# Phoneme-level mispronunciation detection in Quranic recitation using ShallowTransformer

**Mohamed Nadhir DAOUD, Mohamed Anouar BEN MESSAOUD**
Laboratoire Signal, Images et Technologies de l'Information
Université de Tunis El Manar, Tunis, Tunisia
mohamednadhir@gmail.com, anouar.benmessaoud@yahoo.fr

## Abstract

Preserving the integrity of Qur'anic recitation requires accurate pronunciation, as even subtle mispronunciations can alter meaning. Automatic assessment of Qur'anic recitation at the phoneme level is therefore a critical and challenging task. We present ShallowTransformer, a lightweight and computationally efficient transformer model leveraging Wav2vec2.0 features and trained with CTC loss for phoneme-level mispronunciation detection. Evaluated on the Iqra'Eval benchmark (QuranMB.v2), our model outperforms published BiLSTM baselines on QuranMB.v1 while achieving competitive performance relative to the official Iqra'Eval challenge baselines, which are not yet fully documented. Such improvements are particularly important in assisted Qur'an learning, as accurate phonetic feedback supports correct recitation and preserves textual integrity. These results highlight the effectiveness of transformer architectures in capturing subtle pronunciation errors while remaining deployable for practical applications.

## 1 Introduction

Mispronunciation detection and diagnosis (MDD) systems play a key role in computer-assisted pronunciation training (CAPT), helping language learners identify and correct pronunciation errors without human instructors (Neri et al., 2008). The detection component aims to detect pronunciation anomalies, whereas the diagnosis component aims to assign a specific class to each anomaly.

Most of the foundational research and development of MDD systems has been conducted in the context of English. For example, datasets such as L2-Arctic which includes non-native English speech annotated at phoneme level, for substitution, insertion, and deletion errors, have been extensively used to train and benchmark detection algorithms (Jiang et al., 2021).

In contrast, progress in mispronunciation detection for low-resource languages such as Arabic has been slow. The Arabic phonological system contains 28 consonants and 6 vowels (short and long), where complex phonetic structures (for example, uvular and pharyngeal sounds) (Alotaibi and Muhammad, 2010), present unique problems that do not commonly arise in more standardized and highly resourced languages. Moreover, subtle phonetic contrasts, such as between emphatic and non-emphatic consonants, can be difficult to perceive (Alrashoudi et al., 2025).

Furthermore, the diversity of Arabic dialects introduces substantial variability in pronunciation and vocabulary, while code-switching further complicates speech modeling efforts (Besdouri et al., 2024). These factors, along with the absence of short-vowel diacritics in most written text, create unique challenges for both learners and automated pronunciation assessment systems.

Previous research on Arabic mispronunciation detection has relied on either simplistic datasets such as isolated letters (Ziafat et al., 2021) or words (Aly et al., 2021), or on privately collected corpora that are not publicly accessible (Nazir et al., 2019)(Algabri et al., 2022). This reliance on private and limited datasets has prevented the establishment of standardized benchmarks and hindered objective comparison between different methodologies. (El Kheir et al., 2025) recently released an open phoneme annotated Arabic dataset, designed to provide a unified benchmark for Arabic pronunciation assessment. Building on the release of this benchmark dataset, we are positioned to rigorously evaluate advanced mispronunciation detection methods.

We present an end-to-end Arabic MDD model that leverages self-supervised speech representations. Our approach uses a pretrained wav2vec 2.0 encoder (Baevski et al., 2020) to extract robust acoustic features from raw audio, followed



*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 457–463

by a shallow Transformer network (Vaswani et al., 2017) trained with a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) to predict phoneme sequences. This combination enables the system to learn fine-grained phonetic distinctions while avoiding the need for explicit phonetic alignments.

Our contributions are:

- **Model:** A phoneme-level Arabic MDD system combining wav2vec 2.0 acoustic representations with a lightweight Transformer encoder trained via CTC.
- **Dataset:** An evaluation of the proposed approach on the QuranMB.v1 dataset (Kheir et al., 2025).
- **Analysis:** A performance comparison against baseline approaches, including an error-type breakdown to assess diagnostic capabilities for different phonetic categories.

## 2 Related Works

Earlier mispronunciation detection (MDD) methods primarily used the Goodness of Pronunciation (GOP) metric (Witt and Young, 2000), an objective measure of pronunciation quality based on likelihood scores. GOP computes the likelihood of acoustic segments corresponding to each phoneme using a set of Hidden Markov Models (HMMs).

