Computer Science and Application 计算机科学与应用, 2023, 13(3), 510-517
Published Online March 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.133050

Hans汉斯

基于Transformer的多语种字音转换

张亚停，张　寒*，曹少中，姜　丹，肖克晶

北京印刷学院信息工程学院，北京

收稿日期：2023年2月18日；录用日期：2023年3月20日；发布日期：2023年3月29日

摘　要

字音转换(Grapheme-to-Phoneme, G2P)是语音合成前端的重要部分，影响着语音合成的质量。现如今，大多数的字音转换的研究是针对于单一语种的，而在实际应用中，单一语种合成的语音远没有多语种的实用性高。因此，本文利用Transformer架构研究了在文本交叉混合条件下多语种(英、日、韩)的字音转换，使用音素错误率(Phoneme Error Rate, PER)和单词错误率(Word Error Rate, WER)作为评价指标。英文在基于美国英语的CMUDict数据集进行评估，韩语和日语则是先对SIGMORPHON 2021字音转换任务上的韩语及日语数据集进行了数据扩充，并在扩充后的数据集上进行评估。实验结果表明，在文本交叉混合条件下，基于Transformer架构的英、日、韩字音转换在音素错误率和单词错误率方面与基于Transformer架构的英、日、韩三个语言的单一语种相比都大大降低了。

关键词

字音转换，Transformer，多语种，交叉混合

Transformer Based Multilingual Grapheme-to-Phoneme Conversion

Yating Zhang, Han Zhang*, Shaozhong Cao, Dan Jiang, Kejing Xiao

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Feb. 18th, 2023; accepted: Mar. 20th, 2023; published: Mar. 29th, 2023

Abstract

Grapheme-to-Phoneme (G2P) conversion is an important part of the front end of speech synthesis, which affects the quality of speech synthesis. Nowadays, most of the research on G2P conversion is aimed at a single language, and in practical applications, single-language synthesized speech is far

*通讯作者。

文章引用: 张亚停, 张寒, 曹少中, 姜丹, 肖克晶. 基于 Transformer 的多语种字音转换[J]. 计算机科学与应用, 2023, 13(3): 510-517. DOI: 10.12677/csa.2023.133050

less practical than multilingual. Therefore, this paper uses the Transformer architecture to study the G2P conversion of multiple languages (English, Japanese, and Korean) under the condition of text cross-mixing, and uses Phoneme Error Rate (PER) and Word Error Rate (WER) as evaluation indicators. English is evaluated on the CMUDict dataset based on American English, while Korean and Japanese are first expanded on the Korean and Japanese data set on the SIGMORPHON 2021 G2P conversion task, and then evaluated on the expanded data set. Experimental results show that under the condition of text cross-mixing, the phoneme error rate and word error rate of English, Japanese and Korean characters based on Transformer architecture are greatly reduced compared with the single language of English, Japanese and Korean based on Transformer architecture.

Keywords

Grapheme-to-Phoneme Conversion, Transformer, Multilingual, Cross-Mixing

1. 引言

字音转换是指单词从正字法(字母/字符/字素序列)转换为它们的发音(音素序列)的任务，字音转换技术在语音合成中占有重要位置。近年来，随着深度学习方法的多领域应用，开始将深度学习应用于字音转换。同时文献[1]也表明了基于深度学习方法更能有效的降低音素错误率和单词错误率。字音转换质量的提升，可以大大提高语音合成的自然度[2]。

其中，Transformer 模型[3]是完全基于注意力机制[4]的一个深度学习模型，注意力机制有更好的记忆力，能够记住更长距离的信息，同时注意力机制支持并行化计算。从文献[5]中可知在进行英语字音转换时，Transformer 模型相比序列到序列(seq2seq) [6]等深度学习模型大大降低了音素错误率和单词错误率。

针对字音转换的研究目前大多数都是基于单一语种的，对同一个深度学习模型来说，针对不同语言进行字音转换时，均需要对模型进行参数调整，同一套参数并不适用于所有语言，这样就增加了时间成本。在实际应用中，多语种语音合成实用性更强，但是目前多语种语音合成研究较少，因此合成涵盖多语种的语音非常迫切。

