Journal of Network Intelligence ©2024 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **9**, Number **2**, May **2024**

A Deep Learning Assessment Model for Oral Learning in English Online Education

Yao Ji*

School of International Education
Yellow River Conservancy Technical Institute
Kaifeng, Henan 475004, China
jyyrcti@163.com, jyyrcti@outlook.com

*Corresponding author: Yao Ji

Received August 4, 2023, revised December 27, 2023, accepted February 10, 2024.

Abstract. *Existing methods for oral English proficiency assessment such as subjective instructor evaluations are inconsistent, time-consuming, and susceptible to bias. This study aims to address these limitations by developing an automated scoring system called BERTOphone to enhance the accuracy and efficiency of spoken English evaluation. BERTOphone integrates two deep learning architectures - BERT (Bidirectional Encoder Representations from Transformers) and BiLSTM (Bidirectional Long Short-Term Memory). BERT excels at contextual representation learning while BiLSTM captures sequential dependencies in speech. Despite recent advancements, prior studies have not fully exploited the synergies of these models for accurate English speech analysis. This paper makes novel contributions by proposing the first BERT-BiLSTM fusion model tailored to English speech scoring. When evaluated on the TIMIT dataset, BERTOphone achieved 12.6% Word Error Rate and 87.4% accuracy, outperforming state-of-the-art models like LAS and CSS. For assessment, BERTOphone was tested on 100 English speech samples from an oral test. Compared to manual scores, BERTOphone obtained a high Pearson correlation of 0.9345, demonstrating reliable automated scoring.*

Keywords: Deep learning-based assessment model; English speech recognition; Speech recognition techniques; Automatic scoring algorithm; BERTOphone.

1. **Introduction.** In recent years, the field of language assessment has experienced significant growth due to the rapid development of deep learning techniques and the increasing demand for online education. As a result, the accurate assessment of English speaking proficiency has become crucial in evaluating students' language skills. However, traditional methods that rely on subjective instructor assessment are inconsistent and time-consuming [1]. Combining deep learning models, such as BERT and BiLSTM, has been proposed to overcome these challenges as a promising solution to automate and enhance the accuracy of spoken English assessment in online education. Deep learning, a subfield of machine learning, has made notable advancements in various domains, including natural language processing (NLP) and speech recognition. Deep learning models have demonstrated exceptional performance in understanding and processing complex languages, mainly when dealing with large volumes of data [2]. In NLP, based on the Transformer architecture, the BERT model has achieved remarkable results and has been applied in tasks such as text classification and question answering [3]. Leveraging BERT in the assessment of spoken English offers several advantages, including capturing contextual information, semantic relations, and the flexibility to fine-tune the model for specific

tasks [4]. Another deep learning model, BiLSTM, has been widely utilized for modeling sequential data and capturing long-term dependencies. By processing input sequences in both forward and backward directions, BiLSTM effectively captures spoken language's contextual and temporal dynamics [5]. Integrating BERT and BiLSTM in the assessment of spoken English provides significant benefits. The resulting BERTOphone model enables comprehensive speech analysis and delivers accurate and reliable assessments of spoken English proficiency. This research addresses the challenges associated with assessing spoken English in online education. Traditional assessment methods relying on subjective instructor evaluations have objectivity, consistency, and effectiveness limitations.

Previous research has demonstrated the effectiveness of deep learning models in language assessment. For instance, Smith et al. showcased BERT's capability to capture contextual information and semantic relations in text classification tasks [6]. Similarly, Zhang et al. employed BiLSTM for sequential data modeling and time-dependent analysis in speech recognition, highlighting its efficacy [7]. These studies provide a theoretical foundation and guiding framework for developing the BERTOphone model. Furthermore, the BERTOphone model builds upon automatic speech recognition and evaluation advancements. Prior studies have developed deep learning-based speech recognition systems that exhibit high accuracy in recognizing speech [8]. Liu et al. proposed an automatic scoring algorithm for spoken language, demonstrating the feasibility of utilizing deep learning in assessing speech proficiency [9]. The significance of this research lies in its potential to revolutionize the assessment of spoken English in online education. This enables students to receive timely feedback and creates an environment conducive to developing their language skills .

Advancements also inspire this research in automatic speech recognition and speech assessment. Previous studies have demonstrated the effectiveness of the BERT model in text classification tasks and the capability of BiLSTM in modeling sequential data and capturing temporal dependencies . These studies provide the theoretical foundation and guidance for developing the BERTOphone model. The BERTOphone model aims to reduce instructors' workload, provide instantaneous and fair feedback to students, and mitigate biases and inconsistencies by automating the assessment process, ensuring student consistency and fairness. This creates a favorable environment for students to accurately gauge their speaking abilities and effectively develop their language skills. The development of the BERTOphone model builds upon previous research in deep learning, automatic speech recognition, and speech assessment, thereby contributing to the advancement of automated assessment systems in language learning .

This research aims to develop an automated scoring system to enhance the assessment of spoken English proficiency. The scope includes designing a novel deep learning architecture for speech analysis and evaluating its performance on speech recognition and scoring tasks. The key problem is the subjective nature and inconsistency of human-based scoring. The solution is an objective algorithm that leverages state-of-the-art deep learning techniques. This work is significant because it could revolutionize assessment in online English education. By eliminating human errors and biases, the proposed model ensures standardized evaluation to support students' language development. Potential outcomes include improved feedback for learners, reduced workload for instructors, and advancement of speech processing capabilities . This research will demonstrate the feasibility of accurate and reliable automated oral proficiency evaluation using deep learning. The proposed model, BERTOphone, integrates BERT and BiLSTM to set a new benchmark for English speech scoring performance

This paper is structured as follows. Section 2 reviews related work on online English education and oral learning assessment, deep learning in language assessment, and BERT models for language applications. Section 3 then describes the proposed methodology, including the architecture of the BERTOphone model integrating BERT and BiLSTM. Next, Section 4 presents experiments and analysis evaluating BERTOphone on the TIMIT speech recognition dataset as well as on an oral English test. Results demonstrate the model's effectiveness for automated speech recognition and scoring. Finally, Section 5 concludes with a summary of key findings and implications.

2. Related Work.

2.1. Online English Education and Oral Learning Assessment.

The proliferation of online education has facilitated learners' access to diverse educational resources and opportunities, including online English education [10]. Online platforms have become increasingly prevalent as productive means of acquiring language proficiency from a distance [11]. The evaluation of spoken language proficiency in virtual English instruction poses distinctive obstacles in contrast to conventional face-to-face learning environments. The conventional approaches to oral assessment, such as the subjective appraisal by educators, frequently suffer from a lack of uniformity and impartiality. To tackle these obstacles, scholars have investigated novel methodologies for evaluating spoken language proficiency within the framework of virtual English instruction . Technological advancements and a rising need for adaptable learning alternatives have propelled the growth of online education. Digital platforms allow students to avail themselves of superior English language educational materials, establish connections with educators and classmates across the globe, and acquire knowledge at a self-determined rate. The platforms provide diverse courses and programs to improve language proficiency in listening, speaking, reading, and writing . Students can participate in immersive learning experiences that replicate genuine language interactions through interactive exercises, video tutorials, and virtual communication tools. The accessibility and flexibility offered by online English education have rendered it a favored option among learners who aspire to enhance their language proficiency [12].

The evaluation of spoken proficiency is of utmost importance in language acquisition as it indicates learners' capacity to communicate effectively within authentic contexts. Nonetheless, conducting precise and consistent evaluations within a virtual setting poses distinctive obstacles. Conventional approaches to evaluating oral proficiency, such as in-person interviews or oral examinations, may possess subjectivity, require a significant time investment, and be susceptible to personal prejudices [13]. In addition, scaling individual assessments for many online students can pose logistical challenges and require significant resources. Researchers have investigated diverse, innovative methods to tackle the difficulties of evaluating oral proficiency in online English education [14]. The methodologies above utilize technological progressions and educational tactics to augment verbal evaluations' precision, impartiality, and effectiveness. Several significant methodologies comprise Automated Speech Recognition (ASR) [15], Peer assessment [16], Structured performance tasks [17], Video-based assessments [18], Self-assessment [19], Task-based assessments [20], Rubrics and scoring [21], feedback and feedforward [22], and adaptive assessment systems [23]. Integrating various assessment methods and tools can yield a more comprehensive and precise evaluation of learners' oral proficiency. Incorporating self-assessment, peer assessment, and instructor assessment facilitates data triangulation, enabling a comprehensive comprehension of learners' competencies.

2.2. Deep Learning in Language Assessment.

The subfield of machine learning, known as deep learning, has garnered considerable interest in language assessment owing to its impressive capacity for discerning complex patterns and features from extensive datasets. This section delves into implementing deep learning (DL) in language assessment, specifically emphasizing its ability to augment the precision, dependability, and expediency of appraising language aptitude. Researchers and educators can use deep learning algorithms to transform language assessment practices, as Brunfaut [24] and Sarker [25] demonstrated, resulting in more extensive evaluations of learners' linguistic competencies. The domain of language assessment has seen significant research in Automated Essay Scoring (AES), with deep learning techniques exhibiting considerable promise in this field. Neural network models, namely convolutional neural networks (CNN) and recurrent neural networks (RNN) have been utilized to scrutinize and assess the caliber of essays, considering diverse linguistic attributes such as grammar, coherence, and vocabulary [26, 27]. Scholars such as Ge and Chen [28] and Uto et al. [29] have investigated the utilization of deep learning in automated essay scoring (AES), exhibiting encouraging outcomes concerning precision and uniformity.

