Review

The Application and Development of Grapheme-Phoneme Conversion

Yi Shun and Rengaowa Sa *

Posted Date: 24 September 2025

doi: 10.20944/preprints202509.2016.v1

Keywords: grapheme-phoneme conversion; speech synthesis; automatic speech recognition

Review

The Application and Development of Grapheme-Phoneme Conversion

Yi Shun and Rengaowa Sa *

College of Computer Science and Technology, Inner Mongolia Normal University, Sheng Le Economic Park, Hohhot 010022, China

* Correspondence: ciecsrgw@imnu.edu.cn

Abstract

Grapheme-to-phoneme conversion aims to transform written forms into phonetic representations, holding significant application value in fields like speech synthesis and speech recognition. In recent years, methods based on pre-training paradigms and transfer learning frameworks have shown remarkable advantages in areas like low-resource language modeling and multilingual joint modeling. First, the historical development of G2P research is examined, analyzing the paradigm shift from early rule-based models to contemporary neural network models through three dimensions: interpretability modeling, mapping accuracy, and computational efficiency. Next , a horizontal comparison of state-of-the-art G2P methods based on attention mechanisms and multi-task joint learning is presented , highlighting the mapping accuracy of different models on the same public dataset.Then,the research hotspots in this field are systematically reviewed, and a theoretical development path is constructed based on technological evolution.Finally,three future research directions are proposed: integrating multimodal technologies, neural architecture search, and prompt learning paradigms, providing theoretical references to overcome existing technical bottlenecks.

Keywords: grapheme-phoneme conversion; speech synthesis; automatic speech recognition

1. Introduction

In the theory of writing systems, grapheme is defined as the smallest distinctive writing unit in the orthographic system, whose function is to establish the correspondence between phonological representation and visual symbol[1]. Phoneme is a physical speech unit divided based on acoustic features[2].Grapheme-to-Phoneme (G2P) is dedicated to building a mapping relationship between the smallest unit of written text, grapheme, and the sound unit of the speech system, phoneme. Its mapping function is expressed as:

$$f:\ G \rightarrow P \tag{1}$$

where f is the conversion function, which maps the grapheme sequence $G = (g_1, g_2, \ldots, g_n)$ to the phoneme sequence $P = (p_1, p_2, \ldots, p_3)$. This technology plays an important role in the field of speech information processing, and its value is mainly reflected in three aspects: First, G2P is a key part of the front-end text processing module in the text-to-speech synthesis system (TTS), and the mapping accuracy directly affects the naturalness of the synthesized speech. Secondly, in the field of Automatic Speech Recognition (ASR), G2P effectively expands the vocabulary coverage of the ASR system by generating phoneme sequences of new words, proper nouns, and unregistered words, and improves the adaptability of the ASR system to dynamic language environments. Finally, in speech simulation technology, precise G2P technology provides reliable underlying support for the acoustic modeling of speech signals, thereby enhancing the system's ability to represent speech features and the quality

 doi:10.20944/preprints202509.2016.v1

of generation, helping machines to more accurately simulate human speech and improve the understanding of speech signals.

The research on G2P began in the mid-twentieth century, and its research methodology system has undergone three major paradigm shifts: the initial deterministic modeling based on artificial linguistic rules (1950-1980s), the mid-term statistical probability model (1990s-2010s), and the current technical system dominated by deep learning architecture (2010s). Each paradigm shift stems from the limitations of previous methods in processing complex language phenomena. Against this background, in recent years, researchers have used complex neural networks such as Long Short-Term Memory (LSTM) [3] and Convolutional Neural Network (CNN) [4] models to perform G2P conversion and reconstruct the G2P technical path. In 2018, the Google research team proposed the Bidirectional Encoder Representations from Transformers (BERT) [5] model, which pioneered the "pre-training-fine-tuning" transfer learning paradigm. This method is a milestone in the G2P field. It effectively alleviates the three major problems faced by traditional methods in multilingual speech synthesis systems: the modeling dilemma of irregular phonology, the complexity of traditional TTS steps, and the generalization bottleneck in multiple languages. This paradigm of transfer learning marks that G2P has entered a new stage of scalability and generalization.

This paper aims to sort out the application fields and research method development in the field of G2P, and finally focus on the innovative breakthroughs of neural network architecture and pre-training models. The content is arranged as shown in Figure 1. First, from the perspective of speech engineering, the important role of G2P technology in the two major application fields of TTS and ASR was discussed; then, through a diachronic analysis, the methodological transformation of G2P research from symbolism to connectionism was carried out, and a comparative analysis of G2P methods based on deep learning under the same data set was conducted; finally, the two hot topics of different language dialects and multilingual G2P were discussed, and the future direction of the development of G2P technology was proposed, establishing a theoretical reference system for subsequent research.

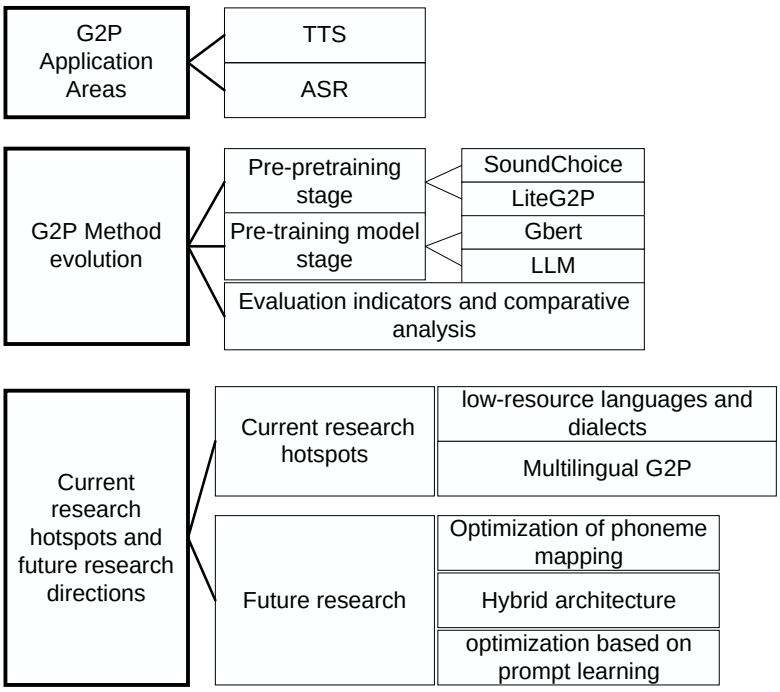

Figure 1. Structure diagram of this review.

2. G2P Application Areas

 doi:10.20944/preprints202509.2016.v1

As an important component of speech information processing technology, G2P plays a key role in the speech technology ecosystem. Its technical value is mainly reflected in the following three application dimensions: front-end processing of TTS, vocabulary expansion mechanism of ASR, and cross-modal alignment of multimodal speech retrieval[6].

2.1. Speech Synthesis System

In the technical framework of TTS, G2P assumes the core function of the grapheme-phoneme interface. Its essence is to provide the underlying representation for the parameter generation of the acoustic model by constructing a deterministic mapping from grapheme sequence to phoneme sequence (Figure 2). With the widespread application of TTS technology in intelligent voice interaction systems (such as virtual digital humans, IoT terminal devices, and barrier-free reading tasks), the performance requirements of the G2P module have evolved from basic intelligibility guarantees to multi-dimensional indicators such as accuracy, rhythmic naturalness, and emotional adaptability.

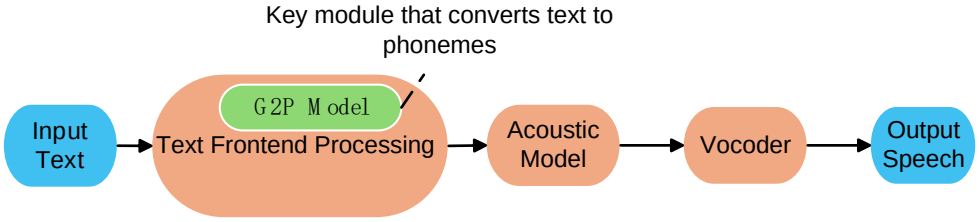

Figure 2. Structure Diagram of a Deep Learning-based Speech Synthesis System.