本文研究了 Transformer 架构在文本交叉混合条件下多语种(英、日、韩)的字音转换，使得到的模型可以适用于多语种情况。将本文的实验结果分别与基于 Transformer 架构的单一语种方法进行了比较。实验结果表明，本文的研究结果在音素错误率(PER)和单词错误率(WER)方面大大降低了，对合成涵盖多语种的语音有着积极的促进作用。

2. 研究方法

编码器－解码器(Encoder-Decoder)是深度学习模型的抽象概念。许多模型的起源都是基于这一架构的。比如卷积神经网络(Convolutional Neural Network, CNN) [7]，循环神经网络(Recurrent Neural Network, RNN) [8]，LSTM (Long Short Term Memory) [9]和 Transformer 等。这些网络架构使用时编码器将输入序列转换为一个向量，解码器则基于学习到的向量表示生成输出序列。Transformer 也是一个完全基于注意力机制的模型。

Transformer 的编码器部分由 N 个编码器堆叠而成，每个编码器由分为多头注意力(Multi-Head Attention)和 Feed Forward Network (FFN)两个部分，两个部分后面都添加了残差连接并进行了层归一化。Transformer 模型中的多头注意力可以用来学习不同表示子空间的信息，Attention 的表现形式为 $\text{output} = \text{Attention}(Q, K, V)$，其中 Q 代表 query，K 代表 key，V 代表 value。多头注意力不是执行单一的注意函数，首先将 Q，K，V 复制 h 次，使用 h 个自注意力进行单独计算，然后将 h 个注意力机制获取的结果进行拼接组合，再经过一个可学习的线性映射得到最终的结果。

Transformer 的解码器部分同样由 N 个解码器堆叠而成，整体结构和编码器部分类似，唯一不同解码器内部多了一个 Masked Multi-Head Attention 层。该层的掩码机制是为了让模型保留了自回归属性，确保预测依赖于已经生成的输出部分。解码器的 Multi-Head Attention 层，用来接受编码器部分的输出，K，V 来自于编码器，而 Q 是上一层解码器的输出。解码器的输出结果后还需通过具有线性激活的全连接层，然后是通过 softmax 层得到每个词向量的概率，选择概率最高的词向量作为输出。

除了编码器部分和解码器部分，Transformer 还添加了位置编码(Position Embedding)用来保留有关序列中单词顺序的信息。使用位置编码有两种方法，第一种是用不同频率的三角函数公式直接计算，第二种是随机初始化，通过学习不断更新，通过实验中得到两种方法效果相近，因此本研究采用了第一种。

本研究以单一英文为例的 Transformer 模型示意图如图 1 所示，输入为字素序列，输出为预测的音素序列。

Figure 1. Schematic diagram of the Transformer model using a single English example
图 1. 以单一英文为例的 Transformer 模型示意图

本文研究的是在交叉混合条件下的基于 Transformer 架构的字音转换，因此需要通过多次训练对模型进行动态调整，使它在多个语言下的 PER 值和 WER 值都较低。多语种交叉混合条件下的整体流程图如图 2 所示，由于该模型针对的是多语种，在利用训练好的模型进行字音转换前，对待转换的文本需要进行语种识别的处理。通过 Python 语言自带的语种字符识别，韩语是在\uac00-\ud7ff 范围内，日语则是在\u30a0-\u30ff\u3040-\u309f 范围内，但是由于日语中可能出现中文字符，为了避免这种情况，日语在在语种识别前将需要将日语文本均转为片假名。然后利用已经训练好可适用于英日韩的 Transformer 模型将处理好的多语种文本进行字音转换，最后输出转换结果。

