Do We Need to Watch It All? Efficient Job Interview Video Processing with Differentiable Masking

Hung Le
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
hungle@jaist.ac.jp

Sixia Li
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
lisixia@jaist.ac.jp

Candy Olivia Mawalim
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
candylim@jaist.ac.jp

Hung-Hsuan Huang
The University of Fukuchiyama
Fukuchiyama, Kyoto, Japan
hhhuang@acm.org

Chee Wee Leong
Educational Testing Service
Princeton, New Jersey, USA
cleong@ets.org

Shogo Okada*
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp

ABSTRACT

With technological advancements in transmitting and storing large video files, more and more organizations are incorporating asynchronous video interviews as part of their personnel selection process. Automatic evaluation of these videos is a challenging machine learning setting because the samples are composed of time series input data but only one overall label is available. It is unclear which segments of the time series input (i.e., videos) are the most important ones for prediction. Not all nonverbal features, spoken words, and utterances contribute equally to the prediction; some segments of the videos might even introduce noise to the model. Processing all multimodal information is therefore inefficient. To address this challenge, we propose a framework to model the content of the answer via the full transcription and the speaking patterns of the interviewee via short clips. Our model learns to automatically select the most informative segment by previewing the acoustic modality using a technique called differentiable masking. The results show that our method outperforms existing approaches while being more efficient since only partial multimodal data are processed, and the interpretability of the model is enhanced.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**; *Machine learning approaches*; • **Computer systems organization** → *Neural networks*; • **Human-centered computing** → *HCI theory, concepts and models.*

KEYWORDS

interview performance prediction; multimodal learning; differentiable masking; deep learning

ACM Reference Format:
Hung Le, Sixia Li, Candy Olivia Mawalim, Hung-Hsuan Huang, Chee Wee Leong, and Shogo Okada. 2024. Do We Need to Watch It All? Efficient Job Interview Video Processing with Differentiable Masking. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685718

1 INTRODUCTION

Interviews have long been the standard method for assessing job applicants. They offer insights into a candidate's social, communication, and improvisational skills that cannot be gleaned from their application materials alone. Traditionally, these interviews are conducted in person, requiring candidates to visit the company's location, which can be costly in terms of both time and money. With technological advancements, asynchronous video interviews (AVIs) have become increasingly popular. AVIs allow companies to conduct interviews more efficiently, leading to the emergence of businesses that specialize in providing these services such as Talview[1] or Hirevue[2].

As the name suggests, an AVI is a video-based interviewing process in which interviewers are absent. Instead of receiving questions from interviewers, candidates receive a list of predetermined questions and record their responses within the specified time limit. The questions could be text-based or be delivered by a virtual agent [4, 6, 26]. To ensure fairness among candidates, most AVI systems allow only a few seconds between questions and candidates' answers. The video responses are stored and subsequently rated, and these ratings are used to select the most suitable candidates according to certain criteria. These candidates are then invited to onsite interviews.

As remote work becomes more common, asynchronous video interviews (AVIs) are gaining popularity among organizations. AVIs

*Shogo Okada is the corresponding author

[1] https://www.talview.com
[2] https://www.hirevue.com

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685718

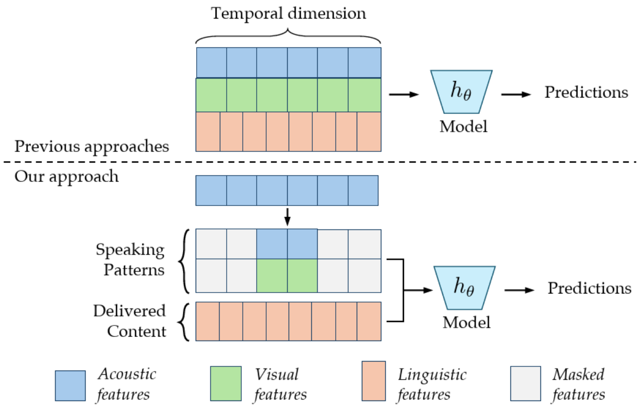

Figure 1: An overview comparing our approach with previous approaches. Instead of using all the information from all the modalities, in our approach, a short clip is selected using acoustic features, and irrelevant video segments are masked before further processing.

are easier and less expensive to conduct than traditional interviews are, allowing a broader range of candidates to apply for positions during the hiring season. However, manually reviewing and rating these videos can be a laborious task that places substantial pressure on Human Resources departments. To address this issue, machine learning (ML) algorithms have been developed over the years to automate this process [37–39, 46].

Some of the most widely used ML algorithms for predicting interview performance are regularized logistic regression (LASSO or ridge), random forest (RF), and support vector machines (SVMs), as noted in [24]. The inputs to these algorithms are features extracted from the videos, including frame-based visual, audio and linguistic features, and language features. There are multiple methods to aggregate temporal features such as using statistic functions (e.g., mean, max, and standard deviation) or the bag-of-words and bag-of-audio [11, 18] representations. While these algorithms are considered less sophisticated and have lower computational costs than other approaches, they cannot model the intra- and inter-modality dynamics [24] in interview videos. Thus, deep learning (DL) architectures (e.g., long short-term memory (LSTM) networks [25], gated recurrent units (GRU) [14], and transformers [48]) are more appropriate for modeling temporal features.

In previous study, several classification models for this video interview rating task have been developed [11, 23, 24, 30]. While such models can be used to distinguish between high- and low-quality candidates, they do not address scenarios where a company might want to select the top *n* percentage of candidates based on the number of applications and positions available. Such scenarios require regression models. Some studies have focused solely on nonverbal cues, which can introduce biases related to ethnicity, age, or attractiveness, potentially overlooking what is often the most crucial aspect of a candidate's application—the content of their answers [32]. To address this issue, other methods consider information from all modalities. However, extracting features across different modalities varies significantly in computational demands. Visual features, for instance, are much more resource-intensive to process than linguistic or acoustic features. Moreover, analyzing every frame in a video can add noise and distract the model from identifying useful signals, as many frames are repetitive. These challenges have led us to explore new approaches.

