UDC 004.5 **DOI: 10.18413/2518-1092-2020-5-4-0-6**

Sorokina S.A.
Soma G.M.

INTELLIGENT ASSISTANCE IN ONLINE INTERVIEWING. EMOTIONAL ROUTING METHOD

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics,
49 Kronverkskiy prospekt, St. Petersburg, 197101, Russia

e-mail: ssofia.sorokina.12@gmail.com, guedes.soma@mail.ru

Annotation
In the era of the coronavirus pandemic, traditional human communication has undergone several significant changes. Most of the usual offline activities went online, which made it necessary to adapt to a completely new environment. Not spared the transition to the online sphere of small and medium-sized businesses. Companies are transferring interviews and negotiations to video conferencing systems, where the ease of perception of the multimodality of information flows is lost. In this paper, methods of helping the interviewer with an online interview were studied and a fundamentally new one was proposed – the method of emotional routing. Emotional routing includes the analysis of the audio stream of speech (intonation and semantics), the video channel (facial expression, look, posture, gestures), as well as the analysis of the context of changes in emotions over time. Based on an intellectual analysis of the context in which psycho-emotional changes in a person's state occurred, the method of emotional routing predicts the success of the outcome of dialogue, determined by parameters interactively set by the user.
Keywords: video-interviewing; psycho-emotional analysis; emotional routing; machine learning.

УДК 004.5

Сорокина С.А.
Сома Г.М.

ИНТЕЛЛЕКТУАЛЬНАЯ ПОМОЩЬ В ОНЛАЙН-ИНТЕРВЬЮИРОВАНИИ. МЕТОД ЭМОЦИОНАЛЬНОЙ МАРШРУТИЗАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Кронверкский пр., д. 49, г. Санкт-Петербург, 197101, Россия

e-mail: ssofia.sorokina.12@gmail.com, guedes.soma@mail.ru

Аннотация
В эпоху пандемии коронавируса традиционное человеческое общение претерпело несколько существенных изменений. Большинство привычных офлайн-активностей были вынуждены трансформироваться в онлайн-формат, что потребовало адаптации к совершенно новой среде. Не обошел стороной переход в онлайн-сферу малый и средний бизнес. Компании переносят интервью и переговоры в системы видеоконференцсвязи, где теряется легкость восприятия мультимодальности информационных потоков. В данной работе были изучены методы помощи интервьюеру в проведении онлайн-интервью и предложен принципиально новый метод – метод эмоциональной маршрутизации. Эмоциональная маршрутизация включает анализ аудиопотока речи (интонации и семантики), видеоканала (выражение лица, взгляд, поза, жесты), а также анализ контекста изменения эмоций с течением времени. На основе интеллектуального анализа контекста, в котором произошли психоэмоциональные изменения в состоянии человека, метод эмоциональной маршрутизации предсказывает успех результата диалога, определяемого параметрами, интерактивно задаваемыми пользователем.
Ключевые слова: видеоинтервью; анализ психоэмоционального состояния; эмоциональная маршрутизация; машинное обучение.

INTRODUCTION

During traditional interpersonal communication, people almost always interact multimodally using verbal and non-verbal channels [1]. Moreover, verbal and non-verbal communication is almost always tightly connected: in real monologues, dialogues and polylogues people combine these two parts in a single whole [2]. However, under the conditions observed in the post-COVID-19 era growth in the volume and intensity of the use of technical means of communication, the latter is functionally limited. Current online conditions do not allow users to regulate the direction of the communicative process, create psychological contact with the interlocutor, enrich the information transmitted by verbal means, guide the interpretation of the verbal text, unambiguously express emotions and reflect the interpretation of the situation they are facing with.

The resulting contradiction is due to several factors. The objective factors are poor quality of the provided communication channels (low data rate) and unfavorable acoustic environment in which the technical means of communication is used. The subjective factors are the transformation of formal-role communication into business communication, in which, along with the exchange of information, the characteristics of the subscriber's personality, their mood, physiological and psycho-emotional states must be taken into account; an increased rate of change in the situation and an increase in the volume of transmitted information, requiring subscribers to increase the effectiveness of their actions.

The presence of these contradictions leads to a decrease in the effectiveness of interpersonal communication – an increase in the time to achieve the goals of such communication, and therefore, to resolve them, it is necessary to develop means (mathematical, methodological, and software) that provide an increase in such efficiency due to the correct interpretation and consideration of the non-verbal component of interpersonal communications.

