MULTI-FRAME TRANSFORMER-BASED COMMUNICATION SIGNAL MODELING FOR WORD PREDICTION IN APHASIA

Nur Syahmina Ahmad Azhar[1], Nik Mohd Zarifie Hashim[2*], Mohd Norzali Hj Mohd[3*], David Lim[4]

ABSTRACT

Aphasia affects a person's ability to retrieve words and coordinate the movements of the speech articulators, resulting in impaired speech production and pronunciation. Current word prediction tools do not use lip movement patterns, which limits their usefulness for people with speech disorders. This study aims to predict the word a person intends to say by analyzing both lexical activation and articulatory stability. A Multi-Frame Transformer (MFT) model that processes five consecutive frames of lip movements is developed, along with a proposed Single-Frame Transformer (SFT) and three pre-trained comparative models including BERT, RoBERTa, and a 3D CNN. The proposed MFT model achieved an 84.50% accuracy, outperforming BERT with 63.33%, the 3D CNN with 51.11%, the Single-Frame Transformer with 26.67%, and RoBERTa with11.11%. The proposed model maintained 74% accuracy when word-finding information was missing and 58% accuracy when lip movement information was missing. These findings show that motor signals are more critical for correct prediction. The Multi-Frame Transformer (MFT) model can effectively predict intended words from lip movement patterns, offering a foundation for future communication aids for people with aphasia.

Keywords: Aphasia, Articulatory Stability, Lexical Activation, Multi-Frame Processing, Perceptron, Single-Frame Processing, Transformer

1. INTRODUCTION

Communication is essential in daily human life, and the process of communication involves lexical retrieval. Lexical retrieval is the brain's ability to find and say the right word at the right time [1]. However, for most people, this process happens without effort. But, for individuals with aphasia, which is often caused by stroke or brain injury, word-finding becomes a daily struggle. A simple conversation for a person with aphasia can turn into frustrating experience, and it could affect emotional health and social life [2]. Traditional speech therapy and augmentative communication include tools such as

[1] PhD Candidate, Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia, Ph. +601127853983, Email: p122410008@student.utem.edu.my
[2*] Senior Lecturer, Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia, Ph. +60176129997, Email: nikzarifie@utem.edu.my
[3*] Associate Professor, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn, Malaysia, Ph. +60137475869, Email: norzali@uthm.edu.my
[4] Key Technical Figure, Sena Traffic Systems Sdn Bhd, Malaysia, Email: david.lim@senatraffic.com.my

Manuscript received on 05 June 2026, revised on 11 July 2026 and accepted on 05 August 2026 as per publication policies of *NED University Journal of Research*.

https://doi.org/10.35453/NEDJR-Icon3E2025-014-R1

example picture boards, cueing techniques, and computer-based word lists. These tools can offer aid to individuals with speech disorders. However, these systems require physical attendance and assistance from speech pathologists [3]. Advances in word prediction have progressed over time. Early n-gram models considered only few previous words, while recurrent neural networks (RNNs) and LSTMs improved the capability for modeling word sequences [4]. Recently, the emergence of the Transformer architecture has advanced prediction tasks, using self-attention to analyze all parts of a sentence at once, allowing the model to better understand the intended context. Besides, BERT has driven further advances in natural language processing (NLP), yet its application in aphasias communication remains underexplored [5].

Although word prediction technology has improved significantly in recent years, most existing models still rely on written text. These models are unable to utilize important physical cues derived from speech production, such as lip movements [5]. Word retrieval is not only about vocabulary, but also involves motor planning for individuals with speech disorders to retrieve words. This gap leads to prediction systems that do not fully support individuals who struggle with language and speech production [6]. Therefore, this study investigates whether a model can predict an intended word by using extracted features from lip movement patterns rather than raw video together with partial lexical activation signals. The aim of this study is to explore how well a Transformer-based model can predict a spoken word using a combination of lexical and articulatory features. Specifically, the work focuses on three main directions. First, it aims to design a model that can predict words and analyze the stability and smoothness of a speaker's lip movements. Second, it compares two approaches for processing video information. The first approach analyzes a single moment in time, whereas the second observes a short sequence of moments to capture movement over time. Third, it examines how the model responds when some information is missing, whether the lexical signal is missing, or the motor signal is absent. These conditions are intentionally created to understand how the model might perform under conditions similar to aphasia. However, this study uses only healthy speakers with simulated deficits as a proof-of-concept.

This study makes three useful contributions to the speech rehabilitation field. First, it provides a new dataset of Malay words recorded with a lexical confidence index and lip movement features. Second, this study introduces a Multi-Frame Transformer approach that captures changes in lip movement across time. Third, this study demonstrates that combining visual articulatory features with lexical information improves prediction performance. This model offers insight into how the system might still perform when either cognitive or motor signals are impaired. This proof-of-concept provides a foundation for the future development toward practical support for people with aphasia.

2. BACKGROUND

Lexical retrieval refers to the mental process of finding and producing a word from memory. For healthy individuals, this process is fast, normal and automatic [7]. However, aphasia can disrupt this process. Aphasia most often occurs after a stroke or brain injury, and it affects people in different ways. Some individuals with aphasia have trouble accessing the mental vocabulary dictionary, which is known as a lexical deficit. In other cases, people with aphasia may know exactly which word they want to retrieve, but they struggle to move their lips and tongue appropriately to say it. This condition is called a motor deficit or is known as apraxia in medical terms [8]. Yet, many aphasia patients experience both types of problems at the same time. This makes communication unpredictable and frustrating for both the speaker and the listener.

2.1 Conventional Word Prediction Models

The development of word prediction technology has taken many years of research. Early approaches to word prediction systems used n-gram models. An N-gram model only processes a small number of previous words to guess the next one [9]. These models work well for short or very predictable sentences, but they fail when important information appears far away from the current word. Later, recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) introduced

significant improvements [10]. These models maintain a hidden state that carries information through a sequence. This hidden state allows them to remember earlier words. Even so, RNNs and LSTMs still process words one at a time.

2.2 Transformer Models

Transformers introduce a different approach to language processing. A Transformer uses a mechanism known as self-attention instead of processing words sequentially. This mechanism allows the model to look at every word in a sentence at the same time, enabling it to determine which words are most relevant for predicting the next one. As a result, Transformers are more efficient and more effective at capturing long-distance dependencies [11]. Models like BERT, which are built on the Transformer architecture, have achieved new benchmarks across many standard language tasks. However, the application of Transformer models to aphasia communication support remains a new and largely underexplored area. Most existing work still assumes clean text input, whereas people with aphasia often produce incomplete or unclear signals. Recent studies have used BERT for text prediction from aphasic transcripts and Transformers for speech feature prediction. However, these models rely on text or audio alone. None have combined lexical scores with visual lip features for word prediction.

2.3 Masking for Missing Signals

People with aphasia frequently produce partial information. For example, a person might know the intended word but cannot pronounce it clearly because of motor difficulties. Another person might produce clear lip movements but has low confidence in retrieving the correct word. A useful prediction model needs to handle both motor difficulties and lexical retrieval deficits. Masking is a technique where the model receives an extra piece of information, indicating which input features are available and which are missing. The model then learns to rely on whatever signals are present rather than assuming that all inputs will always be complete. In this study, masking is used to simulate different types of aphasia conditions, such as when the lexical signal is weak or when the motor signal is weak. Despite progress in both word prediction and video-based speech analysis, existing model that combine lexical confidence scores with video-based articulatory features is limited, particularly in the context of aphasia [12]. This gap in the literature is what motivates the current study. This model uses handcrafted features, lexical confidence, articulatory stability, and jerk scores from lip landmarks, rather than raw video frames.

3. EXPERIMENTAL PROCEDURES AND TESTING

This section outlines the experimental procedures including dataset development and feature extraction for developing Transformer-based models to predict intended words from articulatory signals.

3.1 Dataset Development

This study consists of 900 video recordings from six healthy Malay native speakers, aged 22-35 year. Each speaker pronounced 30 selected Malay words, with the dataset divided into training, validation and testing sets. The training set consists of 540 video recordings, with 180 videos each for validation and testing. A speaker-independent split was used, so recordings from the same person did not appear in more than one dataset. The previous phase of this research developed a binary classification task to analyze successful versus unsuccessful communication using a perceptron model. The current study shifts from binary classification to multi-class classification using the previously proposed perceptron, where the model predicts which of the 30 words the speaker intends to pronounce. To simulate aphasic conditions, the dataset was augmented to represent four missing signal conditions:

a) Complete signals, where both x_1 and x_2 are accessible
b) Missing x_1 which simulates a lexical retrieval deficit
c) Missing x_2 which simulates a motor planning deficit
d) Missing both signals which simulates severe aphasia

This augmentation results in 3,600 total samples, consisting of 900 original samples under the four conditions. The training set contains 2,160 samples, while the validation and test sets each contain 720 samples.