(Harrison et al., 2009) used a GMM-HHM acoustic model to extract phone level representations. Phonological rules are modeled with finite state transducer to create an extended recognition network (ERN). This approach requires modeling correct pronunciation but also common mispronunciations.

(Li et al., 2016) overcame the need to design mispronunciation rules in ERN, by using a deep neural network that predict L2-speaker pronunciation from acoustic features and canonical phonemes, allowing for simultaneous detection and diagnosis of pronunciation anomalies.

CTC-CNN-RNN was introduced in (Leung et al., 2019) to leverage the ability of convolutional neural networks (CNN) to extract features, recurrent neural networks (RNN) to model sequences and CTC-loss to avoid explicit alignment between input frames and target phoneme sequence.

(Wu et al., 2021) used an encoder-decoder type transformer to predict phones from MFCC features and conduct experiments on the CU-CHLOE corpus (Meng et al., 2007).

The success of large language models in natural language processing, not only showed the power of scaling transformer models, but revealed also the importance of self-supervised learning (SSL) as a pre-training technique. This is no different for speech tasks, where it has been proven that transformer models can learn self-supervised speech representations (SSSR) (Mohamed et al., 2022).

Foundation models such as Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) became widely used for SSSR extraction.

(Peng et al., 2021) finetuned Wav2Vec 2.0 on the TIMIT dataset (Garofolo et al., 1993) to then test it on L2-Arctic. While (Wu et al., 2021) used Wav2Vec 2.0 as a backbone to extract SSSR and use it as input to an MLP prediction layer.

MDD for arabic was also influenced by the same trends, wher for example (Algabri et al., 2022) used CNN-RNN-CTC technique on Arabic-CAPT, a private dataset that contains phoneme transcription of Arabic words. Also (Alrashoudi et al., 2025) finetuned Wav2Vec 2.0 and HuBERT on the L2-KSU data set. While (Kheir et al., 2025) uses frozen SSL models as backbones for SSSR extraction, to train a BiLSTM based model.

Our work builds on these trends by employing a Wav2Vec 2.0 encoder for feature extraction and a shallow transformer for phoneme prediction, enabling accurate detection and diagnosis of mispronunciations in Arabic speech while balancing performance with memory efficiency.

## 3 Methodology

We propose a shallow transformer-based approach for Arabic phoneme sequence recognition. Our architecture leverages pre-trained wav2vec2 features and a lightweight transformer encoder. We opt for a shallow transformer to balance accuracy with computational efficiency. This approach makes the model easily deployable on resource-constrained environments, such as mobile applications or embedded systems used by learners. The model is trained end-to-end using CTC loss for automatic alignment.

### 3.1 Datasets

#### 3.1.1 Training and dev sets

The CMV-Ar data corpus, detailed in (Kheir et al., 2025), is derived from the Common Voice Dataset

(Ardila et al., 2019) and enhanced with Quranic recitation samples. It includes a training set of 71,391 utterances (approximately 79 hours of speech) and a development set of 2,588 utterances (3.33 hours of raw audio). Each audio file in the corpus is accompanied by its corresponding spoken phoneme sequence.

#### 3.1.2 Test set

(Kheir et al., 2025) utilized the QuranMB.v1 test set, which contains 2.2 hours of Qur'anic recitation from 18 native Arabic speakers, the majority of whom are female. A more recent release, QuranMB.v2, was made publicly available through the Iqra'Eval challenge (El Kheir et al., 2025). This updated version includes 98 utterances from the same 18 speakers, totaling approximately 2 hours of audio, although the exact differences between the two versions remain unclear. The corresponding labels for QuranMB.v2 are not yet available; however, performance metrics can be obtained by submitting predicted phoneme sequences to an online API.

All sets labels are based on the phoneme dictionary provided by (Halabi and Wald, 2016).

### 3.2 Audio feature extraction with Wav2Vec 2.0

#### 3.2.1 Wav2Vec 2.0

Proposed by (Baevski et al., 2020), Wav2Vec 2.0 is a self-supervised framework for learning speech representations. It learns contextualized audio features from raw waveforms. The model consists of a convolutional feature encoder, which transforms audio signals into latent representations, and a Transformer-based context network, which captures long-range temporal dependencies.

During pretraining, Wav2Vec 2.0 uses a contrastive loss to predict masked latent representations from their surrounding context. This enables the model to learn rich, domain-agnostic acoustic representations without requiring transcriptions. It has been proven that these representations can be fine-tuned for various downstream tasks or used directly as high-quality feature vectors, thereby reducing the need for large datasets.

