



HAL
open science

Slices of Attention in Asynchronous Video Job Interviews

Léo Hemamou, Ghazi Felhi, Jean-Claude Martin, Chloé Clavel

► **To cite this version:**

Léo Hemamou, Ghazi Felhi, Jean-Claude Martin, Chloé Clavel. Slices of Attention in Asynchronous Video Job Interviews. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Sep 2019, Cambridge, United Kingdom. hal-02370820

HAL Id: hal-02370820

<https://hal.science/hal-02370820>

Submitted on 19 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Slices of Attention in Asynchronous Video Job Interviews

Léo Hemamou^{*†‡}, Ghazi Felhi^{*}, Jean-Claude Martin[†] and Chloé Clavel[‡]

^{*}EASYRECRUE, 38 Rue du Sentier, 75002 Paris, France

Email: {l.hemamou,g.felhi}@easyrecrue.com

[†]LIMSI, CNRS, Paris-Sud University, Paris-Saclay University / F-91405 Orsay, France

Email: Jean-Claude.Martin@limsi.fr

[‡]LTCI, Télécom ParisTech, Paris-Saclay University / F-75013 Paris, France

Email: chloe.clavel@telecom-paristech.fr

Abstract—The impact of non verbal behaviour in a hiring decision remains an open question. Investigating this question is important, as it could provide a better understanding on how to train candidates for job interviews and make recruiters be aware of influential non verbal behaviour. This research has recently been accelerated due to the development of tools for the automatic analysis of social signals (facial expression detection, speech processing, etc), and the emergence of machine learning methods. However, these studies are still mainly based on hand engineered features, which imposes a limit to the discovery of influential social signals. On the other side, deep learning methods are a promising tool to discover complex patterns without the necessity of feature engineering. In this paper, we focus on studying influential non verbal social signals in asynchronous job video interviews that are discovered by deep learning methods. We use a previously published deep learning system that aims at inferring the hirability of a candidate with regard to a sequence of interview questions. One particularity of this system is the use of attention mechanisms, which aim at identifying the relevant parts of an answer. Thus, information at a fine-grained temporal level could be extracted using global (at the interview level) annotations on hirability. While most of the deep learning systems use attention mechanisms to offer a quick visualization of slices when a rise of attention occurs, we perform an in-depth analysis to understand what happens during these moments. First, we propose a methodology to automatically extract slices where there is a rise of attention (*attention slices*). Second, we study the content of *attention slices* by comparing them with randomly sampled slices. Finally, we show that they bear significantly more information for hirability than randomly sampled slices, and that such information is related to visual cues associated with anxiety and turn taking.

Index Terms—Attention mechanism, recurrent neural networks, thin-slice, social signals, job interview.

I. INTRODUCTION

The procedure of personnel selection includes gathering data about the potential candidates, for example, in a job interview [32]. Research in Affective Computing can be helpful in many ways with respect to job interviews, for example virtual recruiters can help candidates train their social skills and rehearse [16]. This automatic processing can help recruiters assess candidates. Additionally, it can help researchers and recruiters understand the evaluation process done by recruiters when assessing a candidate. Initially conducted face to face or via phone, job interviews are now often done by online video

conferencing systems or by asynchronous video recordings. An asynchronous video interview is an emergent tool now offered by several companies responding to the needs of initial assessment in personnel selection. The procedure is as follows: the candidate connects to a web platform and answers a sequence of questions predefined by the recruiter while recording a video of himself with his webcam, smartphone or tablet. Later, recruiters connect to the same platform, watch the candidate's answers, rate the answers and then decide whether they want to invite the candidate to a face-to-face interview.

Researchers are already developing systems for automatically predicting hirability based on non verbal cues of candidates in asynchronous video interviews [4], [26]. In this context and in addition to new legislative constraints (General Data Protection Regulation), such automatic systems require interpretability and transparency. With these systems, candidates will be able to improve their non verbal behavioral strategy, and recruiters will be able to assess these decision support models. These models could even help recruiters understand their own biases.

