Human vs. Computer Performance in Voice-Based Recognition of Interpersonal Stance

Daniel Formolo and Tibor Bosse(✉)

Department of Computer Science, Vrije Universiteit Amsterdam,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
{d.formolo,t.bosse}@vu.nl

Abstract. This paper presents an algorithm to automatically detect interpersonal stance in vocal signals. The focus is on two stances (referred to as 'Dominant' and 'Empathic') that play a crucial role in aggression de-escalation. To develop the algorithm, first a database was created with more than 1000 samples from 8 speakers from different countries. In addition to creating the algorithm, a detailed analysis of the samples was performed, in an attempt to relate interpersonal stance to emotional state. Finally, by means of an experiment via Mechanical Turk, the performance of the algorithm was compared with the performance of human beings. The resulting algorithm provides a useful basis to develop computer-based support for interpersonal skills training.

Keywords: Emotion recognition · Voice · Interpersonal stance · Experiments

1 Introduction

Conversation is an important way to transmit information between individuals. However, the content of what is said only determines one-third of the information that is communicated. The rest involves prosody, gestures, facial expressions, and others signals [1], which convey much information of speaker`s emotions [2]. In situations where physical presence is not required, prosody is an important source to make the partners aware of the emotional state of the speaker, what is crucial to follow social conventions and to coordinate the interaction [1].

So, for human beings, prosody can be a useful cue to infer information about the socio-emotional state of others. But also for computers, the ability to recognise this in (human) conversation partners is a useful feature, because systems with this ability may reduce user frustration, facilitate more natural communication, resulting in more effective applications [3]. Examples of applications that may benefit from the ability to recognise emotion in the user's behaviour are learning environments for social skills training, therapeutic applications for autistic patients, and entertainment games.

This research described in this paper was triggered by a larger research endeavour that aims to develop a simulation-based training system for professionals that are often confronted with aggressive behavior, with an emphasis on public transport employees [4]. For such a system, being able to recognise socio-emotional cues in the user's voice is very

© Springer International Publishing AG 2017
M. Kurosu (Ed.): HCI 2017, Part I, LNCS 10271, pp. 672–686, 2017.
DOI: 10.1007/978-3-319-58071-5_51

important, as the tone of voice is an important factor in successful aggression de-escalation [5].

Specifically, when it comes to aggression de-escalation, the notion of *interpersonal stance* plays an important role. Interpersonal stance can be seen as the relative position speakers take in relation to the ongoing conversation [6]. To successful de-escalate aggressive behaviour, the stance of the de-escalator should depend on the type of aggression shown by the aggressor. Here, the difference between reactive and proactive aggression is very important: if the interlocutor shows reactive aggression (i.e. aggression caused by frustration of the person's own goals), the best solution is to show an *empathic* response. Instead, if the interlocutor shows proactive aggression (i.e., aggression used as a means to satisfy one's own goals), the best solution is to show a more *dominant* response [5]. Hence, for an effective training tool, it is important if the system is able to distinguish empathic from dominant features in the user's behaviour.

In the current paper, this is realised by developing an algorithm that detects empathy and dominance in vocal signals, based on the OpenSmile toolkit [7]. This was done by creating a database with 1383 samples from 8 speakers from different countries. In addition to creating the recognition algorithm for interpersonal stance, these samples are analysed in more detail, to gain more insight into the nature of empathic and dominant speech (and in particular, in their relation to emotion). Finally, by means of an experiment via Amazon's Mechanical Turk, the performance of the algorithm is compared with the performance of human beings.

The remainder of this article is structured as follows. Section 2 discusses some background information about aggression de-escalation as well as the concepts 'empathy' and 'dominance'. Section 3 describes the algorithm to detect interpersonal stance in vocal signals, as well as an analysis of its performance. Section 4 presents the experiment that was conducted to compare the performance of the algorithm with the performance of human beings. Section 5 presents a conclusion of the research.

2 Background

2.1 Aggression De-escalation

Aggressive behaviour may be caused by a variety of factors. Hence, in order to de-escalate aggression, it is important to recognize the type of aggression that the interlocutor is showing, and subsequently, to show the appropriate communication style (or interpersonal stance) in your own behaviour.

With respect to the type of aggression, two main categories are distinguished in the literature: aggression can be either *reactive* (or *emotional*) or *proactive* (or *instrumental*) [8]. In case of reactive aggression, the aggressive behaviour is typically a response to a negative event that frustrates a person's desires [9]. Such a person is likely to become angry with respect to whatever stopped him or her from achieving his goal. For example, a client may become very aggressive against a desk employee who tells him that their product is sold out, even if this employee cannot do anything about this.

