

Masked Temporal Alignment for Automated Speech Scoring

Akshay Badola¹ and Ikkyu Choi²

¹ETS Research Institute, Educational Testing Services, India

²ETS Research Institute, Educational Testing Services, Princeton 8541, New Jersey, USA

Correspondence: {abadola, ichoi001}@ets.org

Abstract

Automated speech scoring evaluates spoken responses without human intervention. Audio-only models often miss semantic discrepancies, while text-only approaches overlook critical prosodic and acoustic cues. To resolve this, we propose a multi-modal framework fusing raw audio with ASR transcripts and timestamps. We introduce a Masked Temporal Alignment model with a temporal pooling module to explicitly align audio-semantic representations while masking non-verbal segments. This prevents noise and silences from biasing attention-based scoring. We evaluate our model against three multi-modal encoder variants using 200,000 unconstrained spoken responses from non-native English speakers. Across Quadratic Weighted Kappa (QWK), Pearson’s r , Spearman’s ρ , and accuracy, our model consistently outperforms baselines. These results demonstrate that masking non-verbal temporal features significantly improves both semantic alignment and acoustic modeling in automated speech assessment.

1 Introduction

The assessment of spoken responses is a critical part of language assessment, diagnostic interviewing, and automated customer service feedback and assessment. Traditionally, this process has relied on human raters who evaluate speech using predefined rubrics, considering dimensions such as content relevance, fluency, and pronunciation, and assign ordinal scores (Chen et al., 2018b). However, human scoring is costly and difficult to scale, which motivated the development of automated speech scoring systems. Traditionally, such systems have relied on expert-designed features and statistical models calibrated against human ratings (Zechner et al.,

2009), (Bernstein et al., 2010), (Chen et al., 2018b). These systems use a combination audio and textual features to assess the fluency of speech and content delivery. (Chen et al., 2018b) note that most automated speech scoring systems contain three components: automatic speech recognition (ASR), natural language processing (NLP), and a scoring model. While models with manual features are simpler and more explainable, their overall performance is constrained by the limited coverage of the target constructs, the inaccuracy of individual components, and the lack of explicit audio-semantic alignment with end-to-end neural models outperforming them (Chen et al., 2018b).

The renaissance of deep neural networks (also referred to as deep learning) in the last decade or so, has ignited interest in end-to-end automated speech scoring systems which can be trained on labeled speech data as a whole. Early work in using these methods (Chen et al., 2018a), (Qian et al., 2019) used CNN or BiLSTM networks as the feature extracting backbone and combined features of multiple modalities with respective backbones to predict a final score. Since the emergence of transformer-based models (Vaswani et al., 2017), researchers have also explored transformer-based speech encoders for end-to-end automated speech scoring systems (e.g. (Bannò and Matassoni, 2022), (Nebhi and Szaszák, 2023), (Oh et al., 2025b)). More recently, multi-modal transformer models (Park and Ubale, 2023), (Oh et al., 2025a), (Oh et al., 2025b) have emerged that incorporate both speech audio and corresponding text transcripts generated by ASR systems, thereby better integrating semantic information.

In this work, we develop a multi-modal

transformer based architecture with a novel masked temporal pooler (MTP) module which explicitly masks non-verbal segments in the audio data before pooling overlapping frames with the tokens. In the remainder of this paper, we call this model the masked temporal alignment (MTA) model. We apply the proposed MTA model to a large dataset of spoken responses by non-native English speakers. We compare the results of training on this dataset with two baselines multi-modal methods: 1) **Concatenation**: A simple baseline merging audio and text embeddings, 2) **Cross-Attention**: A mechanism allowing text tokens to attend to the full acoustic sequence.

In our experiments, the MTA model outperformed the baselines in terms of agreement and correlation with human rater scores. To our knowledge, this is the first proposal for a masked audio-semantic alignment method for automated speech scoring, and our findings demonstrate its efficacy.

The remainder of the paper is organized as follows: Section 2 discusses related work, Section 3 describes our motivation for the MTA model, the dataset, and the two baselines, Section 4 presents the proposed MTA model in detail, and Section 5 contains the experiment details and the results. We conclude with implications and future work in Section 6.