Classical approaches in social computing consist of building machine learning models and interpreting the importance of the features [25], [29], [37]. These approaches are not able to bring forward and stress the effect of unexpected and influential social cues. We previously proposed a deep learning model trained only with recruiter's decision and videos [15]. Our model is able to consider temporality and influential slices of video-based asynchronous interviews, due to the use of an attention mechanism running on top of a recurrent neural network. A sequence of features is processed by a recurrent neural network, and the attention mechanism aims at learning a different weight for each time step to enhance the performance of the classification task. Overall, such techniques could be useful for understanding human behaviour, as they aim to separate task relevant time steps in a sequence from irrelevant time steps [39]. However, research in this area is limited to studying and highlighting just a few examples of peaks of attention curves [22], [39]. Hence, a consistent validation is needed in order to ascertain the usefulness of the system's output for predicting hirability as rated by recruiters.

With this in mind, this article describes three experiments

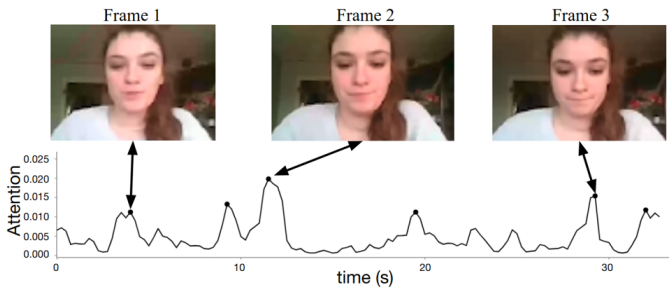


Fig. 1: Example of attention curve and salient moments detected with peaks in HireNet.

we conducted to understand whether slices of video interviews highlighted by the attention model do carry information that is useful to recruiters. In section IV, we propose a methodology to automatically extract thin slices where attention values are high. In section V, we test whether the non verbal behavior occurring during these slices is different from behavior occurring in randomly picked slices. Finally, in section VI, we evaluate whether the extracted slices are more informative with regards to the hirability of a candidate.

II. RELATED WORKS

A. Job interview and Non Verbal Behaviour

Non verbal visual and audio cues have been studied in order to predict interview performance [11], anxiety [10], personality of the candidates [7] or deception [33]. Numerous visual cues such as physical attractiveness, hand gestures, smiling, eye contact, nodding, head movement, body orientation, facial cues, leg movements have been used throughout experiments. For example torso movement, face touching, leg fidgeting, [10] neutral expression and less smiling have been found [13] to negatively correlate to interview performance, whereas eye contact, hand gesture [10], head movement [33] correlates positively with interview performance. Moreover, these cues could have a different impact on interviewer evaluations depending on interview structure, job position (blue collar vs white collar) or settings of the interview (such as telephone, computer-mediated video chat and asynchronous video interview) [12].

Putting aside efficiency, annotating every segment of a video is time consuming. One common approach to deal with the task of annotations is to only annotate a part of the job interview. In fact, it has been shown that, using only a short amount of information, people can infer correctly personal characteristics, traits or states of an individual [2], [24]. This approach is called thin slice analysis and has already been used in social interactions study [24], first impressions [2], public speaking [6], or job interviews [28]. Another advantage of this method is that it highlights brief, non verbal behavior with respect to perceived impressions. Nonetheless, the duration and sampling strategy for thin slices remain an open question. Previous studies focus on sampling thin slices randomly [6], using the structure of the job interview (slices based on questions and answers) [28], or at the beginning and end

of the interactions [7]. Automatic methods based on social signal processing could give way to a better selection of thin slices and their duration, by selecting regions that carry more information.

B. Automatic methods for understanding human behavior in job interviews

Recent advances drastically reduce time spent in manually coding behavioral cues. Tools are now available to automatically code vocal [9] or visual [1] cues. Recent studies use social signal processing and machine learning to understand the links between non verbal cues and hirability. These studies have been applied to different job interview settings: face to face interviews [27], [28], asynchronous video interviews [4] and computer-mediated video chat [29].

Among investigated traits in job interviews (communication skills [29], personality [4], etc), hirability remains the most studied one. Usually, two methods are used to understand which extracted features are important for hirability: correlation analysis and feature importance analysis conducted on a trained machine learning model. However, feature importance analysis is highly dependent on a machine learning pipeline. To the best of our knowledge, only traditional machine learning (SVM, Lasso, Ridge, etc) has been used so far. In a previous work, we investigated the use of deep learning techniques, and established their superiority in terms of predictive capabilities [15]. However no in-depth analysis about the feedback returned by the model was conducted.