Similarly, the authors of [20] argued that processing all frames in a long untrimmed video may be unnecessary and even counterproductive. We tend to agree with this argument, especially in the case of monologue videos, as the scene and camera angle do not change over time. We hypothesize that a model does not have to scan the whole video to make accurate predictions. The hypothesis is grounded in both psychology research, which discovered that thin slices (i.e., short clips) offer insights into "social and interpersonal relations" [3], a finding also echoed in automated interview performance research [13, 39]. Our approach aims to teach the model what the candidate says (i.e., the delivered content) from the full text extracted from the video, and how the candidate presents themselves (i.e., the speaking patterns) from the short clips. We propose a framework for automatically selecting the most informative short clip, and multimodal features are extracted and processed from only this short clip. To learn which part of a video is the most informative segment, we rely on a technique called differentiable masking [15]. As both local features (from the short clip) and global features (from the whole video) are considered, the method takes into account the multifaceted characteristics of the videos.

Our proposed framework has several advantages. First, during the training phase, the model must still scan the whole video to learn how to select the most informative segment, whereas in the testing phase, only a small portion of the visual features is extracted and processed, saving computational power and time. Second, since a large part of the visual features is masked, model weights can be used to focus on the most useful information. Third, since we know which part of the video is focused on by the model, we can easily examine this short clip to verify the model's predictions. The ability to determine what the model selects as important improves the interpretability of the model and provides us with additional analysis information. Figure 1 shows the difference between our proposed framework and other methods.

The contributions of this work can be summarized as follows:

- We propose a framework for modeling the speaking patterns and the delivered content of interviewees (Section 3).
- Based on the proposed framework, we construct a concrete multimodal model and run experiments on a large corpus of job interview performance dataset (Section 4). Our model achieves a new state-of-the-art performance on this dataset.
- We conduct ablation studies to verify the effectiveness of the delivered content and the speaking pattern encoders, as well as provide our analysis regarding different choices for the length of the short segment (Section 5.3).

2 RELATED WORK

2.1 Computational Inference of Hireability

There are two lines of research related to the assessment and selection of candidates [5]. The first line of research treats this process as a multiple-criteria decision-making process. The second line of research aims to develop methods to fully automate the selection process using technology. Our work falls into the latter line of studies, so we describe related work that aims to automatically assess hireability. Initial research, such as that by [38], demonstrated that it is feasible to predict hireability scores using nonverbal behavioral features. In subsequent studies, such as [42], frameworks were developed to extract features for inferring personality traits, leadership, and communication skills, improving classification accuracy significantly over previous methods. Research using online video resumes [40] also highlighted the potential of audio and visual cues in predicting hireability. The release of the ChaLearn First Impressions dataset in 2016 [44] spurred further methodologies for assessing personality traits and hireability. More recent work, including [23] and [41], explored regression models and hierarchical attention models that consider contextual and multimodal information to enhance hireability predictions. Additionally, studies like [36] investigated how candidates' reactions to asynchronous video interviews might affect their performance. Collectively, these studies underscore the potential for computational approaches to reliably predict hireability in job applicants.

In some studies, however, verbal information was not considered during the model design process, which led to some inevitable pitfalls. The First Impressions dataset, consisting of only 15-second clips, provides limited verbal content, causing annotators to potentially rely on perceived attributes such as gender, age, and attractiveness, as discussed by [28]. [8] emphasized that verbal information is the most accurate and unbiased predictor, with additional modalities only marginally enhancing prediction accuracy. This finding supports earlier findings by [11], in which verbal information was identified as the most predictive feature. Additionally, [37] analyzed the impact of speech patterns, with the results suggesting that applicants who use fewer filler words and more unique words are more likely to be successful than other candidates.

That said, research has continually shown that both verbal and nonverbal behaviors are significant in predicting job interview outcomes. [39] reported that even brief interactions provide predictive insights into hireability, although not as effectively as full interactions do. [13] observed that audio-visual features from short video clips are closely related to those from full videos in public speaking assessments. Additionally, in [29], presentation videos were divided into 1-minute segments, and a hidden Markov model was used to analyze these segments, revealing that the final segment of a presentation is most correlated with the overall evaluation score.

In this work, we continue to combine both verbal and nonverbal behaviors to predict job interview performance. However, unlike previous work where the combination was made based on features either from holistic videos or only thin slices, our proposed method allows the combination of linguistic features from the entire video and multimodal features from short clip segments in a hybrid approach that is both efficient and effective.

2.2 Monologue Video Processing with Deep Learning

Recent advancements in ML have been driven primarily by deep neural networks (DNNs), which have been effectively used to learn representations from multiple modalities. For instance, [12] introduced a novel DL architecture, the gated multimodal embedding LSTM with temporal attention (GME-LSTM(A)) network, which enables modality fusion at the word level for more accurate sentiment analysis. Their model demonstrated superior performance with the CMU-MOSI dataset, highlighting the effectiveness of its temporal attention layer in managing noisy audio and visual data for sentiment prediction. [52] redefined multimodal sentiment analysis by focusing on both intra-modality and inter-modality dynamics and introduced the tensor fusion network that learns these dynamics end-to-end. Designed to address the complexities of spoken language, gestures, and voice in online videos, their model surpassed leading models in both multimodal and unimodal sentiment analysis. Additionally, [21] developed a hierarchical attention strategy using only audio and text modalities for classifying sentiment at the utterance level, and provided visual interpretability through its synchronized attention across different modalities.

Building upon the concept of attention mechanisms, [48] introduced the transformer network, which was originally designed for neural machine translation but was proven to be versatile enough for applications in areas like computer vision and audio processing. Inspired by this architecture, [47] introduced the multimodal transformer (MulT) which resolves issues related to the non-alignment of multimodal data across modalities due to differing sampling rates and long-range dependencies. MulT operates in an end-to-end manner with directional pairwise cross-modal attention, facilitating interactions and adaptation among different modalities. Their extensive experiments demonstrate that MulT significantly outperforms existing methods, effectively capturing correlated crossmodal signals through its cross-modal attention mechanism. Leveraging these developments, we developed a new architecture that obtains state-of-the-art performance in predicting the outcomes of job interviews.

2.3 Selection of Clips and Adaptive Computations

There is a line of research that aims to select the most salient clips from untrimmed videos [2, 49, 51]. This approach usually requires the use of supervised labels (ground-truth temporal boundaries). In [20], researchers attempted to overcome this requirement by previewing the audio to select the most salient video segments. Distinct from their approach, in which knowledge distillation is used to transfer knowledge from an expensive teacher model, we used differentiable masking in our approach, as this method does not require training an expensive video model.