The evolution of TTS technology can be divided into three stages. The first stage is the synthesis stage based on unit selection and waveform splicing (1980s-2000s) [7,8] (such as the Diphone system). Its idea is to rely on a large-scale pre-recorded speech database to splice the target phoneme units in the time domain through a dynamic programming algorithm. This method has defects such as obvious splicing traces and exponential growth of database storage in synthesized speech, so it has been replaced [9]. The second stage is the speech synthesis technology based on statistical parameters, which has achieved the transformation of speech generation models [10–13]. Its technical path can be decomposed into three layers: first, in the front-end text analysis layer, the G2P module outputs the phoneme sequence and generates prosodic features through context-dependent modeling; then, in the acoustic parameter generation layer, the acoustic parameters such as Mel-Frequency Cepstral Coefficients (MFCC) are synthesized based on the maximum likelihood parameter generation algorithm [14]; finally, in the waveform reconstruction layer, the time domain signal is reconstructed through the Griffin-Lim algorithm [15], such as the typical Hidden Markov Model (HMM)-based speech synthesis system (HMM-based Text-to-Speech, HTS) [14]. Third, under the speech synthesis technology driven by deep learning, the end-to-end architecture based on neural networks is used to achieve direct mapping from text to waveform (2010s to present). The end-to-end architecture represented by WaveNet [16] and Tacotron [17] has completely reconstructed the TTS technology stack. Its system architecture is usually composed of three core modules: linguistic front-end processing module, neural acoustic model and neural vocoder [9], as shown in Figure 2. The front-end processing module is responsible for text normalization, among which G2P, as a key submodule, needs to complete complex functions such as word boundary detection and polyphone disambiguation. Acoustic models (such as FastSpeech2) [18] map phoneme sequences to mel-spectrogram features through an encoder-decoder architecture. Vocoders (such as WaveGlow) [19] convert mel-spectrograms into 24KHZ high-fidelity waveforms through a reversible flow model

 doi:10.20944/preprints202509.2016.v1

(Glow-based Model), and their signal-to-noise ratio is improved by 14.6dB compared to the traditional Griffin-Lim algorithm.

2.2. Automatic Speech Recognition

As an interdisciplinary field of computational linguistics and speech signal processing, the essence of ASR technology is to construct a mapping function $f: X \to Y$ from a time-domain speech signal $X(t)$ to a discrete symbol sequence $Y = \{y_1, y_2, \ldots, y_n\}$ where X is the speech signal space and Y is the text space consisting of a finite character set. The engineering application of ASR technology has penetrated into multiple dimensions of human-computer interaction, such as the virtual assistant Siri, electronic medical record transcription in the medical field, and automatic TOEFL oral scoring.

The technical evolution of ASR can be divided into four iterative stages, and its methodological transformation reflects the coordinated development of computational linguistics and machine learning theory. Early ASR systems were based on dynamic time adjustment algorithms and used fixed template matching strategies to achieve isolated word recognition (1950s-1980s). Representative systems include IBM's "shoebox" [20]. The contribution of this stage is the establishment of a basic framework for speech time domain alignment, which laid the foundation for subsequent probabilistic modeling. In the statistical modeling stage (1990s-2010s), the joint architecture of HMM and Gaussian Mixture Model (GMM) [21] became the mainstream paradigm. The mathematical form can be expressed as:

$$P(O|W) = \prod_{t=1}^{T} \sum_{s \in S} a_{s_{t-1}s_t} \cdot N(o_t; \mu_s, \Sigma_s) \quad (2)$$

where $a_{s_{t-1}s_t}$ is the state transition probability, and N represents the GMM emission probability. Limited by the ability of GMM to represent nonlinear acoustic features, the rise of the deep learning paradigm was born. In the neural hybrid architecture stage (early 2010s), the deep neural network (DNN) replaced GMM to construct the HMM-DNN hybrid model [22]. Its hidden layer activation function $h_l = \sigma(W_l h_{l-1} + b_l)$ significantly improved the acoustic modeling capability. From the late 2010s to the present, it is in the end-to-end architecture stage. The end-to-end model based on sequence-to-sequence (seq2seq) modeling abandons the traditional acoustic-language model separation architecture for direct modeling, see Equation 3.

$$P(Y|X) = \prod_{t=1}^{T} P(y_t|y_{1:t-1}, X) \quad (3)$$

where X is the speech feature, Y is the target text [23], and $P(y_t|y_{1:t-1}, X)$ represents the conditional probability of generating y_t at the current time under the condition that the generated prefix sequence $y_1, y_2, \ldots, y_{t-1}$ is input into X. Nevertheless, G2P still has independent research value,

The necessity stems from the following technical demands: First, the data-intensive training requirements of the end-to-end ASR system conflict with the low-resource language scenario. The G2P model can help alleviate this dependence on large-scale data. Second, for the black-box characteristics of the end-to-end system, G2P can provide more fine-grained control and debugging. Third, for fields with special G2P rules such as medical terminology and scientific and technological vocabulary, the G2P model can be optimized.

Previously, self-supervised pre-training models (such as Wav2Vec 2.0[24] and HuBERT[25]) achieved speech representation learning by contrasting the learning objective function. Thanks to the self-attention mechanism and pre-training technology, ASR efficiency and accuracy have been further improved, which points out a new direction for the ASR field. In the end-to-end ASR architecture (Figure 3[26]), the technical implementation path can be formally described as:

 doi:10.20944/preprints202509.2016.v1

$$\hat{Y} = f_{post}(f_{dec}(f_{enc}f_{feat}(X))) \quad (4)$$

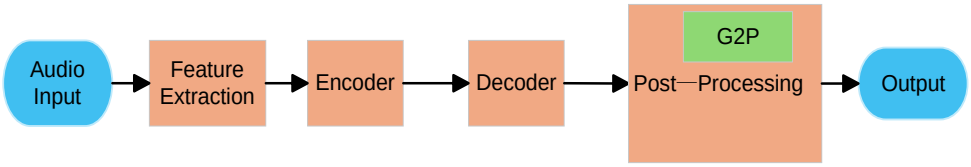

Figure 3. End-to-end speech recognition flow chart.

The input of the system $X \in R^T$ is a continuous speech signal, usually a time-series audio waveform. It is converted into a time-frequency domain representation by the feature extraction module, and is typically implemented as a Mel-filterbank spectrogram [27]. This process is implemented through short-time Fourier transform [28] and Mel-scale mapping, and the expression is:

$$F_{d,t} = \sum_{k=0}^{K-1} |X_t[k]|^2 \cdot M_d \ (k) \quad (5)$$

where M_d (k) is the number of the dth Mel filter group. The encoder f_{enc} extracts context-sensitive representations of speech through hierarchical feature abstraction. Mainstream implementations include convolutional temporal encoders [29] and Transformer encoders [30]. The decoder f_{dec} implements the probabilistic mapping from acoustic features to character sequences. Technical paths include the Connectionist Temporal Classification (CTC) decoding method: solving the optimal alignment path through dynamic programming [31]. The post-processing enhancement module f_{post} improves the output quality (such as correction, word segmentation, and punctuation) by integrating linguistic constraints. The G2P component in its module can convert the output graphemes of ASR into phoneme sequences to verify whether they meet the expected pronunciation (such as polyphones and proper nouns). It can also infer the pronunciation of unregistered words through G2P to assist in generating more reasonable text.

3. G2P Method Evolution

The G2P methodology system has undergone three major paradigm shifts: modeling based on artificial linguistic rules (1950-2000s), modeling based on statistical probability (2000s-2010s), and the current technology system dominated by deep learning architecture (2010s). The brief history of its development is shown in Figure 4.

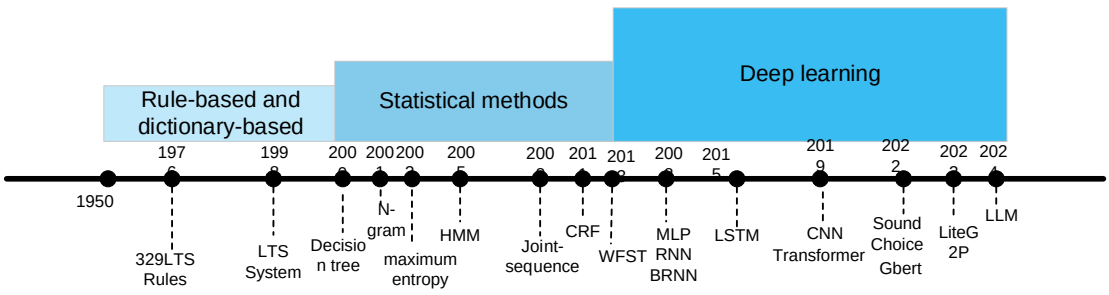

Figure 4. A brief history of G2P technology development.

3.1. Pre-Pretraining Stage

Early G2P research was based on generative phonology theory and focused on the symbolic encoding of linguistic rules (1950s-1990s), such as the English Letter-to-Sound (LTS) rule system [32], alignment through manual seed algorithms, and decision tree technology [33]. The rule system

 doi:10.20944/preprints202509.2016.v1

approach is limited by the completeness of linguists' prior knowledge. The expansion of speech databases (such as CMUdict) and the maturity of Bayesian statistical theory prompted G2P research to enter the stage of statistical methods (1990s-2010s), such as the G2P mapping method based on decision trees [34], the bidirectional G2P conversion method (G↔P) based on the joint n-gram model [35], the G2P model based on conditional maximum entropy and joint maximum entropy n-gram [36], the G2P conversion method based on HMM [37], the G2P conversion method based on the joint sequence model (Joint-sequence) [38], the G2P method based on conditional random field (CRF) [39], and the joint n-gram model conversion method based on failure transition (φ-transitions) [40].