The advent of deep learning has brought about a significant transformation in speech recognition and pronunciation assessment, as it has facilitated the creation of highly resilient and precise models. The application of deep neural networks (DNN) and recurrent neural networks with attention mechanisms (RNN-AM) in the transcription and evaluation of spoken language has been explored by researchers such as Takai et al. [30], Xu [31], and Wang et al. [32]. The models above can accurately detect pronunciation errors by capturing even the most delicate nuances of pronunciation. The integration of deep learning methodologies in pronunciation evaluation instruments has demonstrated the potential to furnish learners with comprehensive evaluations of their pronunciation abilities, thereby facilitating their enhancement in spoken language aptitude. The integration of DL with NLP methodologies has enabled the examination of language competency across diverse linguistic domains. The utilization of LSTM for sentiment analysis, semantic analysis, and syntactic parsing has been investigated by Brunfaut [24], Zhang et al. [33], and Gao [34]. These methods have been applied to evaluate learners' linguistic complexity, comprehension, and fluency. LSTM networks and transformer models have effectively captured semantic and syntactic structures.

Scholars such as Muñoz et al.[35], Lou et al.[36], and Talaghzi et al.[37] have employed DL algorithms to scrutinize students' academic achievements, pinpoint their strengths and weaknesses, and customize educational materials to cater to their individual requirements. Adaptive learning systems can offer learners personalized feedback and recommend appropriate learning resources and approaches to improve their language proficiency by utilizing deep learning methodologies. DL techniques have also facilitated the amalgamation of diverse forms of data, including textual, auditory, and visual cues, to enable a comprehensive evaluation of language proficiency. Scholars such as Suendermann Oeft et al.[38] and Dukic et al. [39] have investigated integrating textual analysis with facial expression recognition and gesture analysis. These studies have yielded significant findings regarding learners' levels of engagement, emotional states, and non-verbal communication abilities. Utilizing a multimodal approach in language assessment amplifies the depth and genuineness of the evaluation, thereby enabling a holistic appraisal of the learners' overall language competency.

Deep learning has shown promising results in advancing language assessment tasks. Relevant studies utilizing deep learning are summarized in Table 1.

2.3. **BERT in Language Applications.** BERT is a pre-trained deep learning model that has gained significant attention in various NLP tasks. BERT utilizes transformer architecture to learn contextual representations of words, enabling it to capture complex linguistic patterns and nuances. BERT has been successfully applied in several NLP

Table 1. Relevant studies utilizing deep learning in language assessment tasks are summarized

Author	Method	Contributions	Limitations
Uto et al. [29]	Hybridizes handcrafted essay-level features with a DNN for improved AES.	The method combines essay-level features with a DNN, enhancing automated essay scoring.	It lacks an in-depth exploration of feature types and comprehensive analysis of f eature importance.
Xu et al. [31]	Applies deep learning for English speech recognition using MFCC	Introduces deep learning to English speech recognition, proposing novel pronunciation quality indicators.	Offers limited model architecture details, uses a single small dataset,and lacks comprehensive analysis of indicators and benchmark comparisons.
Gao [34]	Develops AGAN-DSVM for fluency prediction through a trace-oriented approach.	Achieves high accuracy in fluency prediction, outperforming other methods.	Sparse model details, small dataset evaluation, no comparisons, or feature analysis. Solely focuses on Chinese EFL learners, potentially limiting generalization.
Muñoz et al. [35]	Conducted a systematic literature review per PRISMA guidelines focusing on adaptive learning technology in higher education.	Synthesized 112 studies to unveil trends and key aspects related to adaptive learning in design, implementation, and assessment.	Hindered synthesis due to result heterogeneity, incomplete data in included studies, restricted database scope to English-only studies.
Dukic and Sovic Krzic [39]	Employed Inception-v3 and ResNet-34 for real-time facial expression recognition during a classroom robotics workshop.	Developed a real-time facial expression recognition system using deep learning to analyze emotions, gender, and tasks statistically.	Depended on limited emotion datasets , potential participant biases, limited participant diversity, and suboptimal camera positions affecting face recognition.

applications, such as text classification, question-answering, and language modeling. In recent years, researchers have explored the potential of BERT in language assessment and proposed various models that leverage BERT's strengths to enhance the accuracy and efficiency of language proficiency evaluation. Several studies have explored the applications of BERT in language assessment, including writing assessment and vocabulary evaluation. For instance, Yang et al. [40] proposed a BERT-based model for automated essay scoring, which achieved high performance in scoring essays based on various linguistic features. Similarly, Hu et al. [41] utilized BERT to evaluate the vocabulary proficiency of Chinese learners of English, demonstrating its efficacy in accurately classifying learners' vocabulary levels. BERT has also been employed to evaluate learners' comprehension, fluency, and linguistic complexity. Dai et al. [42] proposed a BERT-based model for assessing reading comprehension, achieving high accuracy in predicting learners' performance on comprehension questions. Moreover, Mageed et al. [43] utilized BERT to evaluate the linguistic complexity of learners' writing, demonstrating its potential to provide a detailed analysis of the syntactic and semantic structures of the text. In addition to these applications, BERT has also been utilized in adaptive language learning systems, which provide personalized feedback and instruction to learners based on their performance. Mohsen [44] and Xu [45] have employed BERT-based models to analyze learners' performance and provide targeted feedback and instructional content to enhance their language skills.

In this study, the author proposed a model that utilizes BERT in language assessment named BERTOphone, designed to evaluate spoken language proficiency. BERTOphone is a novel deep learning model for spoken language proficiency assessment combining BiLSTM and BERT. The BiLSTM component of the model is designed to capture temporal dependencies in speech sequences, while BERT is used to extract contextualized representations of speech segments. The model can analyze and evaluate the quality of spoken language based on various linguistic features, such as pronunciation, intonation, and fluency. The BERTOphone model can be trained using a large speech data corpus to learn spoken language's intricate patterns and features. During training, the model learns to predict the proficiency levels of learners based on their spoken responses to various tasks. The model can be fine-tuned to specific tasks and adapted to different languages, making it a versatile tool for language assessment.

Moreover, the model's ability to generate fine-grained scores for spoken language proficiency, such as pronunciation and fluency, provides valuable feedback to learners and educators, aiding in improving spoken language skills. Deep learning techniques such as BiLSTM and BERT represent a significant field advancement in spoken language assessment. These models can capture spoken language's complex patterns and structures, providing a more comprehensive evaluation of learners' language proficiency. The BERTOphone model is an example of the potential of deep learning methods in language assessment, demonstrating how combining different deep learning architectures can lead to improved performance in language proficiency evaluation.

The major BERT-based models for language applications are summarized in Table 2:

Table 2. Studies Utilizing BERTs for Language Assessment Tasks

Author	**Method**	**Contributions**	**Limitations**
Dai et al. [42]	BERT-IDM enhances Chinese idiom comprehension using idiom masking, interpretation expansion, and multi-headed attention.	Improves ChID dataset accuracy by 4.1% over BERT baseline by creatively representing idioms and expanding their semantics.	Limited assessment on ChID dataset, lacks model comparisons, minimal analysis of biases, and unverified generalization to other NLP tasks.
Xu et al. [45]	Created adaptive English vocabulary learning with AdaBoost and BERT based personalized content recommendations	Defined system architecture and modules, measured learner adaptation, adjusted fitness for tailored recommendations, and boosted learning efficiency.	Incomplete model training details, small learner sample, basic method comparison, and limited scope on vocabulary learning.
Hu, et al. [41]	Created 17 new challenge datasets across 4 categories to assess Chinese natural language inference systems base on BERT, focusing on linguistic phenomena and model robustness.	Provided new resources for Chinese NLI system evaluation and analyzed cross-lingual transfer learning strengths and limitations.	Primarily relied on behavioral testing; additional analysis required to understand cross-lingual performance in various linguistic phenomena. Intervention techniques suggested for future work.
Abdul-Mageed, et al. [43]	Curated a vast Twitter dataset annotated for Arabic dialects across 21 Arab countries ,employing GRUs and BERT for supervised models.	Introduced an extensive Arabic dialect dataset and a hierarchical multi-task learning method for dialect identification.	Depends on location tags as a proxy for dialect, requiring further investigation on the correlation between location and dialect in social media data.
Moshen [44]	Evaluated MY Access automated writing evaluation for 6 intermediate EFL students base on BERT, comparing software-only to combined feedback.	Showed improved essay scores with combined feedback, emphasizing the value of automated feedback on language accuracy	Small sample; more research needed with larger groups Focused on one specific software.

3. **Methodology.** Oral English proficiency is essential to language learning, and the ability to accurately assess it is critical. The development of speech recognition and natural language processing technologies has enabled the creation of automated scoring systems for oral English proficiency. The study proposes BERTOphone, a speech recognition model based on the BERT architecture. BERTOphone has three main components: speech signal processing and feature extraction, acoustic model training, and scoring. The author describes each part, including mathematical formulas, and discusses their implementation in the BERTOphone system.