Several studies have also aimed to develop efficient models by dynamically routing and inferring information in neural networks on the fly [1, 50, 53]. In particular, [10] also proposed the use of differentiable masking for learning which layers of the model can be deactivated. Our work proposes the use of differentiable masking in a different manner, that is, to mask the input features instead of the model's components. This idea is more similar to the work of [15], in which the authors aimed to interpret the models by masking out a subset of the sequence input.

3 METHOD

First, we describe the job interview performance evaluation task setup. Given a video V contains K modalities $M = M_1, M_2, ..., M_K$,

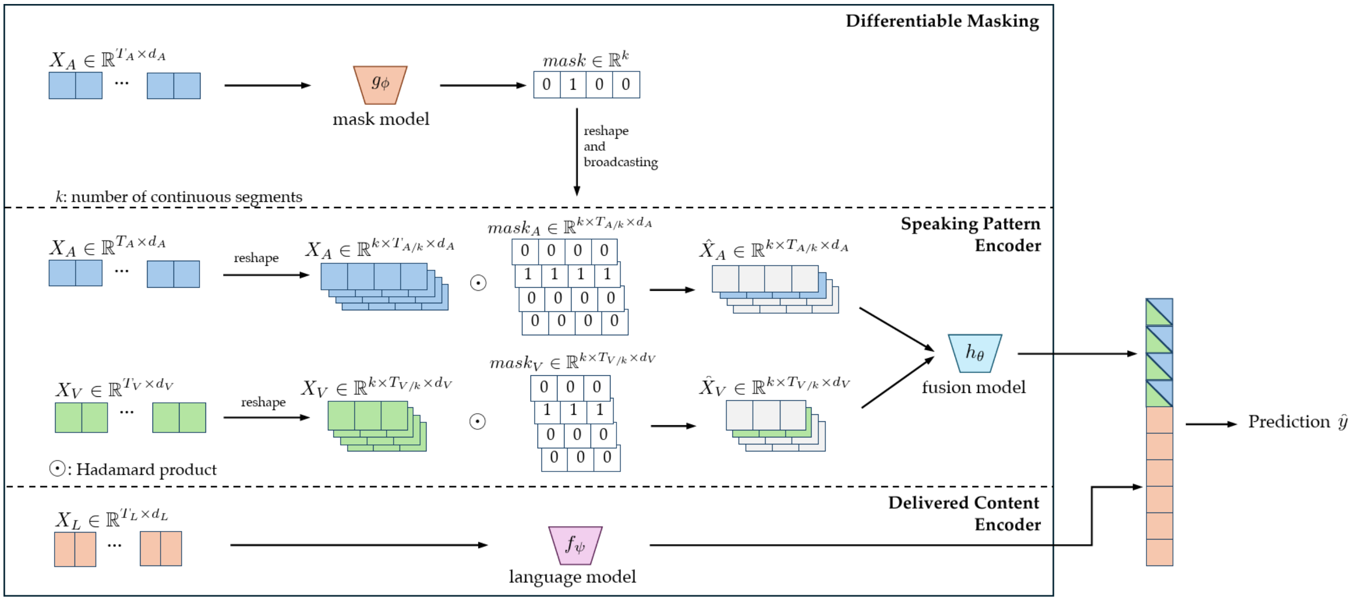
Differentiable Masking
$X_A \in \mathbb{R}^{T_A \times d_A}$
...
g_ϕ
mask model
$mask \in \mathbb{R}^k$
0 1 0 0
reshape and broadcasting
k: number of continuous segments
Speaking Pattern Encoder
$X_A \in \mathbb{R}^{T_A \times d_A}$
reshape
$X_A \in \mathbb{R}^{k \times T_{A/k} \times d_A}$
$mask_A \in \mathbb{R}^{k \times T_{A/k} \times d_A}$
0 0 0 0
1 1 1 1
0 0 0 0
0 0 0 0
$\odot$
$\hat{X}_A \in \mathbb{R}^{k \times T_{A/k} \times d_A}$
h_θ
fusion model
$X_V \in \mathbb{R}^{T_V \times d_V}$
reshape
$X_V \in \mathbb{R}^{k \times T_{V/k} \times d_V}$
$mask_V \in \mathbb{R}^{k \times T_{V/k} \times d_V}$
0 0 0
1 1 1
0 0 0
0 0 0
$\odot$
$\hat{X}_V \in \mathbb{R}^{k \times T_{V/k} \times d_V}$
$\odot$: Hadamard product
Delivered Content Encoder
$X_L \in \mathbb{R}^{T_L \times d_L}$
f_ψ
language model
Prediction $\hat{y}$

Figure 2: Overview of the proposed framework. The process involves initially previewing the audio to create a mask vector, which is used to isolate the most informative acoustic and visual segments of a candidate's speech. These segments are then combined with the content, modeled by a language model, to make the final prediction of the candidate's interview performance.

our goal is to predict the label h corresponding to a score representing the rating of the interviewee. We use the typical modalities which are language (L), visual (V), and acoustic (A) modalities. We denote $X_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ as the input feature sequences corresponding to these modalities. T is the sequence length and d is the embedding dimension.

Inspired by the observation that most humans do not consciously concentrate on all the small nonverbal behaviors of others during communication, we consider that a person's speaking style can be learned without the need to watch the entire video. Accordingly, we can quickly determine "speaking styles" from short clips that contain multimodal information and predict interview performance by combining the speaking style and the entire speech content. In this paper, we refer to the "speaking style" as *speaking patterns* and "the content of the speech" as the *delivered content*. The term *speaking patterns* implies that a person will repeat the same behavior when encountering the same situation during their speech. Some examples of speaking patterns are a person looking to their upper left when they try to recall a memory, a person raising their pitch or nodding their head when they want to emphasize a certain word. In our modeling process, we do not manually define these particular situations. Instead, we rely on the capability of DL models to learn these situations in an end-to-end manner from low-level features. *Delivered content* refers to the verbal information the applicant successfully transmits to listeners.

Our proposed framework consists of three main components: the differentiable masking component, the speaking pattern encoder, and the delivered content encoder. Figure 2 shows the overall framework. We start by extracting the unimodal feature representations $X_{\{L,V,A\}}$; then, we pass these inputs to the components of our network.

