Objective Prediction of Social Skills Level for Automated Social Skills Training Using Audio and Text Information

Takeshi Saga
saga.takeshi.sn0@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

Hiroki Tanaka
hiroki-tan@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

Hidemi Iwasaka
Department of Psychiatry, Nara Medical University
Kashihara, Nara, Japan

Satoshi Nakamura
s-nakamura@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

ABSTRACT

Although Social Skills Training is a well-known effective method to obtain appropriate social skills during daily communication, getting such training is difficult due to a shortage of therapists. Therefore, automatic training systems are required to ameliorate this situation. To fairly evaluate social skills, we need an objective evaluation method. In this paper, we utilized the second edition of the Social Responsiveness Scale (SRS-2) as an objective evaluation metric and developed an automatic evaluation system using linear regression with multi-modal features. We newly adopted features including 28 audio features and BERT-based sequential similarity (seq-similarity), which indicates how well the meaning of users remains consistent within their utterances. We achieved a 0.35 Pearson correlation coefficient for the SRS-2's overall score prediction and 0.60 for the social communication score prediction, which is a treatment subscale score of SRS-2. This experiment shows that our system can objectively predict the levels of social skills. Please note that we only evaluated the system on healthy subjects since this study is still at the feasibility phase. Therefore, further evaluation of real patients is needed in future work.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Applied computing** → **Health care information systems**.

KEYWORDS

Social skills training, Social signal processing, Multimodal analysis

ACM Reference Format:
Takeshi Saga, Hiroki Tanaka, Hidemi Iwasaka, and Satoshi Nakamura. 2020. Objective Prediction of Social Skills Level for Automated Social Skills Training Using Audio and Text Information. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425221

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425221

1 INTRODUCTION

Social Skills are essential for communicating with others. According to Bellack et al., the components of social skills are mainly divided into three groups: expressive behavior, such as eye contact or posture; receptive behavior, which is attention to and the interpretation of the relevant cues for emotion recognition; and interactive behavior, such as response timing or turn taking [3]. We usually acquire these skills in childhood. However, some children fail to learn them for some reasons such as Autism Spectrum Disorders (ASD) [1]. Daily life is quite difficult without such skills since most daily situations are based on the integration of several social skills. Even though social skills are necessary, learning them by ourselves once we've become adults is very complicated.

One solution is called Social Skills Training (SST), which is a type of behavioral therapy for improving social skills. According to Bellack et al., it is based on "conditioned reflex therapy" and "psychotherapy by reciprocal inhibition" and "social learning theory" [2, 13, 18, 27]. Once SST was first introduced, it became widely used in a wide range of areas [3]. SST's minimum setup is one leader and one member [3]. It can also be done in a group with several members. In that setting, a sub-leader might be designated to help the leader. The following is its basic procedure. First, the leader and the members decide the objective skill and the goal of the SST. Second, the leader demonstrates a good example of the skill/goal by acting in the situation himself. Third, a member imitates the example by himself. Fourth, positive and negative feedback are given to the member by the leader. Fifth, based on the feedback, the member repeats his performance and tries to improve it. Sometimes, homework might be assigned to him for further improvement.

Although SST is well-known method, access to it is difficult for the following reasons. One is social stigma attached to those with mental illnesses [8, 10]. Many such people are not completely accepted in certain situations. Patients sometimes fear that they or their family might be abused or shunned by neighbors who learn about such problems. Another reason is the difficulty of training to become an SST leader. To conduct SST effectively, a person must master how it works and learn how to give feedback to maximize the

members' improvements, which is time consuming. Some programs need 8 to 12 weeks to complete [5].

Therefore, virtual agents have been introduced. SimSensei is a virtual therapist that conducts therapy sessions for PTSD patients, whose quality approaches that of face-to-face sessions [6]. However, this virtual agent wasn't tested it in an SST context. Tanaka et al. automated the SST process by a computer system using a multi-modal dialogue system with an embodied conversational agent [20]. Their system goal was to improve user's speaking skill. They used speaking skill score as an evaluation metric to measure user's speaking level. Scores were annotated by several experienced therapists and averaged as groundtruth labels for each bit of data. Although the scores were annotated by therapists, they remain unfair since they are subjective. An objective metric is still needed.

In this paper, we solved the subjectivity problem by utilizing the Social Responsiveness Scale Second Edition (SRS-2) as an objective evaluation metric [11]. To improve the SRS-2 score estimation, we used some new audio and text features, including novel BERT-based sequential similarity.