C. Neural Networks and explainability

Neural networks are able to find more statistical patterns than traditional machine learning methods such as SVMs, logistic regressions or Random Forests. Moreover, specific architectures such as recurrent neural networks allow for modeling temporality by managing sequences. However, the freedom they have to construct intermediate representation comes at the cost of an extreme opacity. This opacity hinders their usability for critical applications such as healthcare, justice, or human resources. Therefore several researchers have tried to propose methods to better explain these networks. First, the visualization of hidden states has been explored to better understand intermediate representations automatically built by the networks, especially in computer vision [38].

A second approach, called knowledge distillation, consists of learning an interpretable model from an already trained complex neural networks [21]. A third method consists of explaining predictions for specific instances (as opposed to explaining the whole model). Some of the attempts build a local boundary for these predictions [30], using sensitivity methods or analyzing integrated gradients for image analysis [34]. Finally, attention mechanisms have recently gained popularity for enhancing performance and interpretability. Specifically in the Social Computing area, the use of attention mechanisms for rapport detection [39] or for the evaluation of job interview performance in asynchronous video interviews [15] has been proposed to extract fine grained information at temporal

Modality	Text	Audio	Video
Train set	6350	6034	5706
Validation set	794	754	687
Test set	794	755	702
Questions per interview (mean)	5.05	5.10	5.01
Total length	3.82 M words	557.7 h	508.8 h
Length per question (mean)	95.2 words	52.19 s	51.54 s
<i>Hirable</i> label proportion	45.0 %	45.5 %	45.4 %

TABLE I: Number of candidates in each set and overall statistics of the dataset.

level using only coarse annotation at the interview level. However, most of the studies restrict the analysis of attention mechanisms to the display of examples and do not conduct an in-depth analysis. Moreover, the validity of the attention curves as an explanation has recently been called into question [17].

III. EXPERIMENTAL SETUP

A. Dataset

As our goal is to evaluate and assess the relevance of attention mechanisms that are already trained, we use the same database previously collected by us [15]. This database contains real French asynchronous video interviews of 7938 candidates applying for 475 sales positions. Each interview of a specific position has the same number of questions predefined upstream by the recruiter. Once the candidates finish answering the set of predefined questions on the web platform, recruiters and managers can connect to this platform, watch these answers, and evaluate the candidate. They can like, dislike, shortlist candidates, evaluate them on predefined criteria or write comments. Based on this information, candidates who have been liked or shortlisted have been labelled "hirable", otherwise they are labelled "not hirable". If candidates received different annotations from multiple recruiters, a majority vote was taken. In case of draw, the candidate is considered "hirable". To the best of our knowledge, this database is the one with the highest number of real applicants assessed by real practitioners for a real position. We extracted verbal content using an automatic speech recognition tool (Google API). These Asynchronous Video Interviews have been recorded in the wild from various devices leading to a wide range of setups. Due to this condition, technical problems could occur such as videos without audio, illumination problems in videos, or failure in the automatic speech recognition. Descriptive statistics for each modality are available in Table I. The dataset can not be made available to the public due to high privacy constraints.

B. HireNet

In a previous article, we proposed HireNet, an attention neural network to infer hirability from structured video interviews. HireNet [15] was conceived to represent a sequence of questions and their answers containing themselves a sequence of social signals. In the following sections, we focus only on

the low level encoder of our model which aims to detect salient social signals. This encoder is a bidirectional Gated Recurrent Unit (GRU) [5] which encodes information from a sequence of low level descriptors. This encoder is followed by an attention mechanism that weights each timestep differently according to its importance. We aim to validate the usefulness of attention mechanisms to automatically extract the most useful slices to predict hirability. For the following study, we define an *attention slice* as a slice selected according to the attention curve. For the sake of simplicity, we decided to focus only on visual features in the first question. The first question is highly linked to self presentation tactics, and initial impressions play an important role in the variance of interview scores [35]. Moreover, our preliminary inspection of attention curves has shown that for the visual modality, attention peaks appear more frequently than for other modalities. These visual features are the position and orientation of the head, and continuous and categorical facial action units activations which have been extracted using OpenFace [1]. Values were smoothed with a time window of 0.5s and an overlap of 0.25s before being fed to our model. This duration is frequently used in the literature of Social Computing [36] and we validated it for our corpus by annotating the duration of social signals in a set of videos.