#### 3.2.2 Featurizer

The authors of (Kheir et al., 2025) provided several pretrained models as baselines, that can be loaded using the S3PRL toolkit (Yang et al., 2024). We used the pretrained `iqra_wav2vec2_base`[1] checkpoint to load the upstream feature extractor , which returns features extracted by 13 layers of the pretrained Wav2Vec 2.0. The default S3PRL featurizer computes a weighted sum of these 13 representations for each frame.

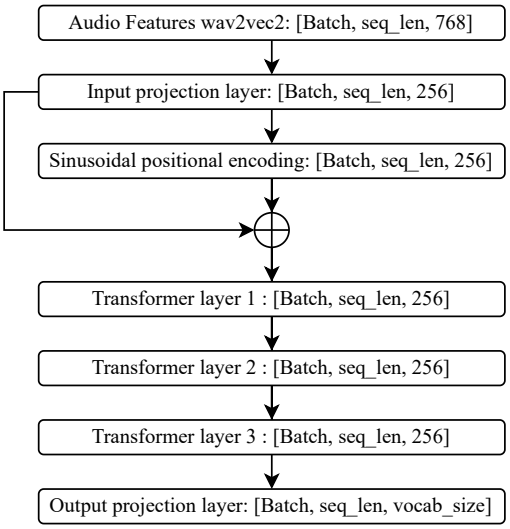


Figure 1: Architecture of the Shallow Transformer.

### 3.3 CTC loss

Introduced by (Graves et al., 2006), CTC loss allows for aligning speech utterances with associated shorter phoneme sequences without requiring explicit alignments. Instead of forcing a one-to-one correspondence between input frames and output labels, CTC loss allows for repetitions and blank symbols in the predicted sequence. This enables the model to handle variations in speaking speed and pronunciation, as well as silence between phonemes. The loss function sums the probabilities of all valid alignment paths that correspond to the true phoneme sequence, effectively allowing the model to learn the most probable sequence without needing pre-segmented data.

### 3.4 Detailed architecture of ShallowTransformer

**ShallowTransformer (ST)**, depicted in Figure 1, incorporates three stacked transformer layers. To optimize training performance and memory consumption, we downsample the audio features from 768 to 256. We augmented the input data with sinusoidal positional encoding. Subsequently, an output linear layer transforms the 256 transformer encodings to match the vocabulary size. The model's out-

[1]https://huggingface.co/IqraEval/Iqra_wav2vec2_base

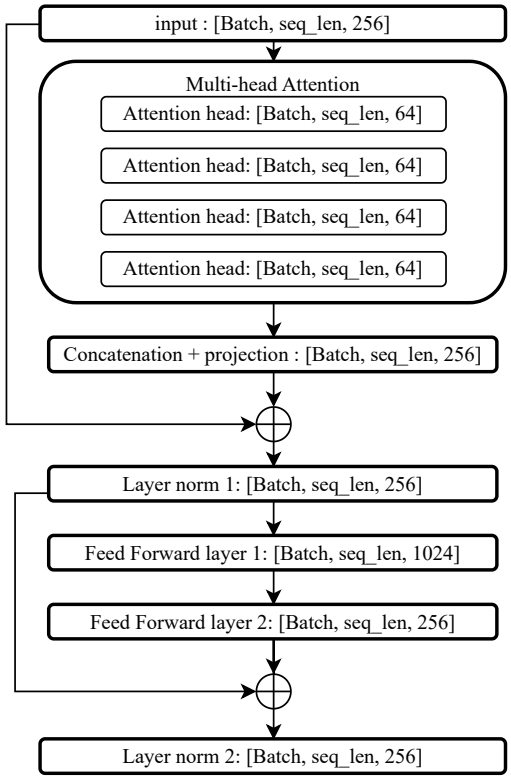


Figure 2: Architechture of transformer layers.

put comprises logits with dimensions [Batch size, sequence length, vocabulary size]. These logits are utilized by the CTC loss for loss computation and by a decoding algorithm to produce the predicted sequence. In order to process speech features in batches, all samples are padded to the length of the sample with maximum length.

As depicted in Figure 2 each transformer layer has 4 attention heads, followed by a shared layer normalization layer and a feed forward network, that projects the 256-dim features to 1024 (4 x 256) and back again to 256.

