments using a deep neural network (DNN) with the encoder–decoder architecture, trained on whole multi-word chunks of texts. The output of the trained model and the output of our current rule and dictionary-based G2P module are compared to the evaluation part of our data, consisting of phrases aligned with their phonetic forms, provided to us and checked by phonetic experts. We point out the strong and weak points of the approach, illustrating them by examples in Section 4.

2. Baseline G2P converter

Our current G2P converter was originally designed nearly 20 years ago within a development of our new text-to-speech (TTS) system *ARTIC* [18] to primarily handle the Czech language, but other languages like Slovak [19, 20], Russian, English and Armenian were added as well during this period [21]. It is based on the combination of rules and simple key–value dictionary, where the dictionary is searched first (to handle pronunciation irregularities not covered or coverable by rules), and when no appropriate key is found, the rules are applied.

The languages we experiment with in this paper are thus handled as follows:

Czech
The dictionary currently contains more than 170,000 items, which are irregular words in (ideally) all their inflected variants. Then there is a set of approximately 100 main rules providing the expected pronunciation plus about 64 additional rules used to handle alternative, yet still correct, pronunciations. All of these were designed manually by phonetic experts, and since Czech is our native language, the G2P module is very well tuned to handle this language.

English
Due to irregular pronunciation of English, the G2P relies primarily on a dictionary containing nearly 300,000 words.

To handle out-of-dictionary words, there are more than 1,000 rules which were automatically derived from the dictionary using the method of phonological rules induction based on the construction of phonological decision tree (enhanced version of algorithm described in [22]). Naturally, the rules must be treated here as a fallback solution only – to reach the ability of transcribing the whole dictionary by induction rules, the number of rules would be similar to the dictionary size. And thus, especially for English irregular pronunciation, the rules may fail for words not seen during their induction.

Russian
The handling of the Russian language is the most complex. The dictionary contains nearly $310,000$ items which are inflected variants of words. Contrary to Czech, the dictionary is the primary source of pronunciation even for regular Russian words, since it contains correct transcripts taking into account the stress placement. Let us note that the dictionary is not very large, considering that the language is inflected.

In case a word is not in the dictionary, there are slightly over $1,000$ rules to estimate the stress position which may influence the final pronunciation of a word. These rules were also derived automatically from the dictionary in the same way as the rules for English, thus showing the same flaws in correct stress estimation. Then, there are 60 main rules plus 4 additional rules to carry out the actual conversion to obtain the sequence of phones to pronounce. These were designed and tuned manually by phonetic experts.

3. Proposed DNN-based G2P converter

For the experiment, we designed a DNN-based model shown in Figure 1. As the input of the network, all the words w_i in the input phrase (as well as the corresponding transcriptions at its output) are transformed to the sequences of symbols having the same length by using a padding symbol $-$. The first layer of our model is the *embedding* layer which transforms each symbol/grapheme to a vector representation. The embeddings are then put to the bidirectional LSTM layer. This layer should, according to our assumption, learn within-word phonetic relations. After that, the next LSTM *encoder* layer creates the word embeddings. At this point, the sequence dimension is changed from character into words since the LSTM encoder generates a fixed length representation for each word. We can now pass these embeddings to the input another layer – either it is biLSTM or convolutional network, based on the experiment we carry out. This biLSTM/conv layer is supposed to learn cross-word-boundary relations. Finally, we decode the fixed length intermediate representations into phonemes by using the LSTM *decoder* followed by a linear projection with *softmax* activation. The decoder generates the sequences of phoneme posteriors p_i from the word representations. It employs the feedback loop, in which the previously generated phoneme is fed back to the next input (see Figure 2).

As mentioned above, we have tested two basic structures of our proposed model (see Figure 1) – marked as **biLSTM**