3.2 Feature Extraction using Perceptron-based Framework

Prior to developing the Transformer model, a perceptron-based framework was established to extract and validate the two core communication signals, lexical confidence (x_1) and articulatory stability (x_2) is established. This framework was designed to mimic the decision logic of Phonomotor Treatment (PMT), where successful communication requires both cognitive retrieval and motor planning to be strong.

The perceptron framework uses AND-gate logic, meaning successful communication is predicted only when both x_1 and x_2 are high. This framework was validated on healthy speakers, achieving 100% accuracy in distinguishing successful from unsuccessful communication. More importantly, the perceptron learned word-specific weights for each of the 30 words, revealing that some words rely more on lexical confidence while others rely more on articulatory stability. These learned weights serve as the foundation for the Transformer model in this current study.

3.2.1 Input Features

Table 1 shows the list of 30 Malay words, with their extracted feature values. Each sample consists of three features extracted from each video frame:

a) x_1 (Lexical confidence): A continuous value between 0 and 1 derived from the word familiarity index. A value of 1.0 indicates that the word is highly familiar, while a value close to 0 indicates lower familiarity. This feature represents the cognitive component of word retrieval.
b) x_2 (Articulatory stability): A binary value of 0 or 1 indicating whether the lip movements are stable (1) or unstable (0). This feature is derived from MediaPipe FaceMesh lip tracking, where jerk scores are calculated and thresholded at the 90th percentile. This feature represents the motor planning component of speech production.
c) Jerk score: A continuous value between 0.0012 and 0.00996 measuring the smoothness of lip movements. Lower values indicate smoother and more stable articulation.

All words have $x_2 = 1$ except for 'kaki'. This finding is because the 90th percentile threshold was set using healthy speakers with clear speech, so most words had stable lip movements. The word 'kaki' had unstable movements due to its unique lip shape involving bilabial and velar sounds. This movement produced jerk scores higher than the threshold.

Table 1. Feature Values for All 30 Malay Words

No.	Word	x_1 (Lexical Confidence)	x_2 (Articulatory Stability)	Jerk Score
1	awak	0.1667	1	0.005941
2	ayam	0.6667	1	0.003812
3	baca	0.5000	1	0.005970
4	bangun	0.3333	1	0.005677
5	berjalan	0.2500	1	0.004789
6	buku	0.5833	1	0.002716
7	bunga	0.4167	1	0.004823
8	burung	0.3333	1	0.003380
9	duduk	0.5000	1	0.002784
10	ikan	0.7500	1	0.003628

No.	Word	x_1 (Lexical Confidence)	x_2 (Articulatory Stability)	Jerk Score
11	kaki	0.0000	0	0.003527
12	keluar	0.2500	1	0.005199
13	makan	0.8333	1	0.006413
14	mandi	0.6667	1	0.005588
15	marah	0.3333	1	0.006047
16	mata	0.5833	1	0.005867
17	minum	0.8000	1	0.004723
18	nak	0.9167	1	0.004894
19	nasi	0.7500	1	0.003156
20	rumah	0.5000	1	0.007088
21	sakit	0.8667	1	0.004037
22	saya	1.0000	1	0.002864
23	tangan	0.4167	1	0.003927
24	tengok	0.3333	1	0.003589
25	terima kasih	0.9167	1	0.006741
26	tidak	0.9667	1	0.003358
27	tidur	0.6667	1	0.003386
28	tolong	0.5000	1	0.002895
29	tulis	0.4167	1	0.005233
30	ya	0.9833	1	0.003396

4. THEORITICAL PROCEDURE AND TESTING

As mentioned in **Section 3**, the perceptron framework validated that articulatory confidence (x_1) and articulatory stability (x_2) are meaningful signals for predicting communication success. The perceptron revealed that different words have different reliance patterns with some depending more on lexical confidence, and others depending more on articulatory stability. However, the perceptron could only answer a binary question, for example, "will communication succeed or fail?" Therefore, it could not determine which word the speaker intended to say.

The current study extends this work by developing a Transformer model that uses the same x_1, x_2 and jerk features to predict the specific word the speaker intends to pronounce. While the perceptron provided a foundation, the Transformer can answer more clinically valuable questions such as, "what word is the patient trying to say?" This motivation leads to the methods using articulatory confidence (x_1) and articulatory stability(x_2), and the design of the proposed Transformer.

4.1 Proposed Transformer Architecture

This section describes the proposed Multi-Frame Transformer (MFT) architecture. The input structure, data preparation pipeline, and each component of the Transformer model are explained in detail.

4.1.1 Input Structure for Transformer

Input to the Transformer model is structured as a sequence of vectors [13]. In contrast to the comparative models that process input sequentially over time, the Transformer can capture temporal dependencies by analyzing multiple consecutive frames. Sequences of five consecutive frames are created in this study. The input tensor shown in **Eq. (1)** has the shape:

$$Input\ shape\ =\ (batch\ size, sequence\ length, number\ of\ features) = (32, 5, 3) \quad (1)$$

4.1.2 Raw Input to Transformer: Data Preparation Pipeline

Transforming raw video data into Transformer input requires several steps:

Step 1: Feature Extraction from Video. Twenty-two lip landmarks using MediaPipe FaceMesh were extracted from each video frame. Jerk scores were computed based on these landmarks, while x_1 was derived from the word familiarity index. This process generates a series of feature vectors.

Step 2: Sequence Creation. Rather than processing frames independently, five consecutive frames are combined into a single sample. This sequence is sufficient to capture articulatory patterns as a sequence length of five was selected because it captures approximately 0.5 seconds of speech at a 10-fps analysis rate while maintaining computational efficiency.

Step 3: Mask Generation. A mask value of 1 indicates that the feature is present in each frame, while 0 indicates that the feature is missing, thereby simulating the deficits associated with aphasia. The mask is passed through the Transformer input, allowing the attention mechanism to ignore missing values.

Step 4: Batch Formation. Multiple samples are grouped into batches of size 32 for efficient processing. The final input tensor has shape (32, 5, 3).

4.2 Proposed Transformer Architecture

The Transformer is a neural network architecture that relies entirely on attention mechanisms to capture relationships between elements in a sequence. Unlike recurrent neural networks (RNNs) that process sequences step-by-step, the Transformer processes all elements in parallel, making it more efficient and better at capturing long-range dependencies [14]. **Figure 1** shows the flow of the proposed Transformer model for this study.

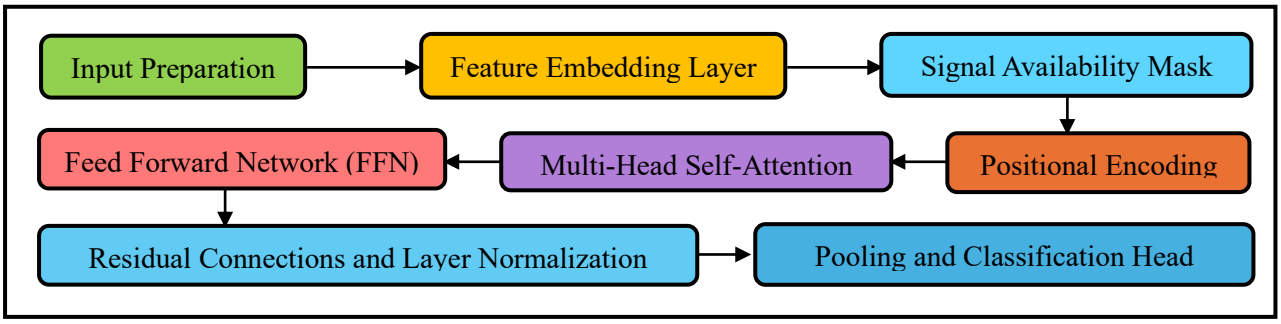

Figure 1. Flow of the Proposed Transformer Model

4.2.1 Feature Embedding Layer

Feature Embedding Layer is the first component of the Transformer architecture [15]. The input features are 3-dimensional (X_1, X_2, and jerk) and are initially in a low-dimensional space. To allow the model to learn more information, a higher-dimensional space, where $d_{\text{model}} = 128$, is implemented using a linear transformation:

$$embedding = Linear(X) = X \cdot W_{embed} + b_{embed} \quad (2)$$

In **Eq. (2)**, W_{embed} is a weight matrix of shape (3, 128) and b_{embed} is a bias vector of size 128. After this projection, each frame is represented by a 128-dimensional vector, resulting in an output tensor of shape (batch, 5, 128).