In contrast, proactive aggression refers to aggressive behaviour that is used 'instrumentally', i.e., to achieve a certain goal. Such behaviour is not a direct response to a negative

event and is less strongly related to heavy emotions. Instead, it typically is a more planned and calculated type of aggression, often in the form of intimidation. For example, a child may start bullying his classmates because he wants to have power over them.

Based on observations in animals, it has been proposed that reactive aggression is 'hot-blooded', and that proactive aggression is 'cold-blooded'. As a result, the difference between both types of aggression can be recognized (besides looking at the context) by closely observing the non-verbal behaviour of the aggressive individual. The reason for this is that reactive aggression is usually paired with a lot of physiological and behavioural arousal indicating negative affect (such as a flushed face, gestures, and fast speech). Instead, proactive aggression comes with fewer signs of anxiety, but with cues like a dominant posture, slower speech, and sometimes even a smile [8].

Because of the different nature of both types of aggression, it takes a very different approach to deal with each of them. When dealing with an emotional aggressor, supportive behaviour from the de-escalator is required, in order to reduce the aggressor's level of arousal. This can be achieved for instance by ignoring the conflict-seeking behaviour, calmly making contact with the aggressor, actively listening to what he has to say, showing empathy, and suggesting solutions to his problems (see [5]).

Instead, to de-escalate instrumental aggression, showing too much empathy will only make the situation worse, as the aggressor will be reinforced in his belief that his deviant behaviour pays off. In such a case, a directive response is assumed to be more effective. This means that it is necessary to show the aggressor that there is a limit to how far he can pursue his aggressive behaviour and to make him aware of its consequences [10].

A summary of the differences between reactive and proactive aggression is shown in Table 1. Although there is some debate in the literature about whether these two types are really disjoint, in practice, it is often useful to treat them as such because it gives clear guidelines on how to act in confrontations with aggressive individuals[1].

Table 1. Reactive versus proactive aggression.

	Reactive aggression	Proactive aggression
Synonym	Emotional aggression ('hot-blooded')	Instrumental aggression ('cold-blooded')
Underlying mechanism	Behaviour influenced by an emotional state, resulting from the frustration of own goals	Planned, learned behaviour, using intimidation as a means to achieve own goals
How to recognize?	Context, flushed face, gestures, fast speech, …	Context, dominant posture, slower speech, smiling, …
How to de-escalate?	Ignore conflict-seeking behaviour, show empathy, help solve problem	Draw a line, confront aggressor with behaviour, point out consequences

[1] For instance, discussions with domain experts in public transport pointed out that the ability to distinguish between reactive and proactive aggression is a key element in their training program.

2.2 Interpersonal Stance

As mentioned earlier, our long term goal is to develop a simulation-based training system for professionals that are often confronted with aggressive behaviour. Inspired by the literature discussed in the previous section, this system is centred around two main learning goals, namely (1) recognizing the type of aggression of the conversation partner (i.e., reactive or proactive), and (2) selecting the appropriate interpersonal stance towards the conversation partner. Regarding this interpersonal stance, guidelines for how to act are summarized in the last row of Table 1. In the remainder of this paper, we will use the terms *Empathic* and *Dominant* to refer to the appropriate styles to deal with reactive and proactive behaviour, respectively.

We envision a training tool that is able to analyse the behaviour of the trainee, and automatically distinguishes Empathic cues from Dominant cues, in particular with respect to the vocal signals of the trainee's speech. This allows the system to provide feedback on the extent to which the trainee shows the appropriate interpersonal stance to de-escalate a particular aggressive situation. As a first step towards this system, the current paper aims to develop an algorithm that detects Empathy and Dominance in a user's voice. The algorithm will make use of the OpenSmile toolkit [7]. To obtain training data, a number of speakers from different countries will be asked to speak sentences by using either an Empathic or a Dominant interpersonal stance.

To define the concepts of Empathic and Dominant in an unambiguous way, they are related to specific points in circumplex [11]. This theory, also known as Leary's Rose, is often used in training interpersonal skills, and assumes that interpersonal behaviour can be represented as a point in a two-dimensional space determined by the dimensions affiliation (positive versus hostile, or 'together versus opposed') and power (dominant versus submissive, or 'above versus below'); see Fig. 1. Based on discussion with experts in aggression de-escalation training, we define the communicative style called *Empathic* as the border of the categories 'Helping' and 'Cooperative', and the style called *Dominant* as the border of the categories 'Competitive' and 'Leading'.