2 Related Work

Automated speech scoring is closely related to, yet distinct from, other automated assessment of linguistic output, such as automated essay scoring (AES) and automated pronunciation assessment (APA). AES shares similar evaluation metrics with automated speech scoring while focusing exclusively on textual proficiency. Similarly, APA focuses exclusively on phonetic aspects of the speech. Automated speech scoring, on the other hand, aims to provide a holistic score taking into account the pronunciation, fluency and prosodic elements, alongwith the linguistic structure and content of the response.

Early work incorporating deep neural networks for automated speech scoring used BiLSTMs and CNNs (Chen et al., 2018a), (Qian et al., 2019). Papi et al. (Papi et al., 2020) used various DNNs on word2vec (Mikolov

et al., 2013) and BERT (Devlin et al., 2019) embeddings of text generated using an external ASR system. (Wang et al., 2021) used a text transformer XLNet (Yang et al., 2019) to score the transcripts of the speech. (Bannò and Matassoni, 2022) used a wav2vec2 model (Baevski et al., 2020) to demonstrate the superiority of large transformer-based acoustic models over methods focusing solely on speech transcripts. (Phan et al., 2025) used a Whisper (Radford et al., 2022) model to generate a combined score for Speak and Improve corpus (Qian et al., 2025).

As the ASR capabilities of deep learning models improved (Baevski et al., 2020), (Chen et al., 2022), (Radford et al., 2022), efforts to incorporate text and other modalities within a coherent deep learning approach accelerated. (Park and Ubale, 2023) used wav2vec2 (Baevski et al., 2020) and BERT (Devlin et al., 2019) as audio and text models respectively and concatenate their embeddings for automated speech scoring, outperforming models based on either audio or text. (Oh et al., 2025b) proposed a multitask learning method with a joint loss which adds a Trait-Aware loss to incorporate lexical traits extracted with speech and syntactic complexity analyzers. (Nebhi and Szaszák, 2023) also used a multitask speech scoring approach with the additional tasks of predicting coherence and prompt relevance. (Gupta et al., 2024) focused on language-independent speech scoring with multi-modal models using audio as well as phoneme and speech transcripts for low resource languages.

A key aspect of multi-modal models is combining the representations of different modalities. Simple approaches for multi-modal fusion include concatenation (Chen et al., 2018a), (Lin and Wang, 2022), (Singla et al., 2021) and feature ensemble (Marchal et al., 2025). A more sophisticated method is to align the intermediate representations via attention, referred to as multi-modal *fusion* (Qian et al., 2018), (Gupta et al., 2024), (Oh et al., 2025b). *Fusion* using cross-attention has been used in speech translation to create rich intermediate representations (Zheng et al., 2021), (Tang et al., 2022). In the audio domain, masking of intermediate representations (Wang et al.,

2020), (Baevski et al., 2020), first popularized in deep networks by (Devlin et al., 2019), has been used effectively for the pre-training of speech encoders.

Recently, automated speech scoring models are increasingly based on transformers. Examples of such models include (Park and Ubale, 2023), (Oh et al., 2025b), (Gupta et al., 2024), (Nebhi and Szaszák, 2023), and (Oh et al., 2025a) which have used multi-modal fusion to align intermediate representations. Although (Grover et al., 2020), (Qian et al., 2019) used fusion, they did not use transformer-based models. To our knowledge, (Gupta et al., 2024) is the only reported use of masking for automated speech scoring, although their use involved a self-supervised learning task similar to wav2vec2 (Baevski et al., 2020).

While built upon existing research, our approach is distinct from those of the prior studies. We use multi-modal fusion for alignment and employ a custom module to mask and align the inter-word spans in speech. This use of masking is distinct from randomized masking of (Gupta et al., 2024).

3 Speech Assessment with Multi-Modal Transformers

In this section, we discuss our motivations, describe the dataset, and the processing pipeline together with the various model configurations used in our study.

3.1 Motivation

The foundation of our acoustic modeling relies on pre-trained self-supervised speech representations. We utilize the wav2vec2 (Baevski et al., 2020) architecture, which learns latent speech representations through a contrastive learning objective on masked acoustic frames. Specifically, we use the self-supervised pre-trained `wav2vec2-large-960h-1v60-self` variant of the wav2vec2 family¹ as the audio backbone for all our multi-modal models. We also conducted an ablation study with on wav2vec2 to determine the best audio length, learning rate and other hyperparameters. Our best audio-only wav2vec2 model achieved a *QWK*

¹<https://huggingface.co/facebook/wav2vec2-large-960h-1v60-self>

of 0.638 on the test set with first 40 seconds of audio.

