Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 6, December 2025, pp. 4732~4742
ISSN: 2302-9285, DOI: 10.11591/eei.v14i6.9539

Enhanced speech recognition in natural language processing

Siu-Hong Chang, Kok-Why Ng, Su-Cheng Haw, Yih-Jian Yoong
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia

Article Info

Article history:

Received Nov 10, 2024
Revised Sep 28, 2025
Accepted Oct 14, 2025

Keywords:

Automatic speech recognition
Bidirectional encoder representations from transformers
GPT-2
Natural language processing
Transformer

ABSTRACT

Speech recognition is crucial for helping individuals with physical disabilities access digital content. However, current systems have significant flaws that hinder user experience and complicate daily tasks. Environmental disturbances can cause misinterpretation, and existing automatic speech recognition (ASR) systems struggle with comprehending acoustic and linguistic nuances and handling diverse speaking styles and accents. To address these issues, a new model integrates bidirectional encoder representations from transformers (BERT) and transformer features with natural language processing (NLP) capabilities. This model aims to consolidate semantic, linguistic, and acoustic information extracted from the Kaldi speech recognition toolkit and improve accuracy by rescoring the list of N-best hypotheses. The innovative approach leverages advancements in NLP to enhance speech recognition's accuracy and robustness across various scenarios. Evaluations on the LibriSpeech dataset show that integrating BERT, transformer encoder, and generative pretrained transformer 2 for rescoring N-best hypotheses significantly improves transcription accuracy. The proposed model achieves a word error rate (WER) of 17.98%, outperforming other models. This development paves the way for advancements in speech recognition technology, offering better user experiences in real-world applications.

Corresponding Author:

Kok-Why Ng
Faculty of Computing and Informatics[, Multimedia University
Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
Email: kwng@mmu.edu.my

1. INTRODUCTION

With the gradual advancement of technology in the past decades, speech recognition is gaining popularity across most of the devices we use nowadays. Speech recognition, which was developed in the early 1900s, is an engineering technology that identifies and converts sound signals into text or commands has significantly changed how we live [1]. However, there still exists some reluctance that the current speech recognition technology is not able to comprehend the spoken words accurately, which results in erroneous commands being executed. Natural language processing (NLP) is a field of artificial intelligence (AI) and Linguistics designed to allow computers to recognize human's spoken words [2], [3], which has been empowering machine translation, automatic summarization, coreference resolution and much more for over 35 years. It is gradually shifting its focus from machine learning to deep learning-based algorithms to tackle with difficult tasks [4]. By applying this knowledge, the accuracy of comprehending audio signals would be improved, increasing the user's satisfaction level.

Contemporary speech recognition systems encounter significant obstacles in accurately transcribing audio inputs to text, primarily due to the intricate variability of human speech patterns and the disruptive influence of environmental disturbances. Malik *et al.* [5] stated that the prevalence of diverse regional dialects and individual speaking styles poses substantial challenges to achieving precise audio-to-text conversion. In addition, environmental factors, including ambient noise, significantly impede the system's ability to discern and interpret speech accurately, particularly in dynamic real-world environments [6]. Moreover, Han *et al.* [7] addressed the issue, rarely considering contextual information, which complicates the selection of the most accurate hypothesis, leading to inaccuracies in generating transcription.

Recent advancements in connectionist temporal classification (CTC) models have significantly improved speech recognition accuracy and efficiency through various innovative approaches [8], [9]. Omachi *et al.* [10] proposed a CTC model integrated with a conditional masked language model (CMLM), utilizing global and type-wise mask-predict algorithms, which showed superior performance on the CSJ and SLURP datasets. Lu and Chen [11] introduced context-aware knowledge transfer techniques, incorporating an enhanced Vanilla wav2vec2.0 and bidirectional encoder representations from transformer (BERT), which improved performance on the AISHELL-1 and AISHELL-2 datasets. Deng *et al.* [12] developed KT-RL and KT-CL models, employing BERT and GPT2 for knowledge transfer, with KT-RL-CIF achieving a 4.2% character error rate (CER) on the AISHELL-1 corpus. Additionally, Salazar *et al.* [13] introduced SAN-CTC, a self-attentional network that predicts tokens concurrently, achieving a 4.7% CER on the eval92 dataset and 2.8% CER on the LibriSpeech dataset. Jiang *et al.* [14] addressed the challenges of transformer-based speech recognition models, such as transcription difficulties and limited labeled data, by proposing an unsupervised pre-training algorithm called masked predicative coding (MPC). MPC utilizes a masked language model (MLM) applied to FBANK input and encoder output directly, aiming to enhance recognition in varied speech patterns and environmental conditions. Testing on the HKUST dataset during pre-training achieved a 23.3% CER, with increased pre-training data further reducing CER by 11.8% compared to the baseline. Nigmatulina *et al.* [15] introduced the ASR-NLP model to enhance automatic speech recognition (ASR) performance in noisy environments and improve callsign identification accuracy. This model first reduces the weights of likely callsign n-grams in the grammar finite-state transducer (G.fst) or decoding lattices and compares NLP-boosting ASR outputs, derived through named entity recognition (NER) based on the BERT model with surveillance data. A significant improvement from 32.1% to 60.4% is shown in recognition accuracy.