3.1. **Speech signal processing and feature extraction.** Speech signal processing and feature extraction are crucial in developing an accurate oral scoring system. These processes involve transforming the raw speech signal into a sequence of feature vectors that can be used as input to the acoustic model. The following are the steps involved in speech signal processing and feature extraction:

Digitization: The speech signal is first digitized using the sound card of a personal computer. The signal is sampled at 8 kHz to capture the frequency range of human speech.

Pre-emphasis: The pre-emphasis process involves flattening the spectrum of the speech signal. This is achieved using a 6 dB/oct high-frequency boosting pre-emphasis digital filter. The following equation represents the pre-emphasized signal:

$$y(n) = x(n) - \alpha x(n-1) \tag{1}$$

where $x(n)$ represents the input speech data, and α is the pre-emphasis coefficient.

Framing: The speech signal is divided into frames of a quasi-steady-state process of 10-40 ms. The following equation represents the framed signal:

$$f(n) = w(n) * y(n) \tag{2}$$

where $f(n)$ is the framed signal, $y(n)$ is the pre-emphasized signal, and $w(n)$ is the window function.

Windowing: The windowing process involves strengthening the speech waveform near the sampling point. This is achieved by multiplying each frame by a window function. The following equation represents the windowed signal:

$$s(n) = f(n) * w(n) \tag{3}$$

where $s(n)$ is the windowed signal, $f(n)$ is the framed signal, and $w(n)$ is the window function.

Feature extraction: The feature extraction process involves transforming the windowed signal into a sequence of feature vectors that can be used as input to the acoustic model. The most commonly used feature extraction method is Mel-frequency cepstral coefficients (MFCCs). The MFCCs are obtained by applying a sequence of transformations to the windowed signal, including the Fourier transform, Mel filterbank, logarithmic compression, discrete cosine transform, and liftering. The resulting feature vectors are then used as input to the acoustic model. The following equation represents the MFCCs:

$$MFCCs = IDCT(log(\mathbf{E} * \mathbf{H})) \tag{4}$$

where $\mathbf{E}$ is the Mel filterbank matrix, $\mathbf{H}$ is the Fourier transform matrix, and $IDCT$ is the inverse discrete cosine transform.

3.2. **Acoustic Model Training.** The acoustic model maps the input feature vectors to a sequence of phonemes or subword units. In the BERTOphone system, I uses a neural network-based acoustic model trained on a large corpus of labeled speech data. The training data consists of pairs of input feature vectors and their corresponding phoneme or subword labels. The network is trained using back-propagation and stochastic gradient descent to minimize the cross-entropy loss between the predicted and actual labels. The architecture of the acoustic model is based on the BERT model, which is a transformer-based neural network architecture that has achieved state-of-the-art performance in various natural language processing tasks. In the BERTOphone system, I adapted the BERT architecture to the speech recognition task by adding a linear layer to the output of the transformer encoder to predict the probability distribution over the phonemes or subword units. The model is trained using a masked language modeling objective, where a certain percentage of the input frames are masked during training, and the network is tasked with predicting the masked frames. Table 1 shows the Phoneme or Subword Label used in the BERTOphone system, which employs a neural network-based acoustic model to map input feature vectors to a sequence of these labels.

Table 3. Phoneme or Subword Label used in the BERTOphone system

Phoneme/Subword Label	Description
AA	vowel (open front unrounded)
AE	vowel (open front unrounded)
AH	vowel (central unrounded)
AO	vowel (open back rounded)
AW	diphthong (open back rounded)
AY	diphthong (front open unrounded)
B	stop consonant (voiced bilabial)
CH	affricate consonant (unvoiced palato-alveolar)
D	stop consonant (voiced alveolar)
DH	fricative consonant (voiced dental)
EH	vowel (mid front unrounded)
ER	vowel (r-colored)
EY	diphthong (mid front unrounded)
F	fricative consonant (voiceless labiodental)
G	stop consonant (voiced velar)
HH	aspirated consonant (voiceless glottal)
IH	vowel (mid central unrounded)
IY	vowel (close front unrounded)
JH	affricate consonant (voiced palato-alveolar)
K	stop consonant (voiceless velar)
L	liquid consonant (voiced alveolar lateral)
M	nasal consonant (voiced bilabial)
N	nasal consonant (voiced alveolar)
NG	nasal consonant (voiced velar)
OW	diphthong (mid back rounded)
OY	diphthong (front open rounded)
P	stop consonant (voiceless bilabial)
R	liquid consonant (voiced alveolar)
S	fricative consonant (voiceless alveolar)
SH	fricative consonant (voiceless palato-alveolar)
T	stop consonant (voiceless alveolar)
TH	fricative consonant (voiceless dental)
UH	vowel (mid back rounded)
UW	vowel (close back rounded)
V	fricative consonant (voiced labiodental)
W	semi-vowel (voiced labial-velar)
Y	semi-vowel (voiced palatal)
Z	fricative consonant (voiced alveolar)
ZH	fricative consonant (voiced retroflex)

3.3. **Scoring.** The final step in the oral scoring system is to use the output of the acoustic model to generate a score for the input speech signal. There are two main approaches to scoring in speech recognition systems: a lattice-based approach and a sequence-based approach. The lattice-based process involves decoding the output of the acoustic model into a lattice, which represents a set of possible word sequences. The lattice is then pruned to select the most likely sequence, and the score is computed based on the likelihood of the selected sequence. However, this approach has some limitations as it may not capture the full context of the speech signal and may result in suboptimal performance.

In contrast, the sequence-based approach directly decodes the output of the acoustic model into a sequence of phoneme or subword labels. This approach is effective in speech recognition systems and is used in the BERTOphone system. The input to the network is a sequence of feature vectors extracted from the preprocessed speech signal. The BiLSTM processes the sequence of feature vectors in both forward and backward directions, allowing the network to capture context information from past and future frames. The output of the BiLSTM is then fed into a transformer network, which further processes the sequence of feature vectors and generates a sequence of phoneme or subword labels. Once the sequence of phoneme or subword labels has been developed, it is compared to the ground truth labels for the input speech signal to compute a score. One common approach is to use the Word Error Rate (WER), which measures the percentage of incorrectly recognized phonemes or subwords relative to the total number of phonemes or subwords in the ground truth transcription.

3.4. **Architecture of BERTOphone.** The architecture of BERTOphone consists of several interconnected layers that process the input speech data to generate accurate scores. Figure 1 illustrates the flow of information through the different stages of BERTOphone's architecture. Firstly, the speech data is provided as input to the system. This data contains the spoken language that needs to be transcribed and evaluated. Next, the input data undergoes a preprocessing step. This step involves various techniques and algorithms to clean and transform the raw speech data into a suitable format for further analysis.

3.4.1. *BERT Layer.* Following preprocessing, the data flows into the BERT layer. This layer learns speech data representations that encode local and global dependencies. BERT stands for Bidirectional Encoder Representation from Transformer, which signifies a model for bidirectional word representation based on the Transformer architecture. It constitutes a sequence-to-sequence model comprising two phases: an encoder and a decoder. Notably, the architecture does not rely on RNN structures, but rather employs attention layers to embed words within sentences. The specific architecture of the model is depicted in Figure 2. The Transformer model consists of two main phases:

- Encoder: This phase consists of six consecutive layers. Each layer encompasses a sub-layer, Multi-Head Attention, combined with a fully-connected layer, as illustrated in the left branch of the diagram. An embedding vector output is obtained for each word after the encoding process.
- Decoder: The decoder architecture also comprises sequential layers. Each layer of the decoder incorporates sub-layers that closely resemble those of the encoder, with the addition of a first sub-layer known as Masked Multi-Head Attention, designed to exclude future words from the attention process. BERT is designed for pre-training word embeddings. A unique aspect of BERT is its ability to balance context in both left-to-right and right-to-left directions. The attention mechanism of the Transformer simultaneously processes all words in a sentence without regard to sentence direction. Thus, the Transformer is considered bidirectional in training, though it is more accurately described as non-directional. This characteristic enables the model to learn word context based on all surrounding words, including those to the left and right.