We trained our model to achieve best results for the mean of F-1 score of the positive and negative class rather than only on the positive class as we did in [15]. We call this average *Mean F1*. As neural networks are subject to various variability sources such as random weights initialization, stochastic gradient descent or dropout, we chose to train five different instances of the model and then averaged the attention values. That way, we aim to capture the more general behaviour of attention mechanisms [18]. Mean performance and confidence interval details on test set are reported in Table II.

TABLE II: Performances of our model on test set for hirability prediction task

Model	F1 Positive Class	F1 Negative Class	<i>Mean F1</i>
HireNet	0.607 ± 0.023	0.628 ± 0.013	0.618 ± 0.008

IV. DO ATTENTION CURVES ACTUALLY EXPOSE DISTINGUISHABLE PEAKS ?

A. Methodology : Extraction of attention slices by unsupervised outlier detection

Attention curves mostly consist of noisy fluctuations with some high value peaks [39] [15]. A typical example is the red curve in figure 2. The first step of our methodology consists of filtering attention curves containing peaks and then extracting where attention rises (*attention slices*). In order to achieve this, we use and adapt an unsupervised outlier detection method already proposed in another study on attention [19]. We sample timescale by randomly selecting samples according to the distribution given by the attention curves. Then, points with higher attention values have a higher chance of being selected. An example of this sampling process is available in figure

2. Once this sampling is done, we use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [8], an unsupervised density based algorithm, which aims to find regions where the density of points drawn is higher. This method proved to be efficient because it manages the noisy values of attention curves, and the number and expansion of regions (duration in our special case of time series) do not need to be specified. A typical result of this algorithm is depicted by blue boxes in figure 2.

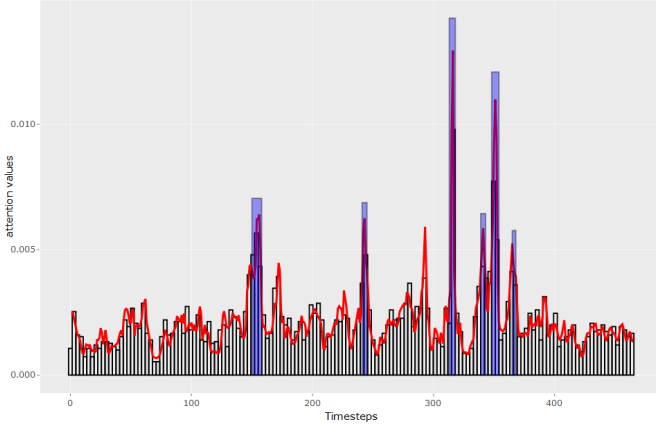


Fig. 2: Attention slice extraction.

The Attention curve is in red, the histogram of points drawn is the result of the sampling procedure, and the detected peaks are highlighted by blue boxes.

B. Results and descriptive statistics about extracted peaks

Table III provides a summary of the data serving as a basis for our study in terms of answers containing peaks. Some attention curves from candidate answers do not have peaks. In fact, some candidates may not display any particularly important moments during the answer to the first question. Figures 3 and 4 describe how long and when the peak with the largest amplitude occurs during an interview. It is interesting to note that the duration of the important slices extracted by the attention mechanism follows a very similar distribution to the duration of facial expressions which typically lasts between 0.5 and 4s [23]. Moreover, it seems they occur more often at the beginning and at the end of an answer. Such cues could indicate that non verbal behaviours occurring at the beginning (turn taking) and at the end of the answer (turn giving) have a strong impact on recruiter’s evaluation as in other face to face interactions [3], [14].