3.1 Unimodal Feature Representations

For preprocessing, we automatically extracted linguistic features, acoustic features, and visual features. We detail the feature extraction process for each of the modalities below.

3.1.1 Linguistic features. To extract linguistic features, we first obtain the transcripts of the videos via an automatic speech recognition (ASR) system. In particular, we used OpenAI's Whisper model to generate the transcripts. The word error rate of this ASR system was found to be less than 6%. Once the transcripts are obtained, we generate the word vector representations using the GloVe 300M model. GloVe is a context-independent model, for each word there exists only one embedding vector [43]. The GloVe model was chosen instead of newer and more powerful large language models (e.g., BERT [16] and RoBERTa [33]) for several reasons. First, since GloVe is a context-independent model, this model is considerably more efficient than context-dependent models and the word vector embeddings can be computed in parallel. Second, one of the goals of this research is to study the interactions between multiple modalities. Linguistic features extracted by a powerful pre-trained language model could dominate features extracted from other modalities, which in turn defeats this goal.

3.1.2 Acoustic features. To extract acoustic features, we used the openSMILE toolkit with the eGeMAPSv02 feature set [18, 19]. This feature set includes standard features for speech processing such as loudness, frequency, bandwidth, and amplitude.

3.1.3 Visual features. To obtain visual features, we extracted action units (AUs) via OpenFace [7]. The AUs represent the actions of facial muscles, and they have been shown to be effective in emotion recognition tasks.

3.2 Differentiable Masking

The goal of the differentiable masking component is to select the most informative segment of each video by previewing the audio modality. For different videos, distinct segments may be dynamically selected as the most informative segment. The input to this module is the embedding X_A, which is not expensive to obtain. In addition, we need to specify a hyperparameter k. k represents the number of continuous segments we want to divide the video into, so it should be a positive integer such that the video can be divided into k equal segments. In a general setting, the length of the untrimmed video and the desired length of each segment should be used as guidelines to tune the hyperparameter k. For example, for 120 second video, if we choose $k = 5$ then conceptually we are dividing the video into 5 continuous segments, each lasting $\frac{120}{5} = 24$ seconds. The desirable output of this component is a one-hot vector of size k. In our framework, the mask model g_ϕ is responsible for taking the input and producing the desirable output. Next, we describe our design choices for the mask model g_ϕ in this paper.

Mask Model g_ϕ: The first layer of g_ϕ is a GRU layer [14]. The GRU was chosen since it can capture the temporal aspect of X_A while having fewer parameters than other sequence models such as LSTM networks or transformers. In the inference phase, g_ϕ is used to decide which part of the video we should extract visual features for further processing, so a compact layer is preferred. Then, we pass the output of GRU's last hidden layer to a fully connected (FC) layer. The output of the FC layer is a vector of size k. This layer is used to map the output of the GRU layer to a vector of size k. The last step is to map this logit vector of size k to a one-hot vector. To obtain this one-hot vector output and incorporate it into an end-to-end model, we need an activation function that can output discrete values while being differentiable so that we can do standard backpropagation. We resolve to the Gumbel-Softmax [27, 34] activation function, in which the Gumbel-Softmax trick is applied.

The Gumbel-Softmax Trick: One way to obtain a one-hot vector of size k is to use the Gumbel-Max trick [22, 35]:

$$mask = one\_hot\left(\arg\max_i \left[G_i + log_{\pi_i}\right]\right) \tag{1}$$

where $G_1, G_2, ..., G_k$ are independent and identically distributed (i.i.d) samples drawn from the Gumbel(0,1) distribution, and $\pi_1, ..., \pi_k$ are the logit outputs of the preceding FC layer. This is the same idea as the *reparameterization trick*, but the stochastic element is sampled from the Gumbel distribution instead of the Gaussian distribution. The arg max function, however, is not differentiable. Therefore, in the Gumbel-Softmax approach, an approximation, the Softmax function, is used instead of the arg max function:

$$mask = \frac{exp((log(\pi_i) + G_i))/\tau)}{\sum_{j=i}^{k} exp((log(\pi_i) + G_i))/\tau)} \tag{2}$$

where τ is a newly added temperature factor. As τ approaches 0, *mask* approaches a one-hot vector but the variance of the gradients is large. Another useful trick is to discretize *mask* in Equation 2 during the forward pass while keeping the continuous version for backpropagation during the backward pass.

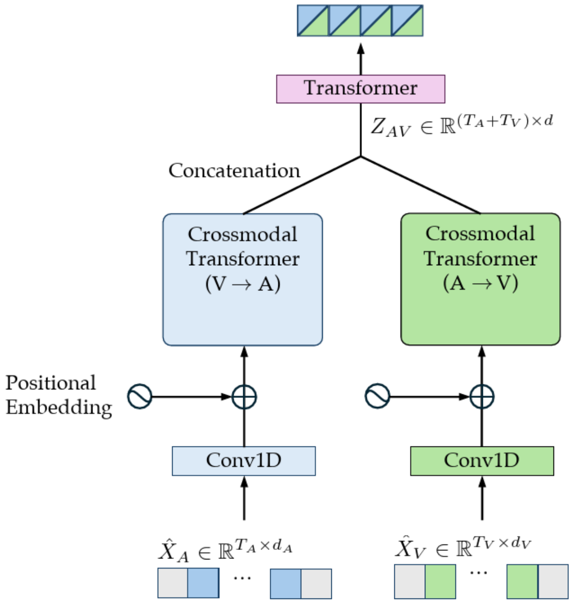
Transformer
$Z_{AV} \in \mathbb{R}^{(T_A+T_V)\times d}$
Concatenation
Crossmodal Transformer (V → A)
Crossmodal Transformer (A → V)
Positional Embedding
Conv1D
Conv1D
$\hat{X}_A \in \mathbb{R}^{T_A\times d_A}$
$\hat{X}_V \in \mathbb{R}^{T_V\times d_V}$
...
...