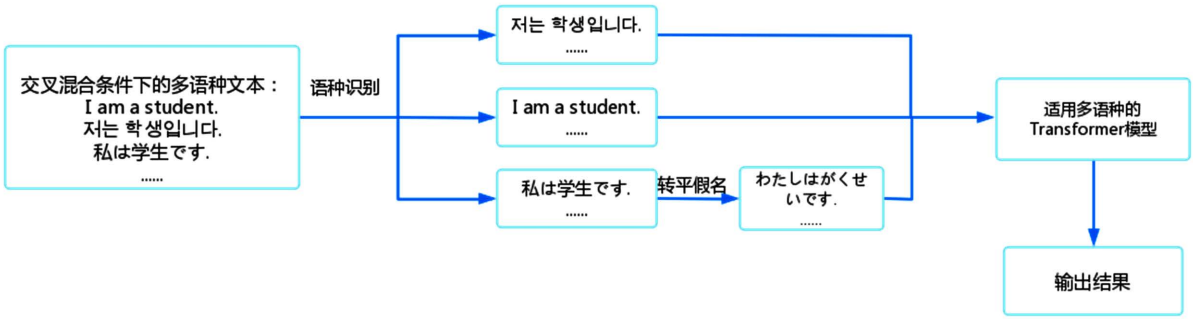

Figure 2. Overall flowchart under multilingual cross-mixing conditions
图 2. 多语种交叉混合条件下的整体流程图

3. 实验结果及分析

3.1. 数据集选取及预处理

对于英语的字音转换，研究者经常选择的数据集是 CMU 发音词典(也称为 CMUDict) [10]、NetTalk 数据集[11]和 Pronlex 数据集[12]。CMU 发音词典是由卡内基梅隆大学(CMU)创建的公共领域发音词典，是为 ASR 而开发的，它定义了从英语单词到其北美发音的映射，通常用于语音处理应用程序。包含超过 134,000 个单词形式及其发音。NetTalk 数据集是在 1986 年创建的一个基于神经网络的翻译系统，是为了研究书面英语(以字素或字母为单位)和口语(以音素为单位)之间的翻译过程。该数据集是根据字母对音素手动对齐的，包含 20,008 个对齐的字母和带重音的语音表示。Pronlex 数据集主要是为语音识别而开发的，最新版本的 Pronlex 数据集包含 90,988 个词条，涵盖 WSJ30、WSJ64、Switchboard 和 CALLHOME 英语。(WSJ30K 和 WSJ64K 是从最近的 ARPA 连续语音识别语料库中使用的几年华尔街日报文本中选择的单词列表。Switchboard 是一个 300 万字的关于各种主题的电话通信语料库。)

本文在对这三个数据集进行选择时，对比了利用不同模型对三个数据集在进行字音转换时的评估结果。如表 1 所示，可以看到无论哪一种模型，在 CMU 发音词典上的 PER 值和 WER 值都低于 NetTalk 数据集和 Pronlex 数据集，因此本文选择了 CMU 发音词典。

Table 1. Results on CMUDict, NetTalk, and Pronlex
表 1. 在 CMUDict、NetTalk 和 Pronlex 上的结果

数据集	模型	PER (%)	WER (%)
CMUDict	Joint sequence model [13]	**5.88**	**24.53**
	Bi-directional LSTM [14]	**5.45**	**23.55**
	Transformer 4 × 4 [5]	**5.23**	**22.1**
NetTalk	Joint sequence model [13]	8.26	33.67
	Bi-directional LSTM [14]	7.83	30.77
	Transformer 4 × 4 [5]	6.87	29.82
Pronlex	Joint sequence model [13]	6.78	27.33
	Bi-directional LSTM [14]	6.51	26.69

对于韩语和日语，SIGMORPHON 2021 G2P 任务[15]在来自于维基百科的韩语及日语数据集上取得的不错的效果，但是这两个数据集在 Transformer 模型上评估时，效果并没有比在 SIGMORPHON 2021 G2P 任务上有所改善。因此，本文研究对 SIGMORPHON 2021 G2P 任务中的韩语、日语数据集进行数据扩充。

SIGMORPHON 2021 G2P 任务使用 1 万条字音数据对用于训练，验证和测试，本文每增加 1000 条数据对就进行一次对比试验，韩语数据集的 PER 值平均每次降低 0.1%左右，WER 值平均每次降低 0.4%左右；日语数据集的 PER 值平均每次降低 0.05%左右，WER 值平均每次降低 0.15%左右。韩语数据集增加了接近 4300 条字音数据对，日语数据集增加了接近 1 万条字音数据对，此时模型效果最好，再增加数据对时，PER 值和 WER 值不降反增。