The transition of the two methods presents the following significant features: first, the methodology changes from symbolism to statistics; second, from decision trees, n-grams to HMM, CRF, WFST, the model gradually enhances the modeling capabilities of long-distance dependencies, many-to-many alignment, and context sensitivity; finally, the core contradiction changes from being limited by expert knowledge to data sparsity (such as unregistered words and low-frequency words), and further to model efficiency and robustness. This technological change has promoted breakthroughs in downstream tasks such as speech synthesis and multi-language processing, and provided a historical mirror for the development of subsequent Neuro-Symbolic AI hybrid models.

The deep learning-based G2P technology has gradually established its dominance since 2010. Its core advantage lies in effectively capturing the nonlinear laws of letter-sound mapping through deep learning models. This technology can be traced back to 2008 when Beatrice Bilcu used three types of neural networks: multilayer perceptron (MLP), recurrent neural network (RNN), and bidirectional recurrent neural network (BRNN) to perform G2P tasks [41]. In 2015, Kanishka Rao et al. achieved significant progress in the G2P task by combining the Deep Bidirectional Long Short-Term Memory - Connectionist Temporal Classification (DBLSTM-CTC) model with CTC [3]. In the same year, Kai Yao's team systematically explored the performance differences of three neural network architectures, encoder-decoder LSTM, unidirectional LSTM and bidirectional LSTM (Bi-LSTM), in the G2P task. In 2019, S. Yolchuyeva's team successively proposed a hybrid architecture based on deep residual convolution [4] and a G2P method based on the Transformer structure. Although deep learning methods are significantly better than statistical models in terms of accuracy, they still face challenges such as homophone disambiguation and a surge in model parameters. To address these two challenges, researchers have proposed the following two model methods.

3.1.1. SoundChoice

In 2022, Alexey Ploujnikovt's team proposed a new G2P model called SoundChoice[42], with a structure as shown in Figure 5[42]. The grapheme-level features are encoded into continuous vectors through a trainable lookup table (green in the lower left corner of Figure 6), and are concatenated with the word-level semantic embedding of the pre-trained BERT (blue in the upper left corner of Figure 6) (pink in the middle left corner of Figure 6) to form a composite representation that combines character details and semantic context. The LSTM-based encoder (light green in the lower middle corner of Figure 6) extracts multi-level contextual information through bidirectional temporal modeling, and combines the gated recurrent unit (GRU) decoder and attention mechanism (light blue in the upper middle corner of Figure 6) to achieve accurate phoneme sequence generation. In terms of training strategy, a hybrid loss function was innovatively constructed. The encoding end uses the CTC loss function (middle right of Figure 6) to enhance the alignment capability, while the decoding end combines the negative log-likelihood loss with the homonym-specific loss to significantly improve the semantic disambiguation capability. Finally, a hybrid beam search mechanism using CTC and final prediction was adopted (bottom right of Figure 6). By integrating BERT word embedding and designing a new loss function, the model uses sentence context to improve the disambiguation accuracy of homographs, thereby improving the pronunciation accuracy in the speech synthesis system.

 doi:10.20944/preprints202509.2016.v1

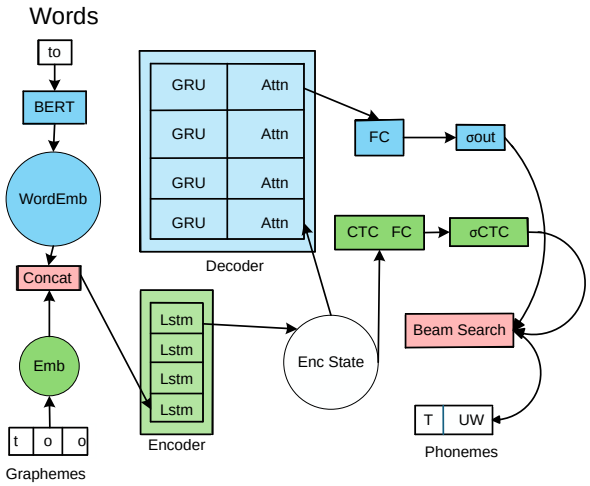

Figure 5. SoundChoice Encoder-Decoder structure diagram.

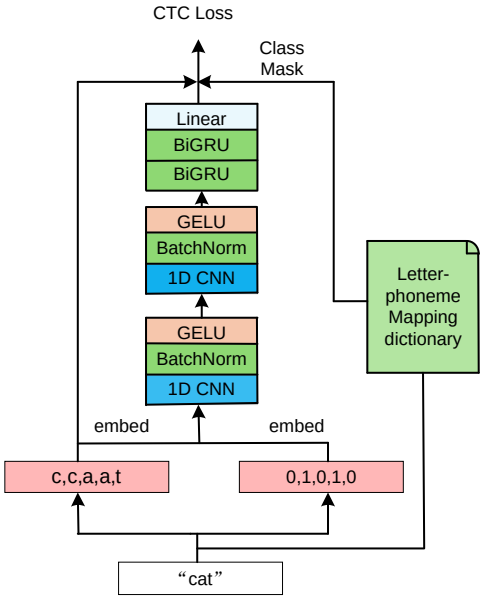

Figure 6. LiteG2P model structure diagram.

3.1.2. LiteG2P

In 2023, ZChen Wang et al. proposed a new G2P conversion model called LiteG2P[43], see Figure 6[43]. This model has a breakthrough design of a dual feature enhancement mechanism: first, the input letters are processed by phoneme length expansion, and each letter is expanded to the corresponding maximum phoneme mapping length to ensure the complete mapping of the phoneme sequence; secondly, the local position embedding technology GRU (pink in the lower right corner of Figure 6) is introduced to effectively solve the positioning ambiguity problem of repeated letters. After that, the letter position encoding vector (embed in the lower right corner of Figure 6) and the expanded letter embedding vector (embed in the lower left corner of Figure 6) are concatenated as a new feature vector and input into the subsequent feature extraction module (CNN and BiGRU in Figure 6). A binary mask matrix (0 or 1) is generated from a lightweight expert dictionary to suppress unreasonable phoneme mappings, and a pure CTC loss function is used to optimize the alignment process, which improves the speed and accuracy of the G2P task while keeping the model lightweight, making it suitable for cloud and mobile devices. Experiments were conducted on the CMUdict dataset, and the results showed that the LiteG2P model outperforms existing CTC-based methods

 doi:10.20944/preprints202509.2016.v1

with 10 times fewer parameters, and is similar to the Transformer-based Seq2Seq model with fewer parameters and less computation.

3.2. Pre-Training Model Stage

With the innovative breakthroughs in the Transformer architecture and its derivative models (such as BERT[5] and GPT), the transfer learning mechanism based on pre-training and fine-tuning has injected new development momentum into G2P technology. This technical paradigm captures deep phonological rules across languages through self-supervised pre-training, and then achieves knowledge transfer through task-oriented fine-tuning, which significantly improves the generalization ability and data utilization efficiency of the model. For example, in 2021, Y. Jia proposed the BERT model (Phonemes and Graphemes PnG BERT)[44], which takes the phoneme and grapheme representation of the text as input. By extending the BERT architecture and innovating the pre-training strategy, the model can learn the joint representation of phonemes and graphemes on a large-scale text corpus and fine-tune it in the TTS task. M. Řezáčková et al. used the pre-trained Text-to-Text Transfer Transformer (T5) model for G2P conversion in 2021 [45]. The T5 model was fine-tuned on English and Czech to adapt to the G2P tasks in these two languages. The T5 model achieved excellent performance on the G2P tasks in both English and Czech, especially in dealing with long words and homographs. Pre-training models have become the mainstream technical paradigm for current natural language processing and speech tasks due to their advantages such as low core data support, improved training performance, and strong cross-task and cross-language generalization capabilities. The paper focuses on the following pre-training methods

3.2.1. GBERT

In 2022, Dong Lu et al. proposed a BERT-inspired pre-trained grapheme model (Grapheme BERT, GBERT) [46], which uses only grapheme information and is trained on a large-scale specific vocabulary through self-supervised learning. The authors proposed two methods to integrate GBERT into the existing Transformer-based G2P model. The first method is to replace GBERT with the encoder in the original Transformer model and train the new model in an end-to-end manner. The second method, as shown in Figure 7 [46], integrates GBERT into the Transformer-based G2P model through an attention mechanism. In each encoder and decoder layer, additional GBERT-Enc and GEBRT-Dec attention modules are added to adaptively control the interaction between these layers and the GBERT representation. A dropout network is used in the structure to regularize network training. Experimental results show that under medium-resource and low-resource data conditions, the G2P model based on GBERT can improve performance, and the fusion of GBERT through the attention mechanism can reduce WER and PER compared with the Transformer model.