Since SRS-2 is originally developed to create the most official treatment plan for clinical practice, this study has potential to accelerate in clinical situation as well [11]. Since this study is still in feasibility phase, we evaluated the effectiveness on healthy subjects as the first step. Evaluation on real patients will be next step in the near future.

2 SOCIAL RESPONSIVENESS SCALE SECOND EDITION (SRS-2)

SRS-2 is an objective evaluation metric comprised of 65 questions. Although SRS-2 was originally designed to assess potential Autism Spectrum Disease (ASD) sufferers, it is also capable of differentiating a variety of mental diseases. Furthermore, its effectiveness was investigated not only on disabled people but on healthy people as well [11]. Therefore, this can be used for healthy people too.

In this paper, we utilized the SRS-2 overall score and one of its treatment sub-scale scores called Social Communication (22 items out of 65) as ground-truth labels for training. Since the SRS-2 overall scores include and can be affected not only by social communication skills but also such information as living style, we also decided to use SRS-2 Social Communication scores, which provides clearer information about the user's social communication skills than the overall scores. Especially, it indicates a physical aspect of social interaction, which is more feasible for future SST automation because it is intuitively understandable and objectively observable.

Social communication score and SRS-2 overall score were highly correlated with its coefficient 0.92. On the other hand, their correlation coefficients with subjective speaking score, rated by experts in previous research, was -0.19 and -0.29 respectively [19].

3 MULTI-MODAL FEATURES

Correctly using every appropriate skill in every situation is very difficult since social skills are constructed with multiple communication modalities. Therefore, multi-modal features should be used to estimate the levels of social skills. In this section, we introduce our predictor's input features for each modality and our basic idea. The features used in our proposed model are shown in Table 1.

Table 1: Audio and text features

Feature name	Description
Energy	Mean spectral energy
F0, F1, F2, F3 Mean	Mean frequency of F0, F1, F2, F3
F0, F1, F2, F3 SD	Standard deviation of F0, F1, F2, F3
F0 Min	Minimum F0 frequency
F0 Max	Maximum F0 frequency
F0 range	Difference between F0 MAX and F0 MIN
F1, F2, F3 BW	Average bandwidth of F1, F2, F3
F2/F1, F3/F1 Mean	Mean ratio of F2-F1 and F3-F1
F2/F1, F3/F1 SD	Standard deviation of F2/F1 and F3/F1
Int mean	Mean vocal intensity
Int Min	Minimum vocal intensity
Int Max	Maximum vocal intensity
Int range	Differences between max and min intensity
Int SD	Standard deviation of vocal intensity
Jitter	Irregularities in F0 frequency
Shimmer	Irregularities in intensity
Unvoiced %	Percentage of unvoiced regions
Breaks %	Average percentage of breaks
WPM	Number of words per minute
Six plus	Number of words over six letters
Fillers	Number of fillers
Vocabulary size	Number of unique words
Seq-similarity	BERT-based sequential similarity

3.1 Audio Features

Previous research identified the importance of audio information to estimate people's speaking skills [20, 21]. Both of these works used four different audio features. We enhanced their research by applying 28 audio features that were originally used to analyze and predict the performances of job applicants during interviews [15]. We decided to apply it to this research since job interviews also use multi-modal information like SST.

We calculated the audio features with Praat [25], which is open-source software for prosody analysis.

3.2 Textual Features

We utilized three features from Tanaka's work as basic features [20]. Word per minutes (WPM in short) indicates user's speaking rate, which is reported it correlates with interview skills [9]. Word more than six letters (called six plus in this paper) indicates how the user

uses complicated or unexpected words. It is reported that individuals with social difficulties use more complicated or unexpected words, and deficits of social difficulties affect inappropriate usage of words [17]. Specifically, words structured by more than six letters may be related to complicated words [16]. Fillers indicates how frequently user uses fillers while talking, which disturbs listener's focus on the content of the speech.

In this paper, we extended them by adding new features, vocabulary size and BERT-based sequential similarity based on the following two hypothesizes:

(1) People with high social skills should have a large vocabulary.
(2) The talk by people with high social skills should be semantically consistent.

We calculated the vocabulary size after transforming words into their original grammatical forms using the mecab-ipadic-NEologd Japanese morphological analysis tool [22–24], which is based on the MeCab Japanese morphological analysis tool extended with a customized system dictionary [14].

