Multimodal Analysis of Client Persuasion in Consulting Interactions: Toward Understanding Successful Consulting

Yasushi Amari[1], Shogo Okada[1(✉)], Maiko Matsumoto[2], Kugatsu Sadamitsu[2], and Atsushi Nakamoto[2]

[1] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
{y_amari,okada-s}@jaist.ac.jp
[2] Future Corporation, Shinagawa-ku, Tokyo, Japan
{m.matsumoto.rb,k.sadamitsu.ic,a.nakamoto.kh}@future.co.jp

Abstract. To analyze successful consulting processes using multimodal analysis, the aim of this research is to develop a model for recognizing when a client is persuaded by a consultant using multimodal features. These models enable us to analyze the utterances of highly skilled professional consultants in persuading clients. For this purpose, first, we collect a multimodal counseling interaction corpus including audio and spoken dialogue content (manual transcription) on dialogue sessions between a professional beauty counselor and five clients. Second, we developed a recognition model of persuasion labels using acoustic and linguistic features that are extracted from a multimodal corpus by training a machine learning model as a binary classification task. The experimental results show that the persuasion was 0.697 for accuracy and 0.661 for F1-score with bidirectional LSTM.

Keywords: Social signal processing · Multimodal interaction · Consulting interaction · Persuasion

1 Introduction

Professional consultants need to satisfy the requirements of the clients by providing appropriate advice via interaction. In addition, sufficient educational training is required to obtain sufficient consulting skills. To support the skill training for counseling and consulting, this paper focuses on the analysis of the counseling skills of professional counselors. In recent years, many studies have focused on social signal processing [26] (SSP) for developing recognition models of physiological aspects or cognitive performance based on the integration of findings of social physiology and multimodal sensing technology. Common approaches in this research develop recognition models of skill scores annotated by third party coders or participants in the interaction using multimodal features observed in conversation. Though the advantage on the application side of the approach is

G. Meiselwitz (Ed.): HCII 2021, LNCS 12775, pp. 29–40, 2021.
https://doi.org/10.1007/978-3-030-77685-5_3

to predict skill levels and detect people with highly developed skills by using pre-trained recognition models, it is still difficult to analyze how consultants with highly developed skills persuade clients using multimodal time-series data.

To analyze the successful consulting process using multimodal analysis, the aim of this research is to develop a recognition model for when a client is persuaded by a consultant using multimodal features. These models enable us to analyze the utterances of highly skilled professional consultants that persuaded the client. For this purpose, first, we collect a multimodal counseling interaction corpus including audio and spoken dialogue content (manual transcription) on dialogue sessions between a professional beauty counselor and five clients. In addition, we asked these clients (participants) to annotate the persuasion label per utterance. Second, we developed a recognition model of persuasion labels using acoustic and linguistic features that were extracted from a multimodal corpus by training the machine learning model as a binary classification task. The main contributions of this paper are as follows:

Corpus with Client Persuasion. In this paper, we collected a dialogue corpus on beauty counseling in a real situation. The multimodal corpus with persuasion labels was annotated by clients in the counseling interaction settings and is available for analyzing how counselors persuade clients (Sect. 3).

Recognition Model of Client Persuasion. With the dataset, we focus on developing an utterance-based recognition model of client persuasion. The recognition setting of a client's persuasion is unexplored in previous research (Sect. 6.1).

Analysis of Effective Consultant Linguistic Features for Recognizing Client Persuasion. We clarify which linguistic feature of consultant is effective for recognizing when a client has been persuaded (Sect. 6.3).

2 Related Work

The objective of this research is to develop a recognition model of client persuasion to identify the trigger used by a professional consultant to persuade a client. The overall goal of this research is to extract the essence (skills) of how the consultant persuades the client. Many studies focus on automatic assessment communication skills models. The research can be classified from the viewpoint of communication situations and the types of skills. Studies that focus on communication skills in a monolog situation, including public speaking [2,17,27], and social media [16], have been carried out. Other directions for research studies include modeling communication skills in dyadic interactions, including job interview settings [14], group interactions [9,15,18], and human-computer (including robot and virtual humans) interactions [8,22,23]. As target variables related to communication skills, interpersonal communication skills [14,15,23], empathy skills [9], listening skills [22], leadership [18], and public speaking skills [2,17,27] are used. This study's objective is not to analyze the communication process of a person with highly developed skills but to develop a model to predict the skills of participants.