3.2 Data

We use spoken responses from non-native English speakers as the primary dataset in this study. The responses were elicited using 195 prompts across a wide range of topics. The prompts were designed to elicit opinion-based responses with supporting details. For each prompt, speakers were given up to 45 seconds to respond. The dataset includes 199,988 responses collected over multiple years as part of a large-scale standardized language assessment program. All responses were associated with scores from trained human raters on a five-point ordinal scale ranging from 0 (lowest) to 4 (highest), which we use as the label and call the final score. To support robust evaluation, the data were randomly split into two equal partitions of 99,994 responses each, which we use for training and testing, respectively.

3.3 Processing Pipeline and Multi-Modal Models

Wav2vec2 takes an audio input x , and produces a sequence of latent vectors $Z = \{z_1, \dots, z_T\}$, where T denotes the number of acoustic frames. These vectors capture rich phonetic and prosodic information which can be fine-tuned for the desired task (Baevski et al., 2020). To adapt these representations for the automated speech scoring task, we employ a regression head consisting of a pooling layer, varied across models we considered (as described in this section and the next), followed by a fully connected (FC) layer. This representation is then passed through the FC layer to predict the final score, which we normalize in the interval of $[0, 1]$ for training. This normalization was to ensure more stable training. We use a Mean Squared Error (MSE) objective to train the model, while for evaluation we rescale the outputs back to $0 - 4$ by multiplying followed by rounding to the nearest integer for discretization.

The multi-modal models follow the same data pipeline except we add a DeBERTaV3 (He et al., 2023) model as a text encoder. We used the `Deberta-V3-Xsmall` variant of DeBERTaV3 for all the models in our study. We

found that using larger variants did not result in any appreciable improvement. For generating the text transcripts, along with word level hypotheses and timestamps, we employ the `parakeet-tdt-0.6b-v2` model² which is a variant of fast conformer architecture (Rekesh et al., 2023).

As mentioned in Section 1, we experimented with two baselines of the multi-modal architectures along with the proposed MTA model. All of them use the same fine-tuned wav2vec2 encoder and pre-trained DeBERTaV3 but differ in how the speech and text features are aligned, as described in detail in Section 4. For the multi-modal models, the final speech and text embedding tensors differ in their dimensions. Therefore, we employ attention pooling to align them before passing them through an FC layer.

4 Aligning Text and Audio

Here we describe the three models we have used in this study in detail.

4.1 Concatenation

The first approach employs embedding concatenation of both backbones (text and audio) and provides a simple baseline for comparison. In this configuration, the utterance-level acoustic embeddings $\mathbf{Z}$ derived from wav2vec are concatenated with text embeddings $\mathbf{W}$ with the aid of a projection matrix in ($\mathbf{P}_{(.)}$) case the dimensions differ. The text embedding $\mathbf{W}$ is generated by passing the reference transcript through a DeBERTaV3 encoder and pooling the output using a 3 layer MLP, also known as a *context* pooler. The combined feature vector $\mathbf{C} = [\mathbf{P}_z \mathbf{Z}; \mathbf{P}_w \mathbf{W}]$ is then fed into an FC layer. The output from the FC layer is a scalar in $[0, 1]$ and optimized with MSE loss.

This is the simplest form of a multi-modal architecture. For the concatenation models, we used both a frozen audio backbone fine-tuned on the data and a trainable one without prior fine-tuning. While useful as a baseline, this method suffers from a fundamental limitation: it treats the acoustic signal as a monolithic block. It assumes that the entire acoustic sequence contributes equally to the seman-

tic meaning of every text token. Consequently, any acoustic noise, filler words, or silence that does not correspond to the reference text may dilute the acoustic representation.