Figure 3: Architecture of the multimodal fusion model h_θ

3.3 Speaking Pattern Encoder

The goal of the speaking pattern component is to model the way a person speaks from multiple modalities (specifically, audio and vision data). We hypothesize that the speaking style of a person does not change significantly during a short time span, so a short segment of the video is sufficient to model the speaking pattern. As shown in Figure 2, given the acoustic features X_A, we first reshape them into k continuous segments, each of which has a length of $T_{A/k}$, resulting in a new tensor in $\mathbb{R}^{k\times T_{A/k}\times d_A}$. We then multiply this new tensor with the mask tensor using the Hadamard product. This multiplication is made possible thanks to the reshaping and broadcasting functions in PyTorch. As a result, the mask layer has the same shape as the new X_A. The same procedure is applied to the X_V tensor. After the Hadamard product operation, we reshape the result back to its original shape and obtain two masked inputs $\hat{X}_A$ and $\hat{X}_V$ where $k-1$ segments are masked with zeros. These inputs are then passed to a fusion model h_θ. We opted for a slightly modified version of the crossmodal transformers, which was first introduced by [47].

Fusion Model h_θ: The architecture of the fusion model is shown in Figure 3. This architecture is very similar to the architecture of MulT [47]. Given two masked inputs $\hat{X}_A \in \mathbb{R}^{T_A\times d_A}$ and $\hat{X}_V \in \mathbb{R}^{T_V\times d_V}$, we pass them though a 1D-convolution layer. This layer is expected to capture the local structure of the sequence and project the features of the two different modalities to the same dimension d. We then add the positional embeddings and pass them to the two crossmodal transformers and concatenate the results to obtain the embeddings $Z_{AV} \in \mathbb{R}^{(T_A+T_V)\times d}$. For more details on the positional embedding operation and the crossmodal transformer module, we refer readers to [47]. The embeddings Z_{AV} are then passed to a self-attention transformer [48], and the last element of the sequence model is passed to the next stage.

Table 1: Summarized statistics about the dataset

Variable	Number	
Number of interviewees	260	people
Number of questions per interviewee	8	questions
Number of videos	1891	videos
Length of each video	2	minutes
Training set	1211	samples
Validation set	308	samples
Test set	372	samples

3.4 Delivered Content Encoder

To model what the interviewee said, we propose the use of a language model f_ψ. The input to the language model is the whole sequence since we want this model to capture everything the interviewee said. Specifically, our language model f_ψ consists of a 1D convolution layer followed by a self-attention transformer. We use the last element of this sequence model as the input to the next stage.

3.5 Prediction Layers

As the final step, we concatenate the outputs of the speaking pattern and the delivered content components and pass them to two FC layers. A rectified linear unit (ReLU) activation function is used as the nonlinear layer between the two FC layers. The output of the last FC layer is the prediction $\hat{y}$.

4 EXPERIMENTS

4.1 Dataset

In this work, we conducted experiments on the large-sized corpus of video interview judgments collected by [11]. This dataset contains a total of 1891 monologue videos. Each video is a 2-minute record of one of the Mechanical Turk workers (Tuckers) located in the United States answering one of eight predefined questions. In total, videos of 260 Tuckers are included in the dataset (as some videos were corrupted, not all Tuckers answered all eight questions). The questions were designed to evaluate four important social skills (two questions for each skill) required in any job: (a) communication skills, (b) interpersonal skills, (c) leadership skills, and (d) persuasion and negotiation skills.

The dataset includes "Hiring Recommendation" (i.e., job interview performance) labels and five personality trait labels. In this work, we omitted the personality trait labels and used only job interview performance labels. Five experts in rating essays and video performance were asked to make holistic judgments about the video responses separately. The raters were given the assumption that the responses were for an entry-level office position. To train our model and evaluate the performance of our approach, we separated the dataset into a training set, a validation set, and a test set. The sets were separated so that no applicant appears in the same set and the label distributions among the sets are similar. For the test set, we used the same set proposed by the original authors. In Table 1, we summarize the statistics of the dataset.

Table 2: Model hyperparameters

Hyperparameter	Value
Batch Size	64
Learning Rate (LR)	$1e-4$
LR scheduler	Cosine Annealing
Optimizer	Adam
Number of Epochs	60
Number of crossmodal Blocks d	4
Number of crossmodal Attention Heads	5
Textual Embedding Dropout	0.25
Crossmodal Attention Block Dropout	0.3
Gradient Clip	0.8
Loss Function	MSELoss

4.2 Baselines

To evaluate the performance of our proposed method, we conduct experiments with three baseline models. The first two baseline models are classification models (the labels were binarized into two classes before training), so their performance cannot be compared directly with that of our method; however, we still provide their results here since the experiments with these models were conducted with the same dataset as used in our performance evaluation.

4.2.1 *Support Vector Machine (SVM).*

In [11], the authors transformed the labels of the interview dataset into two classes (high-quality and low-quality) and compared the performance of multiple classical ML models. The SVM model was found to produce the best classification results, and only the language modality was found to be effective for classification.

4.2.2 *RoBERTa-FNN.*

Inspired by [11], the authors of [30] investigated the effect of only linguistic features when different ASR systems, feature extraction methods, and models were combined. A model that is a combination of a language model (*RoBERTa*) and a customized feedforward neural network achieved a new state-of-the-art performance based on this dataset.

4.2.3 *Multimodal Transformer (MulT).*

In [47], the authors proposed the MulT architecture to model human multimodal affection recognition. They introduced the idea of using crossmodal attention to learn the interactions among multimodal data at different time steps. In our work, we adopted crossmodal attention to learn the latent representations of two modalities in the speaking pattern component.

4.3 Implementation Details and Hyperparameters

Implementation Details: For linguistic features, we set the maximum sequence length to 512, so X_L has a size of 512 × 300. For the acoustic features, we calculated the set of 88 functional features for every 400 milliseconds, with a sliding window of 200 milliseconds. The final result X_A is a vector of size 600 × 88. For the visual features, 35 AUs were extracted at a rate of 5 frames per second, so X_V is a vector with a size of 600 × 35.