 doi:10.20944/preprints202509.2016.v1

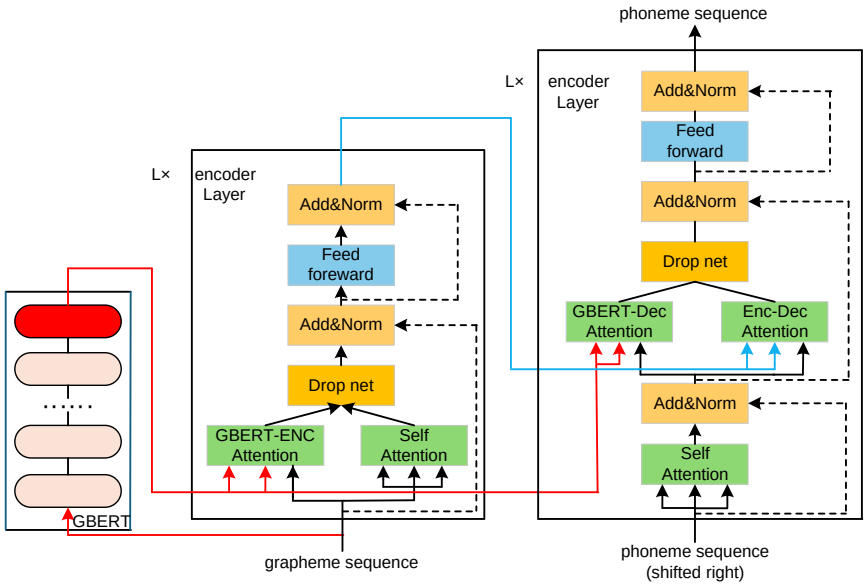

Figure 7. The architecture of GBERT-fused model.

3.2.2. LLM

Large Language Model (LLM) has become a research hotspot in the field of natural language processing due to its massive knowledge reserve, powerful language understanding and generation capabilities, and continuously optimized adaptive characteristics.

M.F. Qharabagh (2024) conducted a study on Persian and compared the phoneme transliteration performance of 9 large language models (LLMs), including LLaMA 3.1 and GPT-4 [47]. The study found that the optimal model LLaMA-3.1 (version 405B) achieved an accuracy of 54.00% in polyphone transliteration, significantly better than traditional models (such as Epitran's PER of 47.53%). In addition, the author systematically explored a multi-stage prompting strategy: starting from the naive method of directly generating the International Phonetic Alphabet (IPA) (PER 31.61%), the author gradually introduced Finglish (Latinized) intermediate conversion, contextual learning examples, and dictionary prompts (embedded word phonetic options) and other techniques, successfully reducing the PER to 8.30%. The study also tested various combinations of these prompting strategies. For Persian, the most effective strategy is to combine dictionary prompts with LLM-based correction methods, see Figure 8. First, input a piece of text, and find the words in the dictionary to clarify the basic information of each word. Then give the operation instructions, ask to return the pronunciation content of the Persian text, and list the candidate pronunciations of some words, so that the model can choose the correct pronunciation. Input these contents into the big model, and the model processes them according to the instructions and candidate information. Finally, the model outputs the mapping result of graphemes to phonemes, completing the conversion process from input text to specific voice-character mapping.

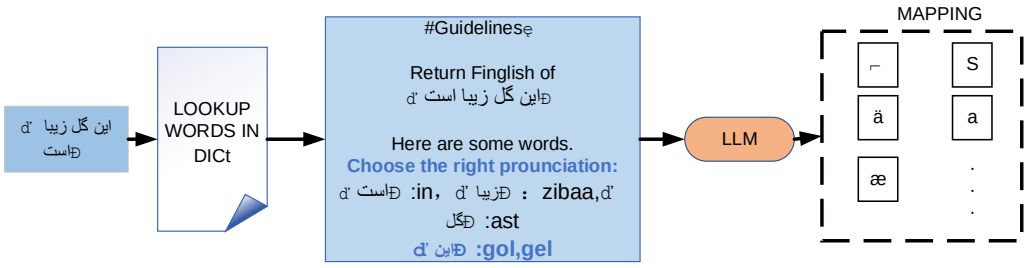

Figure 8. Schematic of the combined methods.

 doi:10.20944/preprints202509.2016.v1

In 2024, Dongrui Han et al. studied the problem of homonym disambiguation in G2P conversion [48] and proposed two prompt learning-based methods, namely one-shot prompt and in-context knowledge retrieval (ICKR). The experiment directly used GPT-4's one-shot prompt on the Librig2p dataset with a weighted average PER of 8.7%. Through fine-tuning models (such as Llama2-7B-chat and Gemma-2B-it), PER was significantly reduced to 5.5% and 5.2%. The latter ICKR system uses GPT-4 as an "AI linguist" to analyze the target word and its context in the sentence through prompts and retrieve the most matching pronunciation in the dictionary[48]. It achieves the lowest weighted average PER (4.9%) and the highest homograph accuracy (95.7%) on the Librig2p dataset, which shows that the latter method is superior to the former.

The ICKR algorithm flow is shown in Figure 9. The core is to dynamically match the correct pronunciation of polyphones in a specific context through semantic analysis of structured dictionaries and GPT-4. First, the input sentence is split into word sequences, and then the target word is detected to see if it is in the predefined homograph dictionary (the yellow color in the lower left of Figure 9). If it is not in the dictionary, the designed prompt is used to let the LLM directly generate the phoneme sequence of the word in combination with the sentence context of the target word (the dotted line on the left of Figure 9); If it is in the dictionary, it will continue to detect whether it is a homonymous word (yellow in the lower right corner of Figure 9). If it is a homonymous word, it will enter the LLM semantic analysis module (GPT-4), otherwise it will directly retrieve the unique phoneme (dashed line on the right of Figure 9). The LLM semantic analysis module (green in Figure 9) passes the target word and the sentence it is in to GPT-4, requesting analysis of its contextual meaning, part of speech, and related semantic information, and compares the output semantic information with the dictionary entries one by one to find the most matching entry. Each word in the homonym dictionary contains multiple sub-entries, each of which corresponds to a unique meaning, part of speech, and related semantic information. Finally, the target phoneme sequence is output.

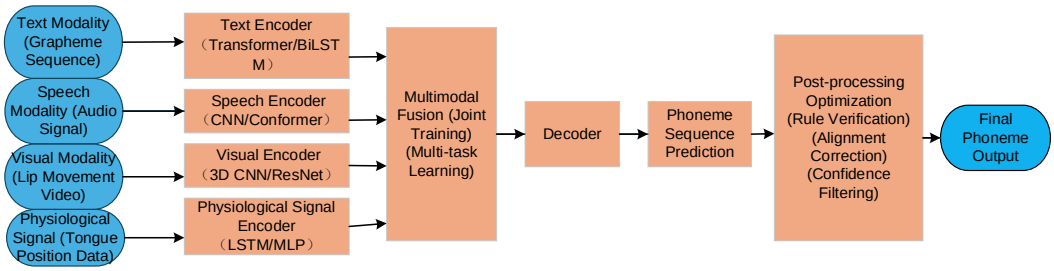

Figure 9. Multimodal G2P system structure diagram.

3.3. Evaluation Indicators and Comparative Analysis

3.3.1. Evaluation Indicators

Phoneme Error Rate (PER) is a quantitative indicator for evaluating the performance of G2P systems. Its theory originates from the sequence alignment algorithm in the field of speech recognition. Its mathematical definition is the minimum edit distance (MED) between the predicted phoneme sequence and the reference phoneme sequence, also known as the normalized ratio of the Levenshtein distance to the length of the reference sequence [49]. It is formally expressed as:

$$\mathrm{PER} = \frac{\mathrm{D}\ (\mathrm{P_{pred}}, \mathrm{P_{ref}})}{\mathrm{N_{ref}}} \tag{6}$$

where P_{pred} represents the output phoneme sequence of the model, P_{ref} is the reference phoneme sequence annotated by experts, and $D\ (P_{pred},\ P_{ref})$ is the editing operation function based on Levenshtein distance. This function calculates the minimum number of editing operations required

 doi:10.20944/preprints202509.2016.v1

to achieve sequence alignment through a dynamic programming algorithm, where normalization processing enables PER to be applicable to phoneme sequence evaluation of different lengths.