The role of counseling is similar to that of consultation. To develop an automatic analysis model of empathy with behavioral signal processing [28] and implementing counselor agents that are used for physiological health care, many studies have focused on analyzing the behavior of counselors. The skills for listening to patient stories with empathy are required for professional counselors. Therefore, analyzing the multimodal behavior for representing empathy is the central challenge in this research. Xiao et al. proposed a prediction model of counselor empathy measures in motivational interviewing [29]. DeVault et al. [4] presented a virtual human interviewer system designed to create an engaging face-to-face interaction where the user feels comfortable talking and sharing information. Using a corpus collected using the agent system [4,24] presented an analysis of behavioral cues that indicate an opportunity to provide an empathetic response by using a multimodal deep neural network. The common objective in many studies is to model the empathetic response and empathy in the communication process. The objective is different from that of this research because the objective of this paper is to model the persuasion of clients on consultations for face makeup.

3 Data Corpus

3.1 Dyadic Consultant-Client Dialogue Setting

To analyze consultation dialogue by a professional consultant, a female beauty consultant with professional experience of more than 60 years and who engaged in beauty consultant education participated in this experiment. 5 female participants were recruited as clients for the consultation. These participants were counseled for the first time by the consultant. Consultation dialogue data of a total of 5 sessions were recorded using wearable microphones (AmiVoice Front PM01). The average session time was 12.9 min. The objective of the consultation session was to propose the appropriate makeup methodology to clients through counseling based on the client's personality and preference. Following the consulting session, clients are wearing makeup by the consultant. In particular, the main goal of this consultation was to provide the clients with a unique and new point of view on makeup application based on the expertise of the consultant. The makeup method proposed by the counselor was not always accepted by the clients as appropriate, and the consultant had to convince the client with explanations based on evidence and past experience.

3.2 Annotation

We asked clients to annotate when they were persuaded after the consultation session by watching recorded speech data with manual transcription. In this study, to annotate the label to the accurate segment interval, utterances of the consultant and client were segmented manually, and dialogue contents were scripted by two coders. When a short pause with less than 2 s existed between

consequent utterances of the consultant, we merged these consequent utterances into a segment. Utterances of the client were also chunked as one segment in the same manner when such a short pause was observed. We defined the chunked segment as the subject of annotation (a sample for machine learning), and coders (clients) annotated whether they were persuaded by the utterances of the consultant. We define the segment (the unit of sample) as a exchange, which consists of client utterance followed by consultant utterance. The number of persuaded scene labels annotated by each client is shown in Table 1. The exchange-level annotation is available as labeled data to develop a recognition model of the persuaded scene, to analyze the linguistic and acoustic features used in the persuasion. The minimum number was 64, the maximum number was 142, the total number of persuasion labels was 223 and total number of samples was 525.

Table 1. Number of persuaded scene labels per client

Client	Convinced scene	Total	Convinced/Total
A	23	92	0.250
B	14	64	0.219
C	43	101	0.426
D	81	126	0.643
E	62	142	0.437
Total	223	525	0.425

4 Multimodal Feature Extraction

We extracted the acoustic features from audio data, which were segmented manually. We also extracted the linguistic features based on manual transcription. All feature values were normalized using the Z-score method. We summarize the details of the feature set in Table 2.

4.1 Acoustic Features

The acoustic features were extracted from the speech signal of an utterance segment. We used the audio processing tool OpenSMILE[1] to extract the acoustic features. The features of 1,582 dimensions included in the INTERSPEECH 2010 Paralinguistic Challenge feature set [19] were extracted using this tool. These features included the loudness as the normalized intensity, mel-frequency cepstral coefficients 0–14, the logarithmic power of the mel-frequency bands 0–7 (distributed over a range from 0 to 8 kHz), the 8-line spectral pair frequencies computed from 8 LPC coefficients, the envelope of the smoothed fundamental frequency contour, and the voicing probability based on the fundamental frequency.