We also utilized BERT-based sequential similarity (seq-similarity) as a feature. It was inspired by a previous work [4]. Bertola et al. proposed a method to calculate embedding similarity by averaging the cosine similarity between the 1st and 2nd words, the 3rd and 4th words, etc. They calculated word embedding using the tf-idf method. Recently, it was reported that BERT works better than previous methods in many different research topics [7]. BERT is a type of neural network based on transformer architecture, which can output word embeddings by taking the adjacent context into account. Therefore, we utilized BERT to create word embeddings. We used a pre-trained Japanese BERT model with whole-word-masking provided by a library named "Transformers" [26]. Once word embeddings were extracted, its average of cosine similarity scores between adjacent pair of words was calculated as the final feature value.

4 EXPERIMENTS

4.1 SRS-2 score prediction from extracted features

We trained a linear regression model using the multi-modal features shown in Table 1. We didn't use regularization in this paper since its result was worse in preliminary experiment. To compare our model with a previous work, we also trained Tanaka's model as a baseline [20]. We used the following as input features of the baseline model: words per minute, words over six letters, number of fillers, vocal amplitude mean, coefficient of F0 variation, pause percentage, spectral tilt between F1 and F3 (H1A3), smile ratio, and head poses (pitch, yaw, roll).

4.1.1 *Training data.*

For model training and evaluation, we used Japanese one-minute speaking videos taken from the experiments of previous research [19].

The following is its data collection procedure. First, the participants talked for one minute about a recent positive event through a laptop screen to the virtual agent, which was created using MMDAgent (http://www.mmdagent.jp/). While the participant is talking, his face and voice were recorded. Its transcription was created by human annotator afterwords. Second, they filled out a 65-item questionnaire that explored their living style, communication style, how they usually feel when they are communicating with others, etc. The dataset included the speaking data of 27 healthy subjects.

Table 2: Our model's coefficients for overall score prediction (left) and social communication score prediction (right)

Name	Coefficients	Name	Coefficients
Seq-similarity	-595.7	Seq-similarity	-277.6
Energy	74.0	Energy	23.6
F2/F1 Mean	32.5	F2/F1 SD	-13.5
F2/F1 SD	-30.2	F2/F1 Mean	5.8
Int SD	13.6	F3/F1 Mean	4.8

4.1.2 *Outlier elimination.*

To create the feature vectors used in both the baseline and our model, we extracted multi-modal features from each video. As we extracted the visual features of the baseline, we found an outlier caused by a failure of the head pose estimation. We eliminated these data from the dataset and trained with the remaining data of the 26 participants.

4.1.3 *Experiment details.*

All the features were normalized with subtracting the mean and dividing by the l2-norm before inputting them to the model. To address the problem of the dataset size, we trained and evaluated it with a leave-one-subject-out cross validation technique.

4.2 Results

The overall score prediction respectively resulted in the following results for the baseline and our method: correlation coefficients, 0.28 and 0.35; RMSE, 32.13 and 27.34; and the standard deviation of the absolute error, 24.08 and 14.38. Figures 1 shows scatter plot with ground-truths and predictions of the SRS-2 overall scores for our model.

The social communication score prediction respectively resulted in the following results for the baseline and our model: correlation coefficients, 0.45 and 0.60; RMSE, 11.33 and 9.36; and the standard deviation of the absolute error, 8.17 and 4.59. For the social communication score prediction, both the baseline model and ours were significantly correlated ($p<0.05$) by a test of no correlation. Our model's result was significantly correlated at a 0.01 p-value threshold. Our model's RMSE was 1.97 smaller than the baseline. Figures 2 shows scatter plot with ground-truth and prediction of social communication scores for our model.

The left of Table 2 shows the top-5 coefficients of the models when we predicted the SRS-2 overall scores and the social communication scores. The coefficients were averaged across the 26 cross-validation trials. We confirmed that no critical multicollinearity was found since its coefficients were either stable across all the trials or were too small to affect the results.