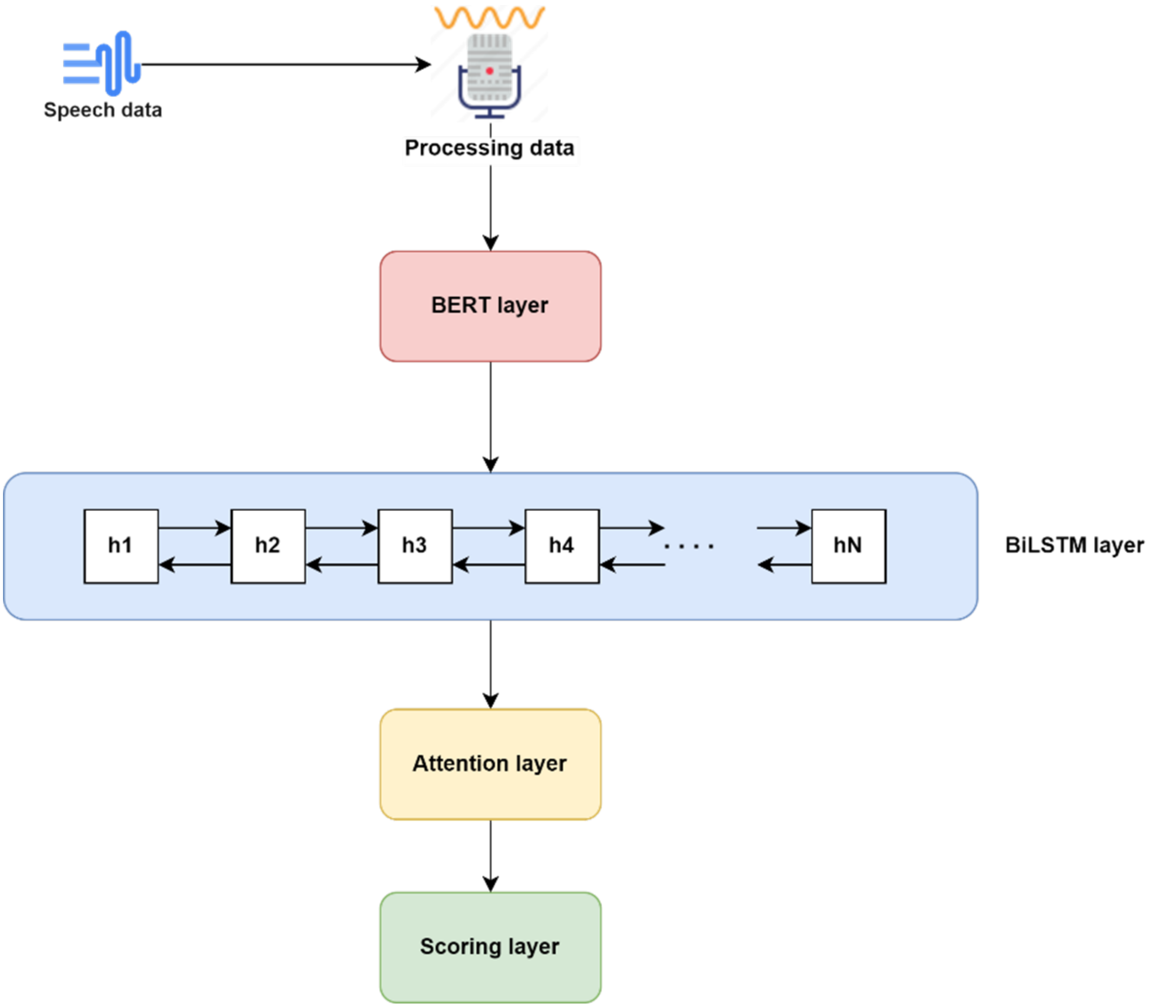

Figure 1. Architecture of BERTOphone

A distinctive feature of BERT, not present in previous embedding models, is the capacity for fine-tuning the training results. The output layer can be customized for the training task. Figure 3 illustrates a similar architecture employed for both the pre-trained and fine-tuned models. The same pre-training parameters initialize models for various downstream tasks. Throughout the fine-tuning process, all transfer learning layer parameters are fine-tuned.

For tasks involving input as a pair of sequences, such as question and answering, tokens [CLS] and [SEP] are added at the beginning and between sentences, respectively. The fine-tuning process unfolds as follows:

Step 1: Embedding of all tokens in the sentence pair using pre-trained word embedding vectors. Token embeddings include both [CLS] and [SEP] tokens to mark the question's start and the sentence separation point. These tokens are predicted in the output to determine the output sentence's Start/End Span portions.

Step 2: The resultant embedding vectors are then input into a multi-head attention architecture comprising multiple code blocks (typically 6, 12, or 24 blocks, depending on the BERT architecture). An output vector is obtained at the encoder.

Step 3: To predict the probability distribution for each word position in the decoder, the encoder's output vector and the decoder's input embedding vector are used to calculate encoder-decoder attention at each time step. This is followed by projection through

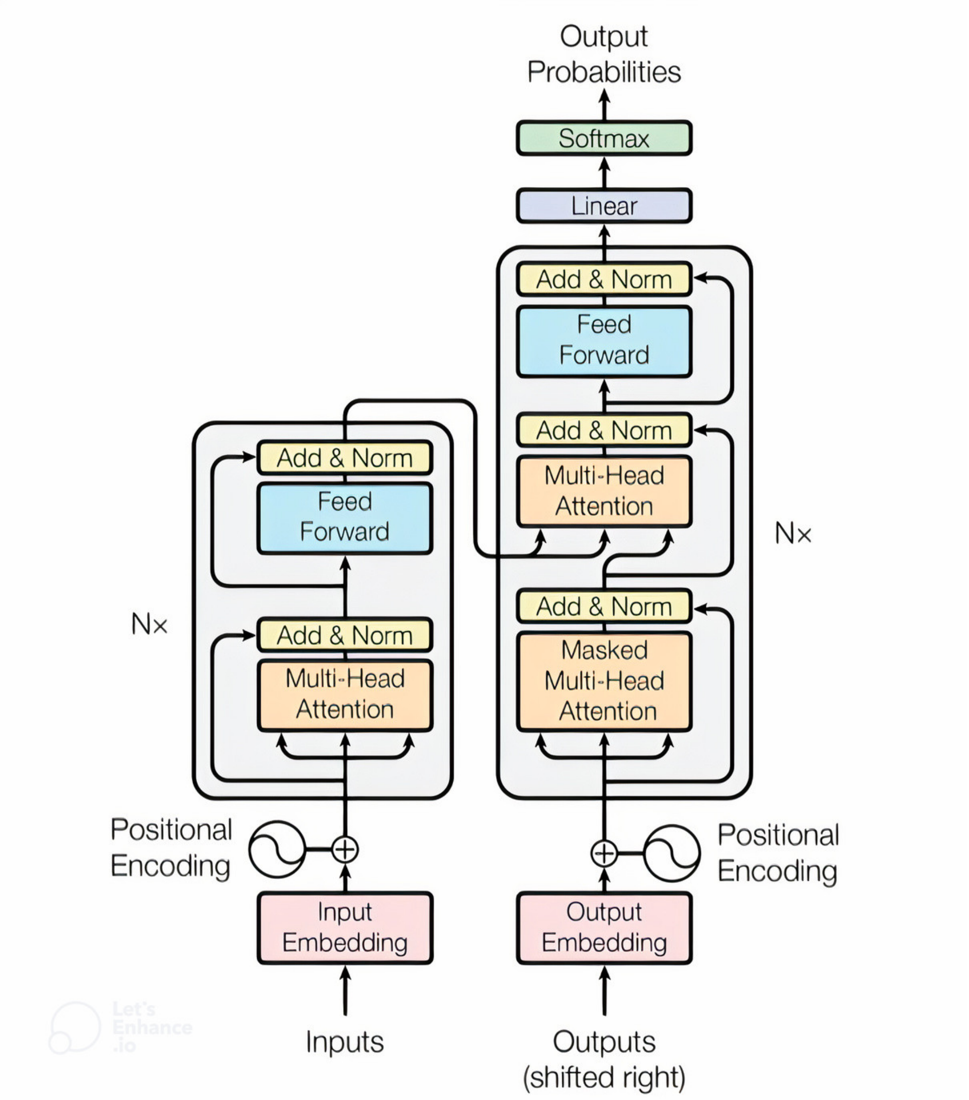

Figure 2. Transformer Architecture

a linear layer and softmax to derive the probability distribution for the corresponding output at the time step.

Step 4: The output of the Transformer contains the fixed result of the question, aligned with the input Question. The remaining positions constitute the expanded Start/End Span components, corresponding to the answer found in the input sentence.

3.4.2. *BiLSTM Layer.* The output from the BERT layer then moves to the BiLSTM layer. The BiLSTM model is a composite model consisting of a forward LSTM and a backward LSTM. The LSTM utilized in this model is a variant of the recurrent neural network (RNN). To address the issue of gradient vanishing in conventional RNN models, researchers have proposed integrating a gating unit into the LSTM architecture. This addition enhances the LSTM's capacity to capture long-term dependencies and empowers the RNN to identify and exploit dependencies present in distant data points more effectively. The process of the BiLSTM layer is elaborated in Figure 4, along with the mathematical expressions (5) - (10).