The word error rate (WER) is the core evaluation indicator for measuring the prediction performance of the G2P model. It reflects the overall performance of the system mapping by calculating the percentage of words that do not completely match the predicted phoneme sequence with the reference pronunciation [49]. The mathematical definition is:

$$\mathrm{WER} = \frac{\mathrm{N_{errors}}}{\mathrm{N_{words}}} \times 100\% \quad (7)$$

N_{errors} is the number of incomplete matches between the predicted phoneme sequence and the reference pronunciation sequence, and N_{words} is the total number of words in the reference pronunciation. Compared with the PER indicator at the phoneme level, WER pays more attention to the accuracy of the overall vocabulary mapping. For example, when a phoneme error occurs in word prediction ("cat" [kæt] → [kæp]), the PER is 33.3%, and WER will be directly judged as 100%.

3.3.2. Comparative Analysis

This paper collects the standardized evaluation results (PER/WER) of the G2P field on the CMUdict benchmark dataset in recent years, and conducts a longitudinal comparative analysis according to the time sequence of technology evolution, as shown in Table 1. The "N/A" in the table indicates that the author of the original document did not provide the corresponding indicator measurement results of the model. The data in the table show that the performance of the model architecture based on deep learning (LSTM, Transformer, GRU) significantly exceeds that of traditional statistical methods. The integrated architecture of 5 global attention encoder-decoder models [50] achieves breakthrough performance of WER=20.24% and PER=4.69%, which is a relative improvement of 18.1% and 20.5% respectively compared with the optimal traditional model (WER=24.7%, PER=5.9%). This performance gain is due to the probabilistic fusion of multi-model decision space and the coordinated optimization of the attention mechanism.

The single-layer LSTM architecture is limited by the unidirectional constraint of time series modeling and cannot effectively capture reverse contextual dependencies, and its performance is usually poor. For example, the PER of uni-directional LSTM[51] is 8.22% and the WER is 32.64%. However, by stacking bidirectional LSTM layers to build a deep time series structure, the optimization effect of PER=5.45% and WER=23.55% is achieved in a 3-layer configuration[51], indicating that deeper network structures can better capture contextual information and thus improve performance. Transformer models also performed well, especially Transformer 4x4[49]. As can be seen from the table, after the introduction of the attention mechanism into the model based on the Encoder-Decoder architecture (such as Encoder-Decoder Bi-LSTM with attention layer), its performance has been significantly improved, with lower values, which means that the attention mechanism has indeed effectively improved the performance.

In addition, the use of a multi-model ensemble architecture [50] further improved the performance, indicating that ensemble learning is an effective method in G2P tasks. The Token-Level Ensemble Distillation model [50] performed relatively well, indicating that this ensemble distillation framework, combining the advantages of multiple models and unsupervised learning, can break through the limitations of traditional G2P methods and improve robustness. In addition, the LiteG2P-medium model [43] achieves WER = 24.3% with only 1.2M parameters, proving that the efficiency bottleneck of traditional models can be broken through through lightweight architecture innovation.

Table 1. Comparison of methods on CMUdict.

Method	PER (%)	WER (%)
Joint n-gram mode[38]	7.0	28.5
Joint n-gram mode[38]	5.9	24.7

 doi:10.20944/preprints202509.2016.v1

Joint maximum entropy (ME) n-gram mode[36]	5.88	24.53
Failure transitions for joint n-gram models and g2p conversion[40]	8.24	33.55
encoder-decoder LSTM[51]	7.54	29.21
encoder-decoder LSTM (2 layers) [51]	7.63	28.61
uni-directional LSTM[51]	8.22	32.64
uni-directional LSTM (window size 6) [51]	6.58	28.56
bi-directional LSTM[51]	5.98	25.72
bi-directional LSTM (2 layers) [51]	5.84	25.02
bi-directional LSTM (3 layers) [51]	5.45	23.55
DBLSTM-CTC 128 Units [36]	N/A	27.9
DBLSTM-CTC 512 Units[36]	N/A	25.8
LSTMwithFull-delay[3]	9.1	30.1
DBLSTM-CTC512Units[3]	N/A	25.8
8-gramFST[3]	N/A	26.5
DBLSTM-CTC512+5-gramFST[3]	N/A	21.4
Many-to-many alignments with deep BLSTM RNNs[52]	5.37	23.23
Ensemble of 5 [Encoder-decoder + global attention] models[50]	4.69	20.24
Encoder-decoder with global attention[50]	5.11	21.85
Encoder-decoder + local-m attention[50]	5.39	22.83
Encoder-decoder + local-p attention[50]	5.04	21.69
Deep Bi-LSTM with many-to-many alignment[52]	5.37	23.23
Combination of sequitur G2P and seq2seq-attention and multitask learning [53]	5.76	24.88
Encoder-Decoder GRU[54]	5.8	28.7
CNN with NSGD[55]	5.58	24.10
Transformer 3x3[49]	6.56	23.9
Transformer 4x4[49]	5.23	22.1
Transformer 5x5[49]	5.97	24.6
Encoder-Decoder LSTM[4]	5.68	28.44
Encoder-Decoder LSTM with attention layer[4]	5.23	28.36
Encoder-Decoder Bi-LSTM[2]	5.26	27.07
Encoder-Decoder Bi-LSTM with attention layer[4]	4.86	25.67
Encoder CNN, decoder Bi-LSTM[4]	5.17	26.82
End-to-end CNN (with res. connections)[4]	5.84	29.74
Encoder CNN with res. Connections,decoder Bi-LSTM[4]	4.81	25.13
Token-Level Ensemble Distillation with unlabeled source words[59]	4.60	19.88
Joint multi-gram + CRF[55]	5.5	23.4

 doi:10.20944/preprints202509.2016.v1

MTL(512×3,λ=0.2)[56]	5.26	22.96
r-G2P(adv)[57]	5.22	20.14
LiteG2P-medium[43]	N/A	24.3

4. Current Research Hotspots and Future Research Directions

4.1. Current Research Hotspots

The research hotspots in the G2P field have transitioned from rule-based and statistical models to technologies based on deep learning and pre-trained models, and are moving towards multilingual and cross-lingual directions. Pre-training and transfer learning, multilingual data sharing, and unsupervised learning will drive further innovation and development in this field in the future.

4.1.1. Challenges of Low-Resource Languages and Dialects

Factors such as language complexity, differences in phonetic rules, and scarcity of data resources have caused low-resource languages to face many challenges in the G2P field,as shown in Table 2. Many languages have complex many-to-many mappings between graphemes and phonemes, such as agglutinative languages such as Turkish and Finnish, and polysynthetic languages such as Inuktitut. To address this problem, Novak et al. proposed the WFST and n-gram grammar model in 2016 [58]. The use of WFST can efficiently handle complex input-output alignments, demonstrating its advantage in handling languages with complex morphological changes. The challenge of low-resource languages is particularly typical in the case of Mongolian, where annotated data is relatively scarce and it is difficult to train powerful neural network models. In response to this, Sun et al. proposed an integrated distillation method in 2019 to effectively compress multi-model knowledge into a small model[59], effectively improving model performance under data-scarce conditions. In addition, it is also possible to consider using a pre-trained model combined with transfer learning and fine-tuning on limited data, such as the PnG BERT model proposed by Yunhui Jia in 2021[44]. For languages with irregular mapping relationships such as Thai, Yamasaki used a neural regression model in 2022 to improve conversion accuracy[60]. Some graphemes have different pronunciations in different words, and phonemes cannot be inferred from spelling rules at all. This is also a major problem in languages where polyphones are common (such as English, Chinese, and Vietnamese). This requires the G2P model to accurately predict phonemes based on context. In this regard, the model proposed by Kim, Hwa-Yeon et al. in 2023 improved the phoneme prediction of polyphones by utilizing neighbor word information [61]. These research points indicate that future research on G2P conversion needs to consider language specificity and the application of model generalization capabilities under low-resource conditions.

Table 2. Research Methods for Some Low-Resource Languages and Dialects.

Language type	Existing problems	Method
Agglutinative laguages (Turkish, Finnish	Complex many-to-many mapping relationships	WFST + n-gram model
Polysynthetic language (Inuktitut)	Complex many-to-many mapping relationships	WFST + n-gram model
low resource language	Data resources are scarce	Pre-trained Model + Transfer Learning
Thai	Irregular mapping relationship	Neural regression model
Polyphonetic languages (English, Chinese, Vietnamese)	Different pronunciations in different contexts	context modeling

4.1.2. Multilingual G2P

 doi:10.20944/preprints202509.2016.v1

In the context of globalization, developing a G2P system that supports multiple languages and improving the performance of models on data-scarce languages through transfer learning and other means have become important research directions. In recent years, researchers have been committed to building cross-language representation models through transfer learning and other technical means, promoting the practical application of multilingual speech recognition and synthesis systems, and improving the phoneme conversion capabilities of low-resource languages. Thanks to the deep collaboration between the technology of large-scale pre-trained language models and the transfer learning technology system, this field has made significant progress. The representative research results are summarized as follows.