TABLE III: Descriptive table of number of answers containing peaks

Set	Percentage of answers containing peaks
Train and Validation	63.8% (3644 answers kept)
Test	57.4% (403 answers kept)

V. ARE SOCIAL SIGNALS DURING ATTENTION SLICES DIFFERENT FROM THOSE IN RANDOM SLICES?

A. Method : Supervised classification between attention slices and random slices

In this section we study the relationship between the values fed to our model and the *attention slices*. Attention values could heavily depend on the context and on the model’s memory, and depend very little on the time-frame they point to. Our model uses Contextual Attention learned on top of a bidirectional GRU. GRU is a sequence modelling component that outputs vectors that depend on the current timestamp as well as its previous output. We hypothesize that a rise in attention is due to a change of behaviour within an answer.

To ensure that attention mostly stems from what is happening in the concerned time-steps, we construct a binary classification task. For our task, we take as one class the most important *attention slices* (the slice containing the peak with the highest amplitude) extracted in each candidate answer. As the other class, we take four moments with the same duration sampled in the candidate’s answer according to a distribution proportional to $(1 - attention_values)$. *attention_values* is the average of the output values of the attention mechanisms of the candidate’s answer. Through this sampling, we aim to select moments of varying importance, and not only the most unimportant ones, while still avoiding the most important moments. As our goal is to understand if *attention slices* are different, we decided to use traditional classifiers with which a methodology to detect important features is well established. Thus, classifiers we used for this task are Lasso (linear and transparent model) and Random Forest (non linear model). As these classifiers take as input fixed vector, we summarize the features of the selected moments’ time-windows through the use of the following functions: mean, mean of positive gradients and mean of negative gradients.

We use these functions here as an attempt at capturing temporal dependencies while keeping an explainable set of features. As the attention mechanisms are trained on top of GRUs, they capture temporal variations. Our gradient functions have been used successfully in a previous behavioral classification work [31]. These functions are applied on the same feature set used in subsection III-B to train HireNet, the unique difference remains on a preprocess step of Z-normalization regarding the whole answer. For this experiment, we keep the same training, development, and test sets as we did before in [15] to prevent any sort of data leaks.

B. Results and analysis

TABLE IV: Classification results between random slices and *attention slices*

Model	F1 Positive Class	F1 Negative Class	Mean F1
Random Baseline	0.286	0.614	0.450
Majority Class	0	0.888	0.444
Lasso	0.812	0.955	0.884
Random Forest	0.760	0.945	0.852

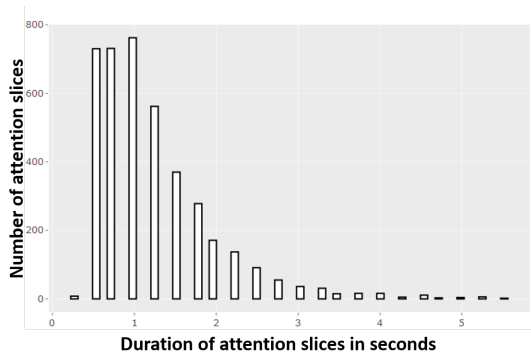


Fig. 3: Histogram of the duration of *attention slices*

As shown in Table IV the classifier’s performance is significantly above the random baseline, proving that despite the influence of sequence modelling and the use of context information, the importance of a moment is still mainly defined by the events occurring in it. It shows that specific moment where peaks of attentions occur are distinguishable from others slices of the same answer.

C. Non verbal features importance analysis

An experiment about features’ importance has also been done in order to highlight the features that contribute the most when identifying *attention slices*. Such analysis provides useful knowledge about what lets a slice be selected. In order to obtain this feature analysis, we inspect coefficients of Lasso Model and rank them according to their magnitude. Concerning, Random Forest importance, we run a permutation importance analysis through the use of Boruta Package [20]. The result table for top twenty features of both methods is available in table V. Blinking (AU45), lip stretcher (AU20), jaw drop (AU26) and lip tightener (AU23) are considered by both analyses to be the top 4 features with the most importance. Based on the sign of the coefficients of the Lasso model, we notice that *attention slices* are induced by: i) Eyes closed longer than usual, ii) The activation of lip stretcher and lip tightener, iii) The non-activation of jaw drop. These cues could indicate that moments when a candidate is not talking (absence of jaw drop) or when he/she displays social cues of anxiety (lip stretcher and lip tightener) are considered more important by the attention mechanism [10]. Frames with chin raiser (AU17) are also considered important. Also, outer brow raiser (AU2) and brow lowerer (AU4) appear in both features importance analyses. Coefficients of Lasso and directions of gradients support that moments when candidates raise and keep outer brow raised are also judged more important. Another interesting features is the use of the depth position (T_z). This analysis seems to indicate that movements back and forth could also be detected as important moments. Finally it’s interesting to highlight the confidence of OpenFace and negative gradients of OpenFace’s confidence are selected by Random Forest permutation analysis.