MATERIALS AND METHODS

Today, there are ways to detect emotions from video channels [3, 4], services that recognize spoken speech and sounds [5, 6], as well as written text analyzers [7, 8]. Moreover, there are methods that allow people to recognize the psycho-emotional state of a person based on a multimodality: some of them combine video and audio, mostly in videoclips [9, 10], others work with acoustic and text data [11, 12]. However, most of the currently available methods have a high computational load, which leads to long-time runs. If someone is working with a high-quality image, the speed of its analysis, even on powerful computers, is not high enough to work with the current video stream in real-time mode. Considering that the target audience to deal with the online limitations in the post-COVID era consists mostly of small and medium business owners, it is impossible to implement the existing methods of multimodal analysis.

The most obvious way to deal with the speed limitations in the case of video streams is reducing the image quality. However, that option could be considered only if the process is running on a pre-recorded data, and not in real-time mode. Another disadvantage of this approach is the loss of analysis quality. Such defects could be crucial, especially since the result of the analysis should be just the same selection of the factors that will not be noticed by the naked eye of a person.

The current methods of analyzing audio data flow are not perfect either. Speech recognition technologies are at a high level, and the current results help us analyze the semantic part of speech [13]. However, the intonation components have not yet been covered properly. Since oral speech consists of two main modules – sounding (including intonations) and semantic [14], only in a complex analysis would it be possible to distinguish the emotional state of speech. In real-life communications people analyze not only the external emotions expressed on the face, but also the internal mood of a person: analyze speech for the presence of passive aggression, identify sarcastic and/or ironic expressions, and correlate non-verbal signs (knocking on the table, frequent changes of poses, etc.) with emotional manifestations of the person.

However, the highest limitation of the existing methods is that they only help to observe a current emotion: no attention is paid to the context of its appearance. In daily life people always face the

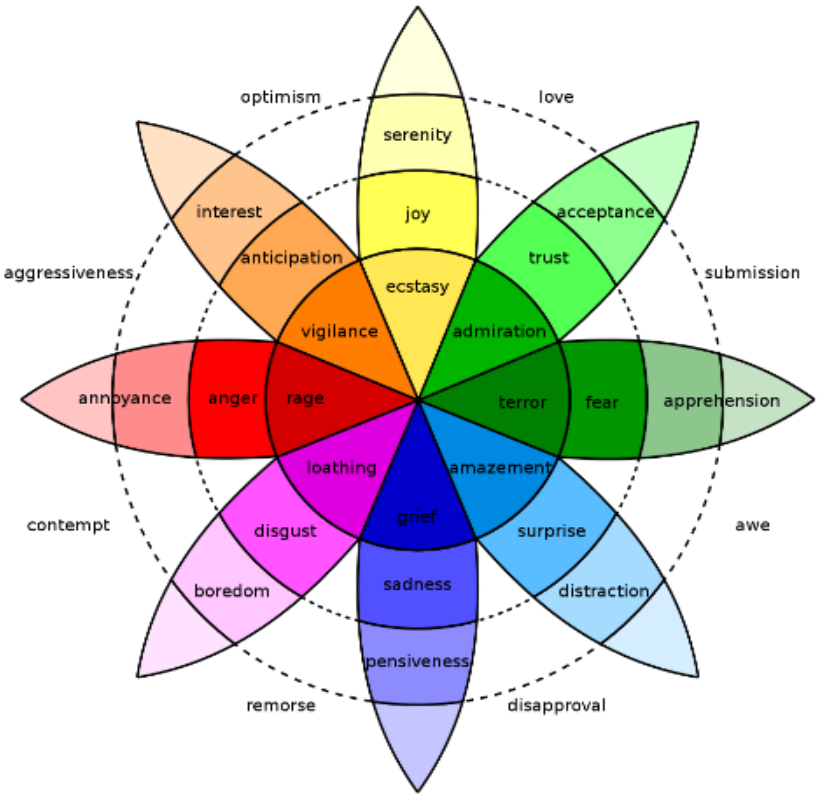
situations when it is more important to analyze not what emotion the person expressed, but what provoked it. That is why the method of emotional routing, which is suggested in this paper, includes the analysis of all the mentioned above characteristics: real-time data from video and acoustic channels, recording of reactions (emotions) during the whole communication on a timescale, as well as the interpretation of changes in emotional state and prediction of the successful outcome of the dialogue.

As the suggested method is not only related to computer science and computations but also to psychology, high attention in the research was paid to the ways of grading and classifying the emotions. In that part of the study two fundamental ways of emotional clustering are presented.