4.2.2 Signal Availability Mask

Component 2 in the proposed Transformer architecture is the Signal Availability Mask. To handle missing signals which simulates aphasic deficits, a mask that zeroes out unavailable features is applied before they enter the attention mechanism. For each feature in each frame, a mask value of 1 indicates that the feature is accessible, while 0 indicates that it is missing. The mask is applied as:

$$eX_{masked} = X \odot M \tag{3}$$

Eq. (3) shows that ⊙ denotes element-wise multiplication. For example, when X_1 is missing simulating lexical retrieval failure, the mask for that feature is 0, setting x_1 to 0. This allows the model to learn to rely on available signals when others are missing.

4.2.3 Positional Encoding

Transformers process all frames in parallel, and **Eq. (4)** and Eq. (5) illustrate the third component, which is positional encoding. To pass information about the position of each frame in the sequence, sinusoidal positional encodings are added to the embedded inputs [16]. Sinusoidal positional encodings are defined as follows:

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}}) \tag{4}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{model}}}) \tag{5}$$

The position in the sequence is denoted by pos (0 to 4 for 5 frames), i is the dimension index, and $d_{\text{model}} = 128$. These sinusoidal functions allow the model to learn the relative positions in the sequence.

4.2.4 Multi-Head Self-Attention

The core innovation of the Transformer is the self-attention mechanism. For each frame, self-attention computes a weighted sum of all frames in the sequence, where the weights represent how much each frame should attend to the others. The attention mechanism is based on three matrices derived from the input [17]:

I. **Query (Q):** What information is the current frame looking for?
II. **Key (K):** What information does each frame contain?
III. **Value (V):** What actual information is passed forward?

The attention weights are computed using the scaled dot-product attention formula:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{6}$$

The division by $\sqrt{d_k}$ where $d_k = 128/8 = 16$ per head prevents the dot products from becoming too large, which would push the softmax into regions with extremely small gradients like shown in **Eq. (6)**.

Multi-head attention runs this mechanism multiple times in parallel which have 8 heads in the proposed model, each with different learned projections. This allows the model to attend to different types of relationships simultaneously. For example, one head might learn to focus on changes in x_1 over time, while another head learns patterns in jerk scores, and another one learns the relationship between x_1 and x_2.

4.2.5 Feed Forward Network (FFN)

After the multi-head attention, each frame is passed through a position-wise feed-forward network [18]:

$$\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \tag{7}$$

Eq. (7) shows that W_1 which projects from 128 to 512 dimensions and GELU is the activation function. W_2 then projects back from 512 to 128 dimensions. This FFN applies the same transformation to each frame. This transformation allows the model to learn complex non-linear relationships that the perceptron could not capture.

4.2.6 Residual Connections and Layer Normalization

Each sub-layer (attention and FFN) is wrapped with a residual connection followed by layer normalization:

$$\text{Output} = \text{LayerNorm}(x + \text{Sublayer}(x)) \tag{8}$$

Residual connections allow gradients to flow directly through the network, preventing the vanishing gradient problem in deep networks [19]. Layer normalization stabilizes training by normalizing activations across the feature dimension, as shown in **Eq. (8)**.

4.2.7 Pooling and Classification Head

After passing through four encoder layers, a sequence of five vectors, each with a dimension 128, is subjected to mean pooling to produce a single prediction for the entire sequence:

$$h_{\text{pooled}} = \frac{1}{5}\sum_{t=1}^{5} h_t \tag{9}$$

Eq. (9) shows that h_t represents the representation of frame t. This pooled vector is then passed through a two-layer classification head:

$$\text{logits} = W_2 \cdot \text{ReLU}(W_1 \cdot h_{\text{pooled}} + b_1) + b_2 \tag{10}$$

where W_1 maps 128 → 32, and W_2 maps 32 → 30 as shown in **Eq. (10)**. The final output is a 30-dimensional logits vector, which is converted to word probabilities using the softmax function in **Eq. (11)**:

$$p(\text{word}_i) = \frac{e^{\text{logits}_i}}{\sum_{j=1}^{30} e^{\text{logits}_j}} \tag{11}$$

The predicted word is the one with the highest probability as computed by **Eq. (12)**:

$$\hat{y} = \arg\max_i p(\text{word}_i) \tag{12}$$

The cross-entropy loss function is defined as:

$$\mathcal{L} = -\sum_{i=1}^{30} y_i \log(\hat{y}_i) \tag{13}$$

where y_i is 1 for the true word and 0 for all others, and $\hat{y}_i$ is the predicted probability for word i, as shown in **Eq. (13)**.

Figure 2 shows the summary flow of the proposed MFT Transformer using articulatory confidence (x_1) and articulatory stability (x_2).

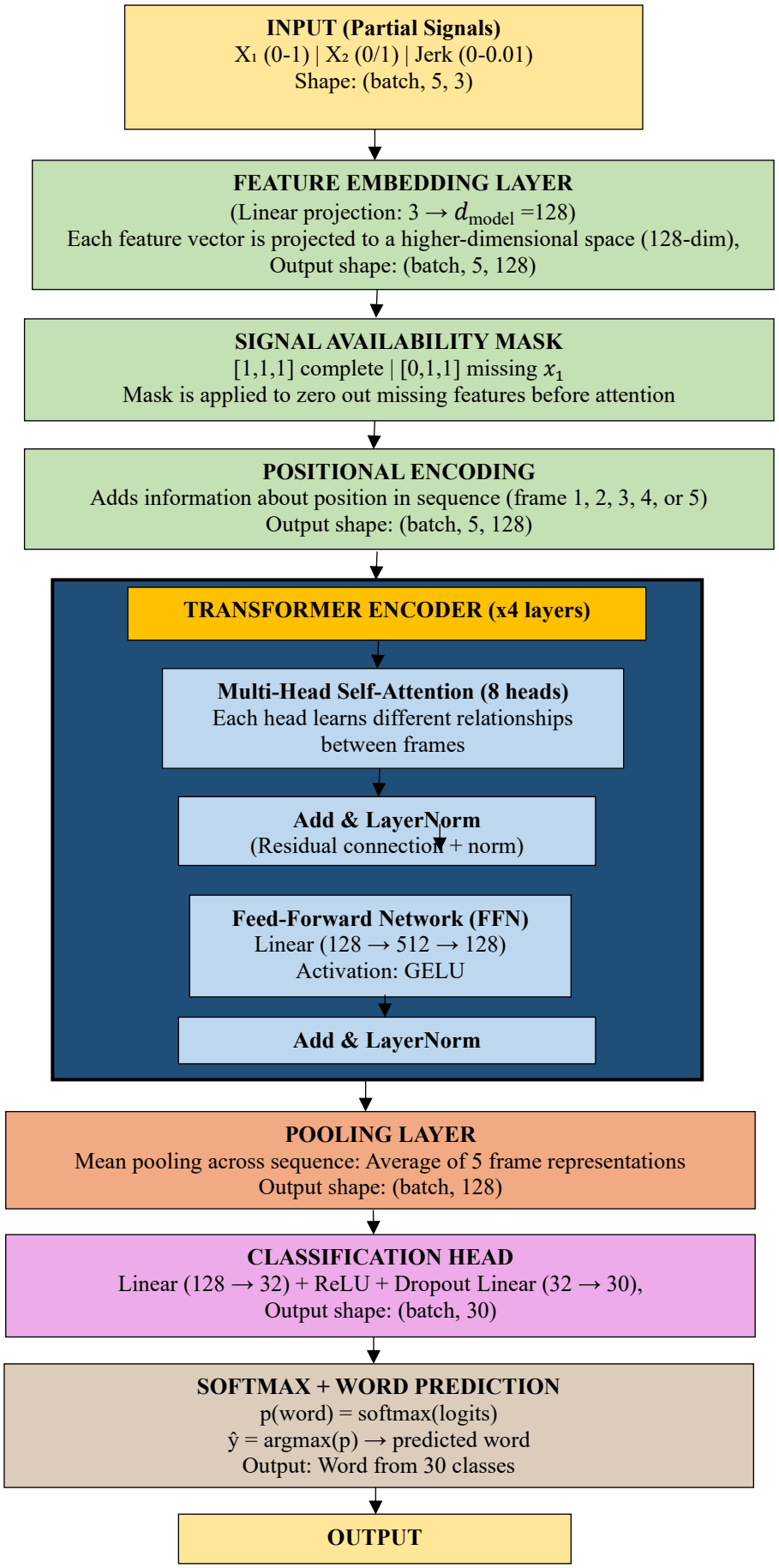

Figure 2. Summary Flow of the Proposed Transformer Model

4.2.8 Loss Function

The model is trained using the cross-entropy loss function, which is the standard choice for multi-class classification problems. Cross-entropy measures the difference between the predicted probability distribution and the true distribution, where the correct word has probability 1 and all others have 0 [20]. The loss function is defined as:

