You Sound Like Your Counterpart: Interpersonal Speech Analysis

Jing Han[1], Maximilian Schmitt[1], and Björn Schuller[1,2]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] GLAM – Group on Language, Audio & Music, Imperial College London, U. K.
jing.han@informatik.uni-augsburg.de

Abstract. In social interaction, people tend to mimic their conversational partners both when they agree and disagree. Research on this phenomenon is complex but not recent in theory, and related studies show that mimicry can enhance social relationships, increase affiliation and rapport. However, automatically recognising such a phenomenon is still in its early development. In this paper, we analyse mimicry in the speech domain and propose a novel method by using hand-crafted low-level acoustic descriptors and autoencoders (AEs). Specifically, for each conversation, two AEs are built, one for each speaker. After training, the acoustic features of one speaker are tested with the AE that is trained on the features of her counterpart. The proposed approach is evaluated on a database consisting of almost 400 subjects from 6 different cultures, recorded in-the-wild. By calculating the AE's reconstruction errors of all speakers and analysing the errors at different times in their interactions, we show that, albeit to different degrees from culture to culture, mimicry arises in most interactions.

Keywords: Affective computing · Conversation analysis · Computational paralinguistics.

1 Introduction

Research in psychology has shown that people unconsciously mimic their counterpart in social interaction, which can be operationalised in varying ways including mimic posture, facial expressions, mannerisms, and other verbal and nonverbal expressions [5]. Moreover, the automatic detection of temporal mimicry behaviour can serve as a powerful indicator of social interaction, e. g., cooperativeness, courtship, empathy, rapport, and social judgement [12].

The previous works focus on automatically detecting mimicry behaviours particularly from head nod and smile, i. e., visual cues[3,23]. In this work, we focus on the acoustic side, given that in social interaction, people mimic others not only by body language, but also in their speech. To the best of our knowledge, this is the first time that identical behaviour is analysed from speech over different cultures in empirical research, though previous works exist where similar topics have been studied in theory [4]. As there is limited related works in this specific topic, we first utilised low-level descriptors (LLDs) such as log-energy, and pitch, and measured the similarities over each conversation turn, but hardly found any obvious trend in these descriptors.

Thus, we propose an autoencoder-based framework to leverage the power of machine learning. Specifically, for each interaction, two autoencoders (AEs) are trained on speech from two subjects A and B, respectively. Then, once the training procedure is done, the instances are exchanged and fed into the two autoencoders again, i. e., A is evaluated on the AE trained by data from B while B is evaluated on the AE trained by data from A. This goes under the hypothesis that, when a subject tends to behave similarly to her counterpart, the reconstructed features from the AE trained with her counterpart's data should have a decreasing error along time.

In the following Section, the related work is summarised both from a sociological and a technical perspective. In Section 3, we describe the data and acoustic features used in our research. In Section 4, we explain the experiments and present the results, before concluding in Section 5.

2 Related Work

Mimicry behaviour can be categorised into two different groups: *emotional mimicry* and *motor mimicry* [13]. While the first describes mimicry in the underlying affective state, such as, *happiness* or *sadness*, the latter considers only imitation of physical expressions, such as, e. g., raising an eye-brow or nodding the head. As can be expected, motor mimicry is much easier to detect than emotional mimicry, given that physical expressions can be classified quite objectively by a human observer and also by automated tools. In the late 1970s, Friesen and Ekman proposed the 'Facial Action Coding System' [11] based on so-called *facial action units* (FAUs). FAUs describe 44 different activations of facial muscles, resulting in a certain facial expression, e. g., 'raising eye brow', 'wrinkling nose', or 'opening mouth'. However, several FAUs can be combined and be active at the same time. Ekman and Friesen have also shown that, there is a strong relationship between FAUs and affective states [8] and that those relationships are largely universal despite there are some differences between cultures [7]. FAUs and head movements can be robustly recognised with state-of-the-art tools, such as OPENFACE [2].

Motor mimicry is a means of persuasion in human-to-human interaction, by conforming to the other's opinions and behaviour [13]. Humans are susceptible to mimic behaviours through both the audio and the visual domain [16]. Although mimicry is usually found in interactions both when subjects disagree with each other and when they do not, there are more mimicry interactions where people agree [23]. Moreover, it has been shown that there is usually a tendency to adopt gestures, postures, and behaviour of the chat partner over time during the conversation [5,6].