GRADATION OF EMOTIONS

The psycho-emotional state of a person is a multilevel characteristic, the dominant role in the formation of which is played by the main emotion of reaction to a specific event [15]. Gradation of emotion is the basis of building a qualified emotional routing. The first concept which is proposed to use is Plutchik's wheel of emotions, which is a fundamental theory in psychology of emotions proposed in 1980 [16]. Plutchik's wheel became the basis for a significant amount of further research on emotional levels and classes.

In the suggested emotional routing method, the users have the right to choose the exact trajectories they want to see and analyze, as well as combine the chosen trajectories into secondary and tertiary connections. All the possible emotions and their combinations are presented in Figure 1.



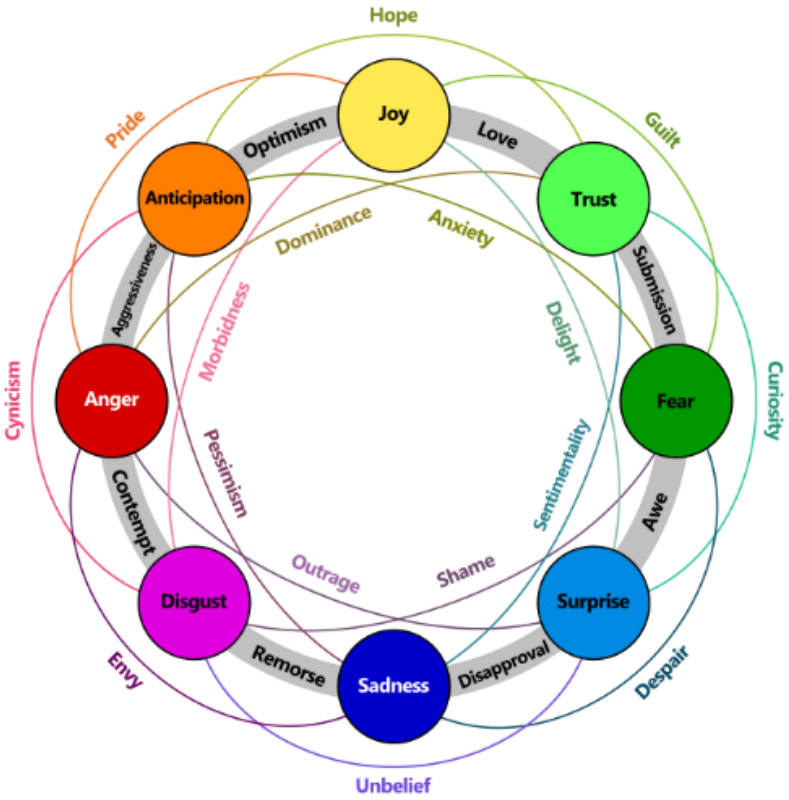


Fig. 1. Plutchik's wheel of emotions. 'Basic' emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) are at the left. At the right primary, secondary and tertiary connections between the basic emotions can be observed [17, 18]

The second theory which is implemented in the research introduces 6 opposite axes concepts. It helps to classify the behavior (as well as the psycho-emotional state) of a person based on an assessment from -1.0 to +1.0 in six different directions which are presented in Table [19]. In that case of 6 opposite axes concepts, users also have the rights to choose the exact axes to analyze and exclude some parts which play no important role for their interviewing sessions.

Table

Six opposite axes of emotions [19]

Axis	-1.0	-0.5	0	0	+0.5	+1.0
Anxiety/Confidence	Anxiety	Worry	Discomfort	Comfort	Hopeful	Confident
Boredom/Fascination	Ennui	Boredom	Indifference	Interest	Curiosity	Intrigue
Frustration/Euphoria	Frustration	Puzzlement	Confusion	Insight	Enlightenment	Epiphany
Dispirited/Encouraged	Dispirited	Disappointed	Dissatisfied	Satisfied	Thrilled	Enthusiastic
Terror/Enchantment	Terror	Dread	Apprehension	Calm	Anticipatory	Excited
Humiliation/Pride	Humiliated	Embarrassed	Self-conscious	Pleased	Satisfied	Proud

EMOTIONAL ROUTING: RESULTS AND INTERPRETATION

The main part of emotional routing and its advantage relates to the recording of the emotional changes in time with respect to the context of the discussion. The changes in the psycho-emotional state can be fixed using the video stream and acoustic channel (both flows are reachable through the systems of videoconference systems). In this paper we would not stop on the technical aspects of the emotional recognition in detail, more attention will be paid to the routing itself.