[1] https://www.audeering.com/opensmile/.

Table 2. Summarize the details of the feature set for each modality

Modality	Features
Acoustic	PCM loudness
	MFCC [0–14]
	log Mel Freq. Band [0–7]
	LSP frequency [0–7]
	F0 by sub-harmonic sum
	F0 envelope
	Voicing probability
	Jitter local
	Jitter DDP
	Shimmer local
Linguistic	Word count per PoS type
	Number of letters in words
	Word repetition
	Content of the vocabulary
	Word sentiment
	Linguistic features of previous utterance

4.2 Linguistic Features

We extracted linguistic features from manual transcription data. The linguistic features were extracted from the transcriptions using the Japanese morphological analysis tool MeCab [12] because all of the participants were Japanese and spoke in Japanese. First, the sentence was segmented into word sets by the tool. Second, the PoS type and filler were automatically annotated to each word in the word set. The same linguistic feature set was extracted for the consultant and clients. The linguistic features are as follows:

Word Count Per PoS Type: We counted the spoken words for each grammatical construction or PoS: “noun”, “proper noun”, “verb”, “conjunction”, “adjective” and “interjection”. In addition, the word “filler” was counted. The number of features was 7.

Number of Letters in Words: To capture the length of the spoken words, we counted the characters (called “Hiragana”) in words. The number of characters captures how many words with many characters participants spoke.

Word Repetition: Word repetition is effective for delivering important information. When a word appeared in an utterance more than once during the session, the number of occurrences of the part-of-speech for that word was counted. The parts of speech counted were nouns, verbs, and adjectives. The number of features was 3.

Content of the Vocabulary: We extracted the features based on a word-embedding method to analyze the content of the vocabulary in the consultation. The word-embedding model was trained on the Japanese Wikipedia corpus [21]. The feature extraction procedure was as follows: Vector V_{w_i} in the (embedded) vector space of word w_i by using the model was summed for all words in an utterance, where symbols, particles, and auxiliary verbs were removed from the word set. The average and variance in each element in V for all spoken utterances were calculated as features. The dimension of the vector space was set to 200, and thus, the number of features was 400 (mean and variance of 200 elements in a vector).

Word Sentiment: In consulting sessions, consultants often advise with positive language expression. From this background, the sentiment of words is an effective feature for capturing the positive or negative words used in a consultation. For this purpose, we used the Japanese Sentiment Polarity Dictionary [6,11] of words annotated with their semantic orientation to analyze the sentiment orientation of words in spoken dialogue. We extracted the word sentiment features as follows: We counted the number of word w_i in the utterance, which was found in the dictionary word set by matching word w_i with the word set. Let the number of words with a positive orientation be N_{pos} and that with a negative orientation be N_{neg}, the sentiment feature l_s was calculated as $(N_{pos} - N_{neg})/(N_{pos} + N_{neg})$. If words in the dictionary were not included in utterance, $l_s = 0$. In addition, N_{pos} and N_{neg} are also used as features.

Linguistic Features of the Previous Utterance: We extracted all feature sets except the word2vec feature from the previous utterance $(t-1)$ as the feature set of the current utterance (t) considering the time-series dependency of the spoken utterance.

5 Experiment

Three research questions are addressed in order to analyze multimodal features that were effective in recognizing whether a client was persuaded during counseling.

[**RQ1:**] Which features of consultant or client are effective to improve the recognition performance?
[**RQ2:**] Does temporal context in dialogue contribute to estimate the persuasion label?
[**RQ3:**] Which linguistic features of consultant contribute to detect client persuasion?

5.1 Preprocessing and Evaluation

To preprocess for classification modeling, we reduced the number of dimensions of features using PCA. PCA was performed for specific feature groups with a high

dimensional vector. The specific feature groups were composed of (i) acoustic features, (ii) the mean of word2vec, and (iii) the variance in word2vec. PCA was performed for a total of six feature groups (each feature (i)–(iii) extracted from the consultant and clients). We set a cumulative contribution rate of 95% as the threshold for dimension reduction. After dimension reduction, the number of acoustic features was 232 for counselor and 248 for clients. The number of word2vec features was 183 for counselor and 110 for clients. In addition, the feature selection with the Kolmogorov-Smirnov (KS) test was conducted with a significance level of 5% in all datasets. To evaluate the model, cross-validation testing by leave-one-client-session-out was conducted. We report the accuracy and F1-score of the test results as classification accuracy.