From the methodological point of view, for the automatic detection of behavioural mimicry, a temporal regression model has been proposed by Bilakhia et al. predicting audio-visual features of the chat partner using a deep recurrent neural network [3]. An ensemble of models has been trained for each class (mimicry / non-mimicry) and the ensemble providing the lowest reconstruction error determined the class. *Mel-frequency cepstral coefficients* have been employed as acoustic features and *facial landmarks* as visual features.

Compared to motor mimicry, emotional mimicry has been studied much less. However, it has been found that the tendency to mimic others' behaviour is much less valid from the emotional perspective [14]. The extent of emotional mimicry highly depends on the social context and emotional mimicry is not present if people do not like each other or each other's opinion. Scissors et al. found the same analysing the linguistic behaviour [21]. They observed that in a text-based chat system, within-chat mimicry (i. e., repetition of words or phrases) is much higher in chats where subjects trusted each other than in chats with a low level of trust. Furthermore, it was found that linguistic mimicry has a positive effect on the outcome of negotiations [24].

3 Dataset and Features

Our experiments are based on the SEWA corpus of audio-visual interaction in-the-wild[3]. Hand-crafted acoustic features have been extracted on the frame-level from the audio of all chats.

3.1 SEWA Video Chat Dataset

In the SEWA database, 197 conversations have been recorded from subjects of six different cultures (Chinese, Hungarian, German, British, Serbian, and Greek). Table 1 summarises the number and total duration of conversations for each culture. The number of subjects is always twice the number of conversations. In these conversations, each lasting up to 3 minutes, a pair of subjects from the same culture discussed about an advertisement they just watched beforehand on a web platform. Figure 1 illustrates a screenshot of one dyadic conversation. The commercial seen beforehand was a 90 seconds long video clip advertising a (water) tap.

All subjects were recorded in an 'in-the-wild setting', i. e., using the subjects' personal desktop computers or notebooks and recording them either at their homes or in their offices. The chat partners always knew each other beforehand (either family, friends, or colleagues) and were balanced w. r. t. gender constellations (female-male, female-female, male-male). Subjects with an age ranging from 18 to older than 60 are included in the database. The dialogues had to be held in the native language of the chat partners, but there were no restrictions concerning the exact aspects to be discussed during their chat about the commercial. Conversations showed a large variety of emotions and levels of agreement/disagreement or rapport. The SEWA corpus has been used as the official benchmark database in the 2017 and 2018 Audio-Visual Emotion Challenges (AVEC) [17,18].

3.2 Acoustic Features

We use the established ComParE feature set of acoustic features [9]. For each audio recording, we capture the acoustic low-level descriptors (LLDs) with the openSMILE toolkit [10] at a step size of 10 ms. The ComParE LLDs extracted on frame-level

[3] https://sewaproject.eu/

Table 1. SEWA corpus: Number of conversations and subjects and total duration for each culture.

Index	Culture	# Conversations	# Subjects	Total duration [min]
C1	Chinese	35	70	101
C2	Hungarian	33	66	67
C3	German	32	64	89
C4	British	33	66	94
C5	Serbian	36	72	98
C6	Greek	28	56	81
Sum		197	394	530

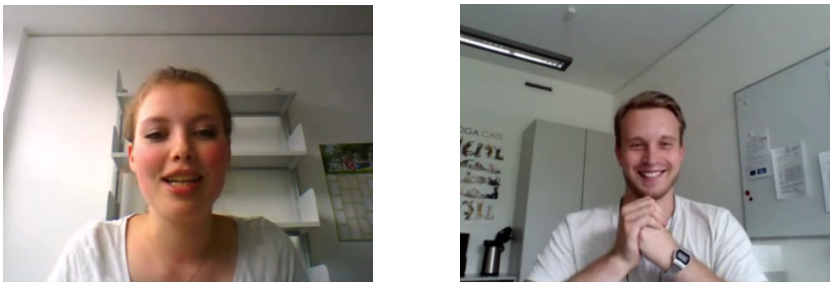

Fig. 1. SEWA corpus: Screenshot taken from a sample video chat with one female and one male German subject.

have been introduced at the *Interspeech 2013 Computational Paralinguistics Challenge (ComParE)* [20]. However, the *functionals* defined in the feature set, i. e., the statistics summarising the LLDs on utterance level, are not applied in this work, as we are interested in the time-dependent information on frame-level. COMPARE comprises 65 LLDs summarised in Table 2, covering spectral, cepstral, prosodic, and voice quality information, extracted from a frame with a size of 20 ms to 60 ms length. In addition, the first order derivatives (deltas) are computed, resulting in a frame-level feature vector of size 130 for each step of 10 ms.