Ben Peters' team[62] took the lead in building a multilingual encoder-decoder framework based on the attention mechanism, proving that the shared parameter model can effectively utilize cross-language similarities to provide solutions for low-resource languages. David R. proposed the large-scale multilingual G2P system Epitran[63] in 2018. High-precision pronunciation conversion is achieved through rule-driven mapping files and post-processing technology, supporting 61 languages. Experiments show that in ASR tasks in 11 languages, Epitran significantly reduces WER (for example, Amharic from 58.6% to 57.2%), proving its effectiveness in low-resource scenarios.

Sokolov et al. (2020)[64] proposed a multilingual G2P model based on a bidirectional LSTM encoder-decoder, which distinguishes language rules by language ID and distribution vector, and achieved an average PER reduction of 7.2% on 18 languages (including low-resource languages), without degrading the performance of high-resource languages. In 2020, Mingzhi Yu et al. proposed a multilingual G2P model based on the attention mechanism Transformer architecture[65], using bytes as input representation to adapt to different grapheme systems. The authors used data from multiple European and East Asian regions to evaluate the performance of character-level and byte-level G2P. The experimental results show that the model using byte representation has a relative word error rate improvement of 16.2%-50.2% over the corresponding character-based model. In addition, the size of the byte-level model is reduced by 15.0%–20.1%.

In 2022, Jian Zhu et al. trained a large-scale multilingual G2P model based on ByT5 using a G2P dataset in about 100 languages[66]. Experiments show that in terms of multilingual G2P, ByT5 running on byte-level input significantly outperforms the token-based mT5[67] model. ByT5 can handle more than 100 languages, including Latin letters and East Asian languages (such as Chinese and Japanese), and improves the pronunciation prediction accuracy of low-resource languages through multilingual joint training. In the future, as the demand for low-resource languages grows and the generalization ability of models improves, research in this field will continue to deepen and further promote the inclusiveness of speech technology.

4.2. Future Research Directions

4.2.1. Optimization of Phoneme Mapping for Low-Resource Languages

The development of Seq2Seq models and pre-trained language models based on deep learning frameworks has significantly improved the performance of G2P systems in multilingual scenarios. However, due to the scarcity of labeled data and complex mapping rules, the performance improvement of low-resource languages still faces the following bottlenecks. First, the complexity of language morphology, agglutinative languages (such as Turkish) and polyglot languages (such as Inuktitut) have irregular mapping relationships between graphemes and phonemes; second, the lack of training data for low-resource languages limits the generalization ability of the model.

The following development strategies are proposed to address the above bottlenecks. In recent years, multilingual pre-trained models such as XLM-RoBERTa [68], mT5 [67], and ByT5 [66] have demonstrated excellent performance in NLP tasks. Their deep semantic representation space can effectively capture the phonological correlation characteristics between languages. Future research can explore the migration of such multilingual pre-trained models to low-resource language G2P tasks, further benefit low-resource G2P by performing zero-shot predictions on unseen languages, or

 doi:10.20944/preprints202509.2016.v1

provide pre-trained weights for fine-tuning, which helps the model converge to a lower PER than randomly initialized weights.

Applying multimodal fusion technology to low-resource language G2P can also optimize the phoneme mapping of low-resource languages. The introduction of auxiliary modalities (such as audio) can provide additional information and enhance the robustness of G2P. For example, as early as 2019, James Route and other scholars proposed that relying solely on IPA characters to represent pronunciation may not capture the subtle differences in speech. They innovatively introduced audio data as an auxiliary supervision signal and dynamically optimized the intermediate representation of source language characters through a multi-task learning mechanism to improve the accuracy and cross-language generalization ability of the G2P model [69]. Experimental results show that the multimodal model reduces PER to 2.46%, which is more than 65% lower than the unimodal model (7.05%), fully verifying the technical advantages of the multimodal model in enhancing robustness. In 2023, Manuel Sam Ribeiro et al. proposed using speech data to improve the G2P model for low-resource languages[70]. The authors first trained a baseline G2P model with a small amount of manually annotated pronunciation dictionaries, then used the model to generate pronunciation hypotheses for the target language speech corpus and combined it with multilingual data to train the phoneme recognition system. By decoding the target language audio, the forced alignment algorithm is used to find word boundaries and learn the pronunciation of new words. Finally, the learned pronunciation dictionary is combined with the original annotated data to retrain the G2P model. Experiments show that this method effectively improves the PER of the G2P system across languages and available data. Most current research focuses on the integration of text and audio modalities. In the future, this technology can integrate more modalities, such as text, speech, vision (such as lip movement) and even pronunciation physiological signals (such as tongue position, vocal cord vibration) to improve the model's ability to model pronunciation rules. See Figure 9. Each modality extracts features through an independent encoder, integrates features using joint training and multi-task learning strategies, and generates phoneme sequences through a decoder. In the future, it is also possible to solve the ambiguity or low resource problems of a single modality by building a multimodal dataset for low-resource languages.

4.2.2. Hybrid Architecture Innovation

The end-to-end model demonstrates high accuracy and strong generalization capabilities through a unified learning framework, especially when processing complex language mappings. However, it also has shortcomings, such as large data requirements, poor interpretability, and excessive reliance on hardware. Traditional G2P models such as the HMM model [37] and CRF [39] are mainly characterized by strong regularity and interpretability, and good adaptability to specific languages. However, they also have limitations, such as difficulty in processing complex languages, the need for manual design of features and rules, and high development costs.

The hybrid architecture that integrates the end-to-end model and the traditional rule-based approach can effectively break through the limitations of a single approach. In the Mongolian G2P conversion task, Zhinan Liu and other scholars innovatively proposed a collaborative mechanism of "rule priority + LSTM supplement", as shown in Figure 10[71]. This strategy first processes graphemes and phonemes with clear mapping relationships through the rule system, and then leaves the complex cases that the rule system cannot cover to the LSTM model for learning. Experimental data show that compared with the single LSTM model, the hybrid model achieves a significant decrease of 2.5% and 0.6% in the WER and PER indicators, respectively, which fully verifies the synergistic advantages of the high efficiency of the traditional rule system and the strong representation ability of the deep learning model.

 doi:10.20944/preprints202509.2016.v1

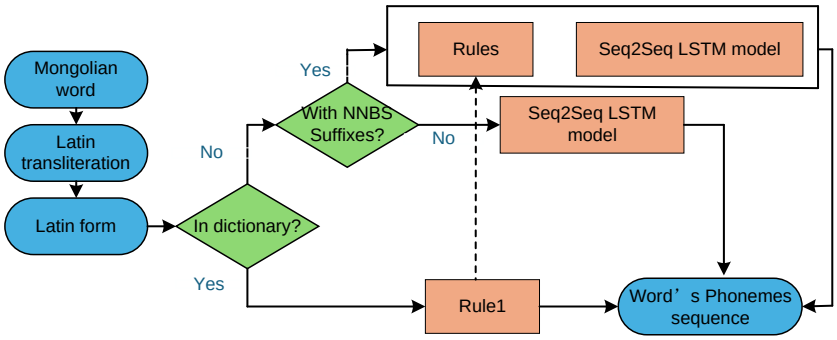

Figure 10. The structure of combination rules and Seq2Seq LSTM model.

The application of Neural Architecture Search (NAS) [72] technology in the G2P field is also worthy of special attention. This technology allows the system to dynamically select the optimal processing path based on the input features: for regular word-sound mapping, the traditional model is called; for complex nonlinear mapping, the end-to-end model is selected, and the performance of the selected model is evaluated and then the structure search is iterated. The process is shown in Figure 11. Currently, NAS has shown its application potential in NLP tasks such as text classification (Auto-Keras [73]) and question-answering systems [74], but it has not yet formed a complete theoretical system in the G2P field. Especially in low-resource scenarios, automatically optimizing the model structure through NAS to improve the accuracy and robustness of pronunciation prediction will become an important direction for technological breakthroughs. In summary, the deep integration of traditional models and end-to-end models has multi-dimensional value: first, it improves data utilization efficiency through complementary advantages, especially in scenarios where labeled data is scarce; second, the interpretability of the rule system and the strong fitting ability of the neural network form a benign complement; third, the hybrid architecture can enhance the model's adaptability to different language features. These characteristics collectively point to a research paradigm worthy of in-depth exploration - building an intelligent G2P system that is efficient, interpretable, and has strong generalization capabilities.