BERT-based models have significantly improved speech recognition by addressing various challenges [16]. Chaudhari *et al.* [17] utilized a BERT model to identify and correct speech recognition errors in radiology reports, achieving 75% accuracy. Baevski *et al.* [18] showed that fine-tuning BERT with transcribed speech data in the vq-wav2vec model reduces word error rate (WER). The proposed RescoreBERT [19] includes a MLM and discriminative loss functions, achieving a WER of 4.36 on the LibriSpeech dataset. Song *et al.* [20] introduced learning-to-rescore (L2RS), integrating BERT for text feature extraction and ESPnet for acoustic features, achieving a WER of 13.41% on the TED-LIUM dataset. Chuang *et al.* [21] highlighted BERT's improved performance with contextual word embeddings in ASR. Shin *et al.* [22] developed biSANLM, incorporating BERT for n-best list rescoring, achieving lower WER on the LibriSpeech task. Bai *et al.* [23] proposed LASO, a non-autoregressive model using BERT for token sequence generation, showing 50 times faster speed and low CERs. Fohr and Illina [24] suggested $BERT_{sem}$ and $BERT_{alsem}$ models for rescoring ASR hypotheses, with $BERT_{alsem}$ combined with ac./GPT-2 achieving the best performance. Illina and Fohr [25] improved $BERT_{alsem}$ with $BERT_{alsem-fg}$ and $P - BERT_{alsem}$, reducing WER by 1-3%. Chiu and Chen [26] proposed TPBERT, combining BERT with unsupervised topic modeling for N-best hypothesis reranking, achieving a WER of 20.49% on the AMI dataset. BERT-ASR, proposed by Nguyen *et al.* [27], utilizes whole word masking for efficient next word classification, showing lower perplexity and CER on the AISHELL-1 dataset. Finally, Yu *et al.* [28] introduced NAR-BERT-ASR, combining pretrained LM benefits with non-autoregressive capabilities, achieving the lowest CERs on the AISHELL-1 dataset with significant speed improvements.

Transformer-based models have shown significant improvements in ASR. Hrinchuk *et al.* [29] proposed a transformer ASR correction model, achieving an average WER of 14% on LibriSpeech datasets by applying transformer-based encoder-decoder architecture to a deep-convolutional E2E model called Jasper. Zhang *et al.* [30] developed a transformer-based spelling correction model that achieved a CER of 3.41% on a Mandarin dataset. Chen *et al.* [31] introduced a transformer with a directional decoder (STBD), achieving a CER of 5.8% on AISHELL-1. Wang *et al.* [32] highlighted that transformer-based models outperform BLSTM models in hybrid acoustic modeling. Li *et al.* [33] modified the self-attention decoder of the transformer by integrating it with DACS, achieving a WER of 5.5% on WSJ and a CER of 7.4% on AISHELL-1. Finally, Kim *et al.* [34] proposed Squeezeformer, which uses Temporal U-Net structure and depth-wise down-sampling, achieving the lowest WER of 2.27 on LibriSpeech.

RNN-based models have shown promising results in improving ASR. Klosowski [35] proposed an RNN model with an embedding layer, two LSTM hidden layers, and dense layers, which achieved increased accuracy from 0.402 to 0.936 and decreased loss from 2.779 to 0.265 after training for 500 epochs using Polish text data. Oruh *et al.* [36] applied an LSTM RNN model to address bandwidth limitations in ASR, achieving 99.36% accuracy on the Pannous dataset. Hori *et al*. [37] developed an RNN language model (RNN-LM) with a look-ahead mechanism, achieving 5.1% WER on WSJ and 5.4% WER on LibriSpeech. CNN-based models have been effectively combined with ASR systems to enhance topic detection accuracy. Sun *et al*. [38] proposed a multi-stream CNN framework using two ASR systems, HMM-BiLSTM and CTC, which process word embeddings. This model, tested on the Japanese CTS dataset, outperformed both an unsupervised model and a CNN with single-stream input. Additionally, Aitoulghazi *et al*. [39] introduced DarSpeech, a model based on deep speech with two CNN layers for feature extraction, demonstrating increased accuracy with larger input data sizes.

To overcome the issues identified earlier, this paper primarily aims to develop an advanced algorithm that can accurately refine the hypothesis in the presence of background noises, thereby mitigating the impact of environmental disturbances. This research aims to enhance the existing algorithm's capacity to comprehend diverse human speech patterns, accents, and linguistic nuances by integrating NLP methodologies to generate contextually relevant transcription. Lastly, to boost the accuracy of the output generated by NLP-based speech recognition, the existing algorithm will be refined to select the hypothesis that contains the lowest WER or CER, ensuring the highest accuracy score, from N-best hypotheses list.

2. METHOD

In this research work, a proposed model incorporating NLP techniques in enhancing speech recognition is introduced to increase the reliability and efficiency of ASR. The experimental setup and implementation of the proposed model of this research work are shown in the following subsections in providing a comprehensive understanding. As the proposed NLP-enhanced speech recognition is used to process the audio data consisting of acoustic signals carrying diverse semantic information (SI), some key presumptions must be adhered to achieve higher accuracy and efficiency in the proposed algorithm. One of the most important presumptions is that the contextual information of the audio data should be clearly defined in the presence of minimal background noise. Moreover, English language audio data should be used without long pausing between utterances, containing only one speaker, and with accurate transcriptions of test data. Lastly, only the first ten best hypotheses generated will be considered.

2.1. Dataset

LibriSpeech corpus, particularly the test-other subset, is chosen to be the primary dataset in evaluating the models due to its extensive collection of English-language audiobook recordings and the presence of more background noises, diverse accents, and other acoustic variations, which could increase the reliability of the results. This dataset is a widely recognized benchmark in the domain of ASR systems.