Table 3: Comparison of the results of the baseline models and our models

Model	MAE ↓	Corr ↑	Acc7 ↑	Acc2 ↑	F1 ↑
SVM	-	-	-	66.40	66.38
RoBERTa-FNN	-	-	-	70.33	70.29
MulT	0.7274	0.4913	43.27	61.67	62.63
Ours	**0.5201**	**0.6722**	**57.53**	**74.57**	**74.73**

Our model was implemented in PyTorch. For the GRU layer in the mask model g_{ϕ}, the hidden unit size was set to 64. The following FC layer had an input size of 64 and an output size of k. The dimensions d of the 1D convolutional layer in the fusion model h_{θ} and the language model f_{ψ} were set to 30 and 120, respectively. The lengths of the convolutional neural network (CNN) filters were calculated based on the input modality feature size and the output dimension d. The kernel size and stride were set to 1, and the padding was set to 0. For the prediction layers, the sizes of the input and hidden layers were both 150, and the size of the output layer was 1. The training process was completed within 20 minutes with one NVIDIA A40 GPU.

Hyperparameters: The full list of the other hyperparameters is reported in Table 2.

5 RESULTS AND ANALYSIS

In this section, we present the results of the experiments and the analysis of the results. We evaluate the performance of the models using five different metrics: the mean absolute error (MAE), Pearson correlation (Corr - %), accuracy 7 (Acc7 - %), accuracy 2 (Acc2 - %), and F1 score (%). The Acc7 scores are accuracy scores calculated by rounding the regression predictions to the nearest integers, and the Acc2 and F1 scores are accuracy scores calculated when the median score is used as the threshold for separating regression predictions into two classes. The five metrics were previously used in [47], and we used the same binary threshold as used in [13, 30].

5.1 Comparison with Baselines

We first evaluate the performance of the proposed method and that of prior approaches based on the same dataset. Table 3 shows a comparison of the results of the baseline approaches with the results of our method. Across all the metrics, the proposed method shows significant improvements compared with the scores of the baseline models, especially when multiple modalities are considered. In terms of binary classification, the F1 score of the proposed method is improved by 3.75% compared with the previously reported single-task result.

5.2 Choosing an Appropriate k Value

To understand the trade-off between accuracy and efficiency, we conducted experiments with multiple values of k. In theory, using a larger value of k could speed up the inference time, but the use of larger k values may lead to reduced accuracy because of insufficient information. Table 4 shows that when k is large (i.e., $k = 10$, $k = 15$, and $k = 20$), the model does not have sufficient information to learn the speaking patterns of the interviewees. In addition, when $k = 1$, the performance of the model is slightly worse than that obtained with $k = 5$. This finding supports our original hypothesis that using all visual features may be counterproductive.

Table 4: Results from the multimodal experiments with different values of k.

k	MAE ↓	Corr ↑	Acc 7 ↑	Acc 2 ↑	F1 ↑
1	0.5374	0.6644	55.91	73.33	73.39
5	**0.5201**	**0.6722**	**57.53**	**74.57**	**74.73**
10	0.5332	0.6590	55.65	72.12	72.31
15	0.5498	0.6491	54.30	72.28	72.31
20	0.5357	0.6585	55.65	72.63	72.85

Table 5: Results of the ablation studies when one of the two main components was removed ($k = 5$)

Description	MAE ↓	Corr ↑	Acc 7 ↑	Acc 2 ↑	F1 ↑
Delivered content	0.5348	0.6635	54.03	72.14	72.31
Speaking patterns	0.6376	0.4595	52.69	66.95	67.47
Full	**0.5201**	**0.6722**	**57.53**	**74.57**	**74.73**

5.3 Ablation Studies

To better understand the effects of the components in the proposed method on model performance, several ablation studies are conducted. First, we conduct experiments in which only the delivered content or speaking pattern components are used. Table 5 shows the results of these ablation studies.

The results show that when the speaking pattern component is deactivated, the F1 score decreases by more than 2% compared to the F1 score of the model in which this component is used. In addition, the correlation coefficient decreases by nearly 1% (from 0.6722 to 0.6635). This result corroborates previous studies, in which verbal behavior alone was shown to be highly predictive [11].

On the other hand, when only the speaking pattern component is activated, model performance significantly declines. The correlation coefficient decreases by 21.27% (from 0.6722 to 0.4595), and the F1 score decreases to 67.47% from 74.73%. This drop in performance is expected because the speaking pattern component uses only one-fifth of the video. It is rather surprising that with only a 24-second segment, the model can still perform better than random guesses.

It is clear from Table 5 that linguistic features have a huge contribution to the performance of the model. To further study the contribution of the speaking pattern module, in Table 6, we conducted the experiments with different values for k while deactivating the delivered content component. The results show that the model performance metrics decrease as k increases. This is to be expected since as k increases, the model gets to see a smaller segment of the video.

Table 6: Results for different values for k when textual features are not included.

k	MAE ↓	Corr ↑	Acc 7 ↑	Acc 2 ↑	F1 ↑
1	**0.6319**	**0.4871**	**52.69**	**71.23**	**71.24**
5	0.6376	0.4595	52.69	66.95	67.47
10	0.6577	0.4480	52.42	66.00	66.67
15	0.6687	0.3812	49.73	64.00	64.78
20	0.6829	0.3551	52.42	66.95	67.47

5.4 Most Informative Segments across Videos

Given that our approach produces a mask vector identifying the most informative segment in each video, plotting this statistic can provide an overview of the most informative segments across videos. In Figure 4, the X-axis indicates the timespans (with $k = 5$, resulting in each timespan lasting 24 seconds) and the Y-axis represents the number of samples in the test set where the model chose a specific timespan to capture the speaking pattern. The figure shows that for most videos, the model selects the middle segments. This finding supports the assumption presented in [45], in which the authors suggested that interviewees stopped speaking before the end of the allotted time, resulting in less information in those segments.

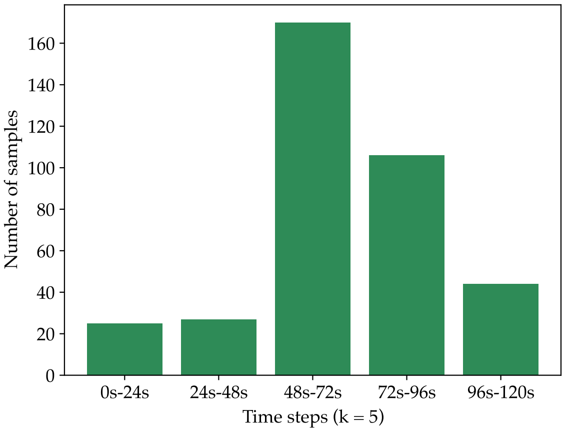

Figure 4: Histogram of informative segments selected the model in different videos. For most videos, the model focuses mainly on the middle part of the video and only occasionally selects the beginning and end segments for some videos.