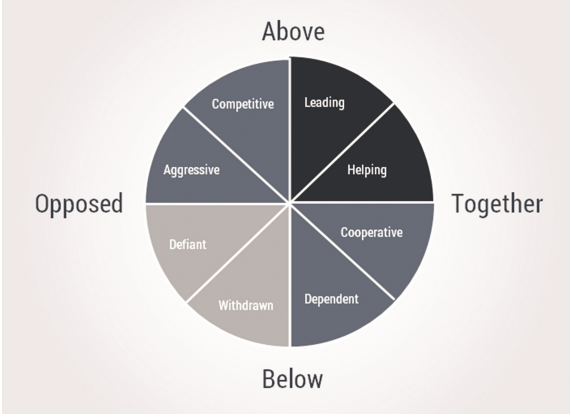

Fig. 1. Leary's Rose.

In addition to developing the recognition algorithm, we aim to gain more insight in the very nature of 'Empathic' and 'Dominant' behaviour, respectively. Even though the distinction seems to be evident based on the explanation above, it may be the case that different individuals interpret these terms rather differently, and show large differences in behaviour when they are asked to produce Empathic and Dominant speech. So, another goal we have is to investigate whether we can identify some common features in people's speech when they are asked to produce Empathic and Dominant behaviour, respectively. To this end, we will relate the Empathic and Dominant samples to the Arousal-Valence circumplex of affect[2] [12], which is an accepted theory in the literature on emotion. This theory views emotions as states that can be represented as points within a continuous space defined by two dimensions, namely valence (i.e., the level of pleasure) and arousal (i.e., a general degree of intensity).

Next, we are interested in the performance of the interpersonal stance recognition algorithm compared to the performance of humans: are computers better in distinguishing Empathic from Dominant behaviour in vocal signals than human beings are? And related to this, is Empathic behaviour more difficult to recognize than Dominant behaviour? To this end, an experiment via Amazon's Mechanical Turk will be set up, in which users are asked to listen to the same samples as used to train the recognition algorithm and classify them as either Empathic or Dominant.

Finally, recognizing the style (the 'how') of an utterance may interfere with processing the content (the 'what') of what is said, which may or may not be consistent with the style. Therefore, an interesting question is whether it makes a difference if people understand what is being said. To this end, we also investigate if people's performance correlates with whether or not the spoken fragments are in a language that they understand.

To summarize, this paper aims to address the following main questions:

1. Can we develop an algorithm that automatically distinguishes Empathy and Dominance in a user's vocal signals?
2. Is it possible to relate the concepts of 'Empathic' and 'Dominant' to specific positions in the Arousal-Valence circumplex?
3. How does the performance of the algorithm compare with the performance of human listeners?
4. Are Dominance and Empathy equally difficult to recognize (both for humans and computers)?
5. Is the performance by human listeners influenced by whether or not they understand the language of the spoken fragments?

3 Automated Recognition of Interpersonal Stance

In order to answer the questions described above, we developed an algorithm to identify Dominance and Empathy in vocal signals. The algorithm is pluggable into a variety of

[2] Note that there is an important difference between the interpersonal circumplex and the arousal-valence circumplex: the former is assumed to address interpersonal stance, whereas the latter addresses individual emotions (independent of stance).

human-computer interaction applications, enabling a system to classify the user's voice in real time. In the following sub-sections, the approach to develop the algorithm is explained, as well as its results. In Sect. 4, its performance is compared with that of humans.

3.1 Training and Test Dataset

The data to train and test the algorithm were collected from 8 different people, in 2 different languages. Four of the participants (2 male and 2 female) are native Dutch speakers from The Netherlands, and the other four (2 male and 2 female) are native Portuguese speakers from Brazil. One Dutch male and one Dutch female were professional trainers from the public transport company in Amsterdam, and had expertise in aggression de-escalation. Because the languages Dutch and Portuguese origin from different roots, they have different prosody characteristics, as described in [13]. This was done to prevent the algorithm from overfitting to one single style of prosody, and to check if there are differences between languages in terms of how people produce Dominant and Empathic speech.

For each participant, at least 50 Dominant voice samples and 50 Empathic voice samples were recorded. The first participants recorded 100 samples of each type, and after a preliminary validation, it was noted that the algorithm is able to learn the required patterns with less than 50 samples. Hence, it was decided to record 50 samples for the rest of the participants. All voices were recorded using the same microphone, following the same procedure. Before the start, our interpretation of the terms Empathy and Dominance was explained to all participants (i.e., their positions in Leary's Rose as explained in Sect. 2), and examples of sentences were provided. These sentences addressed situations in public transport in which either an Empathic or Dominant stance could be used (e.g., 'You are not allowed to bring hot coffee in this tram'). All participants started recording the sentences with an Empathic style, after which they recorded the same sentences with a Dominant style.