4.2 Multi-modal Fusion

The second method involves using cross-attention mechanism to align the text and audio representations. Similar to (Vaswani et al., 2017), (Devlin et al., 2019), we use a multi head attention (MHA) layer followed by an FC layer. In this model, the text tokens serve as *queries*, while the sequence of acoustic frames $\mathbf{Z}$ acts as *keys* and *values*, with the *queries*, *keys* and *values* having the same meaning as in (Vaswani et al., 2017). The attention mechanism is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \mathbf{V} \quad (1)$$

where $\mathbf{Q}$ corresponds to the text token embeddings, and $\mathbf{K}, \mathbf{V}$ are derived from the acoustic latent space. We use 4 MHA blocks with 8 attention heads in each block with embedding dimension of 512. The embeddings from the fusion modules are passed through an FC layer and model is optimized with MSE loss.

4.3 Masked Temporal Alignment

The masked temporal pooler (MTP) module forms the key component of the proposed MTA model. The MTP module integrates frame-level audio embeddings from wav2vec2 with token-level text embeddings from DeBERTaV3 by constructing binary overlap masks to align audio frames with text tokens using timestamp annotations from an ASR system. This process effectively filters out non-verbal segments, such as silence and background noise, by ensuring that only audio features temporally overlapping with spoken words contribute to the final representation. For fine-grained speech scoring, the model may learn to correlate silences or noise in the non-word gaps with the scores, while humans would explicitly filter out such noise even while keeping in mind major disfluencies and large gaps in the speech signal.

The module proceeds by:

²<https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2>

1. Segmenting the continuous acoustic sequence $\mathbf{Z}$ into non-overlapping windows based on the provided timestamps t_{start}, t_{end} for each text token $\mathbf{w}_i$.
2. Extracting for each token $\mathbf{w}_i$, acoustic frames within its corresponding time window and aggregating using a temporal pooling operation (weighted average) to produce a token-specific acoustic representation $\mathbf{h}_i$, with frames in regions of significant silence or noise being weighted down or excluded.

Formally, for $\mathbf{Z}_{\text{audio}} \in \mathbb{R}^{n_f \times d}$ being the hidden representations from the audio backbone, n_f is the number of audio frames, and d is the audio embedding dimension.

For the j -th token let $[s_j, e_j)$ be the start and end times in seconds. Define a binary overlap mask $\mathbf{M} \in \{0, 1\}^{n_t \times n_f}$, where n_t is the number of tokens such that an element $M_{j,i}$ indicates whether the i -th audio frame temporally overlaps with the j -th token:

$$M_{j,i} = \begin{cases} 1 & \text{if } s_j \leq t_i < e_j \\ 0 & \text{otherwise} \end{cases}$$

The aligned audio representation $\mathbf{Z}_{\text{aligned}} \in \mathbb{R}^{n_t \times d}$ is obtained by mean-pooling the audio frames that overlap with each token. For the j -th token, the aligned representation $\mathbf{z}_{\text{aligned}}^{(j)} \in \mathbb{R}^d$ is computed as:

$$\mathbf{z}_{\text{aligned}}^{(j)} = \frac{\sum_{i=1}^{n_f} M_{j,i} \cdot \mathbf{z}_{\text{audio}}^{(i)}}{\max(1, \sum_{i=1}^{n_f} M_{j,i})},$$

where $\mathbf{z}_{\text{audio}}^{(i)}$ is the i -th row of $\mathbf{Z}_{\text{audio}}$. The denominator is clamped to a minimum of 1 to ensure numerical stability, resulting in a zero vector if no frames overlap with a given token.

In matrix notation, this operation can be expressed as:

$$\mathbf{Z}_{\text{aligned}} = \mathbf{M}\mathbf{Z}_{\text{audio}} \oslash \mathbf{c},$$

where $\mathbf{c} \in \mathbb{R}^{n_t \times 1}$ is the count vector with elements $c_j = \max(1, \sum_{i=1}^{n_f} M_{j,i})$, and the operation $\oslash \mathbf{c}$ is performed to normalize $\mathbf{Z}_{\text{aligned}}$.

A visual depiction of the alignment module is provided in Figure 1.