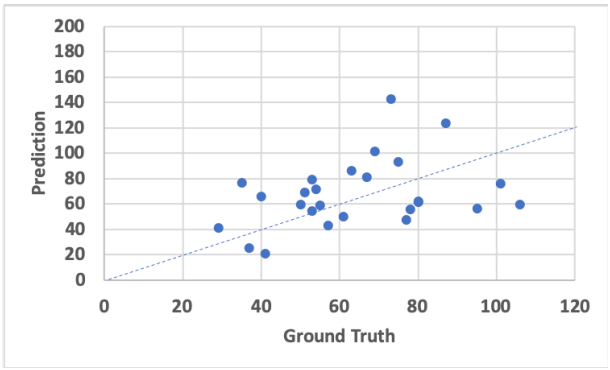

Figure 1: Prediction of SRS-2 overall score with our model: correlation coefficient: 0.35

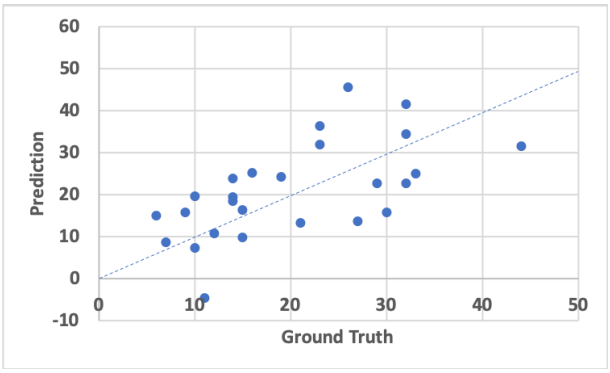

Figure 2: Prediction of SRS-2 social communication scores with our model: correlation coefficient: 0.60

5 DISCUSSION

From the result of SRS-2 overall score prediction, our proposed model outperformed the baseline for both the SRS-2 overall scores and social communication skill scores. Especially in the social communication skill score prediction, our model's predictions were more precise and more accurate since the resulting correlation coefficient, RMSE, and standard deviation of the absolute error were 0.60, 9.36, and 4.59, respectively.

To determine which features most affected the results, we investigated the trained model's feature coefficients. Since every input feature was normalized to a value range from -1 to 1, we treated each feature coefficient as the importance of the feature. From the left of Table 2, note that the Seq-similarity produced by BERT was much more dominant compared to others. We found a similar tendency on the right side of Table 2. Hence, our second hypothesis "The talk by people with high social skills should be semantically consistent" was proved; the first one "People with high social skills will have large vocabularies" was rejected.

We also tried to determine the meaning of the coefficients. Note that since the SRS-2 score was originally developed to assess the severity of mental diseases, a low SRS-2 score denotes better social skills. For the overall score prediction, F2/F1 mean and Energy and Int SD had positive coefficients; Seq-similarity and F2/F1 SD had negative ones. Since the F2/F1 mean is intuitively difficult to understand, we need to scrutinize it in this section.

Kagomiya reported that the long vowels of /i/, /e/, /u/ have lower F2/F1 ratios than their short vowels [12]. Although the long vowels of /a/, /o/ have higher F2/F1 ratios than their short vowels, we believe that the latter can be ignored because the former is dominant (3 out of 5 vowels). Therefore, from this theory about the relation between short/long vowels and F2/F1 ratio, with the support of the our results, we found that using longer vowels in conversation improved social skills. This is intuitively understandable because we sometimes feel that people speaking without long vowels in Japanese are too systematic or too dry and hence less friendly.

In summary, we believe that the following points are critical to improve social skills. Note that each point corresponds to each feature on the left of Table 2.

- Maintain semantic consistency while you are talking.
- Do not talk too loudly.
- Use long vowels instead of short ones.
- Talk in a clear, distinguishable voice.
- Maintain an equal level of voice intensity.

We also identified a similar tendency in social communication scores (the right of Table 2). That was because the social communication score was a sub-scale of the SRS-2 overall score, and both of them were highly correlated with its coefficient of 0.92.

At the same time, the social communication score prediction is more accurate and precise than the overall score estimation because the latter is more difficult since the overall score includes more complex user traits. We believe that more critical features must be included to more accurately estimate the overall scores.

6 CONCLUSION AND FUTURE DIRECTION

We proposed a model to predict the SRS-2 overall score and its sub-scale score (social communication score) with a novel set of multi-modal features. Our model achieved a correlation coefficient 0.35 for the overall score and 0.60 for the social communication score.

Although we showed the effectiveness of audio and text features to predict social skills level, there are some possibility to improve. First, in terms of data collection, we need to investigate the effect of virtual agent. Results might be different when the human therapist handles the session instead of virtual agent. Second, since all participants for this experiments were healthy subjects, therefore, it is needed to test this system with people who has real social difficulties for further evaluation. Third, since we didn't use any visual information, there is opportunity to improve by adding such information. Our team is mainly working on second and third one, and it will be published soon.

ACKNOWLEDGMENTS

Funding was provided by the Core Research for Evolutional Science and Technology (Grant No. JPMJCR19A5) and the Japan Society for the Promotion of Science (Grant No. JP17H06101 and JP18K11437).