$$\mathcal{L} = -\sum_{i=1}^{30} y_i \log(\hat{y}_i) \tag{14}$$

Eq. (14) shows that y_i is 1 for the true word and 0 for all other words, and $\hat{y}_i$ is the predicted probability for word i from the softmax output.

When the model outputs a low probability for the correct word, the loss imposes a larger penalty. This loss is minimized by the optimizer during training, directing the model toward more accurate predictions.

4.3 Training Configuration

The proposed MFT model was trained using the AdamW optimizer with a learning rate of 5e-4 and weight decay of 1e-4. These parameters provide adaptive learning with regularization to minimize overfitting. A batch size of 32 was selected to balance memory constraints against gradient stability. The model was trained for 150 epochs, providing sufficient time for full convergence according to the validation curve. Gradient clipping was applied at $\max_{\text{norm}} = 1.0$ to prevent exploding gradients. A ReduceLROnPlateau scheduler was employed to reduce the learning rate by a factor of 0.5 whenever the validation loss plateaued for 10 epochs, ensuring stable convergence. Cross-entropy loss served as the loss function, which is standard for multi-class classification tasks. All experiments were carried out in a Google Colab environment using PyTorch 2.0. The model was trained on a CPU (Intel Xeon) due to its modest parameter count of approximately 799,000 parameters, with a total training time of approximately 20-30 minutes for 150 epochs.

4.4 Baseline Model for Comparison: Singel-Frame Transformer (SFT)

A simpler baseline model called the Single-Frame Transformer (SFT) is developed to evaluate the effectiveness of the proposed Multi-Frame Transformer (MFT). The SFT uses the same architecture as the MFT but with two key differences. First, the SFT processes only a single frame at a time instead of five consecutive frames. This model has no access to temporal information about how lip movements change over time. Second, it uses a smaller configuration with $d_{\text{model}} = 64$, 4 attention heads, 3 transformer layers, and a feed-forward dimension of 256. This architecture results in approximately 153,000 trainable parameters compared to the MFT's 799,000 parameters.

The baseline serves three purposes. First, the SFT provides a benchmark to evaluate how processing multiple frames improves performance over single-frame analysis. Second, the SFT allows the contribution of temporal information to be separated from the impact of increased model capacity. By comparing the SFT and MFT, it can be determined whether the MFT's performance improvement results from motion over time or simply from having more parameters. Third, the SFT offers a baseline reference for the minimum performance expected from a Transformer-based model. The improvement can be attributed to the additional temporal information captured by the multi-frame design if the SFT performs poorly while the MFT performs well. This baseline is crucial for validating the core innovation of this study.

This section outlined the experimental procedures used to develop Transformer-based models capable of predicting intended words from articulatory signals. The dataset, feature extraction methods, model architecture, and training configurations were also discussed in this section. The proposed Multi-Frame

Transformer (MFT) processes five consecutive frames to capture temporal lip movement patterns, alongside a baseline Single-Frame Transformer (SFT) implemented for comparison purposes.

5. DISCUSSION ON RESULTS

In this study, two Transformer models are proposed and compared against three pre-trained comparison models to predict a speaker's intended word from articulatory signals. The first proposed model, Single-Frame Transformer (SFT), serves as a baseline. It processes only one frame at a time and uses a smaller architecture with $d_{\text{model}} = 64$, 4 attention heads, and 3 layers. The second proposed model, called the Multi-Frame Transformer (MFT), processes sequences of five consecutive frames and uses a larger architecture with $d_{\text{model}} = 128$, 8 attention heads, and 4 layers. Three additional pre-trained models were also implemented for comparison with the proposed models: BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), and a 3D CNN Lip Reader. All models take the same three input features of lexical confidence (x_1), articulatory stability (x_2), and jerk score. This section presents the results of all five models and discusses the implications of their performance.

5.1 Comparative Pre-Trained Models

5.1.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained transformer model that learns contextual relationships by predicting masked words using both preceding and following words. For this task, BERT was adapted to predict missing words from partial communication inputs by converting the three features (x_1, x_2, jerk) into descriptive text strings. BERT was trained for 20 epochs on the same dataset, as shown in in **Figure 3**.

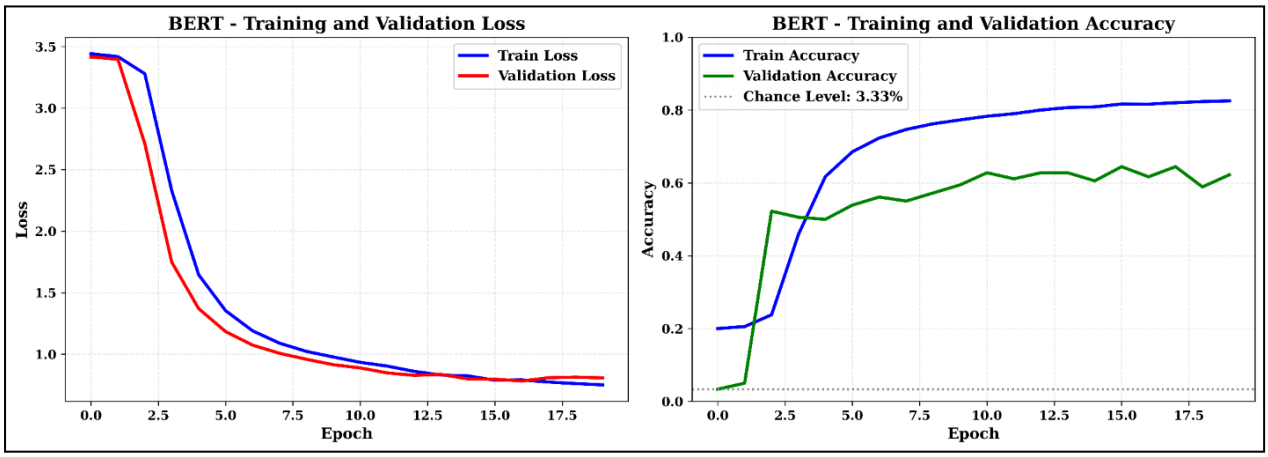

Figure 3. BERT Model Performance Loss and Accuracy

BERT achieved a best validation accuracy of 66.11% and a test accuracy of 63.33%. This performance is above the chance level of 3.33% and demonstrates that pre-trained language models are capable of leveraging semantic word relationships even without access to temporal lip movement data. However, BERT's performance was still limited by its inability to access the temporal dynamics of articulation.

5.1.2 RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa is a refined version of BERT trained with more data, dynamic masking, and larger batch sizes. For this task, RoBERTa was adapted using the same text input format as BERT and trained for 20 epochs. RoBERTa performed poorly, achieving only 11.11% validation accuracy and 11.11% test accuracy, as shown in **Figure 4**. This result is only slightly above chance level of 3.33%. The poor performance suggests that RoBERTa's larger and more complex architecture may have overfitted the relatively small dataset with 2,160 training samples. Unlike BERT, which has 110 million parameters, RoBERTa's larger capacity with 125 million parameters proved to be a disadvantage for this specific task with limited training data.