2.2. Data preprocessing

Kaldi voice recognition toolbox, an open-source toolkit utilizing a dynamic neural network capable of performing data preprocessing and extracting features required by our proposed model, is selected to generate the features required by our model. Kaldi must be installed in a Unix-like environment to allow the operations and audio test data in LibriSpeech to be converted into .wav format. Figure 1 depicts the flowchart of operations performed by Kaldi in generating the features required. To produce N-best hypotheses list, vectors are extracted and utilized in recognizing speakers and digitization tasks. This decoding process creates a lattice file that contains information about the probability of various sequences of words given the input.

The lattice file generated is converted into a list of N-best hypotheses, which consist of linear sequences of words, as shown in Figure 2. In addition, the lattices will be transformed into human-readable texts, representing the most likely word sequence, as shown in Figure 3. In extracting acoustic features from audio datasets, Mel-frequency cepstral coefficients (MFCCs) are computed, which provide the details of the signal's spectral envelope that assist in configuring vocal tract. These coefficients are essential for further analysis and model training. The MFCCs computed will be required to compute the cepstral mean and variance normalization (CMVN) statistics, which normalize the characteristics, ensuring consistency across different utterances and speakers. Prior to processing the data with the proposed model, hypothesis pairs will be created.

2.3. Bidirectional encoder representations from transformers and transformer architecture model

Figure 4 summarizes the architecture of the combination of BERT and transformer model. To create a format that the pre-trained BERT model can accept, the text of the hypothesis pair will be tokenized into individual units using BERT Tokenizer. The pre-trained BERT model used will accept the tokens of words from the transcript of hypothesis pair and produce embedding vectors that carry the contextualized representations of each token owing to its ability to capture contextual information from the tokens in bidirectional (left and right) simultaneously.

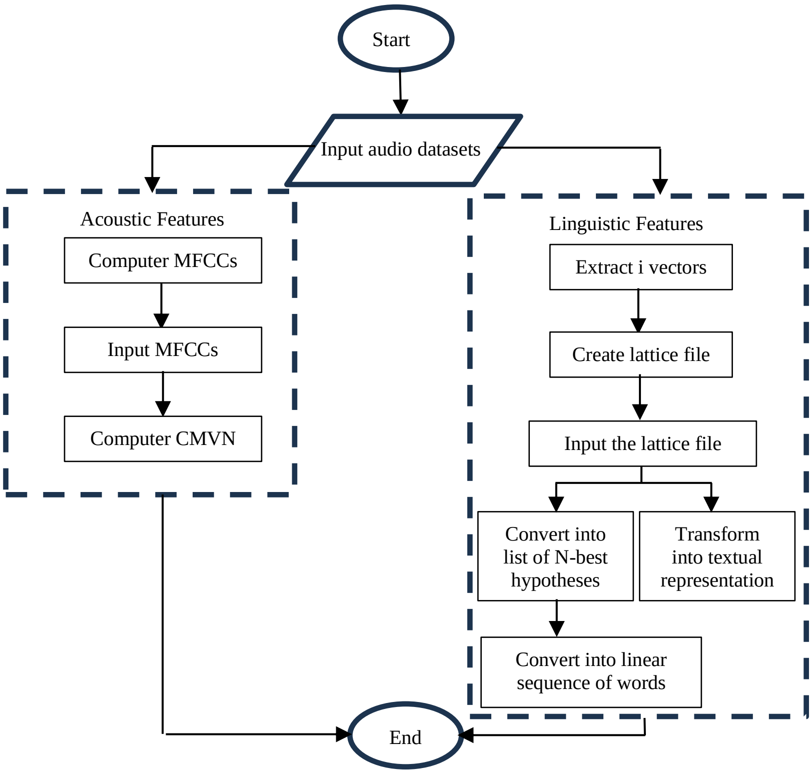
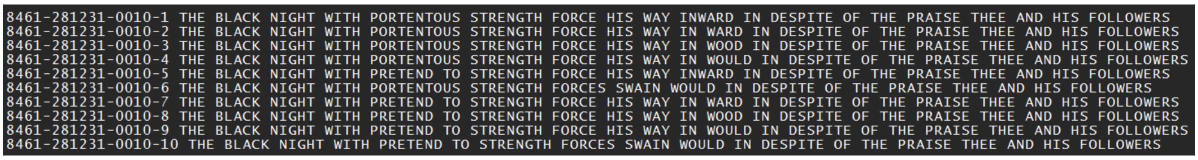
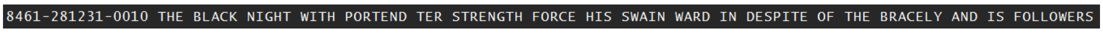


Figure 1. Flowchart of feature extraction with Kaldi



```
8461-281231-0010-1 THE BLACK NIGHT WITH PORTENTOUS STRENGTH FORCE HIS WAY INWARD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-2 THE BLACK NIGHT WITH PORTENTOUS STRENGTH FORCE HIS WAY IN WARD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-3 THE BLACK NIGHT WITH PORTENTOUS STRENGTH FORCE HIS WAY IN WOOD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-4 THE BLACK NIGHT WITH PORTENTOUS STRENGTH FORCE HIS WAY IN WOULD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-5 THE BLACK NIGHT WITH PRETEND TO STRENGTH FORCE HIS WAY INWARD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-6 THE BLACK NIGHT WITH PORTENTOUS STRENGTH FORCES SWAIN WOULD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-7 THE BLACK NIGHT WITH PRETEND TO STRENGTH FORCE HIS WAY IN WARD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-8 THE BLACK NIGHT WITH PRETEND TO STRENGTH FORCE HIS WAY IN WOOD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-9 THE BLACK NIGHT WITH PRETEND TO STRENGTH FORCE HIS WAY IN WOULD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
8461-281231-0010-10 THE BLACK NIGHT WITH PRETEND TO STRENGTH FORCES SWAIN WOULD IN DESPITE OF THE PRAISE THEE AND HIS FOLLOWERS
```