5.5 Limitations

Our framework, while effective, has certain limitations that need to be acknowledged and addressed. One significant shortcoming is that it does not account for biases that may arise from the model itself or from the annotations used in training. Addressing bias is crucial for ensuring fairness in the personnel selection process, a matter of great importance and urgency. While not the focus of this paper, it is important to note that we have actively engaged in efforts to mitigate bias within our models in a separate line of work. In another paper [17], we introduced a series of approaches aimed at reducing bias in the multimodal, multi-class prediction of a behavioral construct. There, our study evaluated the performance enhancements achieved by these bias mitigation techniques using the same dataset. The results indicated that by adjusting the loss function to account for perceived races, genders, accents, and ages, our multimodal models significantly outperformed the unmitigated baselines. These findings, along with other implications for automated feedback on the construct prediction, are discussed in detail in this prior publication.

Another limitation of our method is that applicants are rated on a fixed scale from 1 to 7. This restricted range may not capture the full nuances of an applicant's capabilities and performance. Moreover, the system does not provide feedback to applicants, which is a critical gap. Providing constructive feedback is essential, as it helps applicants understand their performance and identify areas for improvement. Furthermore, our method suffers from low interpretability. This means that the model does not offer clear insights into or reasons for the scores assigned to applicants. Such transparency is vital for companies to understand and justify their hiring decisions. Without this information, companies cannot provide candidates with reasons for their scores, which can lead to perceptions of unfairness and lack of transparency in the hiring process. This effort to provide feedback and model interpretation is addressed in another line of work currently [31].

Addressing these limitations will involve refining the model to incorporate mechanisms for detecting and mitigating bias and enhancing the model's transparency and feedback capabilities. Such improvements are essential for building trust and ensuring fairness that is beneficial for both companies and applicants.

6 CONCLUSION

In this paper, we discovered that utilizing brief segments from job interview videos could enhance model performance. Our method employs differentiable masking to automatically select these short segments, capturing the candidate's speaking patterns and speech content through multimodal information from the clips and the full transcript of the entire video, respectively. With this approach, computational cost and the time required for the inference phase are both reduced. The experimental results confirm the effectiveness of our method, showing improved performance over models using multimodal information from entire videos. This study sets the stage for more in-depth investigations into the efficient use of multimodal information, particularly in analyzing job interview performance.

Looking forward, there are several promising avenues for further research. Instead of a fixed duration, an utterance-based approach could be adopted for learning speaking patterns. Additionally, a more comprehensive analysis could involve using multiple brief clips from different parts of the interview rather than a single segment, selecting the top-k most informative segments. Moreover, leveraging advanced large language models like BERT or GPT-3 [9], which have demonstrated impressive results in various natural language processing tasks, could significantly enhance our ability to analyze the content delivered by applicants.

ACKNOWLEDGMENTS

Hung is supported by the Japanese Government (MEXT) Scholarship. This work was partially supported by JSPS KAKENHI (22H04860, 22K21304, 22H00536, 23H03506), and JST AIP Trilateral AI Research, Japan (JPMJCR20G6).

REFERENCES

[1] Chanho Ahn, Eunwoo Kim, and Songhwai Oh. 2019. Deep Elastic Networks With Model Selection for Multi-Task Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6528–6537. https://doi.org/10.1109/ICCV.2019.00663

[2] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. 2018. Action Search: Spotting Actions in Videos and Its Application to Temporal Action Localization. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IX* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 253–269. https://doi.org/10.1007/978-3-030-01240-3_16

[3] Nalini Ambady, Frank J. Bernieri, and Jennifer A. Richeson. 2000. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In *Advances in Experimental Social Psychology*. Vol. 32. Academic Press, 201–271. https://doi.org/10.1016/S0065-2601(00)80006-4

[4] Keith Anderson, Elisabeth André, T. Baur, Sara Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, Kaśka Porayska-Pomsta, P. Rizzo, and Nicolas Sabouret. 2013. The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews. In *Advances in Computer Entertainment (Lecture Notes in Computer Science)*, Dennis Reidsma, Haruhiro Katayose, and Anton Nijholt (Eds.). Springer International Publishing, Cham, 476–491. https://doi.org/10.1007/978-3-319-03161-3_35

[5] Umut Asan and Ayberk Soyer. 2022. A Weighted Bonferroni-OWA Operator Based Cumulative Belief Degree Approach to Personnel Selection Based on Automated Video Interview Assessment Data. *Mathematics* 10, 9 (2022). https://doi.org/10.3390/math10091582

[6] Pooja Rao S B, Manish Agnihotri, and Dinesh Babu Jayagopi. 2020. Automatic Follow-up Question Generation for Asynchronous Interviews. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, Daniel Sánchez, Raquel Hervás, and Albert Gatt (Eds.). Association for Computational Lingustics, Santiago de Compostela, Spain, 10–20. https://aclanthology.org/2020.intellang-1.2

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[8] Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K. D'Mello. 2021. Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. ACM, Montréal QC Canada, 268–277. https://doi.org/10.1145/3462244.3479897

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[10] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao. 2021. Learning What and When to Drop: Adaptive Multimodal and Contextual Dynamics for Emotion Recognition in Conversation. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 1064–1073. https://doi.org/10.1145/3474085.3475661

[11] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 504–509. https://doi.org/10.1109/ACII.2017.8273646 ISSN: 2156-8111.

[12] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 163–171. https://doi.org/10.1145/3136755.3136801

[13] Mathieu Chollet and Stefan Scherer. 2017. Assessing Public Speaking Ability from Thin Slices of Behavior. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. 310–316. https://doi.org/10.1109/FG.2017.45

[14] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Workshop on Deep Learning, December 2014* (2014).

[15] Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3243–3255. https://doi.org/10.18653/v1/2020.emnlp-main.262

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[17] Andrew Emerson, Arti Ramesh, Patrick Houghton, Vinay Basheerabad, Navaneeth Jawahar, and Chee Wee Leong. 2024. Multimodal, Multi-Class Bias Mitigation for Predicting Speaker Confidence. In *International Conference on Educational Data Mining*.