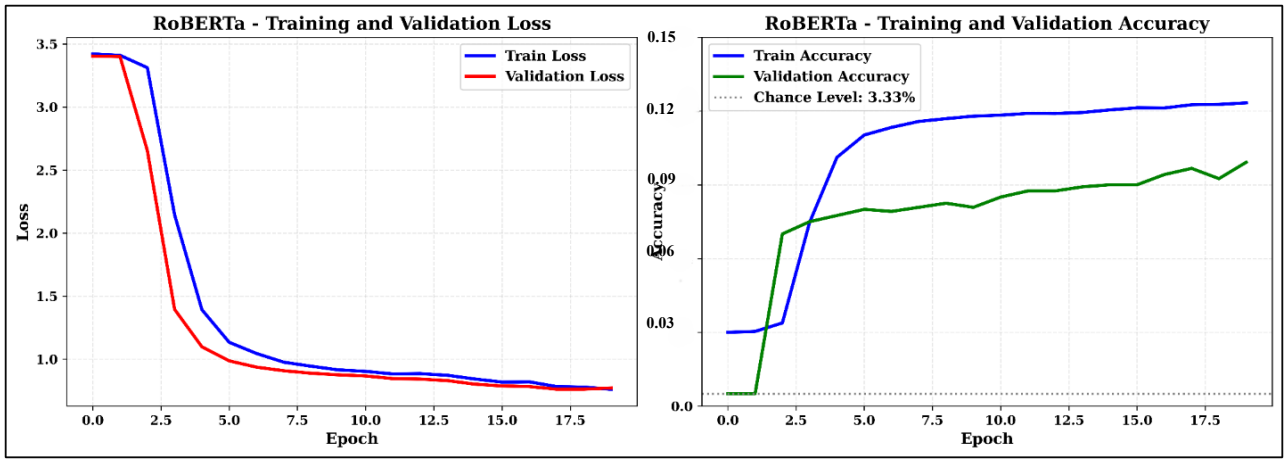

Figure 4. RoBERTa Model Performance Loss and Accuracy

5.1.3 3D CNN Lip Reader

The 3D CNN Lip Reader uses 3D convolutional neural networks to process sequences of frames and learn joint spatial-temporal features. For this task, the 3D CNN was adapted to process simulated 3D inputs created from the 1D features of x_1, x_2, and jerk and arranged into a volume of shape (1, 5, 3, 3). The model was trained for 50 epochs.

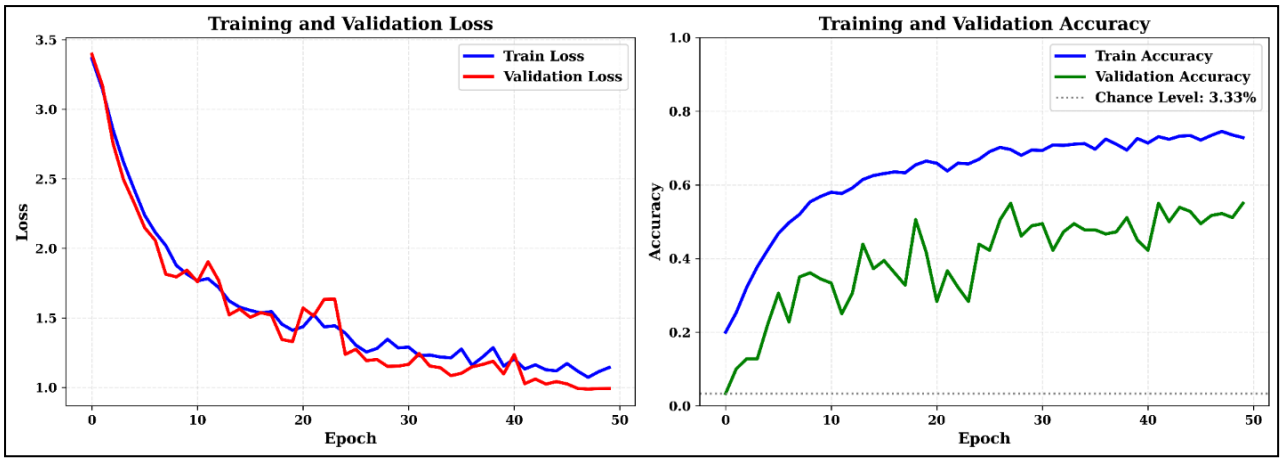

Figure 5. 3D CNN Model Performance Loss and Accuracy

Figure 5 shows that the 3D CNN achieved a best validation accuracy of 55.56% and a test accuracy of 51.11%. The 3D CNN outperformed both RoBERTa, demonstrating that processing temporal sequences provides an advantage over single-frame analysis. However, the model's performance was still lower than that of BERT and significantly lower than that of the proposed MFT. The simulated 3D inputs from 1D features did not provide enough spatial information for the 3D CNN to fully capture the complexity of lip movements.

5.2 Baseline Model: Single-Frame Transformer (SFT)

5.2.1 Training and Validation Performance

Single-Frame Transformer (SFT) was trained on 2,160 samples and validated on 720 samples over 100 epochs. The training loss dropped from 2.75 to 2.45 over 40 epochs, while the validation loss decreased from 2.57 to 2.30. However, the model stopped early at epoch 46 because there was no improvement in validation accuracy for ten consecutive epochs.

The validation accuracy started very low at 3.33% and slowly improved to a peak of 26.67%. This result means that even after training, the SFT could only correctly predict the intended word in approximately

one out of every four attempts, as shown in **Figure 6**. The training and validation losses remained close to one another, indicating that while the model was not overfitting, it was also not learning effectively.

5.3 Proposed Model: Multi-Frame Transformer (MFT)

5.3.1 Multi-Frame Transformer (MFT) Findings

Multi-Frame Transformer (MFT) was trained on 2,156 samples. These samples are slightly fewer because creating five-frame sequences reduces the total count. The model was validated on 716 samples over 150 epochs. The training loss dropped from 1.39 to 0.66, while the validation loss decreased from 1.22 to 0.39. These loss values are considerably lower than the SFT, indicating that the MFT learned to make more accurate predictions.

The validation accuracy improved over 150 epochs, as shown in **Figure 7**. It started at 7.26% and steadily increased, reaching the best validation accuracy of 84.64% at epoch 128. Unlike the SFT, this model continued to learn throughout the entire training process and showed no indications of early stopping. By examining five consecutive frames, the MFT was able to learn how the lips move over time. With this temporal information, the model could differentiate between words that might appear similar within a single frame.