Figure 2. Sample of generated N-best hypotheses list



```
8461-281231-0010 THE BLACK NIGHT WITH PORTEND TER STRENGTH FORCE HIS SWAIN WARD IN DESPITE OF THE BRACELY AND IS FOLLOWERS
```

Figure 3. Sample of best generated hypothesis

The initial step in preparing BERT embeddings is crucial. Transformer encoder, which consists of 6 layers of blocks stacked hierarchically, will refine these outputs based on the attention-weighted relationships between tokens in the sentence, achieved with PyTorch. These transformer encoder blocks, each applying self-attention mechanisms to capture contextual relationships within the input text, learn hierarchical representations of the input text so more contextual information will be acquired, enhancing the contextual information captured in the initial step. As a result, a new set of embeddings will be produced. Acoustic features extracted from the Kaldi voice recognition toolbox will be concatenated with transformer embeddings using linear transformation to match the embedding dimension before being passed to two Bi-LSTM layers and several pooling operations. Bi-LSTM layers identify the dependencies in terms of sequence in bi-directional (forward and backward), which aids in understanding the temporal relationships, while average pooling and max pooling aggregate the data across the sequence and reduce the dimensionality of the data. The pooling operations performed significantly improve computation efficiency, reduce overfitting issues, and generate a more concise representation. The output from average and max pooling will be processed through two fully connected layers with the rectified linear unit (ReLU) activation function to allow the learning of complex relations as well as patterns of the features concatenated to develop more comprehensive representations. The first fully connected layer reduces the dimensionality of the feature representations, while the second fully connected layer produces the final output logits. The integrated feature representation will be passed into a sigmoid activation function to squash the output logits to a range between 0 and 1, facilitating binary classification tasks generating an output that carries the sentence-level SI by combining contextual, acoustic, and sequential information, producing a rich and comprehensive representation.

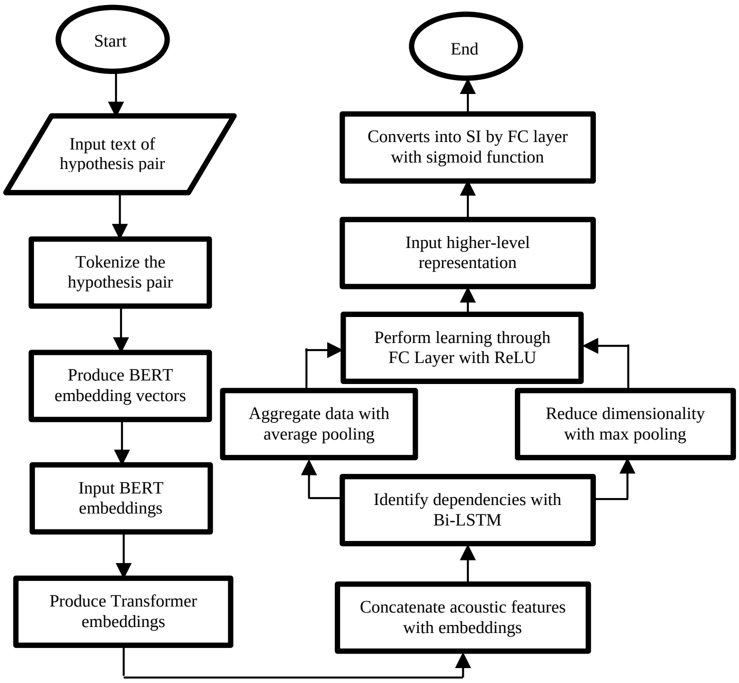

Start
Input text of hypothesis pair
Tokenize the hypothesis pair
Produce BERT embedding vectors
Input BERT embeddings
Produce Transformer embeddings
Concatenate acoustic features with embeddings
Identify dependencies with Bi-LSTM
Aggregate data with average pooling
Reduce dimensionality with max pooling
Perform learning through FC Layer with ReLU
Input higher-level representation
Converts into SI by FC layer with sigmoid function
End

Figure 4. Flowchart of BERT and transformer architecture model

2.4. Fine-graining with GPT-2 model

In generating fine-grained information required to fine-tune the model's understanding of text, GPT-2 model provides embeddings and token probabilities. The pre-trained GPT-2 model allows more flexibility in targeting the most appropriate word tokens [40] with its attention mechanisms. This model takes the hypothesis pair text as input and tokenizes it using GPT-2 tokenizer. This tokenized text is then passed through the GPT-2 model to generate the last hidden states, representing each token's embeddings in the sequence. These embeddings are transformed into probabilities using a SoftMax function, resulting in a probability distribution over the vocabulary for each token in the sentence. This approach aims to generate fine-grained information with additional features applicable in various downstream tasks. The flowchart below depicts the summarized procedures taken in fine-graining.