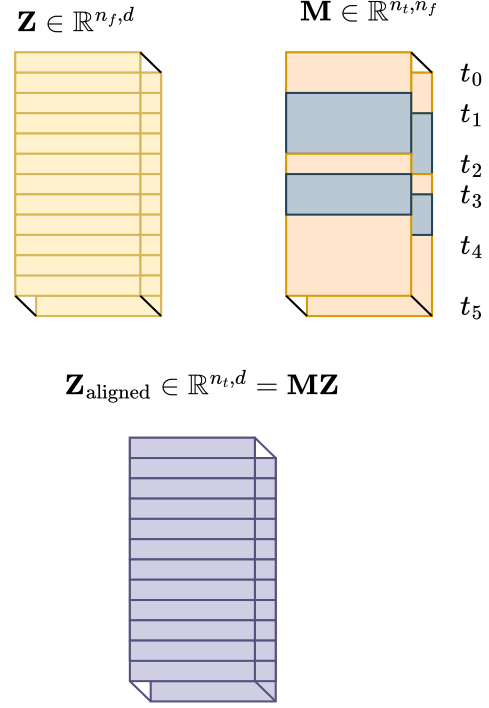


Figure 1: Visual depiction of the Masked Temporal Pooler Module

The output $\mathbf{Z}_{\text{aligned}}$ from the MTP module is then fused with text embeddings $\mathbf{W}$, passing through 4 bidirectional cross-attention blocks in sequence, each with 8 attention heads and dimension of 512. The final result is aggregated via an attention pooling operation and fed into an FC layer with the final output and model training similar to Sections 4.1 and 4.2.

A schematic of the architectural overview of all models is given in Figure 2.

5 Experiments and Results

We conducted experiments by training the models on the 99,994 samples demarcated as the training set. From the training set 10,000 items were randomly selected as a validation set to monitor the training. Due to the imbalanced nature of the data, a balanced sampler was used while training which sampled data randomly with replacement ensuring equal proportions of all scores. All models were trained with identical hyper-parameters except for learning rate which was chosen from $\{10^{-4}, 10^{-5}\}$, with optimizer AdamW (Loshchilov and Hutter, 2017). The training

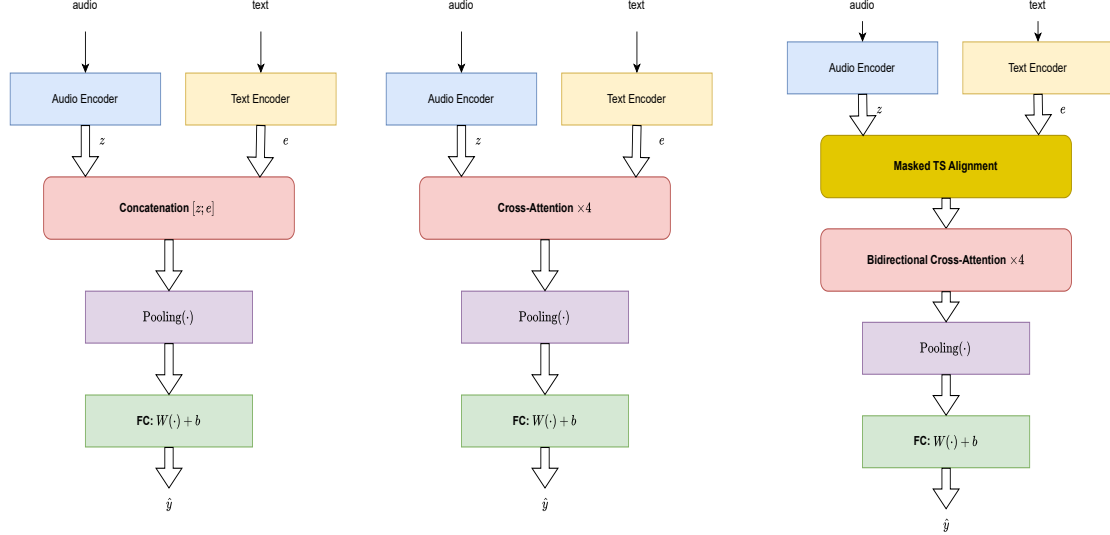


Figure 2: Architectural overview of the three models under consideration. The proposed MTA module is placed right after the audio and text encoders.

was done for 500,000 samples, with the batch size being set to fill the maximum available GPU memory. The results for best 3 models for each variant are given in Table 1.