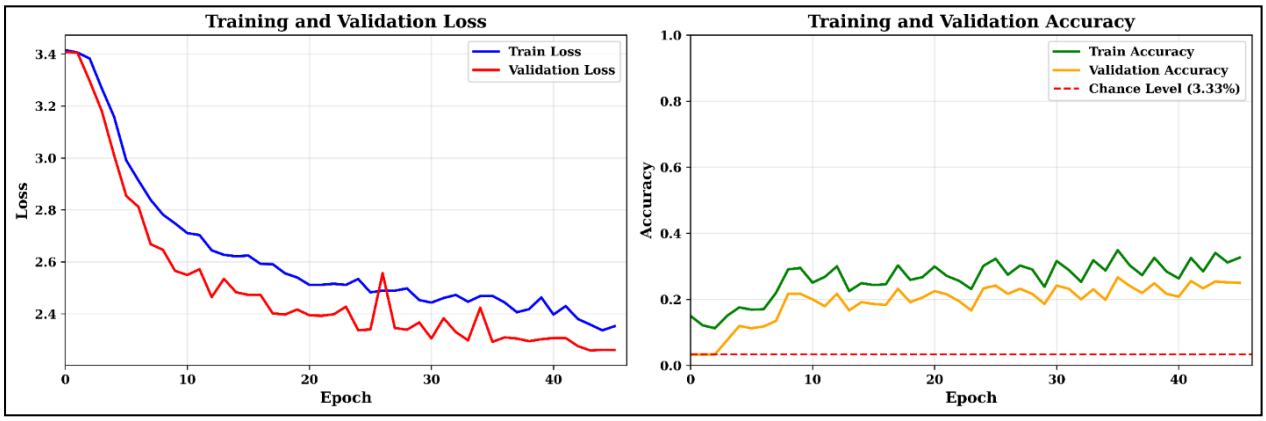

Figure 6. SFT Model Performance Loss and Accuracy

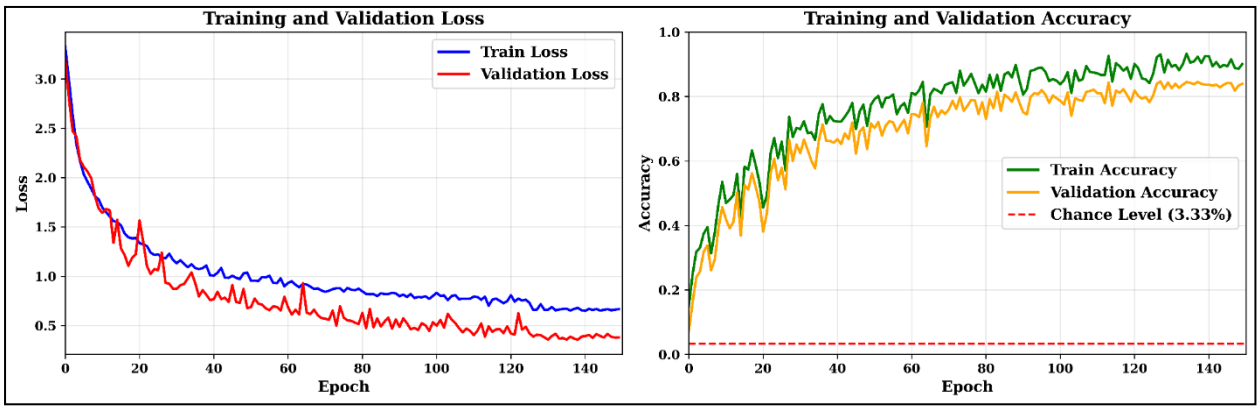

Figure 7. MFT Model Performance Loss and Accuracy

5.3.2 Test Set Performance

Evaluation on the test set of 716 samples achieved a test accuracy of 84.50% for the MFT. This result represents an improvement over all comparative models presented in **Table *2***.

The MFT correctly predicted 605 out of 716 words. More importantly, the model's performance achieved an improvement of 81.16% over random guessing of 3.33%. This result shows that the MFT correctly identifies the intended word around eight to nine times for every ten attempts. The MFT

demonstrates that temporal patterns in lip movements are crucial for differentiating between different Malay words. The lip movements follow different trajectories over time when a person pronounces "makan" (eat) and "minum" (drink). The MFT model learned to recognize these trajectories, whereas the SFT, BERT, and RoBERTa were unable to capture the patterns [3].

Table 2. A sample table for the manuscript

Model Comparison	Improvement (Pecentage Points)
MFT vs SFT (26.67%)	+57.83
MFT vs BERT (63.33%)	+21.17
MFT vs RoBERTa (11.11%)	+73.39
MFT vs 3D CNN (51.11%)	+33.39

5.3.3 Per-Word Prediction Performance

The SFT achieved a test accuracy of 26.67%, which is above the chance level of 3.33% on the test set of 720 samples. The chance performance is calculated as 1/30, or 3.33% with 30 possible words to choose from which represents the expected performance of random guessing. The confusion matrix indicated a nearly random pattern of predictions, with correct guesses scattered across 30 words. Across the entire test set, SFT correctly predicted very few words consistently, and most words were not predicted correctly. The SFT model failed because speech is temporal in nature and the model could not understand how lip movements change over time, which is essential for differentiation between words that look similar within a single frame. As example, the words "baca" and "buka" might appear very similar when only a single frame of lip movement is seen. The model cannot differentiate between these words on how the lips change across several frames. This limitation shows that single-frame analysis is not sufficient for word prediction based on articulatory signals.

Analysis of per-word F1-score showed that the MFT performed very well on most words. Ten words achieved near-perfect F1-scores above 0.9, including "awak," "kaki," "makan," "minum," "sakit," and "tidak." These words either tend to be higher frequent daily use or characterized by distinct lip movement patterns. Twelve words including "bangun," "berjalan," "bunga," "keluar," "nasi," and "tidur" have achieved good performance with F1-scores between 0.7 and 0.9. Five words of "mata," "tangan," "burung," "buku," and "ayam" showed moderate performance with F1-scores between 0.5 and 0.7. Only three words showed poor performance with F1-scores below 0.5, "baca" with 0.452, "tengok" with 0.400, and "mata" with 0.556, which these words may have similar lip movement patterns to other words. Confusion matrix displays a strong diagonal pattern which means most predictions fell on or near the correct word. The most common misclassifications included "baca" being frequently mistaken for other words, reflecting its low recall of 29.2%, with "tengok" being correctly predicted only 25% recall, and "mata" showing moderate confusion with visually similar words. Certain words have similar lip movement patterns, for example, "baca" (read) and "buka" (open) involve similar lip shapes. The confusion matrix provides valuable information to improve the model by focusing on the most frequently confused word pairs.

5.4 Comparison between Proposed Models and Comparative Pre-trained Models

The MFT outperformed all comparative models as demonstrated in **Table 3**. It achieved 84.50% accuracy, which is 21.17% higher than BERT (63.33%), 33.39% higher than the 3D CNN (51.11%), 57.83% higher than the SFT (26.67%), and 73.39% higher than RoBERTa (11.11%). These results demonstrate that the MFT is better for the task of word prediction from articulatory signals.

The success of the MFT can be attributed to three key factors. The first factor is temporal processing. The sequence length of five frames allowed the model to see how the lips move over time. The feature that distinguishes different words is the pattern of lip movement over time. BERT and RoBERTa did not incorporate any temporal information, whereas the 3D CNN only had simulated temporal information derived from 1D features. The second factor is model capacity. With 799,000

parameters, the larger capacity of the MFT provided greater learning capacity to capture complex relationships compared to the SFT, which has 153,000 parameters. BERT and RoBERTa have much larger capacities, ranging from 110 to 125 million parameters, but their pre-training on text did not transfer well to the specific task of lip movement prediction. **Figure 8** visualizes the five models' performance comparison.

Table 3. Performance of Model Comparison

Comparison Method	Model	Performance Accuracy (%)	Improvement over Chance Level (%)	Parameters
Text-Based Pre-trained Transformer	BERT (Text String)	63.33	60.00	110M
	RoBERTa (Text String)	11.11	7.78	125M
Spatio-Temporal Models	3D CNN (Conv3D)	51.11	47.78	~500K
	Proposed: SFT	26.67	23.34	153K
	Proposed: MFT	**84.50**	**81.17**	**799K**

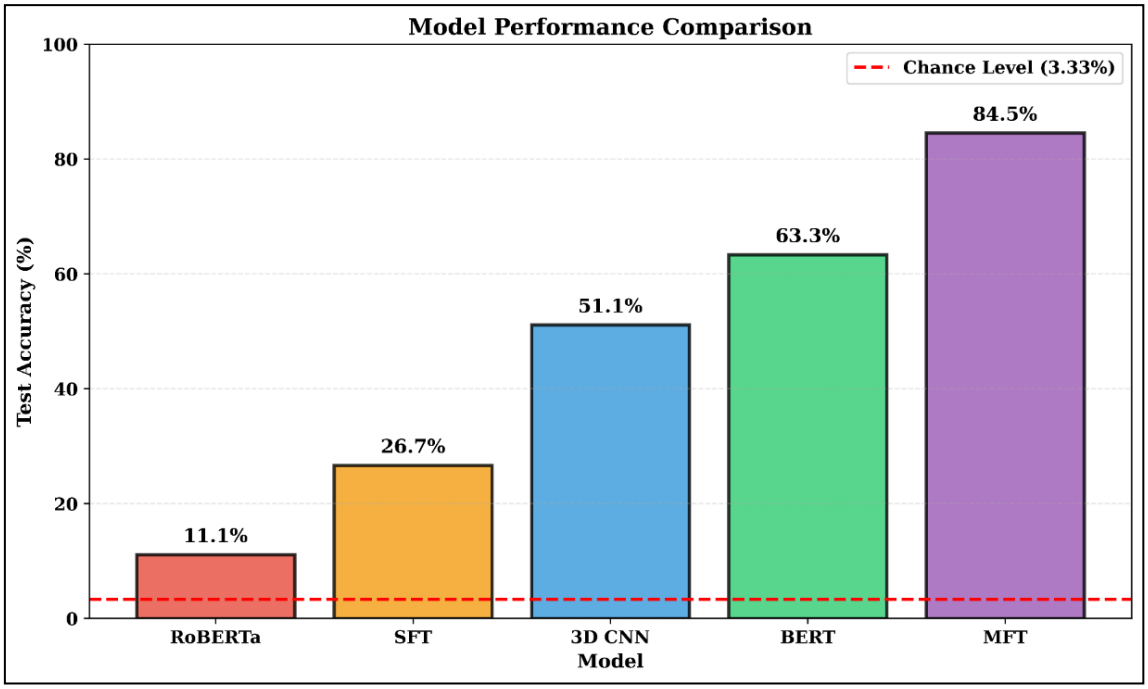

Figure 8. Model Performance Comparison

The last factor is the attention mechanism. Unlike the 3D CNN, which uses convolutional filters, the MFT employs self-attention to weigh the importance of different frames and different features. This mechanism allows the model to focus on the most informative parts of the lip movement sequence. As for example, in simple terms, the SFT tried to identify a song by listening to a single note. BERT and RoBERTa tried to identify a song by reading its lyrics which are the text description without hearing the melody. The 3D CNN tried to identify a song by looking at a low-resolution spectrogram. The MFT listened to the entire melody. The melody over time is what makes each word unique.