### 3.5 Tokenization

Phoneme-level tokenization was performed using the provided phoneme vocabulary. A blank token was added to the vocabulary at index 0, which is necessary for CTC loss. Each phoneme's token corresponds to its index.

### 3.6 CTC decoding and post-processing

The model outputs are processed using argmax at each time step to obtain the most likely token sequence. Before applying CTC decoding rules, predictions are truncated to actual sequence lengths to ignore padding tokens. The CTC alignment is then collapsed by applying two standard rules:

| Model | Recall | Precision | F1-score |
|---|---|---|---|
| BiLSTM (Wav2vec2) | 76.72 | 15.71 | 26.08 |
| BiLSTM (WavLM) | 75.35 | 15.80 | 26.12 |
| BiLSTM (HuBERT) | 74.75 | 15.67 | 25.91 |
| BiLSTM (mHuBERT) | 75.56 | 17.67 | 28.64 |
| **ST (Wav2vec2)** | **84.56** | **22.05** | **34.94** |

Table 1: Recall, precision, and F1-score of the proposed model compared to published baselines on the QuranMB.v1 test set

1. merging consecutive identical non-blank tokens into single occurrences.

2. removing all blank tokens.

This greedy approach provides efficient decoding, making it suitable for real-time phoneme recognition applications.

### 3.7 Metrics

The used metrics follow the established MDD convention defined in (Qian et al., 2010). This approach classifies predictions into four groups: True Accept (TA), False Accept (FA), True Reject (TR) and False Reject (FR). Precision, recall and F1-score are then computed following:

$$\text{Recall} = \frac{\text{TR}}{\text{TR} + \text{FA}} \tag{1}$$

$$\text{Precision} = \frac{\text{TR}}{\text{TR} + \text{FR}} \tag{2}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

### 3.8 Training configuration

The Adam optimizer was employed with a learning rate of 3e-4. A Cosine annealing scheduler was used, setting the minimum learning rate at 1.5e-5. Regularization included Dropout at 0.15 and gradient clipping with a maximum norm of 1.0. Training was conducted for 15 epochs, including 3 warmup epochs, with a batch size of 64 samples.

## 4 Experimental Results and Comparative Analysis

Table 1 reports the performance of our Shallow Transformer (ST) model on the QuranMB.v1 benchmark in comparison with previously published baselines from (Kheir et al., 2025). Across all three evaluation metrics—recall, precision, and

| Model | Recall | Precision | F1-score |
|---|---|---|---|
| baseline 1 (IqraEval) | 77.07 | **30.93** | **44.14** |
| baseline 2 (IqraEval) | 79.08 | 27.15 | 40.42 |
| **ST (Wav2vec2)** | **84.56** | 22.05 | 34.94 |

Table 2: Recall, precision, and F1-score of the proposed model compared to the official Iqra'Eval shared task baselines on the QuranMB.v2 test set.

F1-score—our model outperforms the BiLSTM-based baselines using different SSL feature extractors. The largest improvement is observed in F1-score, where our model achieves 34.94% compared to the best baseline score of 28.64%. Although these results indicate a substantial performance gain, it should be noted that QuranMB.v1 and QuranMB.v2 are not identical. While they are similar in duration and number of speakers, the exact differences are not documented. As such, direct numerical comparison should be interpreted with caution.

Table 2 presents our results on the QuranMB.v2 dataset alongside the baselines provided by the organizers of the Iqra'Eval shared task. These baselines serve as strong reference points for this test set, although they have not yet been officially published or fully documented.

Our model achieves the highest recall (84.56%) among all compared systems, but lower precision and F1 score than both baselines. This suggests that while our model is highly sensitive in detecting relevant phoneme events, further optimization is needed to improve precision and overall balance between recall and precision. Nonetheless, the results confirm the competitiveness of our approach under the same evaluation protocol.

## 5 Conclusion

We presented ShallowTransformer, a lightweight model for automatic phoneme level assessment of Qur'anic recitation pronunciation, leveraging self-attention on SSL-based acoustic features and trained with CTC loss. The model was designed to balance accuracy with computational efficiency, making it suitable for practical deployment. Our results show substantial improvements over published BiLSTM baselines: 10% higher recall, 25% higher precision, and over 22% higher F1-score, while remaining competitive with the official Iqra'Eval challenge baselines.

Although precision remains lower than recall, indicating a higher rate of false rejects, ShallowTransformer demonstrates strong capability in capturing pronunciation patterns. Future work will focus on improving precision through refined decoding, richer data augmentation, and exploring more advanced model architectures.