5.1 Metrics and Evaluation Protocol

While the objective for model training is MSE, the metric of interest in assessment is Cohen’s Kappa (Cohen, 1960) which takes into account agreement by chance. We used the quadratic weighted kappa QWK on the validation set as our primary model selection criterion and QWK along with MSE were monitored on the validation set throughout model training. We also calculated Pearson’s correlation coefficient (r), Spearman’s correlation coefficient (ρ), and accuracy on the validation set and test set to gauge broader model performance.

To optimize computational resources, we implemented an iterative validation strategy. Specifically, 10% of the validation set was reserved for a lightweight evaluation conducted every 1,000 training iterations. Full-scale validation on the entire validation set was triggered only when the QWK score on the 10% subset placed the model within the top 3 performers of the recent checkpoint history. This heuristic significantly reduced the total training time without compromising the ability to track model convergence and prevent overfitting.

While minimizing the squared error $\|y - x\|_2^2$

theoretically minimizes the QWK in the limit, we observed a discrepancy between the two metrics in practice. Consequently, a model may achieve a low MSE but fail to capture the nuanced ordinal distinctions required by the QWK metric, and vice versa. Therefore, we monitor both metrics on the validation set during training. We should note that we also experimented with a Weighted Kappa (Torre et al., 2018) training objective instead of MSE but the results were not encouraging.

5.2 Performance Comparison

The top three models as determined by QWK for for each model variant were selected for final comparison. QWK , Pearson’s r , Spearman’s ρ and accuracy were calculated for these on both validation and test set. Overall The MTA model provided the best results across all the metrics, with a best QWK of 0.6528 on the validation set and 0.6547 on the test set, with an average QWK of 0.6485 and 0.6515 on validation and test set respectively. MTA also had the highest r , ρ and accuracy among the three variants, while concatenation and Cross-Attention variants performed a little worse with similar scores. The performance of the concatenation baseline was comparable to that of using only the acoustic backbone of wav2vec2 (not reported in Table 1).

The results validate our hypothesis that filtering out silence is beneficial for a deep

Model	QWK		Pearson r		Spearman ρ		Accuracy	
	Eval	Test	Eval	Test	Eval	Test	Eval	Test
Concat	0.6358	0.6326	0.6485	0.6463	0.634	0.6368	0.6105	0.6062
	0.6411	0.6436	0.6470	0.6506	0.6291	0.6391	0.6275	0.6286
	0.6397	0.6362	0.6508	0.6478	0.6344	0.6357	0.6116	0.6079
Average	0.6389	0.6375	0.6488	0.6482	0.6325	0.6372	0.6165	0.6142
Cross Attn	0.6392	0.6384	0.6470	0.6457	0.6298	0.6334	0.6278	0.6280
	0.6422	0.6367	0.6470	0.6415	0.6306	0.6270	0.6426	0.6344
	0.6336	0.6353	0.6373	0.6390	0.6180	0.6252	0.6310	0.6330
Average	0.6383	0.6368	0.6438	0.6421	0.6261	0.6285	0.6338	0.6318
MTA	0.6528	0.6547	0.6573	0.6597	0.6396	0.6473	0.6447	0.6464
	0.6468	0.6515	0.6518	0.6574	0.6324	0.6446	0.6444	0.6471
	0.6457	0.6483	0.6507	0.6537	0.6310	0.6403	0.6346	0.6381
Average	0.6484	0.6515	0.6533	0.6569	0.6343	0.6441	0.6412	0.6439
Ablation	0.6169	0.6192	0.6346	0.6372	0.6208	0.6276	0.5896	0.5920

Table 1: Results on Eval and Test set for 3 best models of Cross Attention, Concatenation and Masked Temporal Alignment (MTA). Values in bold in each column except in row *Average* indicate the best (highest) result in that column. A bold value in the row *Average* indicates the best average metric across three best models. The row **Ablation** denotes the model trained without the MTP module.