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{6}$$

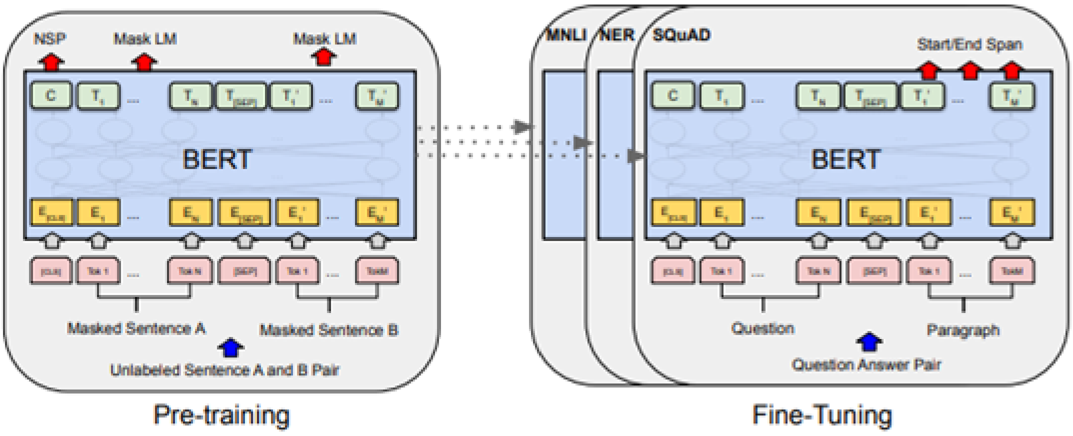

Figure 3. The pre-training and fine-tuning process of BERT

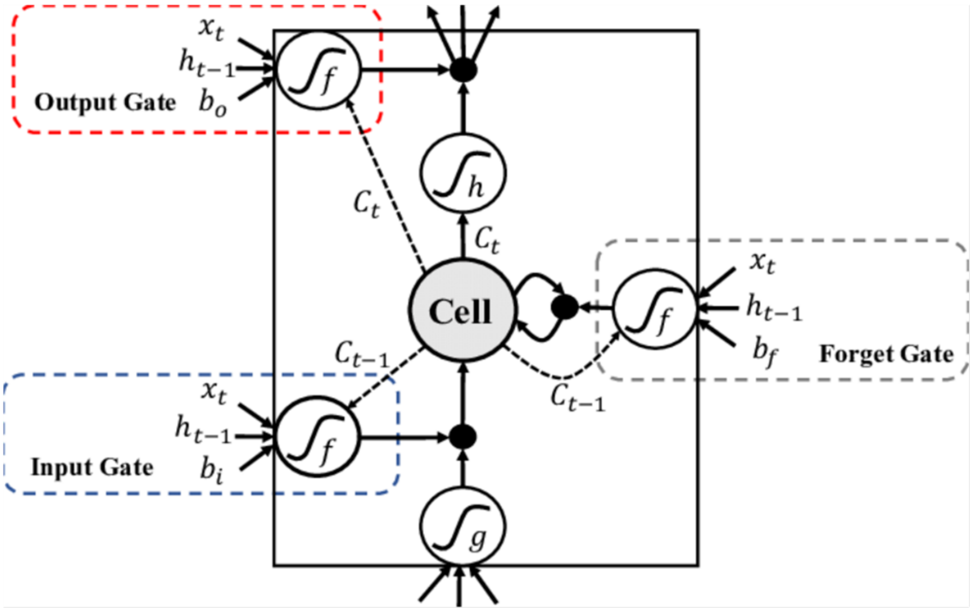

Figure 4. LSTM internal neural

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{7}$$

$$\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{9}$$

$$h_t = o_t \odot tanh(c_t) \tag{10}$$

where, h_t is the hidden state of the BiLSTM layer at time t, x_j is the feature vector at positions j, and v, W_h, W_x, and b are learnable parameters. The attention scores e_{tj} measure the relevance of each feature vector x_j to the currently hidden state h_t.

3.4.3. *Attention layer and Scoring layer.* After the BiLSTM layer, the data passes through the Attention layer. This layer applies attention mechanisms to focus on essential features and relationships within the input data. Attention allows BERTOphone to give more weight to relevant information during the scoring process. The Attention layer computed a context vector for each frame using the Equations (11) to (13):

$$c_t = \sum_j \alpha_{tj}.x_j \tag{11}$$

$$e_{tj} = v^T \; tanh(W_h h_t \; + \; W_x x_j \; + \; b) \tag{12}$$

$$\alpha_{tj} = \; softmax(e_{tj}) \tag{13}$$

where attention weights α_{tj} determine how much weight to give each feature vector in computing the context vector c_t.

Finally, the data reaches the Scoring layer, where the transformed representations are used to generate scores for the input speech data. This layer applies specific scoring algorithms and techniques tailored to the evaluation task, such as assessing pronunciation, fluency, and other linguistic aspects.

4. **Experiment and Analysis.**

4.1. **Dataset.**

For evaluating the performance of BERTOphone, I use the TIMIT dataset, which is a widely used benchmark dataset for speech recognition [46]. The dataset consists of 6300 phonetically balanced sentences spoken by 630 speakers from eight major dialect regions of the United States. The recordings were sampled at 16 kHz and saved as 16-bit linear PCM files. Each sentence is transcribed and labeled with a phonetic transcription, which includes 61 phonetic symbols. I used the standard train/test split provided with the dataset, where 462 speakers were used for training, and 168 speakers were used for testing. The training set consists of 3696 sentences, while the test set consists of 1344. The task is to transcribe each sentence into its corresponding phonetic sequence.

Furthermore, to conduct a comprehensive assessment of BERTOphone's efficacy, the model will be employed to assess the speech recordings of 100 students. Subsequently, the results will be compared with the manual scoring conducted by English teachers. The recordings were collected using a smartphone application that prompted the students to read a set of standardized text passages. The passages were selected to cover a range of phonetic and linguistic features, including vowel and consonant sounds, stress patterns, and sentence intonation. Each student was asked to read three different passages, with each passage presented twice. The first presentation was used as a warm-up, and the second presentation was used for analysis. The students were instructed to speak naturally and at a comfortable pace. The recordings were saved in WAV format and then processed to extract relevant features using MFCCs. The extracted features were then used as input to the BERT-based acoustic model, which was trained to predict phoneme or subword labels.

4.2. **Evaluation Metric.**

I uses Word Error Rate (WER) and Accuracy to evaluate the model's speech recognition performance.

WER is a commonly used metric for evaluating the accuracy of speech recognition models. WER is defined as the percentage of word recognition errors in the predicted transcription. It is calculated as follows:

$$WER \; = \; \frac{S + D + I}{N} \tag{14}$$

where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors, and N is the total number of words in the reference transcription.

Accuracy: In addition to the above error-based metrics, I reports the overall accuracy of the BERTOphone model, defined as the percentage of correctly predicted phones in the test set.

$$Accuracy = \frac{P}{T} \tag{15}$$

where P is number of correctly predicted phones, and T is the total number of phones.

To assess and contrast and manual scoring, I employed the Pearson correlation coefficient and Average difference metrics, as denoted by Formulas (16) and (17), respectively.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X).\sigma(Y)} \tag{16}$$

$$d = |S(BERTOphone) - S(Manual)| \tag{17}$$

4.3. **Experiment Setup.** All experiments were conducted using Python 3.6 on a machine with an Intel Core i7-9700K CPU and 32GB of DDR4 RAM. The BERTOphone model was implemented in PyTorch 1.7.1 leveraging its optimized deep learning modules and computational graph capabilities. Training and inference were performed on an NVIDIA GeForce RTX 2080 Ti GPU with 11GB GDDR6 memory, which provided sufficient resources for batch processing of the speech data. The key BERTOphone training parameters were empirically tuned and are presented in Table 4. The Adam optimizer was used for its adaptive learning rate capabilities, accelerating convergence. A small weight decay of 0.0001 regularized the model to prevent overfitting. The learning rate began at 0.01, enabling fine-tuning of the pretrained BERT weights. Speech segments were truncated/padded to 256 tokens to balance sequence coverage and memory requirements. A dropout of 0.2 was applied on the BiLSTM module during training for regularization. The parameterized setup allowed efficiently training BERTOphone to leverage both BERT's representation learning and BiLSTM's sequence modeling strengths.

Table 4. Training parameters.

Parameter	**Value**
Optimizer	Adam
Weight_decay	0.0001
Learning rate	0.01
Maximum sequence length	256
Dropout	0.2

For the purpose of evaluation, only vehicles in the test set with more than 100 track points recorded were selected for the split.

4.4. **Experimental Results and Analysis.**

4.4.1. *Speech Recognition Results and Analysis.* The TIMIT dataset, a well-established benchmark for speech recognition systems, was utilized to conduct an assessment of the BERTOphone . This dataset consists of speech recordings from various speakers across different dialects and phonetic environments. Standard metrics, including WER, and Accuracy (Acc), were utilized to evaluate the performance of BERTOphone. Table 5 and Figure 5 below illustrate the assessment outcomes conducted on sentences of varying lengths, ranging from one to five.

As we can see from Table 5 and Figure 5, as the number of sentences increases, there is a gradual increase in all four evaluation metrics, with WER increasing while accuracy decreases. This indicates that BERTOphone struggles with longer inputs and that its

Table 5. Performance of BERTOphone on Different Sentence Lengths.