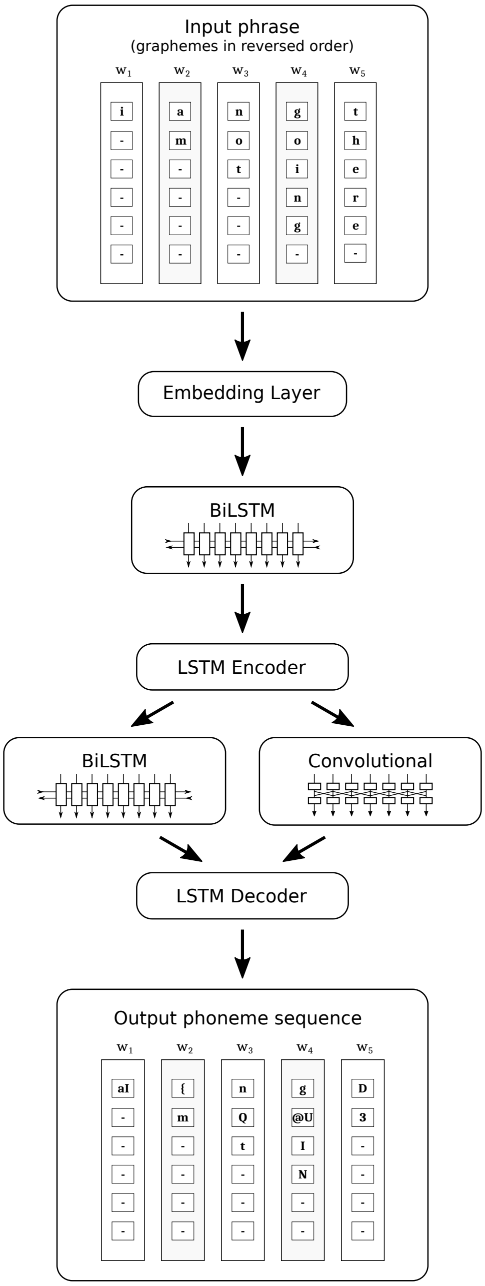

Figure 1: *The structure of our DNN-based converter for G2P, transcribing chunk of text "I am not going there" to phonetic form [*aI {m nQt g@UIN D3*].*

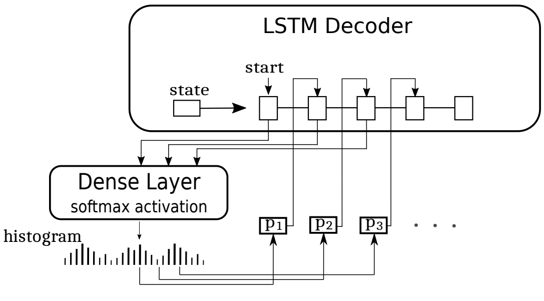

Figure 2: *The structure of the LSTM-decoder layer.*

and **conv** in Section 4. The DNN model with a bidirectional LSTM layer [23] between the encoder and the decoder consists of two neuron sequences, one for a positive time direction and the second for a negative time direction. Due to that, the model is able to learn from past and future states simultaneously. For the **biLSTM** version of our proposed model, we tested 3 different lengths of word sequences presented to the model during the training phase (3, 5 and 10). The second, *conv* DNN model has a convolutional layer [24] between the encoder and decoder. This layer works on a receptive field of 3 neighbouring word vectors (only the preceding and the succeeding word can influence the output), so the model focuses on the direct neighbourhood of the particular word.

4. Results

As mentioned in Section 1, the DNN is trained on proprietary data provided to us by experts in the particular languages. Each of the datasets, English, Russian and Czech, contains several hundred thousands of phrases, each assigned with the corresponding phonetic forms. According to the data owners, the transcriptions were carried out automatically, based on the large dictionary and proprietary expert system responsible for choosing the correct pronunciation form (in case of homographs) as well as for handling cross-word phonetic assimilations. The random selection transcriptions were also regularly checked by phonetic experts through the course of data preparation to tune their expert system.

For each of the languages, naturally trained independently, we have randomly selected 90% of phrases for training and 10% for testing as reference data. The texts from the test set were transcribed by the baseline, **biLSTM** and **conv** trained DNNs, and all the outputs were compared to the reference phonetic forms. As an evaluation measure, we have chosen *word accuracy* since it clearly reflects the correctness of a text transcription, accompanied by *phoneme accuracy* being used in other studies as well.

The overall results are shown in Table 1. In general, the proposed approach is able to outperform the baseline approach for all the languages in both accuracy measures, except for the Czech where the best model's *word accuracy* is nearly equal. Let us emphasize that it is without any language-dependent tuning, adjustments or "hacks", in contrary to the baseline system!

Comparing the results of the two types of our model – **biLSTM** and **conv** – the first one seems to work better on the inflected languages (Czech and Russian) while the **conv** model yields slightly better results on English. For Czech input, we have word accuracy close to 99%, which could be explained by fairly regular phonetic transcription. The best model for English texts achieves 94% word accuracy. For Russian, the results are much lower on the word level; however, the Russian baseline G2P converter also makes many errors (mostly due to wrong stress positions placement for non-dictionary words).

When using only short word sequences of the length 3, the model is not able to achieve the highest accuracy score. On the other hand, when using 5- and 10-grams we obtained very similar results – there is, therefore, no need to work with very long sequences, which seems natural since the cross-word influence hardly crosses more than 5 words. On the other hand, the wider context may help to determine the right pronunciations in cases of homographs (as an alternative to POS tagging).

The detailed analysis showed that the DNN model makes many more mistakes in longer words, with 10 or more graphemes – an average word length in text corpora of all 3

Table 1: *Results of the tested G2P outputs compared to the reference transcription in the test set.*

	English		**Czech**		**Russian**	
	word Acc	phoneme Acc	word Acc	phoneme Acc	word Acc	phoneme Acc
baseline	81.90%	92.13%	**98.74%**	97.79%	77.56%	92.44%
biLSTM-3	86.17%	93.51%	96.35%	98.90%	79.88%	90.71%
biLSTM-5	93.68%	95.90%	**98.69%**	**99.51%**	**86.62%**	95.39%
biLSTM-10	93.58%	95.15%	96.71%	99.05%	86.26%	**96.09%**
conv	**94.17%**	**96.93%**	97.77%	99.26%	81.36%	92.27%

tested languages is about 5 graphemes. Probably, the DNN model is not always able to remember the whole input word correctly if too long – this needs more investigation in our future research. In any case, despite increasing the encoder's/decoder's capacity, the models did not yield better results and the accuracy on longer words was still significantly lower compared to shorter words.