Score	Data	MTA	MTA	CA	CA	Concat	Concat
	Proportion	Accuracy	Proportion	Accuracy	Proportion	Accuracy	Proportion
0	0.002	0.550	0.001	0.550	0.002	0.500	0.001
1	0.025	0.720	0.048	0.624	0.033	0.780	0.061
2	0.320	0.616	0.315	0.726	0.393	0.555	0.289
3	0.542	0.651	0.481	0.592	0.427	0.658	0.496
4	0.111	0.685	0.155	0.655	0.146	0.656	0.152

Table 2: Score distribution and results per score on Eval set for best model each of Masked Temporal Alignment (MTA), Cross Attention and Concatenation. Values in bold indicate the best result for that score across models.

learning multi-modal automated speech scoring model. The details are provided in Table 1. We also conducted an ablation study to determine the effectiveness of the proposed MTP module where the MTP module was removed while the rest of the model architecture remained as is, along with the training hyperparameters. The results for that is given in the row **Ablation** in Table 1.

Table 2 contains the score-wise accuracy of the best model of each variant under consideration, i.e., MTA, Cross Attention and Concatenation. Looking at the distribution of scores, a significant mass is concentrated around scores 2, 3. We can see that MTA gets a higher proportion of correct responses for scores 1, 3, 4, while the accuracy reduces for score 0. MTA also predicts fewer scores 0, even as accuracy for that remains the same

as of Cross Attention. Cross Attention predicts a higher proportion of scores as 2 and has greater accuracy for that as well. Noting the score proportions of outputs of MTA and Cross Attention, we can see that MTA predicts more proportions of rating 3 than Cross Attention and Concatenation. We also note that all the models tend to predict score 4 in greater proportion than in the data.

6 Conclusion and Future Work

This study presents a novel multi-modal Automated Speech Scoring Masked Temporal Alignment model with a Masked Temporal Pooler (MTP) module that utilizes timestamps to synchronize frame-level audio embeddings. By employing a binary overlap mask to mean-pool audio features corresponding to specific text tokens, our method effec-

tively filters out non-verbal segments such as silence and background noise, ensuring that the model attends only to speech-relevant acoustic content. Our experiments demonstrate that explicitly masking non-verbal cues allows the model to predict scores more closely aligned with human judgment.

While the proposed MTA model demonstrates strong performance, the present study is not without limitations. The current study has utilized an internal dataset and while the dataset is sufficiently large, no direct comparison is made with other large spoken response datasets. The proposed MTA model is reliant on an external ASR system to generate transcripts with timestamps and inaccuracies in word boundary detection could potentially misalign audio and text features. Future work could focus on developing methods to refine or correct these timestamps, or integrating a joint learning framework that optimizes alignment boundaries alongside the scoring task, which would further enhance model reliability. The MTA model has also only been developed for unconstrained spoken responses and although promising, it needs to be assessed across other spoken item types. We also leave modeling for additional spoken item types and comparison against other large spoken response datasets to future work.

Additionally, investigating the impact of different ASR models and timestamp granularities (e.g., phoneme-level vs. word-level) could provide further insights into the optimal level of temporal resolution for multi-modal alignment. Another future avenue of work would be to explore the interpretability of the alignment mechanism and predicted scores.

References

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Stefano Bannò and Marco Matassoni. 2022. [Proficiency assessment of L2 spoken english using wav2vec 2.0](#). In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, pages 1088–1095. IEEE.
- J. Bernstein, Alistair Van Moere, and Jian Cheng. 2010. Validating Automated Speaking Tests. *Language Testing*, 27.
- L. Chen, J. Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018a. End-to-End Neural Network Based Automated Speech Scoring. In *IEEE International Conf. on Acoustics, Speech, and Signal Process.*
- Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, Robert Mundkowsky, Chi Lu, Chee Wee Leong, and Binod Gyawali. 2018b. [Automated scoring of nonnative speech using the speechrater v. 5.0 engine](#). *ETS Research Report Series*, 2018(1):1–31.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.
- Manraj Singh Grover, Yaman Kumar Singla, Sumit Sarin, Payman Vafaei, Mika Hama, and R. Shah. 2020. Multi-Modal Automated Speech Scoring Using Attention Fusion.
- Shreyansh Gupta, Abhishek Unnam, Kuldeep Yadav, and Varun Aggarwal. 2024. Towards Building a Language-Independent Speech Scoring Assessment. In *Proc. of the Aaai Conf. on Artif. Intell.*
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: improving deberta Using Electra-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *Proc. of the International Conf. on Learn. Represent.*
- Binghuai Lin and Liyuan Wang. 2022. Pre-Trained Multimodal End-to-End Network for Spoken Language Assessment Incorporating Prompts. In *Asia-Pacific Signal and Inf. Process. Association Annual Summit and Conf.*
- I. Loshchilov and F. Hutter. 2017. Decoupled Weight Decay Regularization. In *Proc. of the International Conf. on Learn. Represent.*