因此将扩充了 4300 条数据对的韩语数据集和扩充了 1 万条数据对的日语数据集用于本文的研究。文献[16]提到将韩国音节字符分解为单音字符，例如，가감→ㄱㅏㄱㅏㅁ，可以显著降低韩语的 WER 和 PER。因此本文中利用 hangul-jamo 将韩语数据集中的音节字符分解为单音字符。

3.2. 实验设置

3.2.1. 激活函数设置

字音转换的研究时大多数都选用 RELU 激活函数，而本文采用 GELU 激活函数(公式如下式(1)所示)。本文在对比了在这两种激活函数的条件下，Transformer 4 × 4 (4 个编码器和解码器层)模型在三种语言下的 PER 值和 WER 值，在表 2 中可以看到，RELU 激活函数条件下的 PER 值和 WER 值相比 RELU 激活函数都有所降低，模型性能也更好。

$$GELU\left(x\right)=0.5x\left(1+\tanh\left(\sqrt{2/\pi}\left(x+0.044715x^{3}\right)\right)\right) \tag{1}$$

Table 2. Results on RELU and GELU
表 2. 在 RELU 和 GELU 上的结果

语种	激活函数	PER (%)	WER (%)
英语	RELU	5.65	23.58
	GELU	**5.46**	**21.91**
韩语	RELU	2.82	17.13
	GELU	**2.74**	**16.80**
日语	RELU	1.65	6.45
	GELU	**1.58**	**6.20**

3.2.2. 模型参数设置

模型的相关参数设置如表 3 所示。本文研究了三种 Transformer 架构，其中有 3 个编码器和解码器层(在下文中称为 Transformer 3 × 3)，4 个编码器和解码器层(在下文中称为 Transformer 4 × 4)，5 个编码器和解码器层(在下文中称为 Transformer 5 × 5)。本文在所有提出的模型中都采用了 h = 4 平行注意层，Q、K 和 V 具有相同的 d_m 维数，因此 $d_k = d_v = d_m = 128$ 和 $d_m/h = 32$。模型训练时使用的损失函数是交叉熵损失函数(Cross Entropy Loss)，使用的优化器为 Adam，进行预测时使用搜索策略为束搜索(beamsearch)，束宽(beamsize)设置为 5。

Table 3. Model parameters
表 3. 模型参数

参数	数值
编码器层数	3/4/5
解码器层数	3/4/5
Dropout	0.2
Batch size	256
学习率	0.001
平滑率	0.1
多头注意力的头数	4
嵌入层隐藏节点数	512

3.3. 实验结果及分析

3.3.1. 评价指标

性能指标方面，使用音素错误率(PER)和单词错误率(WER)来评估文本数据上的 G2P 模型的质量。

音素错误率：用于测量预测的音素序列与参考音素之间的距离除以参考音素中的音素数量。编辑距离(也称为 Levenshtein 距离[17])是将一个序列转换为另一个序列所需的插入(I)、删除(D)和替换(S)的最小数值，可以通过动态规划方法来计算。如果参考的数据中一个单词有多个发音变体，则使用与候选单词的 Levenshtein 距离最小的变体。单词错误率：单词错误的数量除以单词的总数，只有在预测的发音与任何参考发音不匹配时才进行计算。WER 和 PER 越低，代表模型性能越好。

3.3.2. 实验结果

训练模型后，在测试数据集上进行预测。对三个数据集上的评价结果见下表 4。第一列显示了语种的类别，第二列和第三列分别显示了数据集和应用的模型架构，剩余两列分别显示了 PER 值和 WER 值。由结果可以看出，在 Transformer 4 × 4 架构下，三个语种的模型效果最好，无论在 PER 值还是 WER 值方面都低于 Transformer 3 × 3。但是同样可以看到增加编解码器层模型性能却没有得到改善，Transformer 5 × 5 模型下的 PER 值和 WER 值高于 Transformer 4 × 4 甚至比 Transformer 3 × 3 还高，这可能是因为增加编解码器层的数量会导致更多的训练参数。

Table 4. The PERs and WERs using Transformer model in three languages
表 4. 三个语种条件下使用 Transformer 模型的 PER 和 WER