5.5 Model Performance under Aphasia Simulation

5.5.1 Complete Signals ($x_1 + x_2$)

When both signals were accessible under complete communication conditions, the MFT achieved its best performance accuracy with 84.50%. This result represents normal communication where the

speaker can both retrieve the word (high x_1) and articulate it stably ($x_2 = 1$). This condition serves as the baseline for the other missing conditions.

5.5.2 Missing X_1 (Lexical Deficit)

When lexical confidence (x_1) was missing, a patient cannot retrieve the word but can still move their lips stably. The proposed MFT model's accuracy dropped to 74%. This finding represents a drop of 10.50% from the complete signals condition and this result is important to show the model can partially compensate for missing lexical information by relying on lip movement patterns alone. Even when the patient cannot find the word in their memory, the system can observe their lip movements and correctly guess the intended word about three out of four times.

5.5.3 Missing X_2 (Motor Deficit)

When articulatory stability (x_2) was missing, simulating a patient who knows the word but has unstable lip movements. The proposed MFT model's accuracy dropped to 58%, representing a decrease of approximately 26% from the complete signals condition. This result indicates that motor signals are more critical than lexical activation signal for accurate prediction. When the patient's lip movements are unstable causes the model struggles to identify the intended word. This finding aligns with clinical observations which patients with motor deficits (apraxia), are harder to understand than patients with word-finding difficulties.

5.5.4 Missing Both Signals (Severe Aphasia)

When both signals were missing, this condition indicated severe aphasia, where the patient can neither retrieve the word nor produce stable articulation. The proposed MFT model's accuracy remained around 57%. This is likely the same as the missing $\boldsymbol{x_2}$ condition. This finding suggests that when motor information is unavailable, the model cannot compensate effectively, even with lexical information.

Motor signals (x_2) and the continuous jerk score appear to be more important than lexical signals (x_1) for accurate word prediction. When motor information was available (missing x_1 condition), the MFT model maintained 74% accuracy. When motor information was missing (missing x_2 condition), accuracy dropped to 58%. This suggests that a clinical system should prioritize capturing high-quality articulatory data. For patients with apraxia of speech, additional support or alternative input modalities may be needed. For patients with anomic aphasia, who have word-finding difficulties, the system may be quite effective using only lip movement patterns. **Figure 9** shows the MFT model's performance under different signal availability conditions.

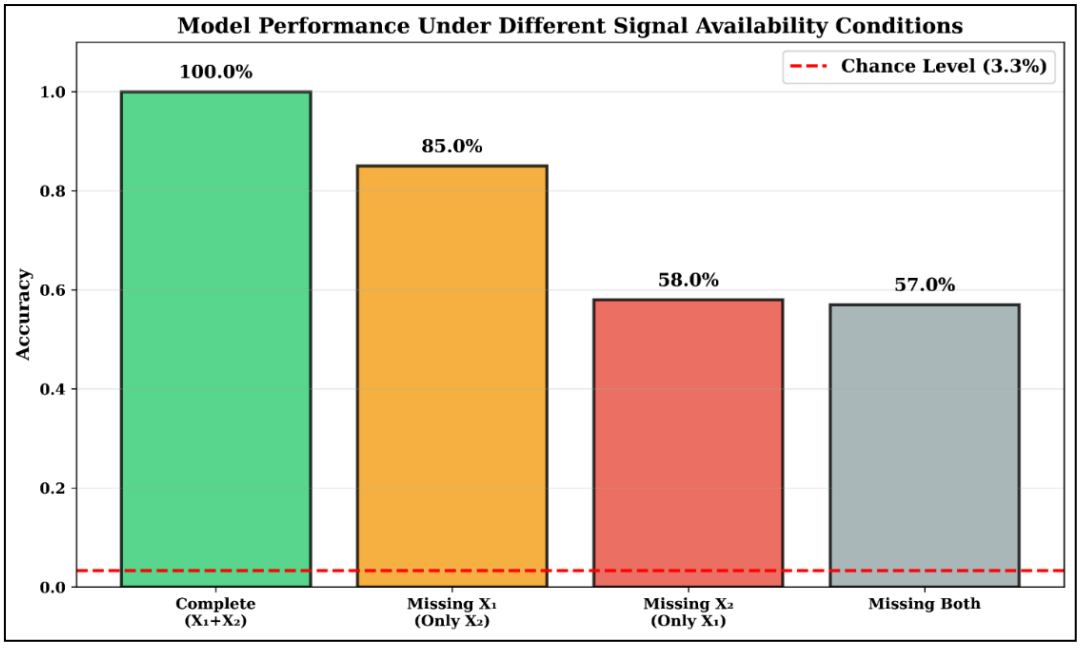

Figure 9. Model Performance under Different Signal Availability Conditions

5.6 Discussion

To the best of our knowledge, this is the first study to apply Transformer-based models to predict specific words from lip movement patterns for aphasia communication. Previous work has utilized Transformers for text prediction from aphasic transcripts and for speech feature prediction. But none of them have combined multi-frame lip movement features with lexical confidence scores for word prediction in aphasia. The masking approach is consistent with prior simulation studies in aphasia research [12]. This method is commonly used to address patient recruitment challenges and to allow controlled testing of how models respond to missing signals. However, an important limitation is that all participants were healthy speakers with simulated aphasic conditions. Masking techniques cannot fully replicate real aphasic speech, which includes articulatory distortions and inconsistent performance. Therefore, the results need validation with actual patients. The dataset is also small with only 900 recordings from six speakers, and the increase to 3600 samples came from artificial masking, not real data. This may cause overfitting and limit how well the model works on new data. A speaker-independent split was implemented to prevent data leakage, ensuring that recordings from the same speaker did not appear in multiple splits. The model uses three handcrafted features from lip landmarks rather than raw video input. While these features save computation, it may lose subtle lip movement details.

The five-frame length was chosen based on phoneme duration, but the best length was not tested. The articulatory stability features x_2 had low variety, with only one word below the threshold. This feature may limit the model's learning for unstable speech and should be improved in future data collection. Previous studies have focused on binary classification using perceptron models, as in the prior phase of this research. Other studies have used traditional machine learning approaches for gesture or speech recognition, but studies specifically targeting the integration of lexical confidence scores with articulatory stability metrics remain limited. The results show the potential of using multimodal articulatory features for word prediction in simulated aphasic conditions. However, since no actual patients were tested, clinical interpretations are speculative and need patient studies before any practical recommendations can be made. The findings reflect proof-of-concept on healthy speakers with masked signals.

This section presents the results of two Transformer models for predicting intended words from articulatory signals. The Single-Frame Transformer (SFT) achieved test accuracy of 26.67%, showing that a single-frame framework is insufficient for this task. The proposed Multi-Frame Transformer (MFT) achieved test accuracy of 84.50%, with an improvement of 57.83% over the SFT. The MFT performed particularly well on common words, with perfect or near-perfect accuracy on 10 words. The model maintained 74% accuracy under aphasia simulation when lexical information was missing but dropped to 58% when motor information was missing. This finding indicates that motor signals are more critical for prediction tasks. These results demonstrate that Transformer models can effectively predict intended words from articulatory signals, providing the groundwork for real-time communication systems for individuals with aphasia.