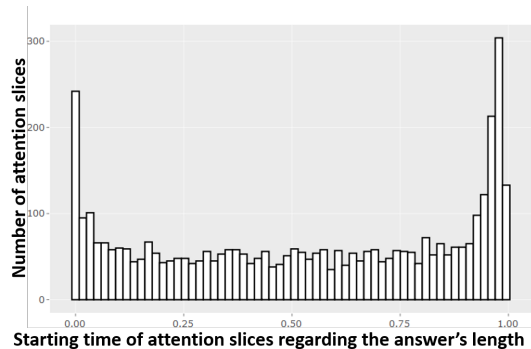


Fig. 4: Histogram of the starting time of *attention slices* relative to the total duration of the answer

VI. ARE ATTENTION SLICES MORE INFORMATIVE WITH REGARD TO HIRABILITY THAN RANDOM SLICES ?

A. Method : Supervised classification of hirability based on random slices or attention slices

We intend to evaluate that the moments highlighted by the attention mechanism carry more useful information than random moments. The following procedure aims at testing that the highlighted moments have superior predictive capabilities compared to the rest of the interview.

We constructed a classification task based on only one slice of the candidate’s answer. We ran two instances of this task: The first one uses the most important moment for each candidate as judged by the attention mechanism, while the second one uses the same sampling as in section V-A. We used the same features as in section V-A. As the input slices are different from the ones required by our model (classification of structured asynchronous video interview), we choose to experiment with non-sequential algorithms. We divide our algorithms into 3 sets. The first set is only composed of Lasso, an L1 regularized linear classifier trained with the same loss as HireNet: a binary cross-entropy. This first set has processing capabilities inferior to those of HireNet. In fact, inspite of their complexity, the GRUs composing our hierachical model process each input with only one non-linearity. Consequently, it is capable of drawing 3 linear separators: 1 for each hierarchy level, and 1 for the final dense layer, and of adding up sequential elements in a learnable fashion. The second set is comprised of only linear-SVM. It is an algorithm with strictly inferior processing capabilities compared to HireNet, trained with a different loss function. The third set comprises SVM with a Radial Basis Function kernel (RBF) and Random Forest, two algorithms with processing capabilities unavailable in HireNet, and with loss functions different than that of HireNet. We choose the Area Under the Curve (AUC) as evaluation metric, as it has the advantage of not requiring any threshold and it is suited to comparing different models. For each of the algorithms used, we performed a bootstrapping procedure as follows; we trained 100 instances, each on a subset of the training set sampled with replacement. We then obtained a set of scores that allowed us to calculate confidence

TABLE V: Feature Importance Analysis

Group	Lasso		Random Forest Permutation
	Positive coefficients	Negative coefficients	
Lower Face	AU20² , AU23⁴ , AU17¹⁰	AU26³ , AU12⁶ , \uparrow AU17¹¹ , \uparrow AU20¹⁶ , \downarrow AU10¹⁹ , AU25²⁰	AU20² , AU26³ , AU23⁴ , AU17⁵ , AU25⁶ , \uparrow AU23⁹ , \uparrow AU17¹⁴ , AU15¹⁵ , AU12¹⁶ , AU14²⁰
Upper Face	\uparrow AU7⁵ , AU2⁸ , AU4¹² , \uparrow AU2¹⁴ , AU1¹⁸	\downarrow AU2¹³ , \downarrow AU7¹⁷	AU4⁷ , AU2¹⁷
Blink and Gaze	AU45¹	\downarrow AU45¹⁵	AU45¹ , \uparrow AU45¹² , gaze_angle _y ¹⁸
Position and rotation of the head	\uparrow T_z⁷ , \downarrow T_z⁹		T_z¹⁰ , T_y¹¹ , R_x¹³
Confidence of OpenFace			confidence⁸ , \downarrow confidence¹⁹