5.2 Machine Learning Models

We used a linear support vector machine (SVM) [3] model as the base model to compare the contributions of multimodal features or features from both participants. To explore an accurate model with high classification accuracy and explore how the temporal context is effective for this task, the linear SVM model was compared with random forest (RF) [1], XGBoost [25], and Feed forward deep neural network (DNN), long-short term memory (LSTM) [7] and bidirectional LSTM (BLSTM) [5]. The hyper parameters of SVM, RF and XGBoost were optimized using a 4-fold cross-validation (CV) scheme in training dataset. The detail of hyper parameter tuning is given in Appendix A.

DNN was composed of feed-forward neural network with multiple fully connected layers and with dropout [20] (rate = 0.5) after each layer. The network architecture is composed of two middle layers, 128 units per layer. We set the batch size to 64 and the number of epochs to 50. For implementation of LSTM and BLSTM, a single LSTM hidden layer with 128 units is used to extract feature from the sequence input data with T exchanges, followed by a dropout [20] (rate = 0.3). The LSTM layer was followed by a fully connected layer to learn the LSTM output, followed by two fully connected layer. A output layer is used for estimating labels. We set the batch size to 128 and the number of epochs to 100. And set the learning rate to 0.004. For DNN, LSTM and BLSTM, binary cross-entropy function is used as a loss function, ReLU [13] for the activation function, and Adam [10] is used as the optimization. The learning rate of Adam is set to 0.001 for DNN and 0.004 for LSTM and BLSTM.

6 Experiment Result

We compared the classification performance of the multimodal features and analyzed the contribution of each modality in recognizing whether a client is persuaded during counseling. “Acoustic” and “linguistic” are represented as A and L, respectively.

6.1 Classification Results

Table 3 shows the binary classification accuracy and F1-score of the SVM trained with the multimodal feature set of both participants. All results with each modal feature per participant (consultant or client) were better than that of the majority baseline. In Table 3, the best accuracy and F1-score for models that were trained using the consultant feature set of 0.670 and 0.575 using acoustic features, respectively. The best accuracy and F1-score for models with the client feature set were 0.653 and 0.597 using multimodal features, respectively. The best accuracy and F1-score of the models with the feature set of consultant and client were 0.670 and 0.608, respectively, which were the best accuracy and F1 score in all models.

Comparing the accuracy between unimodal models, the acoustic features were more effective than linguistic features on both accuracy and F1-score measures in all experiments. This result shows that the acoustic features were more important than the linguistic features to this task. Comparing the accuracy between the unimodal model and multimodal model, multimodal fusion improved the best unimodal accuracy when features from clients were used (0.634 to 0.653) and when features from both participants were used (0.651 to 0.670), respectively. Conversely, the multimodal fusion did not improve the accuracy when features from the counselor were used, and the acoustic model obtained the best accuracy. Comparing the accuracy between the consultant model and client model in unimodal models, better accuracy and F1-score were obtained by the consultant models than those obtained by the clients in both models with acoustic and linguistic feature sets. These results show that although the persuasion label was annotated by the clients, multimodal features observed from the consultant were effective in recognizing the client's persuasion.

Answer to RQ1: The feature set of consultant was more effective than the client. In addition, the recognition performance was improved by fusing the consultant and client feature sets.

Table 3. Binary classification results with the multimodal feature set of counselor and clients. The majority baseline for accuracy was 0.575.

Modality	Consultant			Client			Consultant + Client		
	A	L	A+L	A	L	A+L	A	L	A+L
Accuracy	**0.670**	0.596	0.650	0.634	0.577	0.653	0.651	0.575	**0.670**
F1-score	0.575	0.485	0.551	0.525	0.476	0.597	0.565	0.506	**0.608**

6.2 Comparison of Machine Learning Models

Table 4 shows the binary classification accuracy and F1-score of the SVM, random forest, XGBoost, DNN, LSTM and BLSTM trained with the multimodal feature set that is extracted from both the consultant and client. LSTM and

BLSTM are trained with time-seres features of T exchanges which is defined in Sect. 3.2. The comparison among models enable us to analyze appropriate model for this task and the importance of time-series features (temporal context) to classify the client's persuasion.