语种	数据集	模型	PER (%)	WER (%)
英语	CMUDict	Transformer 3 × 3	5.56	22.98
		Transformer 4 × 4	**5.46**	**21.91**
		Transformer 5× 5	5.63	23.73
韩语	SIGMORPHON 2021 G2P 数据扩充	Transformer 3× 3	2.96	18.27
		Transformer 4 × 4	**2.74**	**16.80**
		Transformer 5 × 5	3.16	18.60
日语	SIGMORPHON 2021 G2P 数据扩充	Transformer 3 × 3	1.63	6.55
		Transformer 4 × 4	**1.58**	**6.20**
		Transformer 5 × 5	1.65	6.75

3.3.3. 实验结果分析

本文将在文本交叉混合条件下基于 Transformer 4 × 4 的模型在 CMUDict 数据集上的结果与单一语种条件下其他深度学习模型在 CMUDict 数据集上的结果进行了比较。如表 5 所示，第一列显示了字音转换所用到的方法，第二列和第三列分别显示了 PER 值和 WER 值。

由结果可以看出，在交叉混合条件下的 Transformer 4 × 4 的模型的 PER 值相比单一语种条件下的 Transformer 4 × 4 的模型的 PER 值略高，但是都低于单一语种条件下其他方法；但是，交叉混合条件下的 Transformer 4 × 4 的模型的 WER 值都低于单一语种条件下的深度学习模型的 WER 值。这是因为在交叉混合条件下模型为了更好的适应多种语言，为了在多种语言的字音转换中都具有更好的转换效果，所以在模型统一参数后结果会有所不同。

Table 5. Results on the CMUDict dataset
表 5. 在 CMUDict 数据集上的结果

方法	PER (%)	WER (%)
Encoder-decoder LSTM [14]	7.63	28.61
Joint sequence model [13]	5.88	24.53
Joint maximum entropy (ME) n-gram model [18]	5.90	24.70
End-to-end CNN [19]	5.84	29.74
Encoder-decoder LSTM with attention [19]	5.68	28.44
Transformer 4 × 4 [5]	**5.23**	22.10
Transformer 4 × 4	5.46	**21.91**

本文将在文本交叉混合条件下的多语种和单一语种情况下 Transformer 4 × 4 方法的结果进行了对比，如表 6 所示。第一列显示了语种的类别，第二列显示了所需要对比的数据集，第三列显示了所需要对比的条件。剩余两列分别显示了 PER 值和 WER 值。由结果可以看出，对于韩语及日语来说，在交叉混合条件下以及对数据集进行数据扩充后，Transformer 4 × 4 模型无论在 PER 值还是 WER 值方面都低于单一语种，说明本研究对多语种语音合成的质量方面有着促进意义。

Table 6. The PERs and WERs in cross-mixed and monolingual conditions
表 6. 交叉混合和单一语种条件下的 PER 和 WER

语种	数据集	条件	PER (%)	WER (%)
英语	CMUDict	单一语种	**5.23**	22.1
		交叉混合	5.46	**21.91**
韩语	SIGMORPHON 2021 G2P	单一语种	3.29	18.87
	SIGMORPHON 2021 G2P 数据扩充	交叉混合	**2.74**	**16.8**
日语	SIGMORPHON 2021 G2P	单一语种	1.99	7.9
	SIGMORPHON 2021 G2P 数据扩充	交叉混合	**1.43**	**5.8**

4. 结论

在本文中，研究了 Transformer 架构在文本交叉混合条件下的多语种(英、日、韩)字音转换，并将其性能分别与基于 Transformer 架构的单一语种的方法进行了比较。对于不同层数的 Transformer 结构，在 CMUDict 以及对 SIGMORPHON 2021 字音转换任务上的日韩数据集进行扩充后的数据集上进行实验，通过评估 PER 和 WER，Transformer 4 × 4 架构的模型都具有较好的性能。在文本交叉混合条件下，基于 Transformer 4 × 4 架构的英、日、韩字音转换在音素错误率和单词错误率方面与基于 Transformer 4 × 4 架构的英、日、韩三个语言的单一语种相比都大大降低了。本文的研究对多语种语音合成的质量以及多语种语音合成的实用性方面都有着促进意义。在未来的研究中，可以实验更多的语种，合成涵盖多语种的语音。