F^i denotes that feature F is ranked in i_{th} position. \uparrow and \downarrow stand respectively for mean of positive and negative gradients. Bold indicates that feature F is significantly different from random slices on the test set based on two tailed t-test $p < 0.0001$

TABLE VI: Result of classification task for hirability prediction from *attention slices* and random slices

Thin slices integrated	AUC			
	<i>Random Forest</i> *	<i>Lasso</i> *	<i>SVM Linear</i> *	<i>SVM RBF</i> *
Random thin slices	0.545 \pm 0.005	0.517 \pm 0.005	0.518 \pm .005	0.528 \pm 0.005
<i>Attention slices</i>	0.554 \pm 0.003	0.550 \pm 0.003	0.543 \pm 0.004	0.537 \pm 0.003

statistical significance is based on two-tailed t-test * $p < 0.01$

intervals for our results to get a sense of their statistical significance.

B. Results and discussion.

Results are reported in Table VI. We observed only based on short slices of 0.5s to 4s, that the prediction of hirability is above random. For all the classifiers, the results show statistically significant differences in the predictive performance of the *attention slices* in comparison to random slices. We can note that the importance of the use of attention slices is clearer (larger performance gap) for linear classifiers compared to non linear classifiers. This highlights an important consideration to keep in mind when using attention mechanisms: The *attention slices* are selected with regard to the learner they are fed to. As shown by our results, processing capabilities vary the obtained importance of thin slices.

VII. CONCLUSION, LIMITS AND FUTURE WORKS

In this paper, we established that moments with peaks of attention are different from randomly picked slices. We described this difference in terms of the input visual social signals. Visual cues seem to relate with anxiety (activation of lip stretcher and lip tightener), blinking and pauses (non-activation of jaw drop). *Attention slices* are more likely to occur during turn taking (at the beginning of the answer) and turn giving (at the end of the answer) as in real face to face interaction [3], [14]. We also study the predictive value of the selected moments in comparison to random moments, and consequently put into perspective the use of the expression "important moment" to qualify an interview slice. In future work, we would like to investigate different types of attention mechanisms [22] in our hirability prediction model, and a larger range of classifiers for the study of *attention slices*. Attention values highlight the importance of moments, but do not include information about whether they have a positive or

negative impact on recruiters' decisions. We aim in our future work to distinguish attention slices that have a positive impact from ones having a negative impact. We plan to expand our work to other questions, modalities and the use of more than the most important peaks. Finally, as our approach is based on a learned model (eg HireNet), one research direction is to improve it in terms of performance and bias control. Next steps of our work will also be dedicated to the design of a procedure to quantify the proportion of important moments. In that sense, we plan to conduct an annotation task and a user study, in order to: *i*) quantify the aforementioned proportion; *ii*) study links between macro cues and micro cues under the light of *attention slices* spotted by our model; *iii*) build an interface that provides useful feedback for candidates, and higher decision transparency for recruiters.

ACKNOWLEDGMENT

This work was supported by the company EASYRECRUE. We would like to thank Jeremy Langlais and Amandine Reitz for their support and their help. We would also like to thank Erin Douglas for proofreading the article.

REFERENCES

- [1] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 59–66, 2018.
- [2] Dana R. Carney, C. Randall Colvin, and Judith A. Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072, 10 2007.
- [3] Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, pages 114–123, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