Number of Sentences	WER	Accuracy
1	9.6	90.4
2	11.7	88.3
3	12.9	87.1
4	13.8	86.2
5	14.8	85.2

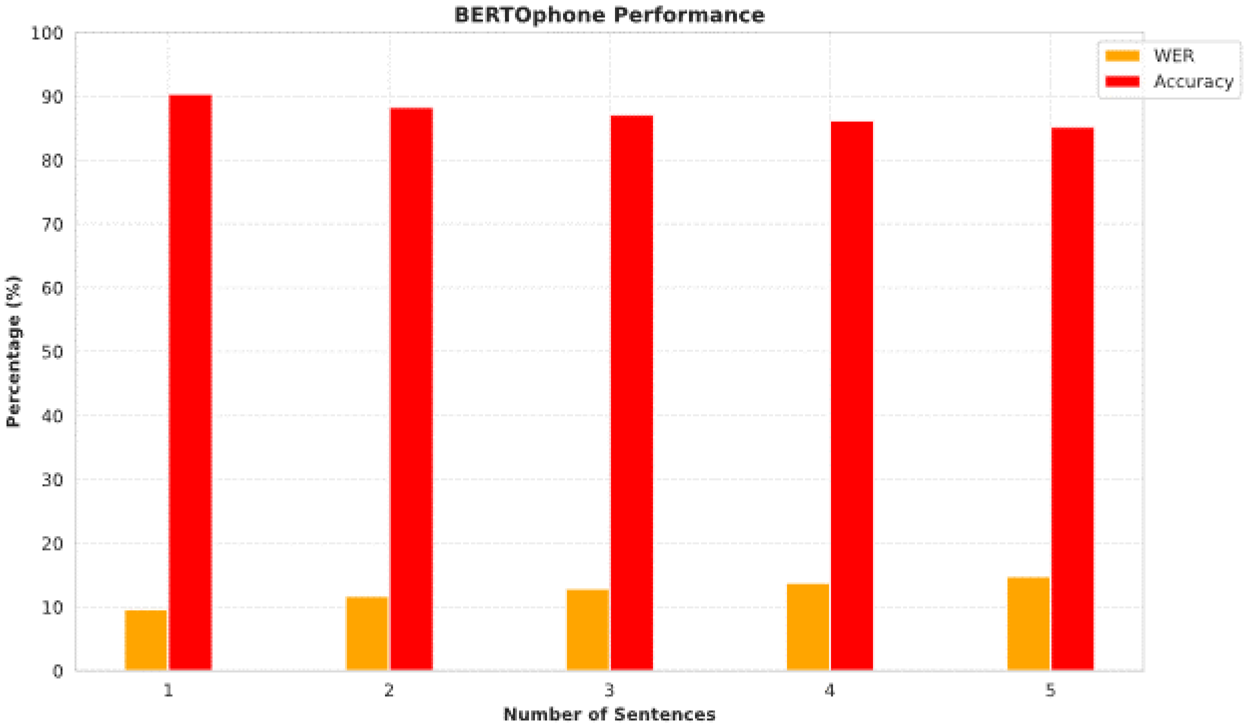

Figure 5. BERTOphone Performance by Sentence Length

performance may be limited for complex, multi-sentence inputs. However, even with longer inputs, the performance of the BERTOphone is impressive, with WER below 15% and accuracy above 85%. These results suggest that BERTOphone is an effective method for speech recognition, particularly for shorter inputs, and can be applied to a wide range of applications, such as language learning, voice-enabled assistants, and automated speech transcription.

Figure 6 presents a thorough evaluation of the classification capabilities of BERTOphone. The Confusion Matrix provided represents the classification performance of the BERTOphone model when applied to the TIMIT dataset. The provided visual representation provides a detailed depiction of the correspondence between the phoneme labels obtained from actual data and the phoneme predictions generated by the BERTOphone model. This comprehensive evaluation offers insights into the performance of the model. The diagonal elements representing accurate predictions are more prominently distinguished by their higher intensity. The identification of deviations from the diagonal reveals the underlying pattern of misclassifications. For example, the rows corresponding to phonemes such as "AE," "IH," and "IY" demonstrate comparatively higher frequencies of misclassification. This observation suggests the presence of inherent phonetic similarities or difficulties in accurately distinguishing these phonemes.

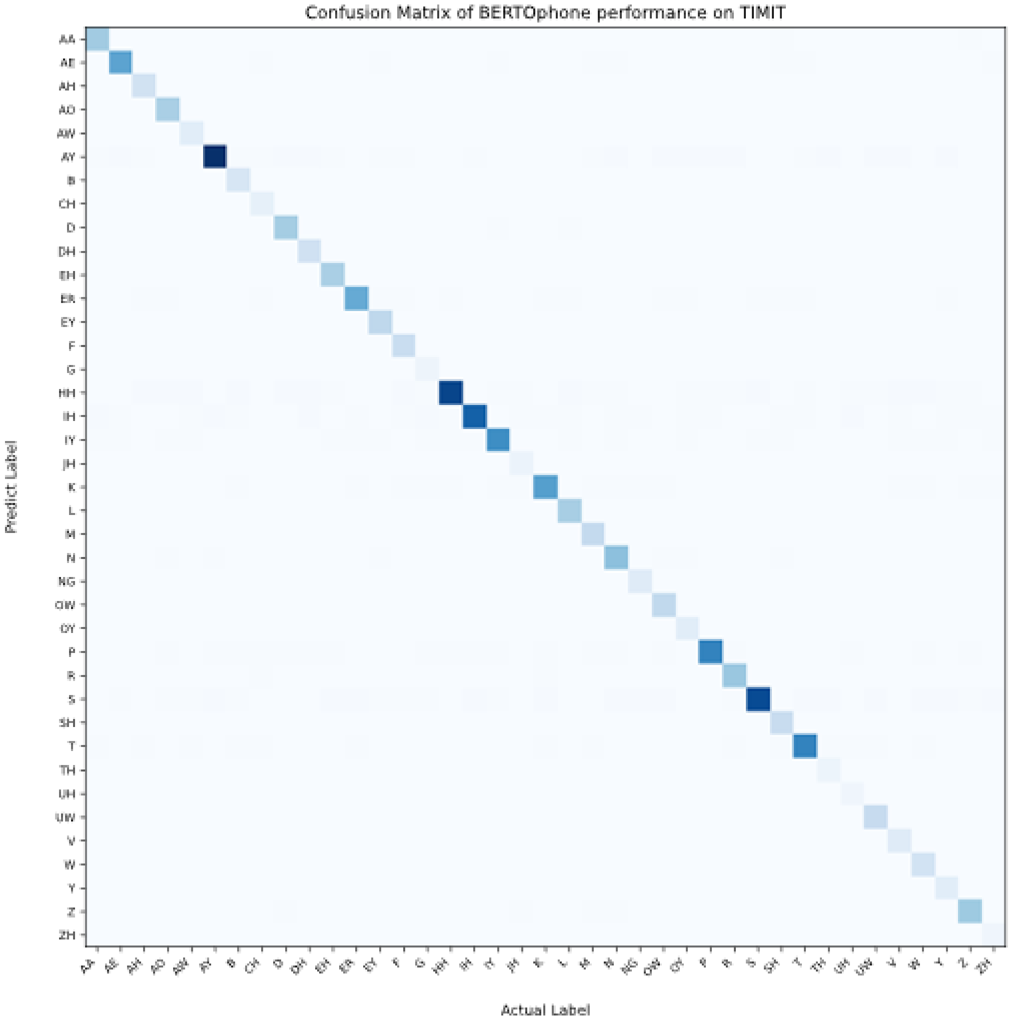

Figure 6. Confusion Matrix of BERTOphone performance on TIMIT

To perform a thorough empirical evaluation and analysis, the results obtained from BERTOphone will be averaged and contrasted with those of two other cutting-edge speech recognition models, specifically Listen, Attend and Spell (LAS) [47] and Convolutional Sequence-to-Sequence (CSS) [48]. The evaluation is based on the metrics of WER, and accuracy, all of which are crucial in assessing the effectiveness of a speech recognition model. The comparative findings are presented in Table 6. The results indicate that

Table 6. Performance Comparison of BERTOphone with LAS and CSS Models

Model	**WER**	**Accuracy**
BERTOphone	12.6	87.4
LAS [47]	14.8	84.8
CSS [48]	13.3	86.7

BERTOphone outperforms the other two models in all four metrics, achieving WER of 12.6% and accuracy of 87.4%. In comparison, LAS achieved a WER of 15.2% and an

Accuracy of 84.8%, while CSS achieved a WER of 13.3% and an accuracy of 86.7%. These results demonstrate the high accuracy and effectiveness of BERTOphone in recognizing speech and highlighting its superiority over the other two models. Such improvements in speech recognition models can have significant implications in various fields, such as online education, speech-to-text translation, and automatic transcription, where accurate and efficient speech recognition is paramount.

4.4.2. *Data Augmentation Techniques for Model Robustness.* This experiment aims to assess the influence of data augmentation techniques on the robustness and performance of the proposed framework. By employing various data augmentation techniques, I aims to enhance the model's ability to handle diverse input variations and improve its accuracy and reliability in evaluating oral English proficiency. Through rigorous evaluation and comparison, the research seeks to identify the augmentation techniques that significantly impact the BERTOphone's performance, thereby providing valuable insights for developing more effective and robust deep learning assessment models in the context of oral English education.