During the real-time analysis, emotional routing records all the reactions of the person in time. While the video conference is running, the user (interviewer) is asking questions, which are also recorded using the speech-recognition modules. The system is also taught to recognize the answers of the interlocutor. When the changes in the emotional state of interviewee occur, the module automatically matches the question of the interviewer with the interlocutor's reaction.

The step before the interpretation of the results includes building the visualization (charts) of the emotional changes. The charts are both for the user and the system, what significantly helps to analyze the obtained picture of the interview. The examples of two emotional trajectories can be observed in Figure 2. Interviewers also get access to the transcription of the dialogue with the exact time, asked questions and given answers.

The final step of the emotional routing is the outcome's prediction. Based on the analyzed emotional series, ANN predicted how successful the rest of the dialogue would be. 'Successful' includes the variable parameters: acceptance, consensus, arguments, and conformity. Now, convolutional neural networks need from 15 to 20 minutes of data to predict the successful outcome with precision 0.75.

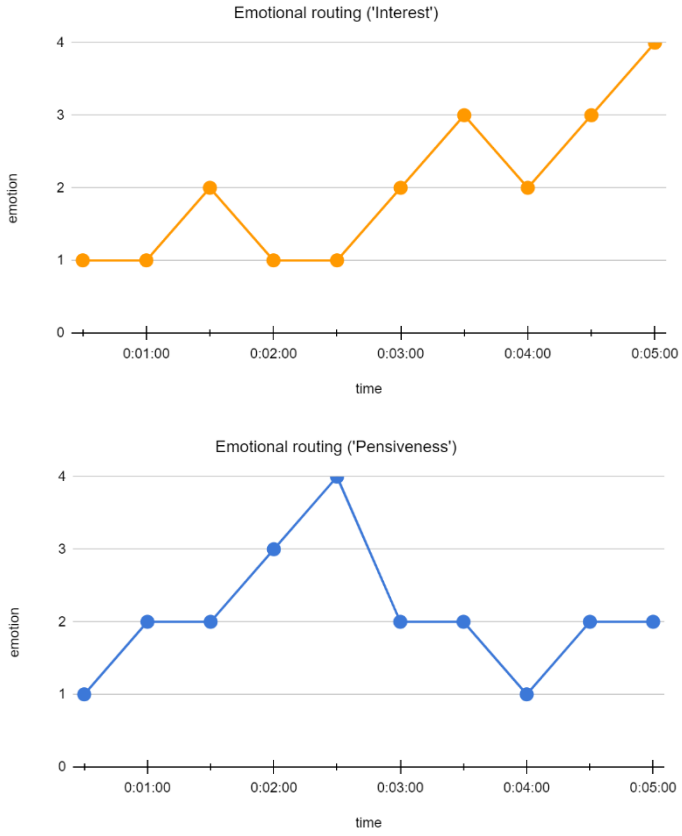


Fig. 2. Example of emotional routing based on a 5-minutes interview for the two groups of emotions (Plutchik's wheel): Interest and Pensiveness. At the top: 1 – Neutral, 2 – Interest, 3 – Anticipation, 4 – Vigilance. At the bottom: 1 – Neutral, 2 – Pensiveness, 3 – Sadness, 4 – Grief

DISCUSSION

As the results of the study, optimized algorithms for working with multimodal information, both verbal and non-verbal, were obtained. We considered two streams: video and audio.

To help predict the outcome of online interviews, we have identified a fundamentally new method – the method of emotional routing. The essence of emotional routing is not just identifying the psycho-emotional interlocutor, but also in fixing this state on the timeline. Thus, it is possible to simply see the confusion or anger of the interlocutor.

The implementation of the emotional routing method is being carried out by using video streams of conferencing systems (the main platform used during this research is Zoom), which significantly reduced the computing power required for use. For the analysis of the audio stream, the real-time mode was also applied. That allowed us to reach a high level of synchronization of playback from two streams.

Synchronization was necessary to establish correlations between sounding and visually observed emotions. Only when these two aspects are combined the results could be used to predict the success of the interview outcome. If this principle is not observed, the analysis of the psycho-emotional state becomes possible only one-sidedly: for example, if the system pays attention exclusively to tapping on the table, but does not read emotions from the interlocutor's face, its conclusion may be very far from the truth.

Based on the recorded emotional route and machine learning methods, the system makes predictions about the potential success of the negotiations. The average time required to record an emotional route, depending on the intensity of the negotiations, varies from 15 to 20 minutes.

RESULTS

This research has a high theoretical and practical importance for several areas of life. Even though initially the product was conceived as an assistant for small and medium-sized business owners who are unable to maintain a full-fledged HR department, the method obtained in the course of the research can be applied in any remote negotiations based on video conferencing methods.