2.5. The main architecture

The selected hypothesis pair will then be passed to the BERT model to generate contextualized token embeddings, followed by a transformer encoder later to refine the embeddings based on attention-weighted relationships between tokens. This generates sentence-level SI. Besides that, fine-graining the pair of hypotheses using the GPT-2 model is conducted to generate representations with more features. Simultaneously, GPT-2 model is implemented in computing the probabilities of linguistic information of the hypothesis pair, which is $P_{lm}(h_i)$. This linguistic probability is multiplied with acoustic probability for each hypothesis, as shown in (1), to integrate both acoustic and linguistic information:

$$P_{ac}(h_i) * P_{lm}(h_i) \quad (1)$$

Concatenation of all the outputs from BERT and transformer architecture model, fine-graining using GPT-2 model, and combination of linguistic and acoustic probabilities from the hypothesis pair is carried out prior to processing it through fully connected layer (FC) with sigmoid activation function to compute the output, v_{ij}. This output will be defined as 1 if the WER of the first hypothesis, h_i, is less than the second in the pair, h_j, and 0 otherwise. The scores for each hypothesis in the pair will be updated based on v_{ij}, as stated in (2) and (3):

$$score_{sem}(h_i) += v_{ij} \tag{2}$$

$$score_{sem}(h_j) += 1 - v_{ij} \tag{3}$$

The cumulated score generated for each hypothesis will be used to select the top N hypotheses, with N representing the total number of hypotheses generated, to compute the pseudo probability, while $P_{sem}(h_i)$, the pseudo probability is multiplied by the linguistic probability and acoustic probability using weighted combinations of $\alpha, \beta, \gamma = 1$, see (4):

$$\widehat{W} = argmax_{h_i} \epsilon H \ P_{ac}(h_i)^{\alpha} * P_{lm}(h_i)^{\beta} * P_{sem}(h_i)^{\gamma} \tag{4}$$

The hypothesis with the best score will be chosen based on the above computations. The WER is computed between the best hypothesis and transcription, while the total duration includes processing and Kaldi durations. This proposed model can learn all the semantic, linguistic, and acoustic information of the audio data, hence empowering the accuracy of ASR in NLP in the recognition of spoken words.

2.6. Experimental setup

The baseline model, namely the Kaldi voice recognition tool, is set up in an open-source Linux-based operating system to allow its seamless integration and utilization within our research framework. This ensures compatibility with the required dependencies and libraries, such as CUDA for GPU acceleration, versions of Python, i.e., Python 3.10, and C++ compilers. A pretrained model is integrated, allowing the ASR to decode audio files using a time-delay neural network (TDNN) acoustic model to generate the word lattices for transcription. Utilizing the transcriptions generated, the recurrent neural network language model (RNNLM) is employed to select the best hypothesis. In experimenting with the proposed model, alpha (α), beta (β), and gamma (γ) are adjusted to 1 for the generation of the best hypothesis. Alpha (α) controls the weight of the acoustic probabilities which affect the acoustic feature alignment during transcription, beta (β adjusts the effect of language modelling on transcription results by modifying the impact of linguistic probability obtained from GPT-2, while gamma (γ) is a parameter provides additional control over the final selection based on combined scores for pseudo-probabilities used in ranking top-N hypotheses. These three parameters are adjusted as shown in (5), which is used to calculate the best hypothesis, as depicted in (4).

$$commbined\ probability = P_{ac}(h_i)^{\alpha} * P_{lm}(h_i)^{\beta} * P_{sem}(h_i)^{\gamma} \tag{5}$$

Besides that, a random selection model is developed to serve as a benchmark for comparison in our assessment. This model does not take into account auditory characteristics or contextual information when choosing words or sentences at random from a predefined vocabulary or corpus. Its inclusion guarantees an equitable comparison between various approaches. With these configurations, we hope to guarantee a fair comparison of the models and a thorough comprehension of the experimental process. The results obtained from the experiments conducted on the proposed model, Kaldi baseline model, and the random selection methodology are analyzed and discussed in the following chapter.

3. RESULTS AND DISCUSSION

In this section, the result of the proposed model is discussed in detail alongside a comparison with the baseline model and random selection methodologies, which are utilized to rescoring the list of N-best hypotheses. Table 1 illustrates the results obtained by evaluating each model with the LibriSpeech dataset, in which the key performance metrics utilized to assess their performance are the WER as well as the execution time. These results will be further elaborated in the subsections.

Table 1. The performance of different models on LibriSpeech dataset

Model	Performance metrics	
	WER (%)	Execution time (s)
Kaldi model	23.3974	157.599918
Random selection	18.5971	157.602507
Proposed model	17.9783	198.089685

3.1. Word error rate

The bar chart, as illustrated in Figure 5, shows the mean WER obtained by the three models used in the experiment. Kaldi's baseline model achieved the highest WER compared to the other two models. By applying random selection to the N-best hypotheses generated by the baseline model, the WER is reduced to 18.5971%. However, using our proposed model to restore the N-best hypotheses generated by the baseline model further reduces the WER to 17.9783%, making it the model with the best accuracy. Moreover, Figure 5 also provides more detailed information on the distribution of WER across the models, showing that the proposed model has the lowest median WER.