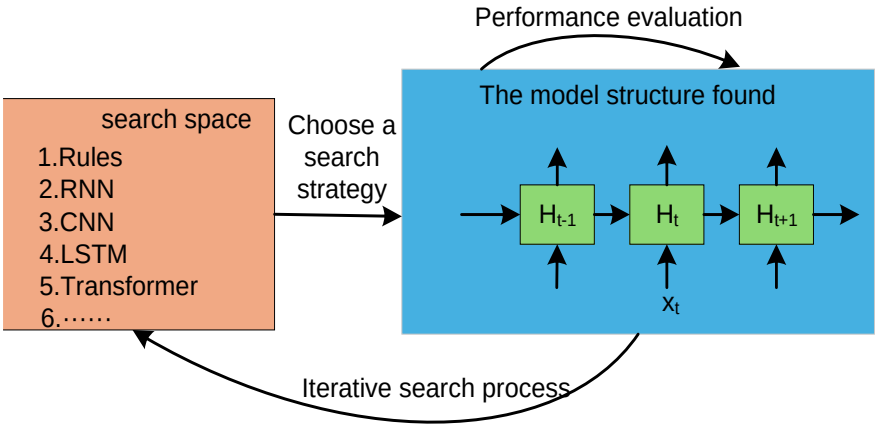

Figure 11. Main flow chart of network structure search.

4.2.3. Mapping Optimization Based on Prompt Learning

Prompt Learning is a cutting-edge method in the field of natural language processing. It guides pre-trained language models (such as BERT, GPT, T5, etc.) to efficiently complete downstream tasks by designing specific forms of natural language instructions or templates, without the need for large-scale structural adjustments or complete fine-tuning of the model. Although this paradigm is still in

 doi:10.20944/preprints202509.2016.v1

the exploratory stage in the G2P field, it has shown breakthrough potential in low-resource language scenarios[47] and polyphone disambiguation[48].For example, the research conclusion of reference [47] shows that LLM is significantly better than traditional models in context-aware tasks, especially in low-resource language scenarios. The experimental results of reference [48] show that the GPT-4 version achieves the lowest weighted average PER (4.9%) and the highest homograph accuracy (95.7%) on Librig2p. The fine-tuned model (such as Gemma-2B-it) performs stably in processing unregistered words, but its semantic understanding ability is still weaker than GPT-4.

The above experimental results show that prompt learning has opened up a new technical paradigm for G2P tasks: (1) Hierarchical prompt engineering can enhance the grapheme reasoning ability of large models; (2) Knowledge-enhanced prompts can narrow the performance gap with fine-tuned models; (3) The fusion application of intermediate representations (such as Finglish) and external knowledge bases provides innovative ideas for low-resource language processing. In the future, we can further explore how prompt learning can enhance the cross-language transfer ability of G2P models and the collaborative optimization path of lightweight models and prompt learning.

5. Conclusions

G2P aims to convert grapheme sequences into phoneme sequences, and its performance directly affects the performance of TTS, ASR and other application fields. This paper systematically reviews the evolution of G2P technology and proposes research hotspots and future directions based on the challenges faced by the G2P field. By combing the evolution of methodology, it is pointed out that the early rule-driven methods relied on manually designed linguistic rules, which were highly interpretable but had limited generalization capabilities; statistical modeling methods significantly improved generalization capabilities through data-driven optimization; deep learning technology (such as end-to-end neural networks) further broke through the performance bottleneck, especially in complex contexts and multilingual scenarios. Current research focuses on multilingual G2P and low-resource language performance improvement. Future directions emphasize the coordinated optimization of technology and applications: on the one hand, it is necessary to improve the robustness of models in low-resource scenarios and explore small sample learning and unsupervised methods; on the other hand, multimodal fusion (such as text-speech joint modeling) technology may promote the application of G2P in edge computing. In addition, prompt learning and hybrid architecture innovation (such as combining rules and neural models) also show potential, providing new ideas for adapting to new languages or dialects.

Author Contributions: Conception, writing —original, Y.S.; review and editing, Y.S., S.R.; supervision, project administration, funding acquisition, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Natural Science Foundation of Inner Mongolia Autonomous Region of China (2023LHMS06002), Research Program of Science and Technology at Universities of Inner Mongolia Autonomous Region (NJZY22568), Basic Scientific Research Business Expenses Project of Inner Mongolia Normal University(2022JBYJ033).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

G2P	Grapheme-to-phoneme

 doi:10.20944/preprints202509.2016.v1

TTS	Text-to-speech
ASR	Automatic Speech Recognition
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
BERT	Bidirectional Encoder Representations from Transformers
MFCC	Mel-Frequency Cepstral Co-efficients
GMM	Gaussian Mixture Model
DNN	Deep neeral network
HMM	Hidden Markov Model
CTC	Connectionist Temporal Classification
CRF	Conditional Random Field
GRU	Gated Recurrent Unit
LLM	Large Language Model
ICKR	In-context Knowledge Retrieval
WER	Word Error Rate
PER	Phoneme Error Rate

References

1. Meletis, D. The grapheme as a universal basic unit of writing. *Writing Systems Research* **2019**, *11(1)*, 26–49.
2. Fromkin, V.; Rodman, R.; Hyams, N. Phonemes: The Phonological Units of Language *An Introduction to Language*, 11th ed.; Cengage Learning: Boston, MA, USA, 2018; pp. 230–233.
3. Rao, K.; Peng, F.; Sak, H.; Beaufays, F.; Schalkwyk, J. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 4225–4229.
4. Yolchuyeva, S.; Németh, G.; Gyires-Tóth, B. Grapheme-to-phoneme conversion with convolutional neural networks. *Applied Sciences* **2019**, *9(6)*, 1143-1162.
5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
6. Sadjadi, S.O.; Greenberg, C.; Singer, E.; Mason, L.; Reynolds, D.A. The 2021 NIST speaker recognition evaluation. *arXiv* **2022**, arXiv:2204.10242.
7. Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, GA, USA, 7–10 May 1996; IEEE: Piscataway, NJ, USA, 1996; Volume 1, pp. 373–376.
8. Black, A.W.; Taylor, P. Automatically clustering similar units for unit selection in speech synthesis. In Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodes, Greece, 22–25 September 1997; ESCA: Grenoble, France, 1997; pp. 601–604.
9. Gao, Z.F. Research and Implementation of End-to-End Chinese Speech Synthesis. Master's Thesis, Beijing University of Posts and Telecommunications, Beijing, China, 2022.
10. Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Istanbul, Turkey, 5–9 June 2000; IEEE: Piscataway, NJ, USA, 2000; Volume 3, pp. 1315–1318.
11. Black, A.W.; Zen, H.; Tokuda, K. Statistical parametric speech synthesis. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, USA, 15–20 April 2007; IEEE: Piscataway, NJ, USA, 2007; Volume 4, pp. 1229–1232.
12. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 7962–7966.

 doi:10.20944/preprints202509.2016.v1

13. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE* **2013**, *101(5)*, 1234–1252.
14. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Communication* **2009**, *51(11)*, 1039–1064.
15. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1984**, *32(2)*, 236–243..
16. Van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
17. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; ISCA: Baixas, France, 2017; pp. 4006–4010.
18. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and high-quality end-to-end text to speech. *arXiv* **2020**, arXiv:2006.04558.
19. Prenger, R.; Valle, R.; Catanzaro, B. WaveGlow: A flow-based generative network for speech synthesis. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, **2019**; pp. 3617–3621.
20. Fox, M.A.; Aschkenasi, C.J.; Kalyanpur, A. Voice recognition is here comma like it or not period. *Indian Journal of Radiology and Imaging* 2013, *23(3)*, 191–194.
21. Xuan, G.; Zhang, W.; Chai, P. EM algorithms of Gaussian mixture model and hidden Markov model. In Proceedings of the 2001 International Conference on Image Processing (ICIP), Thessaloniki, Greece, 7–10 October 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 2, pp. 145–148.
22. Pan, J.; Liu, C.; Wang, Z.; Hu, Y.; Jiang, H. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling. In Proceedings of the 2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 5–8 December 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 301–305.
23. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 4960–4964.
24. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems 2020, 33, 12449–12460.
25. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2021**, *29*, 3451–3460.
26. Jin, X.L. Research and Implementation of End-to-End Speech Recognition Algorithms. Master's Thesis, Lanzhou Jiaotong University, Lanzhou, China, 2023.
27. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1980**, *28(4)*, 357–366.
28. Allen, J.B.; Rabiner, L.R. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE* ***1977***, *65(11)*, 1558–1564.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
31. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML 2006), Pittsburgh, PA, USA, 25–29 June 2006; ACM: New York, NY, USA, 2006; pp. 369–376.