[18] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[19] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, Firenze Italy, 1459–1462. https://doi.org/10.1145/1873951.1874246

[20] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. 2020. Listen to Look: Action Recognition by Previewing Audio. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10454–10464. https://doi.org/10.1109/CVPR42600.2020.01047

[21] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2225–2235. https://doi.org/10.18653/v1/P18-1207

[22] Emil Julius Gumbel. 1954. *Statistical Theory of Extreme Values and Some Practical Applications. A Series of Lectures.* Technical Report PB175818. National Bureau of Standards, Washington, D. C. Applied Mathematics Div. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB175818.xhtml

[23] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019. HireNet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 71, 9 pages. https://doi.org/10.1609/aaai.v33i01.3301573

[24] Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. 2023. Multimodal Hierarchical Attention Neural Network: Looking for Candidates Behaviour Which Impact Recruiter's Decision. *IEEE Transactions on Affective Computing* 14, 2 (April 2023), 969–985. https://doi.org/10.1109/TAFFC.2021.3113159 Conference Name: IEEE Transactions on Affective Computing.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[26] Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: my automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 697–706. https://doi.org/10.1145/2493432.2493502

[27] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rkE3y85ee

[28] Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baró, and Sergio Escalera. 2021. Person Perception Biases Exposed: Revisiting the First Impressions Dataset. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*. 13–21. https://doi.org/10.1109/WACVW52041.2021.00006

[29] Everlyne Kimani, Prasanth Murali, Ameneh Shamekhi, Dhaval Parmar, Sumanth Munikoti, and Timothy Bickmore. 2020. Multimodal Assessment of Oral Presentations using HMMs. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 650–654. https://doi.org/10.1145/3382507.3418888

[30] Hung Le, Sixia Li, Candy Olivia Mawalim, Hung-Hsuan Huang, Chee Wee Leong, and Shogo Okada. 2023. Investigating the Effect of Linguistic Features on Personality and Job Performance Predictions. In *Social Computing and Social Media (Lecture Notes in Computer Science)*, Adela Coman and Simona Vasilache (Eds.). Springer Nature Switzerland, Cham, 370–383. https://doi.org/10.1007/978-3-031-35915-6_27

[31] Chee Wee Leong, Navaneeth Jawahar, Vinay Basheerabad, Torsten Wöertwein, Andrew Emerson, and Guy Sivan. 2024. Combining Generative and Discriminative AI for High-Stakes Interview Practice. In *Proceedings of the 2024 International Conference on Multimodal Interaction*.

[32] Chee Wee Leong, Katrina Roohr, Vikram Ramanarayanan, Michelle P Martin-Raugh, Harrison Kell, Rutuja Ubale, Yao Qian, Zydrune Mladineo, and Laura McCulla. 2019. To trust, or not to trust? A study of human bias in automated video interview assessments. *arXiv preprint arXiv:1911.13248* (2019).

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692 arXiv:1907.11692 [cs].

[34] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=S1jE5L5gl

[35] Chris J. Maddison, Daniel Tarlow, and Tom Minka. 2014. A* sampling. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3086–3094.

[36] Skanda Muralidhar, Emmanuelle Patricia Kleinlogel, Eric Mayor, Adrian Bangerter, Marianne Schmid Mast, and Daniel Gatica-Perez. 2020. Understanding Applicants' Reactions to Asynchronous Video Interviews Through Self-reports and Nonverbal Cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 566–574. https://doi.org/10.1145/3382507.3418869

[37] Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, Ljubljana, 1–6. https://doi.org/10.1109/FG.2015.7163127

[38] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. 2014. Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia* 16, 4 (2014), 1018–1031. https://doi.org/10.1109/TMM.2014.2307169

[39] Laurent Son Nguyen and Daniel Gatica-Perez. 2015. I Would Hire You in a Minute: Thin Slices of Nonverbal Behavior in Job Interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/2818346.2820760

[40] Laurent Son Nguyen and Daniel Gatica-Perez. 2016. Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia* 18, 7 (2016), 1422–1437. https://doi.org/10.1109/TMM.2016.2557058

[41] Tomoya Ohba, Candy Olivia Mawalim, Shun Katada, Haruki Kuroki, and Shogo Okada. 2022. Multimodal Analysis for Communication Skill and Self-Efficacy Level Estimation in Job Interview Scenario. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia (MUM '22)*. Association for Computing Machinery, New York, NY, USA, 110–120. https://doi.org/10.1145/3568444.3568461

[42] Shogo Okada, Laurent Son Nguyen, Oya Aran, and Daniel Gatica-Perez. 2019. Modeling Dyadic and Group Impressions with Intermodal and Interperson Features. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1s (Jan. 2019), 13:1–13:30. https://doi.org/10.1145/3265754

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[44] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. In *Computer Vision – ECCV 2016 Workshops (Lecture Notes in Computer Science)*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, Cham, 400–418. https://doi.org/10.1007/978-3-319-49409-8_32

[45] Wasifur Rahman, Sazan Mahbub, Asif Salekin, Md Kamrul Hasan, and Ehsan Hoque. 2021. HirePreter: A Framework for Providing Fine-grained Interpretation for Automated Job Interview Analysis. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 1–5. https://doi.org/10.1109/ACIIW52867.2021.9666201

[46] Sowmya Rasipuram and Dinesh Babu Jayagopi. 2020. Automatic multimodal assessment of soft skills in social interactions: a review. *Multimedia Tools and Applications* 79, 19 (May 2020), 13037–13060. https://doi.org/10.1007/s11042-019-08561-6

[47] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[49] Zuxuan Wu, Hengduo Li, Caiming Xiong, Yu-Gang Jiang, and Larry S. Davis. 2022. A dynamic frame selection framework for fast video recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2022), 1699–1711. https://doi.org/10.1109/TPAMI.2020.3029425

[50] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, and Rogerio Feris. 2018. BlockDrop: Dynamic Inference Paths in Residual Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8817–8826. https://doi.org/10.1109/CVPR.2018.00919

[51] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. 2678–2687. https://doi.org/10.1109/CVPR.2016.293

[52] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/D17-1115

[53] Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Effective Sentiment-relevant Word Selection for Multi-modal Sentiment Analysis in Spoken Language. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 148–156. https://doi.org/10.1145/3343031.3350987