- Marian Marchal, Joey Stuver, Nafsika Lachana, Max van der Velde, Josine Verhagen, and Joost Kruis. 2025. Towards Explainable Automatic Spoken Language Assessment. In *Slate*.
- Tomas Mikolov, I. Sutskever, Kai Chen, G. Corrado, and J. Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Neural Information Processing Systems*.
- Kamel Nebhi and György Szaszák. 2023. Automatic Assessment of Spoken English Proficiency Based on Multimodal and Multitask Transformers. In *Recent Adv. in Natural Language Process.*
- Sehyun Oh, Minhwa Chung, and Sunhee Kim. 2025a. Multilingual Speech Assessment Using Cross-Attention and Multitask Learning. In *Interspeech*.
- Sehyun Oh, Sunhee Kim, and Minhwa Chung. 2025b. Multimodal and Multitask Learning for Predicting Multiple Scores in I2 English Speech. In *Interspeech*.
- Sara Papi, E. Trentin, R. Gretter, M. Matassoni, and D. Falavigna. 2020. Mixtures of Deep Neural Experts for Automated Speech Scoring. In *Interspeech*.
- Seongjin Park and Rutuja Ubale. 2023. Multitask Learning Model with Text and Speech Representation for Fine-Grained Speech Scoring. In *Automatic Speech Recognit. & Understanding*.
- Nhan Phan, Anusha Porwal, Yaroslav Getman, Ekaterina Voskoboinik, Tamás Grósz, and Mikko Kurimo. 2025. One Whisper to Grade Them All. In *Slate*.
- Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark Gales, and Diane Nicholls. 2025. Speak & Improve Challenge 2025. In *Slate*.
- Yao Qian, P. Lange, Keelan Evanini, Robert A. Pugh, Rutuja Ubale, Matthew David Mulholland, and Xinhao Wang. 2019. Neural Approaches to Automated Speech Scoring of Monologue and Dialogue Responses. In *IEEE International Conf. on Acoustics, Speech, and Signal Process.*
- Yao Qian, Rutuja Ubale, Matthew David Mulholland, Keelan Evanini, and Xinhao Wang. 2018. A Prompt-Aware Neural Network Approach to Content-Based Scoring of Non-Native Spontaneous Speech. In *Spoken Language Technology Workshop*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, C. McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning*.
- D. Rekesh, Samuel Krizan, Somshubra Majumdar, V. Noroozi, He Juang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg. 2023. Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. In *Automatic Speech Recognit. & Understanding*.
- Yaman Kumar Singla, Avykat Gupta, Shaurya Bagga, Changyou Chen, Balaji Krishnamurthy, and R. Shah. 2021. Speaker-Conditioned Hierarchical Modeling for Automated Speech Scoring. In *International Conf. on Inf. and Knowledge Management*.
- Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdel-rahman Mohamed, Michael Auli, and J. Pino. 2022. Unified Speech-Text Pre-Training for Speech Translation and Recognition. In *Annual Meeting of the Association for Computational Linguistics*.
- J. D. L. Torre, D. Puig, and A. Valls. 2018. Weighted Kappa Loss Function for Multi-Class Classification of Ordinal Data in Deep Learning. In *Pattern Recognit. Letters*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. In *Adv. in Neural Inf. Process. Syst.*
- Weiran Wang, Qingming Tang, and Karen Livescu. 2020. Unsupervised Pre-Training of Bidirectional Speech Encoders via Masked Reconstruction. In *IEEE International Conf. on Acoustics, Speech, and Signal Process.*
- Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew David Mulholland. 2021. Automated Scoring of Spontaneous Speech from Young Learners of English Using Transformers. In *Spoken Language Technology Workshop*.
- Zhilin Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Neural Information Processing Systems*.
- Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. [Automatic scoring of non-native spontaneous speech in tests of spoken english](#). *Speech Communication*, 51(10):883–895.
- Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation. In *Proc. of the International Conf. on Mach. Learn.*