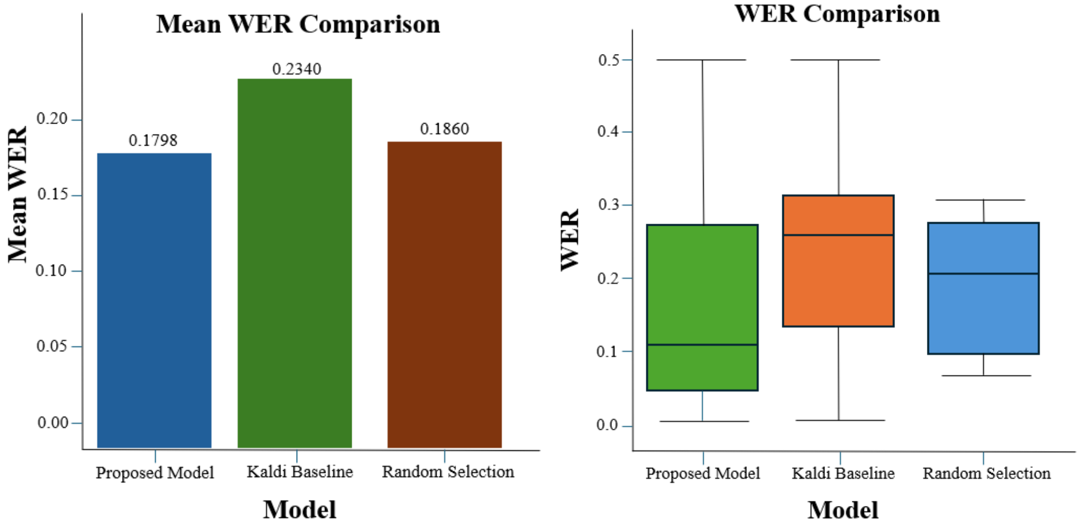


Figure 5. Comparison of WER of each model

3.2. Execution time

On the other hand, the details of the execution time of all the models are depicted in Figure 6. Based on Figure 6, the mean execution time of the proposed model is the longest at 198.09 seconds, which is approximately 40.5 seconds longer than the execution time of the baseline model. The random selection methodology has a difference of less than 0.01 seconds compared to the baseline model. More details on the distribution of the mean execution time are shown in Figure 6, where the Kaldi baseline model and random selection methodology have almost the same median execution time.

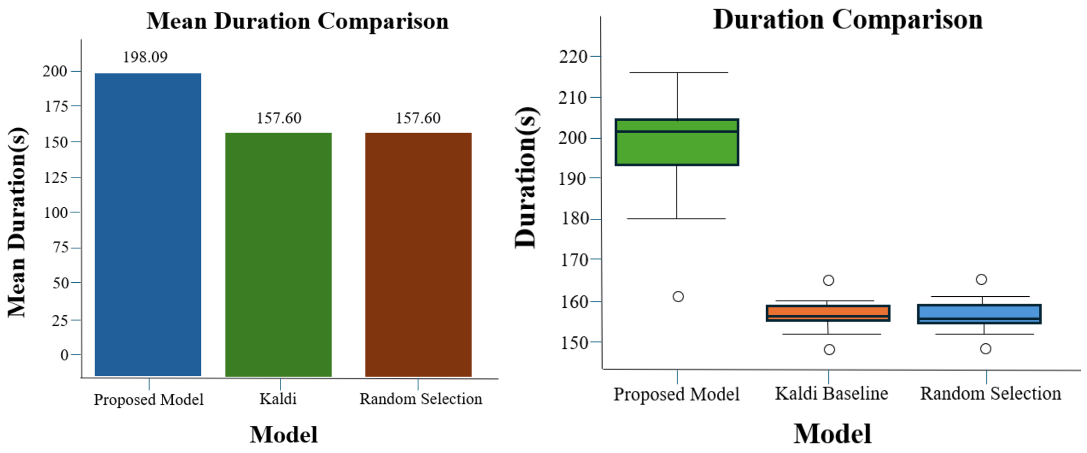


Figure 6. Bar chart of mean duration of each model

Based on the calculation obtained, which can be seen in Figure 7, the analysis shows that the proposed model outperforms both the baseline and random selection models in terms of WER, indicating higher accuracy in transcription. Specifically, it shows a significant WER improvement of 23.16%, despite requiring higher computational resources, as evidenced by the higher execution time. The experimental findings highlight the pivotal role of rescoring techniques in achieving accurate transcriptions, notably by integrating advanced NLP models such as BERT, transformer encoder, and GPT-2. Despite the increased computational demands associated with this approach, the results consistently demonstrate the proposed model's capability to generate highly reliable transcriptions. Its proficiency in refining hypotheses and enhancing transcription fidelity positions the proposed model as the optimal choice for NLP tasks prioritizing accuracy.

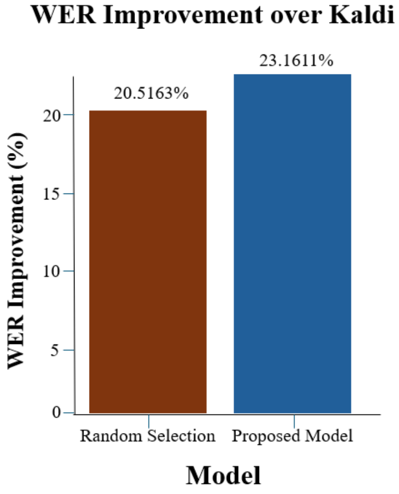


Figure 7. Bar chart of WER improvement shown by each model

4. CONCLUSION

In conclusion, a novel model integrating BERT, transformer encoder, and GPT-2 models is to overcome the limitations of the existing ASR, which include the inability to identify the spoken words in a noisy condition accurately, the presence of diverse speaking patterns leading to wrong transcriptions being produced, and the failure to choose the hypothesis with the most relevant contextual information due to the negligence of linguistic and acoustic features. With the proposed model, a significant improvement of 23.16% in WER, achieving a WER of 17.98% which outperforms other models, has been demonstrated while being evaluated with LibriSpeech, indicating enhanced transcription accuracy. However, a comparative analysis with other state-of-the-art models such as DeepSpeech, Wav2Vec 2.0, Conformer, ESPnet, and Whisper is necessary to highlight the relative performance of our model. To further improve the transcription accuracy of the proposed model, it would be beneficial to incorporate additional audio datasets, including multilingual and multidialectal audio files. This approach could optimize performance by allowing the model to adapt to diverse audio characteristics and nuances, thereby improving the transcription quality across different languages and dialects. Moreover, there exists a significant possibility of optimizing the model's configuration by adjusting its hyperparameters. To determine the settings that optimize performance measures like accuracy, efficiency, and resilience in a variety of real-world scenarios, this procedure may entail more methodical experimentation. Furthermore, while the results from the present small-scale datasets are encouraging, expanding to bigger and more diverse audio datasets would be advantageous. With a richer training environment brought about by this extension, the model could recognize more complex dependencies in audio inputs. Additionally, future work should also focus on assessing the model’s performance in real-time speech applications, considering factors such as efficiency, latency, and scalability for real-world deployment. The computational costs, inference speed, and memory usage of the transformer-based architecture should also be carefully evaluated, especially given the typical computational expense of such models.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING INFORMATION