REFERENCES

[1] [n.d.]. Relationships & Social Skills. https://chadd.org/for-adults/relationships-social-skills/. Accessed: 2020-08-15.

[2] A. Bandura. 1969. *Principles of behavior modification.* Holt, Rinehart and Winston.

[3] Alan S. Bellack, Kim T. Mueser, Susan Gingerich, and Julie Agresta. 2004. *Social Skills Training for Schizophrenia: A Step-by-Step Guide* (2 ed.). Guilford Press.

[4] Laiss Bertola, Natália B. Mota, Mauro Copelli, Thiago Rivero, Breno Satler Diniz, Marco A. Romano-Silva, Sidarta Ribeiro, and Leandro F. Malloy-Diniz. 2014. Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in aging neuroscience* 6 (29 Jul 2014), 185–185. https://doi.org/10.3389/fnagi.2014.00185 25120480[pmid].

[5] Social Skills Co. [n.d.]. Social Skills Co. https://socialskillscompany.com/ Accessed September 17, 2020).

[6] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margot Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014* 2, 1061–1068.

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

[8] Boak et al. 2016. *The mental health and well-being of Ontario students, 1991-2015: Detailed OSDUHS findings.* Technical Report 43. Toronto: Centre for Addiction and Mental Health.

[9] Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My Automated Conversation Coach *(UbiComp '13).* Association for Computing Machinery, New York, NY, USA, 697–706. https://doi.org/10.1145/2493432.2493502

[10] Justin Hunt and Daniel Eisenberg. 2010. Mental Health Problems and Help-Seeking Behavior Among College Students. *Journal of Adolescent Health* 46, 1 (2010), 3 – 10. https://doi.org/10.1016/j.jadohealth.2009.08.008

[11] MD John N. Constantino and PhD Christian P. Gruber. 2012. *Social Responsiveness Scale, Second Edition (SRS-2)Back.* Western Psychological Services.

[12] Takayuki Kagomiya. 2015. Articulatory positions of Japanese vowels as a function of duration computed from a large-scale spontaneous speech corpus. In *ICPhS.*

[13] Alan S. Bellack Kim T. Mueser. 2007. Social skills training: Alive and well? *Journal of Mental Health* (2007).

[14] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Barcelona, Spain, 230–237. https://www.aclweb.org/anthology/W04-3230

[15] Iftekhar Naim, Md Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing* PP (04 2015). https://doi.org/10.1109/TAFFC.2016.2614299

[16] James Pennebaker, Martha Francis, and Roger Booth. 2007. Linguistic Inquiry and Word Count:LIWC2007. (01 2007). Available at https://liwc.wpengine.com/.

[17] Masoud Rouhizadeh, Emily Prud'hommeaux, Brian Roark, and Jan van Santen. 2013. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, Atlanta, Georgia, 709–714. https://www.aclweb.org/anthology/N13-1084

[18] A. Salter. 1949. *Conditioned reflex therapy.* Creative Age Press.

[19] Hiroki Tanaka, Hidemi Iwasaka, Hideki Negoro, and Satoshi Nakamura. 2020. Analysis of conversational listening skills toward agent-based social skills training. *Journal on Multimodal User Interfaces* 14, 1 (01 Mar 2020), 73–82. https://doi.org/10.1007/s12193-019-00313-y

[20] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE* 12, 8 (08 2017), 1–15. https://doi.org/10.1371/journal.pone.0182151

[21] Hiroki Tanaka, Sakti Sakriani, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2016. Teaching Social Communication Skills Through Human-Agent Interaction. *ACM Trans. Interact. Intell. Syst.* 6, 2, Article 18 (Aug. 2016), 26 pages. https://doi.org/10.1145/2937757

[22] Sato Toshinori. 2015. Neologism dictionary based on the language resources on the Web for Mecab. https://github.com/neologd/mecab-ipadic-neologd

[23] Taiichi Hashimoto Toshinori Sato and Manabu Okumura. 2016. Operation of a word segmentation dictionary generation system called NEologd (in Japanese). In *Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL).* Information Processing Society of Japan, NL–229–15.

[24] Taiichi Hashimoto Toshinori Sato and Manabu Okumura. 2017. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing.* The Association for Natural Language Processing, NLP2017–B6–1.

[25] Vincent van Heuven. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 9/10 (2001), 341–345.

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[27] J. Wolpe. 1958. *Psychotherapy by reciprocal inhibition.* Stanford University Press.