Table 4. Binary classification results of SVM, random forest, XGBoost, DNN, LSTM and BLSTM.

Classifier	SVM	Random forest	XGBoost	DNN	LSTM	BLSTM		
					$T = 6$	$T = 4$	$T = 5$	$T = 6$
Accuracy	0.670	0.532	0.686	0.691	0.594	0.651	0.690	**0.697**
F1-score	0.608	0.450	0.604	0.623	0.532	0.642	**0.661**	0.654

Table 4 shows that the best accuracy was 0.697, which is obtained by BLSTM with six exchanges ($T = 6$) and second best accuracy was obtained by DNN (0.691). The best F1-score of persuasion label was 0.661, which was obtained by BLSTM with $T = 5$. It means that deep neural network techniques promise to improve the classification performance. BLSTM with $T = 6$ improved the accuracy of LSTM with $T = 6$ by 0.103, so it is found that using both past and future temporal contexts in dialogue is effective to classify the persuasion label. On the comparison of F1 score between DNN and BLSTM, BLSTM which is trained with temporal contexts ($T = [4, 5, 6]$) improved the F1-score of DNN by 0.019, 0.038 and 0.031, respectively.

Answer to RQ2: The temporal context in advice of the consultant is a key factor whether clients are persuaded.

6.3 Linguistic Feature Analysis of Consultant

To analyze how linguistic features are observed from professional consultant on persuading the client, BLSTM ($T = 6$), which has the highest accuracy in Sect. 6.2 was trained by removing a linguistic feature group of consultant one by one. The groups are linguistic features are "Word count per PoS type", "Number of letters in words", "Word repetition", "Content of the vocabulary", and "Word sentiment". If the evaluation metrics consisting of accuracy and F1-score degraded, the removed feature group was effective for the classification. In contrast, if the evaluation metrics improved, the removed feature group was not effective for classification. Table 5 shows the accuracy and F1-score of the BLSTM trained with linguistic feature of consultant, after each feature group was excluded. "Diff" denotes the difference in accuracy and F1-score between when all feature groups were used and when one feature group was removed. In Table 5, the feature group of the "Content of the vocabulary" was the most effective (accuracy: +0.118, F1-score: +0.088) in recognizing the label. In addition, the feature groups that contributed to both accuracy and F1-score were "Number of letters in words", "Content of the vocabulary" and "Word sentiment".

Moreover, to analyze features in these two groups, we perform the Mann-Whitney U test on "Number of letters in words" and positive words count as a representative of "Word sentiment". As a result, the p-value for "Number of letters in words" was 0.0007 and the p-value for positive words count was 0.0010. The mean value of "Number of letters in words" for the "no persuaded" and "persuaded" was 41.5 and 47.5, respectively. That of positive words count was 0.9 and 1.3, respectively. The results show that clients are more likely to be persuaded when the values of the number of characters in a utterance and positive words count are high. This indicates that clients tend to be persuaded when the consultant often use many words or long words and words with positive polarity. The box plot for these two features are shown in Appendix B.

Answer to RQ3: "Content of the vocabulary" was the most effective features to detect client persuasion. The amount of words and the polarity of words were also important to persuade the client.

Table 5. Contribution of each linguistic feature group for persuasion label classification.

	Accuracy		F1-score	
Consalutant L	0.602		0.481	
Removed features	Accuracy	Diff	F1-score	Diff
Word count per PoS type	0.592	+0.010	0.507	−0.026
Number of letters in words	0.579	+0.023	0.426	+0.055
Word repetition	0.590	+0.011	0.506	−0.024
Content of the vocabulary	0.484	+0.118	0.394	+0.088
Word sentiment	0.587	+0.015	0.451	+0.031

7 Conclusion

We presented a recognition model of client persuasion for analyzing how the client is persuaded in the consultation process by a professional consultant with a novel consultant interaction corpus. The best results were obtained when the BLSTM-trained multimodal features of the consultant and client were used simultaneously with 0.697 for accuracy and 0.661 for F1-score. In addition, we clarified which linguistic feature of consultant is effective for recognizing when a client has been persuade with ablation test. The important future direction is to collect data from consultants with various kinds of experiments and compare how to persuade the clients.