Data augmentation techniques employed in this study include Speed Perturbation, which involves randomly varying the speed of the original recordings by +/- 10% to generate augmented samples. Pitch Shifting techniques are applied to create variations in the pitch of the speech in the original recordings. These techniques aim to enhance the diversity and robustness of the training data, providing the deep learning assessment model with a more comprehensive understanding of different speech patterns and environments in oral English learning. BERTOphone is trained separately using each augmented dataset. This allows the model to adapt and optimize its performance based on the augmented data, capturing the nuances and characteristics introduced by the specific augmentation technique. By training the model on different augmented datasets, the experiment aims to evaluate the impact of each technique on the model's performance and robustness in oral English learning assessment. In this experiment, we consider three models for analysis:

1) BERTOphone: The baseline model, trained on the original data without any augmentation.

2) TempoBoost (BERTOphone -TempoBoost): This model is trained on data with speed perturbation augmentation, where the speed of the original recordings is randomly varied by +/- 10%, creating augmented datasets.

3) PitchFlex (BERTOphone – PitchFlex): This model is trained on data with pitch shifting augmentation, introducing variations in the pitch of the speech in the original recordings.

Based on the conducted experiment, the performance of the BERTOphone and the two augmented models, TempoBoost and PitchFlex, were evaluated and shown in Table 7. In comparison, the TempoBoost model, trained on data with speed perturbation augmentation, demonstrated a lower WER of 12.1% and a higher accuracy of 88.7%. This indicates that the speed perturbation technique enhanced the model's ability to

Table 7. Performance Comparison of BERTOphone with augmented models

Model	**WER**	**Accuracy**
BERTOphone	12.6	87.4
BERTOphone -TempoBoost	12.1	88.7
BERTOphone - PitchFlex	11.2	89.5

handle variations in speech speed, resulting in improved performance in oral English learning assessment. Similarly, the PitchFlex model, trained on data with pitch shifting augmentation, achieved the lowest WER of 11.2% and the highest accuracy of 89.5%. The pitch-shifting technique proved effective in capturing variations in pitch and contributed to enhanced performance in evaluating oral English proficiency. These results highlight the positive impact of data augmentation techniques on the BERTOphone model, providing evidence of improved robustness and effectiveness in oral learning assessment tasks. The augmented models, TempoBoost and PitchFlex, offer valuable insights for developing more accurate and reliable DL assessment models in the context of oral English education.

4.4.3. *BERTOphone Scoring System Results and Analysis.* The scope of the experiments has been extended to assess the automated scoring proficiency of BERTOphone. A sample of 100 English language tests administered to university students was utilized to evaluate the efficacy of BERTOphone scoring. The results were then compared to the manual scores assigned by the English language teacher. The results of the oral English test scoring prediction from BERTOphone and manual scores are depicted in Figure 7. Table 8 reveals the Pearson correlation coefficient, Average different index, Mean absolute error (MAE), Root Mean Squared Error (RMSE) and Concordance Correlation Coefficient (CCC) between BERTOphone and manual scoring. Figure 7 and the Table 8 depict

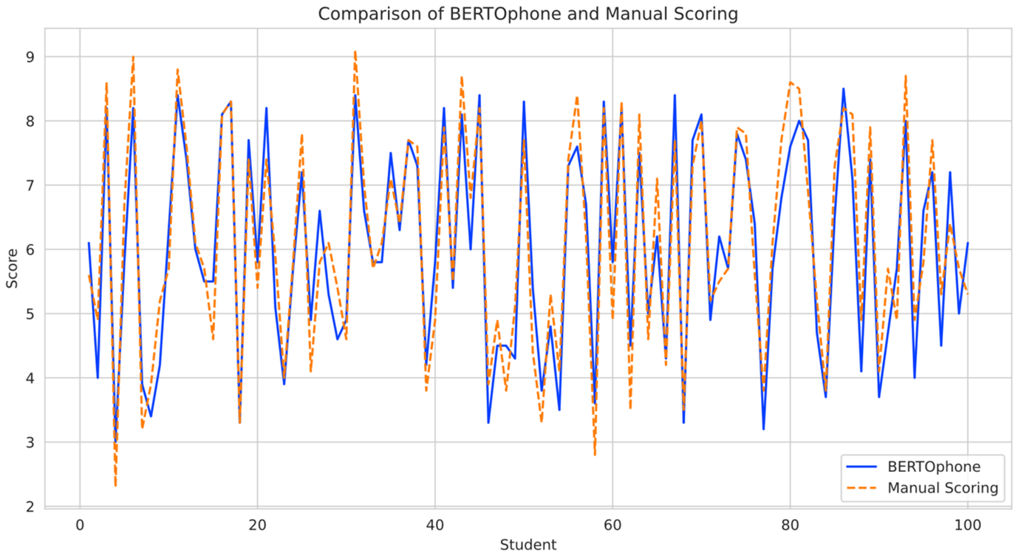

Figure 7. Results of the oral English test scoring prediction from BERTOphone and manual scores

Table 8. Coefficient Metrics for BERTOphone and Manual Scoring.

Metric	**Value**
Pearson correlation	0.9345
Average difference	0.558
MAE	0.531
RMSE	0.6097
CCC	0.9326

the outcomes derived from evaluating 100 students using both BERTOphone and manual scoring methods. These metrics elucidate the divergence and convergence between the two assessment systems. The data showcases a pronounced and positive linear relationship, indicated by the Pearson correlation coefficient of 0.9345, accompanied by an average score difference of 0.558.Additionally, this comparison highlights other metrics essential in understanding the differences between the BERTOphone and manual scoring methods. The Mean Absolute Error (MAE), measuring the average magnitude of errors between the two scoring methods, indicates a value of 0.531. Similarly, the Root Mean Squared Error (RMSE) stands at 0.6097, providing insight into the square root of the average of squared differences between BERTOphone and manual scores. The Concordance Correlation Coefficient (CCC), a crucial measure for evaluating the agreement between two methods, was computed as 0.9326. It indicates not only the correlation between BERTOphone and manual scores but also the precision of their agreement.

The comparative analysis between BERTOphone and manual scoring reveals a noteworthy pattern in how human graders tend to assess student work. Manual assessment often reflects teachers' inclinations to assign higher scores to assessments they consider of superior quality and lower scores to those deemed less impressive. In the manual grading process, English teachers engage subjective judgment, leveraging their experience and interpretative skills to evaluate various aspects such as grammar, vocabulary, sentence structure, and overall coherence.In stark contrast, BERTOphone introduces an objective and automated scoring system that meticulously scrutinizes speech patterns and linguistic attributes. This approach substantially eliminates the subjective elements inherent in manual grading, ensuring a standardized evaluation process. However, BERTOphone's effectiveness is heavily contingent upon the quality of its training data. In this regard, human graders retain their proficiency in offering nuanced understanding, interpreting context, and delivering personalized feedback.

5. **Conclusion.** In this study, I has examined the performance of BERTOphone, an automated scoring system based on machine learning algorithms. BERTOphone demonstrates strong performance in speech recognition tasks using the TIMIT dataset. It achieves a low WER, indicating its accuracy in transcribing spoken language. When com-paring BERTOphone's evaluation of student scores to manual scoring methods, a high Pearson correlation coefficient of 0.9345 demonstrates a strong positive linear relationship between the two approaches. This affirms BERTOphone's reliability and alignment with manual scoring, making it a promising tool for assessing student performance. However, it's important to acknowledge that BERTOphone's performance relies on the quality and diversity of its training data and specific dataset characteristics. Ongoing research and improvements are necessary to address challenges such as accents, speech variations, and noise environments. Overall, BERTOphone offers reliable speech recognition and effective evaluation of student scores, making it a valuable tool in language assessment and speech processing. Continued development will contribute to its integration into educational and speech recognition applications.

REFERENCES

[1] N. Bergen and R. Labonté, "'Everything Is Perfect, and We Have No Problems': Detecting and Limiting Social Desirability Bias in Qualitative Research," *Qual Health Res*, vol. 30, no. 5, pp. 783–792, Apr. 2020, doi: 10.1177/1049732319889354.

[2] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavvaf, and E. Fox, "Natural Language Processing Advancements By Deep Learning: A Survey," *ArXiv*, Mar. 2020, Accessed: May 12, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Natural-Language-Processing-Advancements-By-Deep-A-Torfi-Shirvani/77b91d7607518994d04f75119db4138b23e2eb87

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019, doi: 10.48550/arXiv.1810.04805.

[4] W.-C. Huang, C.-H. Wu, S.-B. Luo, K.-Y. Chen, H.-M. Wang, and T. Toda, "Speech Recognition by Simply Fine-tuning BERT," *arXiv*, Jan. 30, 2021, doi: 10.48550/arXiv.2102.00291.

[5] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61–98, Mar. 2015, doi: 10.1016/j.csl.2014.09.005.

[6] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning–based Text Classification: A Comprehensive Review," *ACM Computing Survey*, vol. 54, no. 3, p. 62:1-62:40, Tháng T 2021, doi: 10.1145/3439726.

[7] D. Efanov, P. Aleksandrov, and N. Karapetyants, "The BiLSTM-based synthesized speech recognition," *Procedia Computer Science*, vol. 213, pp. 415–421, Jan. 2022, doi: 10.1016/j.procs.2022.11.086.

[8] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach.* London: Springer, 2015, doi: 10.1007/978-1-4471-5779-3.