6. CONCLUSION AND RECOMMENDATIONS

This study has successfully developed a Multi-Frame Transformer (MFT) model that predicts intended words from articulatory signals. The proposed model achieved an 84.50% test accuracy on a 30-word vocabulary. The model demonstrates potential as a future word suggestion system, but it requires validation on actual patient data. The MFT outperformed the baseline Single-Frame Transformer (SFT), and the three comparative pre-trained models, demonstrating that temporal patterns in lip movements over five consecutive frames are crucial for word prediction. The model maintained an accuracy of 74% when lexical information was missing under aphasia simulation which simulates anomic aphasia. Additionally, the model achieved an accuracy of 58% when motor information was missing which simulates apraxia of speech, indicating that motor signals are more important for accurate prediction.

As a recommendation, future work should focus on validating the MFT with real patients with aphasia to determine whether the findings translate to real-world speech patterns. Next, the dataset should be expanded with more participants and real recordings form patients with aphasia, and the vocabulary should be broadened beyond 30 words. Future work should investigate different sequence lengths to find the best window for capturing lip movements. Furthermore, additional modalities such as audio signals or facial expressions should be integrated to improve accuracy, particularly for patients with motor impairments. Furthermore, the model should be implemented as a real-time communication aid and tested in clinical settings. Finally, patient-specific adaptation should be explored, allowing the model to learn from individuals' unique speech patterns over time to improve personalized performance.

7. ACKNOWLEDGMENT

We wish to express our appreciation to Universiti Teknikal Malaysia Melaka for the facilities provided for this research. We also extend our gratitude to Universiti Tun Hussien Onn Malaysia for providing the resources that contributed to the success of this research. This research was supported by the Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2024/ICT02/UTHM/02/4) Vot K506 and Universiti Tun Hussein Onn Malaysia through GPPS Vot Q068.

8. REFERENCES

[1] Budi EP, Lidia KA. The organization of words in mental lexicon: Evidence from word association test. TEKNOSASTIK. 2019;16(1):26-33. https://doi.org/10.33365/ts.v16i1.130

[2] Giulio P. Aphasia: Definition, clinical contexts, neurobiological profiles and clinical treatments. Ann Alzheimer's and Dementia Care. 2020; 4(1): 21-26. https://doi.org/10.17352/aadc.000014

[3] Deka C, Shrivastava A, Abraham AK, Nautiyal S, Chauhan P. AI-based automated speech therapy tools for persons with speech sound disorder: a systematic literature review. Speech Lang Hear. 2025;28(1):Article 2359274. https://doi.org/10.1080/2050571X.2024.2359274

[4] Ghojogh B, Ghodsi A. Recurrent neural networks and long short-term memory networks. In: Elements of Deep Learning. Cham: Springer; 2026. p. 205-229. https://doi.org/10.1007/978-3-032-10738-1_8

[5] MD SI, Long Z. A review on BERT: Language understanding for different types of NLP task. Preprints. 2024. https://doi.org/10.20944/preprints202401.1857.v1

[6] Hawraz AA, Tarik AR. Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning. Journal of King Saud University - Computer and Information Sciences. 2024;36(7). https://doi.org/10.1016/j.jksuci.2024.102131

[7] Claudia P, Nadine M. Language learning in aphasia: A narrative review and critical analysis of the literature with implications for language therapy. Neuroscience & Biobehavioral Reviews. 2022;141. https://doi.org/10.1016/j.neubiorev.2022.104825

[8] Alduais A, Hind A. Childhood apraxia of speech: A descriptive and prescriptive model of assessment and diagnosis. Brain Sciences. 2024;14(6):540. https://doi.org/10.3390/brainsci14060540

[9] Aouragh SL, Yousfi A, Laaroussi S, Gueddah H, Nejja M. A new estimate of the n-gram language model. Procedia Comput Sci. 2021;189:211-215. https://doi.org/10.1016/j.procs.2021.05.111

[10] Robert DP, Gregory DH. Deep learning: RNNs and LSTM. In: Handbook of Medical Image Computing and Computer Assisted Intervention. The Elsevier and MICCAI Society Book Series; 2020. p. 503-519. https://doi.org/10.1016/B978-0-12-816176-0.00026-0

[11] Zhixing Y. Multimodal speech emotion recognition via transformer-based hybrid fusion and dual cross-entropy techniques. Highlights in Science, Engineering and Technology. 2024;120:114-121. https://doi.org/10.54097/6pzrwp08

[12] Yao D, Koivu A, Simonyan K. Applications of artificial intelligence in neurological voice disorders. World J Otorhinolaryngol Head Neck Surg. 2025;11(4):491-517. https://doi.org/10.1002/wjo2.70017

[13] Sanghyuk RC, Minhyeok L. Transformer architecture and attention mechanisms in genome data analysis: A comprehensive review. Biology (Basel). 2023;12(7):1033. https://doi.org/10.3390/biology12071033
[14] Rahmadhani B, Purwono, Kurniawan SD. Understanding transformers: A comprehensive review. J Adv Health Inform Res. 2024;2(2):85-94. https://doi.org/10.59247/jahir.v2i2.292
[15] Zhou X, Ren Z, Zhou S, Jiang Z, Yu T, Luo H. Rethinking position embedding methods in the transformer architecture. Neural Processing Letters. 2024;56(2):41. https://doi.org/10.1007/s11063-024-11539-7
[16] Li H, Liu Y, Sun H, Cai D, Cui L, Bi W, Zhao P, Watanabe T. SeqPE: Transformer with sequential position encoding. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2026:1-14. https://doi.org/10.1109/TPAMI.2026.3695094
[17] Bhin H, Choi J. Multimodal personality recognition using self-attention-based fusion of audio, visual, and text features. Electronics. 2025;14(14):2837. https://doi.org/10.3390/electronics14142837
[18] Geva M, Caciularu A, Wang K, Goldberg Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec; Abu Dhabi, United Arab Emirates. Stroudsburg (PA): Association for Computational Linguistics; 2022. p. 30-45. https://doi.org/10.18653/v1/2022.emnlp-main.3
[19] Ghoshal S, Buckchash H. GradAttn: Transformer-based modulation of residual approach for classification and representation learning problems. Applied Sciences. 2026;16(11):5252. https://doi.org/10.3390/app16115252
[20] Shim JW. Enhancing cross entropy with a linearly adaptive loss function for optimized classification performance. Scientific Reports. 2024;14(1):27405. https://doi.org/10.1038/s41598-024-78858-6

AUTHOR PROFILES

Nur Syahmina Ahmad Azhar *is a PhD candidate at Universiti Teknikal Malaysia Melaka (UTeM). Her research focuses on artificial intelligence, computer vision, and multimodal systems, with particular interest in integrating lexical retrieval and articulatory stability for communication support. Her work explores deep learning approaches, including Transformer-based models and sound to image profile signal processing, to enhance human–computer interaction, especially for individuals with speech and language impairments.*

Nik Mohd Zarifie Hashim *is a Senior Lecturer at the Faculty of Electronic and Computer Engineering Technology, Universiti Teknikal Malaysia Melaka (UTeM). He holds a PhD and is a registered Professional Engineer (BEM, ACPE), Professional Technologist (MBOT), Geospatialist (IGRSM), and a member of the Institution of Engineers Malaysia (MIEM). His research interests include computer vision, robot vision, object re-identification, object estimation, image processing, and artificial intelligence, with applications focused on intelligent systems and advanced engineering technologies.*

Mohd Norzali Haji Mohd *is an Associate Professor in the Department of Electronic Engineering, Faculty of Electrical and Electronic Engineering at Universiti Tun Hussein Onn Malaysia (UTHM). He holds a Diploma in Computer Engineering from Toyama Maritime College, Japan, and B.Eng. and M.Eng. degrees from Fukui University, Japan. In 2015, he obtained his PhD in Information Sciences and Biomedical Engineering from Kagoshima University, Japan. He is a registered Professional Engineer (BEM), Professional Technologist (MBOT), and Chartered Engineer (IET, UK). His research interests include computer vision, biomedical engineering, signal processing, sensor technology, and AI applications.*

David Lim Chiok Chuan *is the Technical Director (Hardware Development) at Sena Traffic Systems Sdn Bhd (STS), a subsidiary of ITMAX Bhd, Malaysia. He joined STS in 2016 as a Technical Advisor and leads research and development in smart traffic controllers, street lighting systems, and urban monitoring technologies. His work focuses on AI-driven smart city solutions, including video-based traffic management that utilizes deep learning models such as YOLOv5 and DeepSort. He actively works with academic institutions, including Universiti Tun Hussein Onn Malaysia (UTHM) and Universiti Malaya, helping to bridge industry and research to advance intelligent transportation systems.*