4 Behaviour Similarity Tendency Analysis with Autoencoder

To analyse the interpersonal sentiment and investigate the temporal behaviour patterns from speech, we first standardised (zero mean and unit standard deviation) the 130 frame-level features within the same recordings to minimise the differences between different recording conditions. This procedure turned these LLDs into suitable ranges, as the inputs and target outputs of an autoencoder (AE). Before training the AE, we first segmented the LLD sequences based on the transcriptions provided in the SEWA database, where information on the start and end of each speech segment and the subject ID of the corresponding segment is given. After that, the whole LLD sequences of each recording were divided into two sub-sequences, each including features from only one subject.

Table 2. Interspeech 2013 Computational Paralinguistics Challenge (ComParE) feature set. Overview of 65 acoustic low-level descriptors (LLDs). RMS: Root-Mean-Square, RASTA: RelAtive SpecTral Amplitude, MFCC: Mel-Frequency-Cepstral Coefficients, SHS: Sub-Harmonic Summation.

4 energy related LLD	**Group**
Loudness	Prosodic
Modulation loudness	Prosodic
RMS energy, zero-crossing rate	Prosodic
55 spectral related LLD	**Group**
RASTA auditory bands 1–26	Spectral
MFCC 1–14	Cepstral
Spectral energy 250-650 Hz, 1–4 kHz	Spectral
Spectral roll-off pt. .25, .50, .75, .90	Spectral
Spectral flux, entropy, variance	Spectral
Spectral skewness and kurtosis	Spectral
Spectral slope	Spectral
Spectral harmonicity	Spectral
Spectral sharpness (auditory)	Spectral
Spectral centroid (linear)	Spectral
6 voicing related LLD	**Group**
F_0 via SHS	Prosodic
Probability of voicing	Voice quality
Jitter (local and delta)	Voice quality
Shimmer	Voice quality
Log harmonics-to-noise ratio	Voice quality

Following the above-mentioned separation process, features from one subject were utlised to train an AE, and features from the other subject in the same recording were fed into the trained AE for testing. Furthermore, once all features for testing have been reconstructed with the AE, we calculate the root-mean-squared errors (RMSEs) of the reconstructed features over time, and examine how and to which extent the RMSE varies along time. Consequently, for each recording, two AEs are learnt based on the two subjects involved in the recording, resulting in two one-dimensional RMSE sequences calculated between the input and the output feature sequences during the testing step.

4.1 Experimental Settings

The AE we applied is a 3-layer encoder with a 3-layer decoder. In the preliminary experiments, the number of nodes in each layer has been chosen as follows: 130-64-32-12-32-64-130, where the output dimension is exactly the same as the input dimension. During network training, the network weights were updated by using mean squared error loss and the Adagrad optimizer, and the training process was ceased after 512 epochs. Furthermore, to accelerate the training process, the network weights were updated after running every batch of 256 LLDs for computation in parallel. The training procedure was performed with Keras, which is a deep learning library for python.

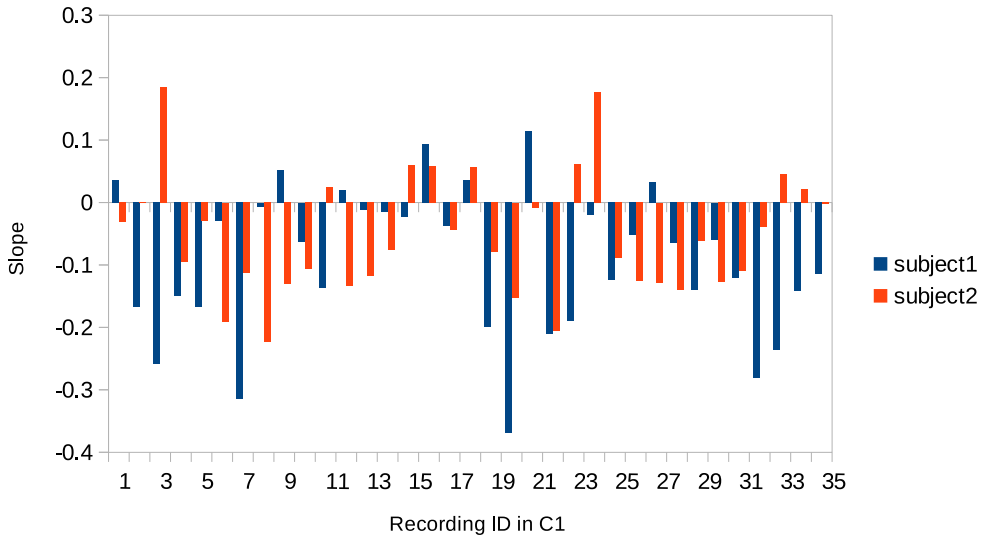

Fig. 2. Slope of RMSE sequences of 70 Chinese subjects from 35 recordings. In each recording, there are two subjects as denoted with blue and red bars, respectively.

After generating the reconstruction errors of the tested subject over time, the resulting sequence is exploited to perform a linear regression, with the assumption that the slope of the learnt line may indicate the changes of the behaviour patterns along time. More specifically, when the slope is negative, it may demonstrate that during the chat session, the tested subject turns to become more similar to the subject who (s)he is talking to. Thence, if the slope is positive, it may imply the opposite. Additionally, the amplitude of the slope can be an indication to denote the level of the similarity or dissimilarity.

4.2 Results and Discussion

We first discuss the results achieved with the data from the first culture, Chinese (C1). From all 35 recordings, the average slope of the RMSE sequences of all 70 subjects is -0.07. From Figure 2, we notice that, most of the slopes (54 of 70) are negative, whereas only a few (16 of 70) are positive. This indicates that, during the recordings, the acoustic LLD features of the tested subjects have a smaller reconstruction error when time passes by. Considering that the AE is trained with the other subject within the same recording, a smaller reconstruction error may reveal a higher similarity between these two subjects. To sum up, a negative slope implies a decreasing reconstruction error along the time and could indicate a similarity increasing among the speakers during the video chat. Interestingly, similar patterns have also been found in all the other five cultures. Nevertheless, the ratio of the negative slopes and the average slope are different from culture to culture.

Given these results, we calculated the average slopes s of all cultures separately, as well as the Pearson correlation coefficients (PCCs) of two slopes obtained from all recordings within the same culture, respectively, with the aim to perceive cultural variation in spontaneous remote conversations. Results are given in Table 3. Note that, a negative slope denotes that the subject shows a more similar speech behaviour in a conversation along the time; the more similar a subject is speaking like his partner, the larger the slope is towards the negative direction. From Table 3, one may notice

Table 3. Average slope of RMSE sequences of all subjects within six different cultures is listed in the upper row, respectively; the correlation coefficient denoted as $pcc\ of\ pairs$ indicates the correlation of behaviours of two subjects and is listed in the last row for each culture (C1: Chinese, C2: Hungarian, C3: German, C4: British, C5: Serbian, and C6: Greek).

	C1	C2	C3	C4	C5	C6
average slope	-0.07	-0.11	-0.10	-0.07	-0.08	-0.12
pcc of pairs	-0.03	0.34	0.15	0.39	0.39	-0.26

that on average, individuals of all six cultures tend to behave more similar during the conversation, given that the average slopes are all negative. However, cultural variation remains, as the most negative slope (-0.12) is obtained for the Greek (C6) culture and the smallest slope (-0.07) is seen for Chinese (C1) and British (C4).

Moreover, taking the PCC into account, we may see the cultural variation from another view. A positive PCC value demonstrates that subjects of a culture tend to converge to a similar state, either both behave like or unlike each other, while a negative PCC may indicate that conversations are more likely to be dominated by one subject. For example, no correlation has been seen in conversations of the Chinese pairs (C1), with a PCC of -0.03, which is close to 0. However, strong linear correlations have been revealed in four cultures, either positive (Hungarian (C2), British (C4), and Serbian (C5)) or negative (Greek (C6)). Besides, a weak positive correlation can be seen in German (C3). These findings need to be verified by literature in sociology, anthropology, and in the anthropologic linguistics domain, particularly in the field of conversation analysis [22], which is, however, out of the scope of this work. Note that, despite that the SEWA database was designed and developed with a control of age and gender of the subjects, discrepancies caused by these or other aspects such as educational background, occupation, and health status cannot be avoided and might still may have an impact on our observations.

5 Conclusion and Outlook

In this work, we have demonstrated that, an autoencoder has a great potential to recognise the spontaneous and unconscious mimicry in the social interaction, by the observation of the reconstruction error using the acoustic features extracted from the speech of a conversational partner. We have given some insights into the synchronisation of vocal behaviour in dyadic conversations of people from six different cultures. Future work

will focus on optimised feature representations, such as *bag-of-audio-words* [19] or learnt features such as *auDeep* [1]. Moreover, we are going to exploit also the linguistic domain through state-of-the-art word embeddings, such as *word2vec* [15]. Lastly, other than the slope of the reconstruction errors, additional evaluation strategies to measure the degree of similarity of similarity between subjects will be explored in the future [6].

Acknowledgement


The research leading to these results has received funding from the European Union's Horizon 2020 Programme under GA No. 645094 (Innovation Action SEWA) and through the EFPIA Innovative Medicines Initiative under GA No. 115902 (RADAR-CNS).

References

1. Amiriparian, S., Freitag, M., Cummins, N., Schuller, B.: Sequence to sequence autoencoders for unsupervised representation learning from audio. In: Proc. of the DCASE 2017 Workshop. Munich, Germany (2017)
2. Baltrušaitis, T., Robinson, P., Morency, L.P.: OpenFace: An open source facial behavior analysis toolkit. In: Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–10. Lake Placid, NY (2016)
3. Bilakhia, S., Petridis, S., Pantic, M.: Audiovisual detection of behavioural mimicry. In: Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII). pp. 123–128. Geneva, Switzerland (2013)
4. Burgoon, J.K., Hubbard, A.E.: Cross-cultural and intercultural applications of expectancy violations theory and interaction adaptation theory. Theorizing about Intercultural Communication pp. 149–171 (2005)
5. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception–behavior link and social interaction. Journal of Personality and Social Psychology **76**(6), 893–910 (June 1999)
6. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing **3**(3), 349–365 (July 2012)
7. Ekman, P.: Universals and cultural differences in facial expressions of emotion. In: Nebraska Symposium on Motivation. University of Nebraska Press (1971)
8. Ekman, P., Friesen, W.V.: Unmasking the face: A guide to recognizing emotions from facial clues. Ishk (2003)
9. Eyben, F.: Real-time speech and music classification by large audio feature space extraction. Springer (2016)
10. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proc. of the 21st ACM International Conference on Multimedia (MM). pp. 835–838. Barcelona, Spain (2013)
11. Friesen, E., Ekman, P.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press (1978)
12. Gueguen, N., Jacob, C., Martin, A.: Mimicry in social interaction: Its effect on human judgment and behavior. European Journal of Social Sciences **8**(2), 253–259 (2009)
13. Hess, U., Fischer, A.: Emotional mimicry as social regulation. Personality and Social Psychology Review **17**(2), 142–157 (2013)

14. Hess, U., Fischer, A.: Emotional mimicry: Why and when we mimic emotions. Social and Personality Psychology Compass **8**(2), 45–57 (2014)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS). pp. 3111–3119. Lake Tahoe, NV (2013)
16. Parrill, F., Kimbara, I.: Seeing and hearing double: The influence of mimicry in speech and gesture on observers. Journal of Nonverbal Behavior **30**(4), 157 (2006)
17. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Çiftçi, E., Güleç, H., Salah, A.A., Pantic, M.: AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In: Proc. of the 8th Annual Workshop on Audio/Visual Emotion Challenge. Seoul, Korea (2018), to appear
18. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M.: AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In: Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge. pp. 3–9. Mountain View, CA (2017)
19. Schmitt, M., Ringeval, F., Schuller, B.: At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In: Proc. of INTERSPEECH. pp. 495–499. San Francisco, CA (2016)
20. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proc. of INTERSPEECH. pp. 148–152. Lyon, France (2013)
21. Scissors, L.E., Gill, A.J., Gergle, D.: Linguistic mimicry and trust in text-based cmc. In: Proc. of the ACM Conference on Computer Supported Cooperative Work. pp. 277–280. San Diego, CA (2008)
22. Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.E., Levinson, S.C.: Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences of the United States of America **106**(26), 10587–10592 (June 2009)
23. Sun, X., Nijholt, A., Truong, K.P., Pantic, M.: Automatic visual mimicry expression analysis in interpersonal interaction. In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 40–46. Colorado Springs, CO (2011)
24. Swaab, R.I., Maddux, W.W., Sinaceur, M.: Early words that work: When and how virtual linguistic mimicry facilitates negotiation outcomes. Journal of Experimental Social Psychology **47**(3), 616–621 (2011)