On the other hand, as will be shown, the proposed model is able to learn common regular transcription rules of the particular language, and the stress positioning in Russian.

The other woman.	[**Di: VD@ wUm@n**]	
The right time.	[**D@ raIt taIm**]	
Дома и окна.	[**dom"a i okn"a**]	*Houses and windows.*
У нас дома.	[**u nas d"oma**]	*In our house.*

As expected, the model can also carry out cross-word assimilation, and it is able to transcribe correctly some words of foreign origin, as can be seen in this example on Czech language:

Bez žen.	[**bez Zen**]	*Without women.*
Bez odezvy.	[**bes ?odezvi**]	*With no response.*
Bez váhání.	[**bez va:ha:Ji:**]	*With no hesitation.*
Bez vztahu.	[**bes fstahu**]	*With no relationship.*
Bez tenisu.	[**bes tenisu**]	*Without tennis.*

Unfortunately, the DNN-based G2P still makes some errors, even in shorter non-foreign words which would not appear if we used the pronunciation dictionary or a set of rules (as in baseline approach).

Cheek by jowl.		[**tSi:k baI dZ@Ul**]
	correct:	[**tSi:k baI dZ*a*Ul**]
Production techniques.		[**pr@dVkS@n tEkn*I*ks**]
	correct:	[**pr@dVkS@n tEkn*i:*ks**]
Любимый цвет.		[**l' "ub' im1j cv'æt**]
Favourite colour.	correct:	[**l' ub' "im1j cv' et**]
Text byl zajímavý.		[**te*k*zd bil zaji:mavi:**]
The text was interesting.	correct:	[**te*g*zd bil zaji:mavi:**]

In any case, despite the errors mentioned above, this first version of DNN-based G2P converter is able to compete the traditional dictionary/rules-based baseline approaches in all tested languages. These very first results are very promising and confirm our decision to use the whole phrases as an input of G2P model.

5. Conclusions

In the present paper, we focused on the building of a unified DNN-based G2P converter which would be able to perform well both on languages with irregular pronunciation (where large pronunciation dictionaries are usually used) and regular-pronunciation languages, easily describable by a set of transcription rules. The results presented in Table 1, together with the examples of correctly transcribed phrases in Section 4, prove the ability of the model to substitute the traditional baseline approaches, as verified on three different languages – English, Russian and Czech.

Of course, there is a need to have a text corpus associated with the correct pronunciation, which may be seen as an inconvenient disadvantage. However, having a text, we can always step back to use a dictionary-only approach and, if necessary, to define simple ad-hoc rules to handle cross-word influences.

In the continuation of this experiment, we plan to add a few more features, such as POS tagging, to the input, since these features are already estimated by our text normalization front end, as they are required by other TTS modules. We expect more robust transcription of homographs, for example. In addition to that, we will continue experimenting with more parameter/structure settings of our DNN-based G2G converter, working on the phrase level.

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.

7. References

[1] 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.

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

[3] H. Kučera, *The phonology of Czech.*, ser. Slavistic printings and reprintings. 's-Gravenhage, Mouton, 1961, vol. 30.

[4] Z. Palková, *Fonetika a fonologie češtiny [Phonetics and phonology of Czech]*, 1st ed. Praha: Univerzita Karlova, Nakladatelství Karolinum, 1994.

[5] P. Machač and R. Skarnitzl, *Principles of Phonetic Segmentation*, ser. Edition erudica. Epocha, 2009.

[6] A. Bičan, "Distribution and combinations of Czech consonants," *Zeitschrift für Slawistik*, vol. 56, pp. 153–171, 2011.

[7] R. Avanesov, *Modern Russian stress*, ser. Commonwealth and international library of science, technology, engineering and liberal studies. Pergamon Press, 1964.

[8] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion." *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[9] J. R. Novak, N. Minamatsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, p. 907–938, 2016.

[10] D. Wang and S. King, "Letter-to-sound pronunciation prediction using conditional random fields," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 122–125, Feb 2011.

[11] S. Jiampojamarn, C. Cherry, and G. Kondrak, "Joint processing and discriminative training for letter-to-phoneme conversion," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008, pp. 905–913.

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

[13] 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.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] 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.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, Montreal, Canada, 2014, pp. 3104–3112.

[17] 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

[18] 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., vol. 1902. Berlin–Heidelberg, Germany: Springer, 2000, pp. 223–228.

[19] J. Matoušek and D. Tihelka, "Slovak text-to-speech synthesis in ARTIC system," in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence, P. Sojka, I. Kopeček, and K. Pala, Eds., vol. 3206. Berlin, Heidelberg: Springer, 2004, pp. 155–162.

[20] J. Matoušek, D. Tihelka, J. Romportl, and J. Psutka, "Slovak unit–selection speech synthesis: Creating a new Slovak voice within a Czech TTS system ARTIC," *IAENG International Journal of Computer Science*, vol. 39, pp. 147–154, 2012.

[21] 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.

[22] 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.

[23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov 1997.

[24] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539