- [4] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 504–509. IEEE, 10 2017.
- [5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [6] Mathieu Chollet and Stefan Scherer. Assessing Public Speaking Ability from Thin Slices of Behavior. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASLAGUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pages 310–316, 2017.
- [7] Timothy Degroot and Janaki Gooty. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology*, 24(2):179–192, 2009.
- [8] Martin Ester, Hans-peter Kriegel, Xiaowei Xu, and D Miinchen. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 1996.
- [9] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.
- [10] Amanda R. Feiler and Deborah M. Powell. Behavioral Expression of Job Interview Anxiety. *Journal of Business and Psychology*, 31(1):155–171, 2016.
- [11] Ray J. Forbes and Paul R. Jackson. Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, 53(1):65–72, 1980.
- [12] Denise Frauendorfer and Marianne Schmid Mast. The Impact of Nonverbal Behavior in the Job Interview. In *The Social Psychology of Nonverbal Communication*, pages 220–247. Palgrave Macmillan UK, London, 2015.
- [13] Robert Gifford, Cheuk Fan Ng, and Margaret Wilkinson. Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, 70(4):729–736, 1985.
- [14] Wells Goodrich. Face-to-Face Interaction: Research, Methods, and Theory. *Family Process*, 18(3):355–356, 9 1979.
- [15] Leo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-claude Martin, and Chloe Clavel. HireNet : a Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. In *AAAI*, 2019.
- [16] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. MACH. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*, page 697, New York, New York, USA, 2013. ACM Press.
- [17] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *North American Chapter of the Association for Computational Linguistics*, 2019.
- [18] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip H S Torr. Learn To Pay Attention. In *International Conference on Learning Representations*, pages 1–14, 4 2018.
- [19] Jinkyu Kim and John Canny. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.
- [20] Miron B Kursa. Feature Selection with the Boruta Package. *Journal of Statistics*, 36(11):1–13, 2010.
- [21] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Improving the interpretability of deep neural networks with knowledge distillation. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2018-Novem:905–912, 2019.
- [22] André F. T. Martins and Ramón Fernández Astudillo. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In *International Conference on Machine Learning*, 2016.
- [23] David Matsumoto and Hyi Sung Hwang. Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, 35(2):181–191, 2011.
- [24] Nora A. Murphy, Judith A. Hall, Marianne Schmid Mast, Mollie A. Ruben, Denise Frauendorfer, Danielle Blanch-Hartigan, Debra L. Roter, and Laurent Nguyen. Reliability and Validity of Nonverbal Thin Slices in Social Interactions. *Personality and Social Psychology Bulletin*, 41(2):199–213, 2 2015.
- [25] Iftekhar Naim, Md. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing*, 9(2):191–204, 4 2018.
- [26] Shruthi Kukal Nambiar, Rahul Das, Sowmya Rasipuram, and Dinesh Babu Jayagopi. Automatic generation of actionable feedback towards improving social competency in job interviews. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education - MIE 2017*, pages 53–59, New York, New York, USA, 2017. ACM Press.
- [27] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia*, 16(4):1018–1031, 6 2014.
- [28] Laurent Son Nguyen and Daniel Gatica-Perez. I Would Hire You in a Minute. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pages 51–58, New York, New York, USA, 2015. ACM Press.
- [29] Pooja Rao S. B, Sowmya Rasipuram, Rahul Das, and Dinesh Babu Jayagopi. Automatic assessment of communication skill in non-conventional interview settings: a comparative study. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*, number November, pages 221–229, New York, New York, USA, 2017. ACM Press.
- [30] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics.
- [31] M. S. Ryoo, Brandon Rothrock, and Larry Matthies. Pooled motion features for first-person videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June(Figure 1):896–904, 2015.
- [32] Neal Schmitt, editor. *The Oxford Handbook of Personnel Assessment and Selection*. Oxford University Press, 3 2012.
- [33] Leann Schneider, Deborah M. Powell, and Nicolas Roulin. Cues to deception in the employment interview. *International Journal of Selection and Assessment*, 23(2):182–190, 2015.
- [34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, 2017.
- [35] Brian W. Swider, Murray R. Barrick, and T. Brad Harris. Initial impressions: What they are, what they are not, and how they influence structured interview outcomes. *Journal of Applied Psychology*, 101(5):625–638, 2016.
- [36] Giovanna Varni, Isabelle Hupont, Chloe Clavel, and Mohamed Chetouani. Computational Study of Primitive Emotional Contagion in Dyadic Interactions. *IEEE Transactions on Affective Computing*, 3045(c):1–1, 2018.
- [37] Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. Multimodal Public Speaking Performance Assessment. *International Conference on Multimodal Interaction (ICMI)*, (March 2016):43–50, 2015.
- [38] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding Neural Networks Through Deep Visualization. In *In ICML Workshop on Deep Learning*, 6 2015.
- [39] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content. *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1743–1751, 2017.