To further expand the scope of the proposed method for implementation, we need to optimize it in such a way that conversations of three or more people can be analyzed without significant quality losses. In that case, it is necessary to study the polylogue format in more detail and propose solutions for optimizing emotional routing in such a way that it includes the possibility of building three or more parallel routes and parallel predictive directions for each of the participants in business negotiations.

Another important aspect of the potential development of emotional routing is the addition of deeper intellectual analysis of the verbal component of communication. For example, the meaning of the speech spoken by both parties to the negotiations is not considered in the emotional routing method in its current form. However, the verbal part can provide a significant amount of necessary information.

CONCLUSION

Within the framework of this study, the existing methods of collecting and analyzing video data were studied and analyzed (mainly for the identification of an emotional state by a dynamic picture). Based on the imperfections of the available methods, requirements were drawn up to optimize the process of analyzing the state of the interlocutor. The following development vectors have been set (adding speech semantics, increasing the speed of composing emotional development), and the development of the method continues to reduce time costs and capacities in the aspect of issuing evaluative and recommendatory feedback on the success of negotiations.

References

1. Zuckerman M., DePaulo B.M., Rosenthal R. Verbal and nonverbal communication of deception // Advances in experimental social psychology. – Academic Press, 1981. – T. 14. – P. 1-59.

2. Jones S.E., LeBaron C.D. Research on the relationship between verbal and nonverbal communication: Emerging integrations // Journal of communication. – 2002. – T. 52. – №. 3. – P. 499-521.

3. Fan Y. et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks // Proceedings of the 18th ACM International Conference on Multimodal Interaction. – 2016. – P. 445-450.

4. Williams J. et al. Recognizing emotions in video using multimodal DNN feature fusion // Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML). – 2018. – P. 11-19.

5. Fayek H.M., Lech M., Cavedon L. Evaluating deep learning architectures for Speech Emotion Recognition // Neural Networks. – 2017. – T. 92. – P. 60-68.

6. Lim W., Jang D., Lee T. Speech emotion recognition using convolutional and recurrent neural networks // 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA). – IEEE, 2016. – P. 1-4.

7. Calefato F., Lanubile F., Novielli N. EmoTxt: a toolkit for emotion recognition from text // 2017 seventh international conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). – IEEE, 2017. – P. 79-80.

8. Batbaatar E., Li M., Ryu K.H. Semantic-emotion neural network for emotion recognition from text // IEEE Access. – 2019. – T. 7. – P. 111866-111878.

9. Liu C. et al. Multi-feature-based emotion recognition for video clips // Proceedings of the 20th ACM International Conference on Multimodal Interaction. – 2018. – P. 630-634.

10. Noroozi F. et al. Audio-visual emotion recognition in video clips // IEEE Transactions on Affective Computing. – 2017. – T. 10. – №. 1. – P. 60-75.

11. Yoon S., Byun S., Jung K. Multimodal speech emotion recognition using audio and text // 2018 IEEE Spoken Language Technology Workshop (SLT). – IEEE, 2018. – P. 112-118.

12. Sailunaz K. et al. Emotion detection from text and speech: a survey //Social Network Analysis and Mining. – 2018. – T. 8. – №. 1. – P. 28.

13. Vryzas N. et al. Speech emotion recognition adapted to multimodal semantic repositories // 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). – IEEE, 2018. – P. 31-35.

14. Jurafsky D. Speech & language processing. – Pearson Education India, 2000.

15. Korenevskiy N. et al. Fuzzy determination of the human's level of psycho-emotional // 4th International Conference on Biomedical Engineering in Vietnam. – Springer, Berlin, Heidelberg, 2013. – P. 213-216.

16. Plutchik R. The emotions. – University Press of America, 1991.

17. https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg#/media/File:Plutchik-wheel.svg

18. https://commons.wikimedia.org/wiki/File:Plutchik_Dyads.svg#/media/File:Plutchik_Dyads.svg

19. Kort B., Reilly R., Picard R.W. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion // Proceedings IEEE International Conference on Advanced Learning Technologies. – IEEE, 2001. – P. 43-46.

Sorokina Sofia Andreevna, student of Saint Petersburg National Research University of Information Technologies, Mechanics and Optics
Soma Guedes Manuel, post-graduate student of Saint Petersburg National Research University of Information Technologies, Mechanics and Optics

Сорокина Софья Андреевна, студент Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики
Сома Гедеш Мануэл, аспирант Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики