Fake Moods: Can Users Trick an Emotion-Aware VoiceBot?

Yong Ma
LMU Munich
Munich, Germany
yong.ma@ifi.lmu.de

Heiko Drewes
LMU Munich
Munich, Germany
heiko.drewes@ifi.lmu.de

Andreas Butz
LMU Munich
Munich, Germany
andreas.butz@ifi.lmu.de

ABSTRACT

The ability to deal properly with emotion could be a critical feature of future VoiceBots. Humans might even choose to use fake emotions, e.g., sound angry to emphasize what they are saying or sound nice to get what they want. However, it is unclear whether current emotion detection methods detect such acted emotions properly, or rather the true emotion of the speaker. We asked a small number of participants (26) to mimic five basic emotions and used an open source emotion-in-voice detector to provide feedback on whether their acted emotion was recognized as intended. We found that it was difficult for participants to mimic all five emotions and that certain emotions were easier to mimic than others. However, it remains unclear whether this is due to the fact that emotion was only acted or due to the insufficiency of the detection software. As an intended side effect, we collected a small corpus of labeled data for acted emotion in speech, which we plan to extend and eventually use as training data for our own emotion detection. We present the study setup and discuss some insights on our results.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

Speech Emotion Detection, Emotion-Aware VoiceBot, Data Acquisition for Training Neural Networks

ACM Reference Format:
Yong Ma, Heiko Drewes, and Andreas Butz. 2021. Fake Moods: Can Users Trick an Emotion-Aware VoiceBot?. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3411763.3451744

1 INTRODUCTION

Recent years have seen an increasing use of voice-based user interfaces and conversational agents (for which we will use the compact term VoiceBots), such as Alexa, Siri and Google Home. VoiceBots often use artificial intelligence (AI) for implementing rather complex HCI systems, such as a voice controlled robot [9] or a mental health VoiceBot [22]. They have become a powerful communication tool between humans and machines. However, developing better voice interfaces for an even more natural conversation with the VoiceBot will eventually require some sort of emotion-awareness, because this is a substantial aspect of communication between humans.

Emotion can be detected in two different ways: Traditionally, a VoiceBot can detect the users' emotion based on standard speech recognition (SR) and natural language understanding (NLU) methods [2]. This approach detects emotion from *what* is being said. Analyzing the users' voice and thereby detecting *how* things are said, is a more recent approach and called Speech Emotion Recognition (SER) [23]. The main processing steps for SER are the extraction of adequate speech features and then the detection of emotion from these features, using traditional machine learning (ML) methods such as Gaussian Mixture Models (GMM), Support Vector Machines (SVM) or Artificial Neural Networks (ANN) [4]. With the recent developments in AI, detecting speech emotion using deep learning architectures [5] has become a feasible alternative.

However, a major challenge for all these ML methods is to obtain accurately labeled data for different speech emotions and to provide a ground truth for learning. Currently, there are two types of speech-emotion databases - acted-emotion datasets and induced-emotion datasets. For an acted-emotion dataset, researchers asked actors to perform different speech-emotions, as in the SAVEE dataset [6, 7]. The alternative approach is to elicit authentic speech emotions. Research in psychology typically induces emotions by showing pictures or videos which can be used to arouse the intended emotions. In the IEMOCAP dataset [3], actors performed selected emotional scripts (acting emotions) and also improvised hypothetical scenarios designed to elicit specific types of true (i.e., non-acted) emotions.

Considering these challenges and the general shortage of training data, we were curious to find out whether regular users (i.e., non-actors) are able to mimic five basic emotions (neutral, happy, sad, anger, fear) and whether they manage to trick emotion recognition into detecting the intended emotion. This is in line with prior research, which investigated how easy it is to fake a personality in conversations with a chat bot [25].

We set up the web page shown in Figure 1 and recruited a small number of participants to record their voice in five basic emotions. The web page provided feedback whether the emotion was successfully recognized. We counted the number of trials until success in each emotion and found that participants were not able to successfully mimic all five basic emotions. However, it was not clear whether this was a failure of the emotion detector we had used or the users' actual inability. As a side effect, our experiment also provided a small labeled data set for acted emotion from regular users, which we were hoping to use for training our own future SER prototypes.

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451744

Figure 1: The web page for collecting voice samples. The emoticons in the upper row can be clicked to select the emotion to enter. In the beginning, all emoticons were gray. After successfully mimicking an emotion, the corresponding emoticon becomes yellow.

2 RELATED WORK

2.1 Affective Computing as the General Background

Our work tries to contribute to the general field of affective computing, which will play a crucial role for conversational agents. It is characterized by recognizing, interpreting, processing, and simulating human affects based on face expression, biometric measurements, linguistics, or speech [21]. NLP and SER are two main techniques to detect emotions in the conversation between users and a VoiceBot and therefore constitute essential building blocks for affective computing.

2.2 Emotion Detection in Voice Dialogs

Currently, voice-interaction systems can detect humans' emotion in a conversation either by analyzing the semantics of what is being said [17] by using linguistic emotion analysis or natural language processing (NLP). For example, certain emotional keywords can point to specific emotional states in the conversation [20]. The alternative is to use the actual voice data to detect how things are being said, for example from time or frequency domain features [1] or by analyzing spectrograms [27]. In our study, we used SER to detect participants' speech emotion.

2.3 Speech Emotion Recognition Technologies

Lee et al. [14] implemented Hidden Markov models (HMM) based on short-term spectral features for recognizing four different emotions. They found that a model trained with the spectral properties of vowel sounds performed better than a model with prosodic features. Other speech features, such as pitch or loudness [13], Mel Frequency Cepstral Coefficients (MFCCs) [24] and Linear Predictive Coefficients (LPC) [24] can also help us analyze basic emotions in speech data.

For the actual learning and recognition, a variety of ML algorithms from SVM [12] to deep learning [15] are used, e.g., the Kernel Extreme Learning Machine (KELM) [11], Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNN) [8, 16]. In our study, we used the SER package, which utilizes one-layer neural network and time-frequency speech features, to detect emotions from voice data.

2.4 Emotion Recognition Across Languages

Usually, humans can recognize another human's emotion reasonably well from their speech, even if they do not understand each other's language. Pell et al. [19] proposed that speech emotion is largely unaffected by language or linguistic similarity. Since it is possible for humans to understand vocally-expressed emotions in any sort of speech [11, 18], we let participants express their emotions in any language they are strongly familiar with.

2.5 Fake Emotion

Fake emotions are intentionally expressed emotions which do not match the actual emotional state of the person who expressed these emotions. People show fake emotions to be conform with social norms or to provoke an intended reaction from others. Fake emotions may eventually play a vital role in detecting users' emotion in a VoiceBot system, because humans may not be willing to express their true emotions, but instead try to act out different ones in order to achieve a certain reaction from the system. Juslin et al. [10] argued that there are reliable differences between spontaneous and posed expression in vocal emotion expression. Therefore, it should not be possible for users to arbitrarily trick emotion-aware VoiceBots. In our study, the majority of participants was able to fake some, but not all basic emotions.

3 EXPERIMENTAL SETUP

We set up the web page in Figure 1 to conduct our study. The page shows five clickable emoticons in a row, plus the currently selected emotion in the center below, as well as the recognition result. Participants can select an emotion from the five given basic emotions and then are asked to say something with the selected emotion. The goal of the study was to find out whether participants would be able to act all basic emotions and in consequence could trick an emotion-aware VoiceBot.

3.1 Apparatus

In our study, we used the SER package OpenVokaturi[1] version 3.4, to detect participants' emotion in the voice samples recorded on our website. OpenVokaturi is a free version of a commercial software package, which offers, in addition to the neutral emotion, only four instead of six basic emotions. According to the manufacturer[2] the product is trained with two different databases, the Berlin Database of Emotional Speech[3] based on voice samples of ten actors and the Surrey Audio-Visual Expressed Emotion (SAVEE) Database[4] based on four actors. The recognition rate is stated with 69% for four emotions (while it would be 76% for human listeners). Vokaturi makes no restrictions regarding the recognition of acted vs. true emotions and it is generally assumed to recognize both equally well. The voice data was collected by the built-in microphone of the participants' own computers and uploaded to our web server. All recorded voice signals had a sample rate of 48 kHz. We used Matlab 2017a for later data analysis.

[1] https://vokaturi.com/
[2] https://developers.vokaturi.com/algorithms/annotated-databases
[3] http://emodb.bilderbar.info/index-1280.html
[4] http://kahlan.eps.surrey.ac.uk/savee/

3.2 Participants

We recruited 26 participants (13 male) from our personal networks to join the experiment. They joined our study after we sent the web page link to them by email, but we did not have any means to connect the collected recordings to the email invitations, as they were free to join whenever and from wherever they wanted.

3.3 Experimental Procedure

We informed participants in our privacy statement that the study was completely anonymous and participants did not even need to provide demographic data. Our study procedure was approved by the local ethics review board of our university. The only instruction for the participants was to select an emotion and then speak with the selected emotion, and so they were free to use any language they liked and to say anything they wanted. There was no particular order and the central Emoji would change to what they had chosen in the top row. When they clicked the "start recording" button, 2 seconds of voice data was recorded and uploaded to our web server. Participants could try as often as they wanted to mimic each emotion. Upon success, the corresponding emoji in the top row would turn from grey to yellow. This meant that they had successfully acted the selected emotion. They could then choose another emotion and continue. If they found certain emotions too difficult to imitate, they could click a "give up" button and end the user study.

4 RESULTS

We only captured the number of trials each participant spent on each emotion. As a derived measure, we calculated the success rate as the inverse of the number of trials until success, or as 0 in case they gave up. We then used a Wilcoxon Signed Rank Test (WSRT) [26] to determine whether there were any significant differences between the different emotions, regarding the success rate and number of trials. Our expectation was that all emotions would be equally well recognized. We only found a significant difference in the number of trials between "neutral" and "happy" (p=0.0284) and between "happy" and "fear" (p=0.0479). Evaluating the success rate, however, revealed a significant difference between "neutral" and all other emotions. Table 1 shows the p-values for "neutral" against all other emotions regarding success rates.

Table 1: Result of a Wilcoxon signed rank test on the success rate between "neutral" and all other emotions

	Neutral	*Happy*	*Sad*	*Angry*	*Fear*
Neutral	-	0.0029	0.0109	0.0354	0.0107

Figure 2 shows the number of trials and the success rates for each emotion. If participants did not try a specific emotion, no success rate was calculated. Not all participants were able to imitate all five emotions equally well. It seemed easy for them to mimic the "neutral" emotion but difficult to act the "happy" emotion. In addition, we found that only few participants chose the "give up" button when they could not imitate a certain emotion. Most participants could not act all five basic emotions, even if they tried many times.

5 ANALYSIS AND DISCUSSION

In our study, most participants did not succeeded in mimicking all five basic emotions. However, we can not simply claim that the SER system OpenVokaturi we used is not good enough for emotion detection. Instead, we argue that it may have failed because the participants were unable to successfully act out each emotion, i.e., to properly fake it. The result of the study hence suggests that it may in general be difficult to 'cheat' the emotion detector, at least OpenVokaturi, with acted emotions. This may be a hint for the detection of true emotion. On the other hand, the detection results vary very much which is not in accordance with the hypothesis that the detector detects the true emotion. The users' emotion should not change too much over the short time of the study. It is possible that the study itself affected the user's emotional state. Success in entering the demanded emotion could have made participants happy, while failure could have made them angry. We were unable to verify such effects as the data set was not big enough. However, even with more data, such effects can only be seen if the emotion detection reports accurately true emotions, which is not guaranteed.

An interesting question for our future research is whether it is possible to build an emotion-in-voice detector, which detects acted emotion and can distinguish it from true emotion. After collecting a bigger corpus of data with the existing system, we will train a neural network with it. If the users will be more successful in mimicking emotions with the new detector, we achieved our goal and can claim to have built a detector for acted emotion. However, there are more general reasons to be skeptical. A 40 millisecond voice sample is probably not enough for humans to judge the emotion in this voice sample. Humans normally need at least a few words to judge the true emotion in a voice. It might even take other methods or longer samples to properly detect acted emotion.

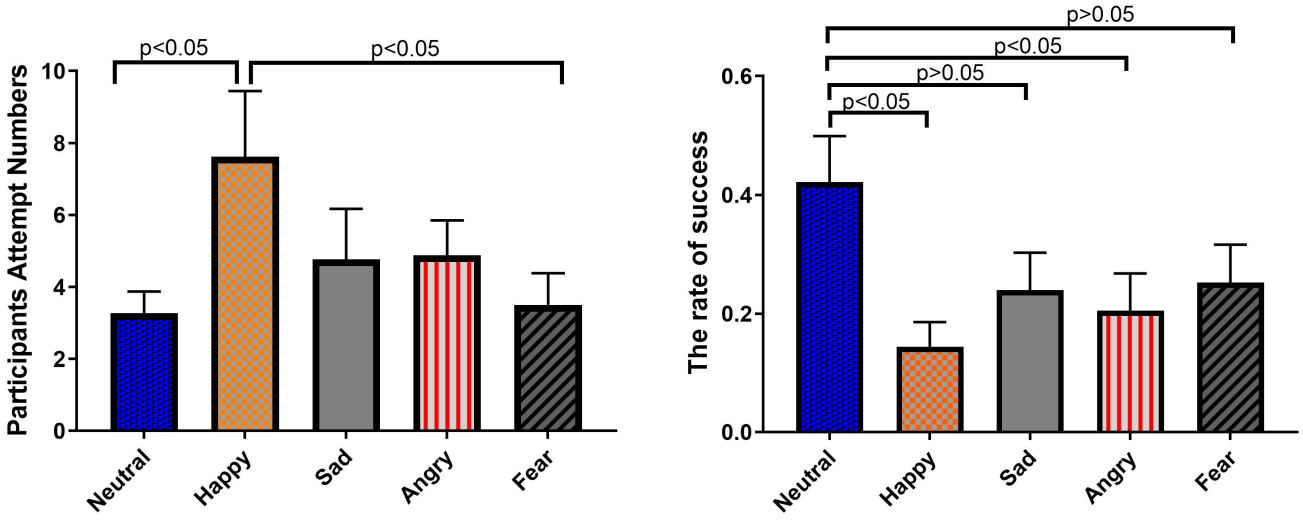
6 LIMITATION AND FUTURE WORK

In our study, not all participants even tried to mimic all basic emotions and some of them only tried one or two. Moreover, certain emotions such as "Happy" or "Angry" were relatively difficult for certain participants to mimic. They tried many times and some of them finally gave up. We simply do not have enough data at this point to reliably trace these problems back to either the SER system we used (OpenVokaturi) or to the participants' inability to act the emotions. The reason might actually be a mix of both.

The long term perspective of our work is to distinguish between true and fake emotion and to be able to reliably detect both. This would enable emotion-aware VoiceBots that can be controlled by fake emotions, as you would do with a dog in training or with small children. As an immediate next step, we will iterate on our study setup and invite a much larger sample of participants, potentially with more well-defined tasks, given texts, or other improvements to the study procedure, in order to produce a much richer and more meaningful corpus of data. In this sense, the submission at hand is truly late-breaking and only the beginning of what we consider an interesting journey.

ACKNOWLEDGMENTS

This work was partially funded by the China Scholarship Council (CSC). We thank our reviewers for their valuable feedback.



(a) Number of trials until success

(b) Success rates

Figure 2: Number of trials and success rates of all 26 participants for all basic emotion they tried to mimic.

REFERENCES

[1] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review* 43, 2 (2015), 155–177.