[9] P. Li, H. Zhang, and S.-B. Tsai, "Design of Automatic Scoring System for Oral English Test Based on Sequence Matching and Big Data Analysis," *Discrete Dynamics in Nature and Society*, vol. 2021, p. e3018285, Oct. 2021, doi: 10.1155/2021/3018285.

[10] J. Tao and X. Gao, "Teaching and learning languages online: Challenges and responses," *System*, vol. 107, p. 102819, Jul. 2022, doi: 10.1016/j.system.2022.102819.

[11] S. Krishnan, H. Norman, and M. Yunus, "Boosted with Online Learning to Improve English Language Teachers' Proficiency," *Arab World English Journal*, vol. 12, pp. 507–523, Sep. 2021, doi: 10.24093/awej/vol12no3.34.

[12] M. Carrier, "Automated Speech Recognition in language learning: Potential models, benefits and impact," *Training Language and Culture*, vol. 1, pp. 46–61, Feb. 2017, doi: 10.29366/2017tlc.1.1.3.

[13] P. Loureiro and M. J. Gomes, "Online Peer Assessment for Learning: Findings from Higher Education Students," *Education Sciences*, vol. 13, no. 3, Art. no. 3, Mar. 2023, doi: 10.3390/educsci13030253.

[14] İ. Y. Kazu and M. Kuvvetli, "A triangulation method on the effectiveness of digital game-based language learning for vocabulary acquisition," *Educ Inf Technol (Dordr)*, pp. 1–27, Mar. 2023, doi: 10.1007/s10639-023-11756-y.

[15] G. Wigglesworth, "Task and Performance Based Assessment," in *Encyclopedia of Language and Education*, N. H. Hornberger, Ed., Boston, MA: Springer US, 2008.

[16] S. Zheng and X. Zhou, "Enhancing Foreign Language Enjoyment through Online Cooperative Learning: A Longitudinal Study of EFL Learners," *International Journal of Environmental Research and Public Health*, vol. 20, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/ijerph20010611.

[17] K. Al-Dosakee and F. Ozdamli, "Gamification in Teaching and Learning Languages: A Systematic Literature Review," *Revista Romaneasca pentru Educatie Multidimensionala*, vol. 13, pp. 559–577, Aug. 2021, doi: 10.18662/rrem/13.2/436.

[18] A. Widayanti and I. Suarnajaya, "Students Challenges in Learning English Online Classes," *Jurnal Pendidikan Bahasa Inggris Undiksha*, vol. 9, p. 77, Jun. 2021, doi: 10.23887/jpbi.v9i1.34465.

[19] F. Lim, W. Toh, and T. Nguyen, "Multimodality in the English language classroom: A systematic review of literature ," *Linguistics and Education*, vol. 69, p. 101048, Jun. 2022, doi: 10.1016/j.linged.2022.101048.

[20] J. Nieminen, M. Bearman, and R. Ajjawi, "Designing the digital in authentic assessment: is it fit for purpose?," *Assessment & Evaluation in Higher Education*, Jun. 2022, doi: 10.1080/02602938.2022.2089627.

[21] F. Honrado and E. Biray, "Error Analysis in Spoken English among Grade 12 Students," *International Journal of Educational Management and Development Studies*, vol. 3, pp. 52–73, Mar. 2022, doi: 10.53378/352863.

[22] J. Fan and X. Yan, "Assessing Speaking Proficiency: A Narrative Review of Speaking Assessment Research Within the Argument-Based Validation Framework," *Frontiers in Psychology*, vol. 11, 2020, Accessed: May 14, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00330

[23] B. Poole, "Corpus-Based Approaches to English Language Teaching," *ELT Journal*, vol. 65, pp. 92–93, Dec. 2010, doi: 10.1093/elt/ccq080.

[24] T. Brunfaut, "Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront," *Language Testing*, vol. 40, no. 1, pp. 15–23, Jan. 2023, doi: 10.1177/02655322221127896.

[25] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[26] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121–2133, Jan. 2022.

[27] K.-K. Tseng, C. Wang, T. Xiao, C.-M. Chen, M. M. Hassan, and V. H. C. de Albuquerque, "Sliding large kernel of deep learning algorithm for mobile electrocardiogram diagnosis," *Computers & Electrical Engineering*, vol. 96, p. 107521, Dec. 2021, doi: 10.1016/j.compeleceng.2021.107521.

[28] S. Ge and X. Chen, "The Application of Deep Learning in Automated Essay Evaluation," 2020, pp. 310-318. doi: 10.1007/978-3-030-38778-5_34.

[29] M. Uto, Y. Xie, and M. Ueno, "Neural Automated Essay Scoring Incorporating Handcrafted Features," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Oct. 2020, pp. 6077–6088. doi: 10.18653/v1/2020.coling-main.535.

[30] K. Takai, P. Heracleous, K. Yasuda, and A. Yoneyama, "Deep Learning-Based Automatic Pronunciation Assessment for Second Language Learners," 2020, pp. 338–342. doi: 10.1007/978-3-030-50729-9_48.

[31] Y. Xu, "English Speech Recognition and Evaluation of Pronunciation Quality Using Deep Learning," *Mobile Information Systems*, vol. 2022, p. e7186375, Apr. 2022, doi: 10.1155/2022/7186375.

[32] K. Wang, F. Li, C.-M. Chen, M. M. Hassan, J. Long, and N. Kumar, "Interpreting Adversarial Examples and Robustness for Deep Learning-Based Auto-Driving Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9755–9764, Jul. 2022, doi: 10.1109/TITS.2021.3108520.

[33] S. Zhang, X. Su, X. Jiang, M. Chen, and T.-Y. Wu, "A Traffic Prediction Method of Bicycle-sharing based on Long and Short term Memory Network," *Journal of Network Intelligence*, vol. 4, no. 2, pp. 17–29, 2019.

[34] S. Gao, "Evaluation Method of Writing Fluency Based on Machine Learning Method," *Mathematical Problems in Engineering*, vol. 2022, p. e1253614, Sep. 2022, doi: 10.1155/2022/1253614.

[35] J. Rivera Muñoz, H. Berríos, and J. Arias-Gonzales, "Systematic Review of Adaptive Learning Technology for Learning in Higher Education," *Eurasian Journal of Educational Research (EJER)*, vol. 98, pp. 221–233, Jan. 2022, doi: 10.14689/ejer.2022.98.014.

[36] T.-L. Luo, M.-E. Wu, and C.-M. Chen, "A framework of deep reinforcement learning for stock evaluation functions," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5639–5649, Jan. 2020, doi: 10.3233/JIFS-179653.

[37] J. Talaghzi, A. Bennane, M. Himmi, M. Bellafkih, and A. Benomar, "Online Adaptive Learning: A Review of Literature," Sep. 2020, pp. 1–6. doi: 10.1145/3419604.3419759.

[38] D. Suendermann-Oeft et al., "A Multimodal Dialog System for Language Assessment: Current State and Future Directions," *ETS Research Report Series*, vol. 2017, no. 1, pp. 1–7, 2017, doi: 10.1002/ets2.12149.

[39] D. Dukić and A. Sovic Krzic, "Real-Time Facial Expression Recognition Using Deep Learning with Application in the Active Classroom Environment," *Electronics*, vol. 11, no. 8, Art. no. 8, Jan. 2022, doi: 10.3390/electronics11081240.

[40] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019.

[41] H. Hu, H. Zhou, Z. Tian, Y. Zhang, Y. Patterson, Y. Li, Y. Nie, and K. Richardson, "Investigating Transfer Learning in Multilingual Pre-trained Language Models through Chinese Natural Language Inference," *arXiv.org*, Jun. 2021, Accessed: May 15, 2023. [Online]. Available: https://arxiv.org/abs/2106.03983v1

[42] Y. Dai, Y. Liu, L. Yang, and Y. Fu, "An Idiom Reading Comprehension Model Based on Multi-Granularity Reasoning and Paraphrase Expansion," *Applied Sciences*, vol. 13, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/app13095777.

[43] M. Abdul-Mageed, C. Zhang, A. Elmadany, A. Rajendran, and L. Ungar, "DiaNet: BERT and Hierarchical Attention Multi-Task Learning of Fine-Grained Dialect," arXiv.org, Oct. 2019, Accessed: May 15, 2023. [Online]. Available: https://arxiv.org/abs/1910.14243v1

[44] M. Mohsen, "The effectiveness of using a hybrid mode of automated writing evaluation system on Efl students' writing," *Teaching English with Technology*, vol. 19, pp. 118–131, Feb. 2019.

[45] Y. Xu, "An Adaptive Learning System for English Vocabulary Using Machine Learning," *Mobile Information Systems*, vol. 2022, Jun. 2022, doi: 10.1155/2022/3501494.

[46] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue., "TIMIT Acoustic-Phonetic Continuous Speech Corpus." *Linguistic Data Consortium*, p. 715776 KB, 1993. doi: 10.35111/17GK-BN40.

[47] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4960–4964. doi: 10.1109/ICASSP.2016.7472621.

[48] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional Sequence to Sequence Model for Human Dynamics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5226–5234, doi: 10.1109/CVPR.2018.00548.