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Siu-Hong Chang	✓	✓	✓		✓	✓		✓	✓	✓	✓			
Kok-Why Ng		✓		✓			✓	✓		✓	✓	✓	✓	✓
Su-Cheng Haw		✓	✓	✓			✓			✓			✓	
Yih-Jian Yoong		✓					✓			✓		✓	✓	

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, NKW, upon reasonable request.

REFERENCES

[1] K. Q. Yip, P. Y. Goh, and L. Y. Chong, “Social Messaging Application with Translation and Speech-to-Text Transformation,” *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 169–187, Jun. 2024, doi: 10.33093/jiwe.2023.3.2.13.

[2] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[3] T. A. Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. M. Su’ud, “Sentiment Analysis using Support Vector Machine and Random Forest,” *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 67–75, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.5.

[4] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing [Review Article],” *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/MCI.2018.2840738.

[5] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9411–9457, Mar. 2021, doi: 10.1007/s11042-020-10073-7.

[6] A. Sriram, H. Jun, Y. Gaur, and S. Satheesh, “Robust speech recognition using generative adversarial networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 5639–5643, doi: 10.1109/ICASSP.2018.8462456.

[7] M. Han *et al.*, “Improving End-To-End Contextual Speech Recognition With Fine-Grained Contextual Knowledge Selection,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 8532–8536, doi: 10.1109/ICASSP43922.2022.9747101.

[8] S. Wang and F. Long, “Robot Human-Machine Interaction Method Based on Natural Language Processing and Speech Recognition,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, pp. 759–767, 2023, doi: 10.14569/IJACSA.2023.0141278.

[9] Q. Q. Wong, K. W. Ng, and S. C. Haw, “Shirt-color recognition for the color-blindness,” *MethodsX*, vol. 13, pp. 1-9, Dec. 2024, doi: 10.1016/j.mex.2024.102866.

[10] M. Omachi, Y. Fujita, S. Watanabe, and T. Wang, “Non-Autoregressive End-To-End Automatic Speech Recognition Incorporating Downstream Natural Language Processing,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 6772–6776, doi: 10.1109/ICASSP43922.2022.9746067.

[11] K. H. Lu and K. Y. Chen, “A Context-Aware Knowledge Transferring Strategy for CTC-Based ASR,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, Jan. 2023, pp. 60–67, doi: 10.1109/SLT54892.2023.10022825.

[12] K. Deng *et al.*, “Improving Ctc-Based Speech Recognition Via Knowledge Transferring From Pre-Trained Language Models,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 8517–8521, doi: 10.1109/ICASSP43922.2022.9747887.

[13] J. Salazar, K. Kirchhoff, and Z. Huang, “Self-attention Networks for Connectionist Temporal Classification in Speech Recognition,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 7115–7119, doi: 10.1109/ICASSP.2019.8682539.

[14] D. Jiang *et al.*, “Improving Transformer-based Speech Recognition Using Unsupervised Pre-training,” *arXiv*, 2019, doi: 10.48550/arXiv.1910.09932.

[15] I. Nigmatulina, J. Zuluaga-Gomez, A. Prasad, S. S. Sarfjoo, and P. Motlicek, “a Two-Step Approach To Leverage Contextual Data: Speech Recognition in Air-Traffic Communications,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 6282–6286, doi: 10.1109/ICASSP43922.2022.9746563.

[16] J. Jayaram, Y. Kulkarni, L. V. Ganesh, P. Naveen, and E. A. Anaam, “Treatment Recommendation using BERT Personalization,” *Journal of Informatics and Web Engineering*, vol. 3, no. 3, pp. 41–62, Oct. 2024, doi: 10.33093/jiwe.2024.3.3.3.

[17] G. R. Chaudhari *et al.*, “Application of a Domain-specific BERT for Detection of Speech Recognition Errors in Radiology Reports,” *Radiology: Artificial Intelligence*, vol. 4, no. 4, Jul. 2022, doi: 10.1148/ryai.210185.

[18] A. Baevski, M. Auli, and A. Mohamed, “Effectiveness of self-supervised pre-training for speech recognition,” *arXiv*, 2019, doi: 10.48550/arXiv.1911.03912.

[19] L. Xu *et al.*, “Rescorebert: Discriminative Speech Recognition Rescoring With Bert,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 6117–6121, doi: 10.1109/ICASSP43922.2022.9747118.

[20] Y. Song *et al.*, “L2RS: A Learning-to-Rescore Mechanism for Hybrid Speech Recognition,” in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2021, pp. 1157–1166, doi: 10.1145/3474085.3481542.

[21] S. P. Chuang, A. H. Liu, T. W. Sung, and H. Y. Lee, “Improving automatic speech recognition and speech translation via word embedding prediction,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 93–105, 2021, doi: 10.1109/TASLP.2020.3037543.