Acknowledgements. We sincerely appreciate Be · Fine Co. ltd. and Ms. Teruko Kobayashi who is the professional beauty counselor.

References

1. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
2. Chen, L., Feng, G., Joe, J., Leong, C.W., Kitchen, C., Lee, C.M.: Towards automated assessment of public speaking skills using multimodal cues. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 200–203 (2014)
3. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. **20**, 273–297 (1995)
4. DeVault, D., et al.: SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems, pp. 1061–1068 (2014)
5. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**, 602–610 (2005)
6. Higashiyama, M., Inui, K., Matsumoto, Y.: Learning sentiment of nouns from selectional preferences of verbs and adjectives. In: Proceedings of the Annual Meeting of the Association for Natural Language Processing, pp. 584–587 (2008)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Hoque, M.E., Courgeon, M., Martin, J.-C., Mutlu, B., Picard, R.W.: MACH: My automated conversation coach. In: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 697–706 (2013)
9. Ishii, R., Otsuka, K., Kumano, S., Higashinaka, R., Tomita, J.: Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 31–39 (2018)
10. Kingma, D., Adam, J.B.: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations, pp. 1–15 (2015)
11. Nozomi, K., Kentaro, I., Yuji, M., Kenji, T.: Collecting evaluative expressions for opinion extraction. J. Nat. Lang. Process. **12**(3), 203–222 (2005)
12. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 230–237 (2004)
13. Nair, V., Hinton, G.E.: Rectified linear units improve Restricted Boltzmann machines. In: Proceedings of the International Conference on Machine Learning, pp. 807–814 (2010)
14. Nguyen, L.S., Frauendorfer, D., Mast, M.S., Gatica-Perez, D.: Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Trans. Multimed. **16**(4), 1018–1031 (2014)
15. Okada, S., et al.: Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 169–176 (2016)
16. Park, S., Shim, H.S., Chatterjee, M., Sagae, K., Morency, L.-P.: Computational analysis of persuasiveness in social multimedia: a novel dataset and multimodal prediction approach. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 50–57 (2014)
17. Ramanarayanan, V., Leong, C.W., Chen, L., Feng, G., Suendermann-Oeft, D.: Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 23–30 (2015)

18. Sanchez-Cortes, D., Aran, O., Mast, M.S., Gatica-Perez, D.: A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Trans. Multimed. **14**, 816–832 (2012)
19. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The interspeech 2010 paralinguistic challenge. In: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 2794–2797 (2010)
20. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
21. Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., Inui, K.: A joint neural model for fine-grained named entity classification of wikipedia articles. IEICE Trans. Inf. Syst. **E101.D**(1), 73–81 (2018)
22. Tanaka, H., Negoro, H., Iwasaka, H., Nakamura, S.: Listening skills assessment through computer agents. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 492–496 (2018)
23. Tanaka, H., et al.: Automated social skills trainer. In: Proceedings of the ACM International Conference on Inteligent User Interface, pp. 17–27 (2015)
24. Tavabi, L., Stefanov, K., Gilani, S.N., Traum, D., Soleymani, M.: Multimodal learning for identifying opportunities for empathetic responses. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 95–104 (2019)
25. Tianqi, C., Carlos, G.: XGBoost: a scalable tree boosting system. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
26. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
27. Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R., Scherer, S.: Multimodal public speaking performance assessment. In: Proceedings of ACM International Conference on Multimodal Interaction, pp. 43–50 (2015)
28. Xiao, B., Imel, Z.E., Georgiou, P., Atkins, D.C., Narayanan, S.S.: Computational analysis and simulation of empathic behaviors: a survey of empathy modeling with behavioral signal processing framework. Curr. Psych. Rep. **18**(5), 1–11 (2016)
29. Xiao, B., Imel, Z.E., Georgiou, P.G., Atkins, D.C., Narayanan, S.S.: "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. PLOS ONE **10**(12), 1–15 (2015)