[2] Rachel Batish. 2018. *Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger.* Packt Publishing Ltd, Birmingham, UK.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[4] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[5] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2017. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks* 92 (2017), 60–68.

[6] Sanaul Haq, Philip JB Jackson, and James Edge. 2008. Audio-visual feature selection and reduction for emotion classification. In *Auditory-Visual Speech Processing (AVSP'08)*. ISCA, Queensland, Australia, 185–190.

[7] Sanaul Haq, Philip JB Jackson, and J Edge. 2009. Speaker-dependent audio-visual emotion recognition.. In *Auditory-Visual Speech Processing (AVSP'09)*. ISCA, Norwich, UK, 53–58.

[8] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. 2019. Streaming end-to-end speech recognition for mobile devices. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*. IEEE, Brighton, 6381–6385.

[9] Brandi House, Jonathan Malkin, and Jeff Bilmes. 2009. The VoiceBot: a voice controlled robot arm. In *The ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, Boston, USA, 183–192.

[10] Patrik N Juslin, Petri Laukka, and Tanja Bänziger. 2018. The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior* 42, 1 (2018), 1–40.

[11] Krishna Mohan Kudiri, Abas Md Said, and M Yunus Nayan. 2016. Human emotion detection through speech and facial expressions. In *The International Conference on Computer and Information Sciences (ICCOINS'16)*. IEEE, Kuala Lumpur, Malaysia, 351–356.

[12] S Lalitha and Shikha Tripathi. 2016. Emotion detection using perceptual based speech features. In *The IEEE Annual India Conference (INDICON'16)*. IEEE, Bangalore, India, 1–5.

[13] Adi Lausen and Kurt Hammerschmidt. 2020. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–17.

[14] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. 2004. Emotion recognition based on phoneme classes. In *The International Conference on Spoken Language Processing(ICSLP'04)*. ISCA, Jeju Island, Korea.

[15] Ruru Li, Dali Yang, Xinxing Li, Renyu Wang, Mingxing Xu, and Thomas Fang Zheng. 2016. Relative entropy normalized Gaussian supervector for speech emotion recognition using kernel extreme learning machine. In *The Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA'16)*. IEEE, Jeju, Korea (South), 1–5.

[16] Wootaek Lim, Daeyoung Jang, and Taejin Lee. 2016. Speech emotion recognition using convolutional and recurrent neural networks. In *The Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA'16)*. IEEE, Jeju, Korea (South), 1–4.

[17] Endang Wahyu Pamungkas. 2019. Emotionally-aware chatbots: A survey. *arXiv preprint arXiv:1906.09774* abs/1906.09774 (2019).

[18] Marc D Pell, Laura Monetta, Silke Paulmann, and Sonja A Kotz. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior* 33, 2 (2009), 107–120.

[19] Marc D Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja A Kotz. 2009. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics* 37, 4 (2009), 417–435.

[20] Isidoros Perikos and Ioannis Hatzilygeroudis. 2013. Recognizing emotion presence in natural language sentences. In *The International conference on engineering applications of neural networks*. Springer, Halkidiki, Greece, 30–39.

[21] Rosalind W Picard. 2000. *Affective computing.* MIT press, Boston, USA.

[22] S Revathy et al. 2020. Health Care Counselling Via Voicebot Using Multinomial Naive Bayes Algorithm. In *The International Conference on Communication and Electronics Systems (ICCES'20)*. IEEE, Coimbatore, India, 1063–1067.

[23] Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 5 (2018), 90–99.

[24] Masaaki Takebe, Kazumasa Yamamoto, and Seiichi Nakagawa. 2016. Investigation of glottal features and annotation procedures for speech emotion recognition. In *The Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA'16)*. IEEE, Jeju, Korea (South), 1–4.

[25] Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, and Andreas Butz. 2020. How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376877

[26] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, New York,USA, 196–202.

[27] Jianfeng Zhao, Xia Mao, and Lijiang Chen. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 47 (2019), 312–323.