[22] J. Shin, Y. Lee, and K. Jung, “Effective Sentence Scoring Method Using BERT for Speech Recognition,” *Proceedings of Machine Learning Research*, vol. 101, pp. 1081–1093, 2019.

[23] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, “Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring from BERT,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 1897–1911, 2021, doi: 10.1109/TASLP.2021.3082299.

[24] D. Fohr and I. Illina, “BERT-based semantic model for rescoring N-best speech recognition list,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISCA: ISCA, Aug. 2021, pp. 966–970, doi: 10.21437/Interspeech.2021-313.

[25] I. Illina and D. Fohr, “Semantic Information Investigation for Transformer-based Rescoring of N-best Speech Recognition,” *Ltc 2023*, 2023, [Online]. Available: https://hal.science/hal-03965397.

[26] S. H. Chiu and B. Chen, “Innovative Bert-Based Reranking Language Models for Speech Recognition,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, Jan. 2021, pp. 266–271, doi: 10.1109/SLT48900.2021.9383557.

[27] T. H. Nguyen, T. B. Nguyen, Q. T. Do, and T. L. Nguyen, “End-To-end named entity recognition for Vietnamese speech,” in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Hanoi, Vietnam, Nov. 2022, pp. 1–5, doi: 10.1109/O-COCOSDA202257103.2022.9997862.

[28] F. H. Yu, K. Y. Chen, and K. H. Lu, “Non-Autoregressive ASR Modeling Using Pre-Trained Language Models for Chinese Speech Recognition,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 1474–1482, 2022, doi: 10.1109/TASLP.2022.3166400.

[29] O. Hrinchuk, M. Popova, and B. Ginsburg, “Correction of Automatic Speech Recognition with Transformer Sequence-To-Sequence Model,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7074–7078, doi: 10.1109/ICASSP40776.2020.9053051.

[30] S. Zhang, M. Lei, and Z. Yan, “Automatic Spelling Correction with Transformer for CTC-based End-to-End Speech Recognition,” *arXiv*, 2019, doi: 10.48550/arXiv.1904.10045.

[31] X. Chen, S. Zhang, D. Song, P. Ouyang, and S. Yin, “Transformer with bidirectional decoder for speech recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Oct. 2020, pp. 1773–1777, doi: 10.21437/Interspeech.2020-2677.

[32] Y. Wang *et al.*, “Transformer-Based Acoustic Modeling for Hybrid Speech Recognition,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6874–6878, doi: 10.1109/ICASSP40776.2020.9054345.

[33] M. Li, C. Zorila, and R. Doddipatla, “Transformer-Based Online Speech Recognition with Decoder-end Adaptive Computation Steps,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, Jan. 2021, pp. 771–777, doi: 10.1109/SLT48900.2021.9383613.

[34] S. Kim *et al.*, “Squeezeformer: An Efficient Transformer for Automatic Speech Recognition,” *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[35] P. Klosowski, “Deep learning for natural language processing and language modelling,” in *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland, Sep. 2018, pp. 223–228, doi: 10.23919/SPA.2018.8563389.

[36] J. Oruh, S. Viriri, and A. Adegun, “Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition,” *IEEE Access*, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[37] T. Hori, J. Cho, and S. Watanabe, “End-to-end Speech Recognition with Word-Based Rnn Language Models,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 389–396, doi: 10.1109/SLT.2018.8639693.

[38] J. Sun, W. Guo, Z. Chen, and Y. Song, “Topic Detection in Conversational Telephone Speech Using CNN with Multi-stream Inputs,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 7285–7289, doi: 10.1109/ICASSP.2019.8682201.

[39] O. Aitoulghazi, A. Jaafari, and A. Mourhir, “DarSpeech: An Automatic Speech Recognition System for the Moroccan Dialect,” in *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco, May 2022, pp. 1–6, doi: 10.1109/ISCV54655.2022.9806105.

[40] A. Khan, K. Khan, W. Khan, S. N. Khan, and R. Haq, “Knowledge-based Word Tokenization System for Urdu,” *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 86–97, Jun. 2024, doi: 10.33093/jiwe.2024.3.2.6.

BIOGRAPHIES OF AUTHORS

Siu-Hong Chang completed his bachelor's degree from the Faculty of Computing and Informatics at Multimedia University. He is currently working at Huawei Malaysia as a data scientist to develop innovative solutions. His research interests are natural language processing (NLP), speech recognition, machine learning model, and GPT-2. He is committed to exploring emerging technologies to drive advancements in artificial intelligence and machine learning. He can be contacted at email: siuhong2075@gmail.com.

Kok-Why Ng is a senior lecturer in the Faculty of Computing and Informatics (FCI) in Multimedia University (MMU), Malaysia. He did his B.Sc. (Math) in USM, Penang, and his M.Sc. (IT) and Ph.D. (IT) in MMU, Malaysia. His research interests are in natural language processing, speech recognition, recommender system, 3D geometric modeling, and animation. He is also active in some research projects related to artificial intelligence, deep learning, and human blood cells. He can be contacted at email: kwng@mmu.edu.my.

Su-Cheng Haw is Professor at the Faculty of Computing and Informatics, Multimedia University, where she leads several funded research projects on the XML databases. Her research interests include XML databases, query optimization, data modeling, the semantic web, and recommender systems. She is also the chief editor for the Journal of Informatics and Web Engineering (JIWE). She can be contacted at email: sucheng@mmu.edu.my.

Yih-Jian Yoong received B.Sc. (Hons) in statistics and a Master of Science in applied statistics from Universiti Putra Malaysia in 1998 and 2000. He is a lecturer in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. His research interest is in applied statistics. He can be contacted at email: yjyoong@mmu.edu.my.