 doi:10.20944/preprints202509.2016.v1

32. Elovitz, H.S.; Johnson, R.; McHugh, A.; Shore, J.E. Letter-to-sound rules for automatic translation of English text to phonetics. *IEEE Transactions on Acoustics, Speech*, and Signal Processing **1976**, *24(6)*, 446–459.
33. Black, A.W.; Lenzo, K.; Page, V. Issues in building general letter to sound rules. In Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Blue Mountains, Australia, 26–29 November 1998; ISCA: Baixas, France, 1998; pp. 77–80.
34. Suontausta, J.; Häkkinen, J. Decision tree based text-to-phoneme mapping for speech recognition. In Proceedings of the Interspeech 2000, Beijing, China, 16–20 October 2000; ISCA: Baixas, France, 2000; Volume 2, pp. 831–834.
35. Galescu, L.; Allen, J.F. Bi-directional conversion between graphemes and phonemes using a joint n-gram model. In Proceedings of the 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, Perthshire, Scotland, UK, 29 August–1 September 2001; ISCA: Baixas, France, 2001; pp. 6–9.
36. Chen, S.F. Conditional and joint models for grapheme-to-phoneme conversion. In Proceedings of the Interspeech 2003, Geneva, Switzerland, 1–4 September 2003; ISCA: Baixas, France, 2003; pp. 2033–2036.
37. Taylor, P. Hidden Markov models for grapheme to phoneme conversion. In Proceedings of the Interspeech 2005, Lisbon, Portugal, 4–8 September 2005; ISCA: Baixas, France, 2005; pp. 1973–1976.
38. Bisani, M.; Ney, H. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication* **2008**, *50(5)*, 434–451.
39. Illina, I.; Fohr, D.; Jouvet, D. Grapheme-to-phoneme conversion using conditional random fields. In Proceedings of the Interspeech 2011, Florence, Italy, 27–31 August 2011; ISCA: Baixas, France, 2011; pp. 2313–2316.
40. Novak, J.R.; Minematsu, N.; Hirose, K. Failure transitions for joint n-gram models and G2P conversion. In Proceedings of the Interspeech 2013, Lyon, France, 25–29 August 2013; ISCA: Baixas, France, 2013; pp. 1821–1825.
41. Bilcu, E.B. Text-to-Phoneme Mapping Using Neural Networks: A Neural Approach, 2nd ed.; Smith, A., trans.; Tampere University of Technology: Tampere, Finland, 2008; pp. 47–75.
42. Ploujnikov, A.; Ravanelli, M. SoundChoice: Grapheme-to-phoneme models with semantic disambiguation. In Proceedings of the Interspeech 2022, Incheon, South Korea, 18–22 September 2022; ISCA: Baixas, France, 2022; pp. 4561–4565.
43. Wang, C.; Huang, P.; Zou, Y.; Cheng, N. LiteG2P: A fast, light and high accuracy model for grapheme-to-phoneme conversion. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
44. Jia, Y.; Zen, H.; Shen, J.; Zhang, Y.; Wu, Y. PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS. *arXiv* **2021**, arXiv:2103.15060.
45. Řeháček, M.; Švec, J.; Tihelka, D. T5G2P: Using text-to-text transfer transformer for grapheme-to-phoneme conversion. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; ISCA: Baixas, France, 2021; pp. 6–10.
46. Dong, L.; Chen, H.; Guo, Y.; Liu, S.; Xu, B. Neural grapheme-to-phoneme conversion with pre-trained grapheme models. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6202–6206.
47. Qharabagh, M.F.; Dehghanian, Z.; Rabiee, H.R. LLM-powered grapheme-to-phoneme conversion: Benchmark and case study. In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), San Francisco, CA, USA, 6–11 April 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
48. Han, D.; Zhang, Y.; Liu, H.; Li, J. Improving grapheme-to-phoneme conversion through in-context knowledge retrieval with large language models. In Proceedings of the 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 1–4 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 631–635.
49. Yolchuyeva, S.; Németh, G.; Gyires-Tóth, B. Transformer based grapheme-to-phoneme conversion. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; ISCA: Baixas, France, 2019; pp. 2095–2099.

 doi:10.20944/preprints202509.2016.v1

50. Toshniwal, S.; Livescu, K. Jointly learning to align and convert graphemes to phonemes with neural attention models. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 76–82.
51. Yao, K.; Zweig, G. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015; ISCA: Baixas, France, 2015; pp. 3330–3334.
52. Mousa, A.E.D.; Schuller, B. Deep bidirectional long short-term memory recurrent neural networks for grapheme-to-phoneme conversion utilizing complex many-to-many alignments. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; ISCA: Baixas, France, 2016; pp. 2836–2840.
53. Milde, B.; Schmidt, C.; Köhler, J. Multitask sequence-to-sequence models for grapheme-to-phoneme conversion. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; ISCA: Baixas, France, 2017; pp. 2536–2540.
54. Arik, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J.; et al. Deep voice: Real-time neural text-to-speech. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 6–11 August 2017; PMLR: Brookline, MA, USA, 2017; Volume 70, pp. 195–204.
55. Chae, M.J.; Park, K.; Bang, L.; Kim, H.K. Convolutional sequence to sequence model with non-sequential greedy decoding for grapheme-to-phoneme conversion. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2486–2490.
56. Wang, Y.; Bao, F.; Zhang, H.; Gao, G. Joint alignment learning-attention based model for grapheme-to-phoneme conversion. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 7788–7792.
57. Zhao, C.; Wang, J.; Qu, X.; Zhang, C.; Li, Y. R-G2P: Evaluating and enhancing robustness of grapheme to phoneme conversion by controlled noise introducing and contextual information incorporation. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6197–6201.
58. Novak, J.R.; Minematsu, N.; Hirose, K.; Kawahara, T. Improving WFST-based G2P conversion with alignment constraints and RNNLM N-best rescoring. In Proceedings of the Interspeech 2012, Portland, OR, USA, 9–13 September 2012; ISCA: Baixas, France, 2012; pp. 2526–2529.
59. Sun, H.; Tan, X.; Gan, J.W.; Liu, H.; Qin, T.; Zhao, S.; Liu, T.Y. Token-level ensemble distillation for grapheme-to-phoneme conversion. *arXiv* **2019**, arXiv:1904.03446.
60. Yamasaki, T. Grapheme-to-phoneme conversion for Thai using neural regression models. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Seattle, WA, USA, 10–15 July 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 4251–4255.
61. Kim, J.; Han, C.; Nam, G.; Lee, K.; Kim, S.; Lee, J. Good neighbors are all you need for Chinese grapheme-to-phoneme conversion. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
62. Eters, B.; Dehdari, J.; van Genabith, J. Massively multilingual neural grapheme-to-phoneme conversion. In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, Copenhagen, Denmark, 7 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 19–26.
63. Mortensen, D.R.; Dalmia, S.; Littell, P. Epitran: Precision G2P for many languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Paris, France, 2018; pp. 2710–2714.
64. Sokolov, A.; Rohlin, T.; Rastrow, A. Neural machine translation for multilingual grapheme-to-phoneme conversion. *arXiv* **2020**, arXiv:2006.14194.

 doi:10.20944/preprints202509.2016.v1

65. Yu, M.; Nguyen, H.D.; Sokolov, A.; Rastrow, A.; Stolcke, A. Multilingual grapheme-to-phoneme conversion with byte representation. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 8234–8238.
66. Zhu, J.; Zhang, C.; Jurgens, D. ByT5 model for massively multilingual grapheme-to-phoneme conversion. *arXiv* **2022**, arXiv:2204.03067.
67. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv* **2020**, arXiv:2010.11934.
68. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3445–3460.
69. Route, J.; Hillis, S.; Etinger, I.C.; Smith, J.; Li, W. Multimodal, multilingual grapheme-to-phoneme conversion for low-resource languages. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Hong Kong, China, 3 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 192–201.
70. Ribeiro, M.S.; Comini, G.; Lorenzo-Trueba, J. Improving grapheme-to-phoneme conversion by learning pronunciations from speech recordings. *arXiv* **2023**, arXiv:2307.16643.
71. Liu, Z.; Bao, F.; Gao, G.; Wang, Y. Mongolian grapheme to phoneme conversion by using hybrid approach. In Natural Language Processing and Chinese Computing; Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H., Eds.; Lecture Notes in Computer Science, vol 11108; Springer: Cham, Switzerland, 2018; pp. 40–50.
72. Bello, I.; Zoph, B.; Vasudevan, V.; Le, Q.V. Neural optimizer search with reinforcement learning. *arXiv* **2016**, arXiv:1611.01578.
73. Jin, H.; Song, Q.; Hu, X. Auto-keras: An efficient neural architecture search system. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019), Anchorage, AK, USA, 4–8 August 2019; ACM: New York, NY, USA, 2019; pp. 1946–1956.
74. Li, K.; Xian, X.; Wang, J.; Zhang, Y.; Zhang, G. First-principle study on honeycomb fluorated-InTe monolayer with large Rashba spin splitting and direct bandgap. *Applied Surface Science* **2019**, *471*, 18–22.