## Acknowledgments

We would like to express our sincere gratitude to the organizers of the Iqra'Eval shared task — Yassine El Kheir (DFKI - Technical University of Berlin), Amit Meghanani (University of Sheffield), Mostafa Shahin (University of New South Wales), Omnia Ibrahim (Alexandria University), Hawau Olamide Toyin (MBZUAI), Nada Al-Marwani (Taibah University), Youssef Elshahawy (HUMAIN), and Ahmed Ali (HUMAIN) — for their invaluable efforts in designing and coordinating this initiative aimed at advancing the automatic assessment of Qur'anic recitation pronunciation.

## References


Mohammed Algabri, Hassan Mathkour, Mansour Alsulaiman, and Mohamed A Bencherif. 2022. Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native arabic speech. *Mathematics*, 10(15):2727.

Yousef Ajami Alotaibi and Ghulam Muhammad. 2010. Study on pharyngeal and uvular consonants in foreign accented arabic for asr. *Computer Speech & Language*, 24(2):219–231.

Norah Alrashoudi, Hend Al-Khalifa, and Yousef Alotaibi. 2025. Improving mispronunciation detection and diagnosis for non-native learners of the arabic language. *Discover Computing*, 28(1):1.

Salah A Aly, Abdelrahman Salah, and Hesham M Eraqi. 2021. Asmdd: Arabic speech mispronunciation detection dataset. *arXiv preprint arXiv:2111.01136*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Fatma Zahra Besdouri, Inès Zribi, and Lamia Hadrich Belguith. 2024. Arabic automatic speech recognition: challenges and progress. *Speech Communication*, 163:103110.

Yassine El Kheir, Amit Meghanani, Hawau Olamide Toyin, Nada Almarwani, Omnia Ibrahim, Youssef Elshahawy, Mostafa Shahin, and Ahmed Ali. 2025. Iqra'eval: A shared task on qur'anic pronunciation assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*. Association for Computational Linguistics.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Nawar Halabi and Mike Wald. 2016. Phonetic inventory for an Arabic speech corpus. pages 734–738, Portorož, Slovenia. European Language Resources Association (ELRA).

Alissa M Harrison, Wai-Kit Lo, Xiaojun Qian, and Helen Meng. 2009. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *SLaTE*, pages 45–48.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Shao-Wei Fan Jiang, Bi-Cheng Yan, Tien-Hong Lo, Fu-An Chao, and Berlin Chen. 2021. Towards robust mispronunciation detection and diagnosis for l2 english learners with accent-modulating methods. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1065–1070. IEEE.

Yassine El Kheir, Omnia Ibrahim, Amit Meghanani, Nada Almarwani, Hawau Olamide Toyin, Sadeen Alharbi, Modar Alfadly, Lamya Alkanhal, Ibrahim Selim, Shehab Elbatal, Salima Mdhaffar, Thomas Hain, Yasser Hifny, Mostafa Shahin, and Ahmed Ali. 2025. Towards a unified benchmark for arabic pronunciation assessment: Quranic recitation as case study. *Preprint*, arXiv:2506.07722.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136. IEEE.

Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

Helen Meng, Yuen Yee Lo, Lan Wang, and Wing Yiu Lau. 2007. Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 437–442. IEEE.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, and 1 others. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Faria Nazir, Muhammad Nadeem Majeed, Mustansar Ali Ghazanfar, and Muazzam Maqsood. 2019. Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for arabic phonemes. *IEEE Access*, 7:52589–52608.

Ambra Neri, Ornella Mich, Matteo Gerosa, and Diego Giuliani. 2008. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5):393–408.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhang. 2021. A study on fine-tuning wav2vec2. 0 model for the task of mispronunciation detection and diagnosis. In *Interspeech*, volume 2021, pages 4448–4452.

Xiaojun Qian, Helen Meng, and Frank Soong. 2010. Capturing l2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (capt). In *2010 7th International Symposium on Chinese Spoken Language Processing*, pages 84–88. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng. 2021. Transformer based end-to-end mispronunciation detection and diagnosis. In *Interspeech*, pages 3954–3958.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, and 1 others. 2024. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899.

Nishmia Ziafat, Hafiz Farooq Ahmad, Iram Fatima, Muhammad Zia, Abdulaziz Alhumam, and Kashif Rajpoot. 2021. Correct pronunciation detection of the arabic alphabet using deep learning. *Applied Sciences*, 11(6):2508.