基金项目

北京印刷学院博士启动基金，北京市教委科研计划资助(KM202110015003)，北京印刷学院校级科研计划项目(Eb202103)。

参考文献

[1] Rao, K., Peng, F., Sak, H., *et al*. (2015) Grapheme-to-Phoneme Conversion Using Long Short-Term Memory Recurrent Neural Networks. 2015 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), South Brisbane, 19-24 April 2015, 4225-4229. https://doi.org/10.1109/ICASSP.2015.7178767

[2] 胡伟湘, 徐波, 黄泰翼. 汉语语音韵律边界的检测和识别研究[C]//第六届全国人机语音通讯学术会议论文集. 北京: 中国中文信息学会, 2001: 39-42.

[3] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30: *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 6000-6010.

[4] Mnih, V., Heess, N. and Graves, A. (2014) Recurrent Models of Visual Attention. *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, Volume 2, 2204-2212.

[5] Yolchuyeva, S., Németh, G. and Gyires-Tóth, B. (2020) Transformer Based Grapheme-to-Phoneme Conversion. 20*th Annual Conference of the International Speech Communication Association*, Graz, 15-19 September 2019, 2095-2099. https://doi.org/10.21437/Interspeech.2019-1954

[6] Sutskever, I., Vinyals, O. and Le, Q.V. (2014) Sequence to Sequence Learning with Neural Networks. *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, Volume 2, 3104-3112.

[7] LeCun, Y., Boser, B., Denker, J.S., *et al*. (1989) Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, **1**, 541-551. https://doi.org/10.1162/neco.1989.1.4.541

[8] Mikolov, T., Karafiát, M., Burget, L., *et al*. (2010) Recurrent Neural Network Based Language Model. *Proceedings Interspeech*, Vol. 2, 1045-1048. https://doi.org/10.21437/Interspeech.2010-343

[9] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[11] Sejnowski, T.J. (1988) The NetTalk Corpus: Phonetic Transcription of 20008 English Words.

[12] Kingsbury, P., Strassel, S., McLemore, C., *et al*. (1997) CALLHOME American English Lexicon (PRONLEX). Linguistic Data Consortium, Philadelphia.

[13] Bisani, M. and Ney, H. (2008) Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, **50**, 434-451. https://doi.org/10.1016/j.specom.2008.01.002

[14] Yao, K. and Zweig, G. (2015) Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion. *INTERSPEECH* 2015, 16*th Annual Conference of the International Speech Communication Association*, Dresden, 6-10 September 2015, 3330-3334. https://doi.org/10.21437/Interspeech.2015-134

[15] Ashby, L.F.E., Bartley, T.M., Clematide, S., *et al*. (2021) Results of the Second Sigmorphon Shared Task on Multilingual Grapheme-to-Phoneme Conversion. *Proceedings of the* 18*th SIGMORPHON Workshop on Computational Research in Phonetics*, *Phonology*, *and Morphology*, August 2021, 115-125. https://doi.org/10.18653/v1/2021.sigmorphon-1.13

[16] El Saadany, O. and Suter, B. (2020) Grapheme-to-Phoneme Conversion with a Multilingual Transformer Model. *Proceedings of the* 17*th SIGMORPHON Workshop on Computational Research in Phonetics*, *Phonology*, *and Morphology*, July 2020, 85-89. https://doi.org/10.18653/v1/2020.sigmorphon-1.7

[17] Levenshtein, V.I. (1966) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, **10**, 707-710.

[18] Galescu, L. and Allen, J.F. (2002) Pronunciation of Proper Names with a Joint n-Gram Model for Bi-Directional Grapheme-to-Phoneme Conversion. 7*th International Conference on Spoken Language Processing*, Denver, 16-20 September 2002, 109-112. https://doi.org/10.21437/ICSLP.2002-79

[19] Yolchuyeva, S., Németh, G. and Gyires-Tóth, B. (2019) Grapheme-to-Phoneme Conversion with Convolutional Neural Networks. *Applied Sciences*, **9**, 1143. https://doi.org/10.3390/app9061143