Table 2 shows an overview of the samples that were recorded. Some people recorded more samples than others and some samples were discarded because they lasted less than 3 s, which makes it difficult to extract features from the voice. This dataset was used to train and test the recognition algorithm, but also in a follow-up experiment to test the performance of human listeners (as explained in Sect. 4).

Table 2. List of vocal samples recorded.

Nationality/Language	Genre	Type	Person 1	Person 2
The Netherlands/Dutch	Male	Empathic	114	100
		Dominant	124	100
	Female	Empathic	96	59
		Dominant	100	59
Brazil/Portuguese	Male	Empathic	111	49
		Dominant	98	50
	Female	Empathic	89	59
		Dominant	101	59

3.2 Recognition Algorithm

The functioning of the recognition algorithm is divided into a training part and a classification part. They run in cycles one after another, starting with the training part. As shown in Fig. 2, the OpenSmile toolkit [7] was used to extract a total of 6552 features from each sample. The extracted features are based on the INTERSPEECH 2010 Paralinguistic Challenge feature set [7]. After that, the non-significant features were removed using the InfoGainAttributeEval algorithm [14]. All features with no information gain were removed, reducing the features to a total of 4959. After that, a Support Vector Machine (SVM) algorithm was trained and a SVM model was generated (through the method described in [15]) to be used for the classification part.

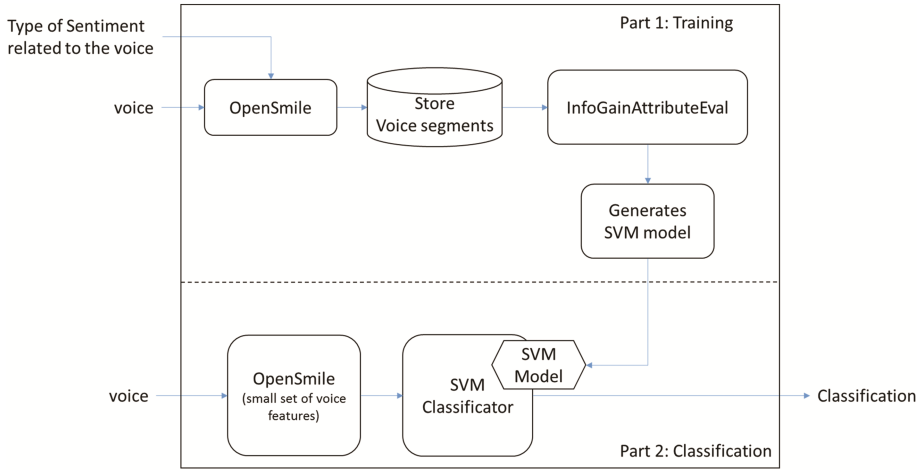

Fig. 2. Block Diagram of the algorithm.

Many algorithms were tested for the task of classifying samples as Empathic or Dominant. Of these algorithms, SVM was selected because it resulted in the best performance. Moreover, it is one of the most frequently used algorithms when it comes to emotion recognition in vocal signals, see for instance [16–18]. The classification part of the algorithm runs for a fixed period of time, classifying and storing the new samples. After that, it returns to training part, in order to train the algorithm again using the old and the new samples and to update the SVM model for a new classification round.

3.3 Performance of the Algorithm

The algorithm was validated with 10-fold cross validation and reached an overall accuracy of 94.58%, with root mean squared error of 0.23. Table 3 shows the results of this evaluation. In addition to the overall accuracy, the accuracy is shown for classifying only the Dominant, Empathic, Dutch, and Portuguese samples. As can be seen, the Empathic samples appeared a bit easier to classify than the Dominant samples. Similarly, the Dutch samples were a bit easier to classify than the Portuguese samples.

Table 3. Computer performance accuracy.

		All samples	Empathic samples	Dominant samples	Dutch samples	Portuguese samples
Algorithm	#samples	1383	823	560	755	628
	Accuracy	94.58%	96.80%	91.60%	96.82%	92.04%

In addition to the above, Table 4 shows a confusion matrix based on the 10-fold cross validation.

Table 4. Confusion Matrix for 10-fold cross validation.

Actual	Predicted		
	Dominant	Empathic	Total
Dominant	**534**	**49**	583
Empathic	**26**	**774**	800
Total	560	823	1383

Finally, the performance of the algorithm was tested by gradually increasing the size of the training set. Figure 3 illustrates how the accuracy increases with a larger number of samples. Notably, there is a substantial increase in accuracy until about 40 samples (reaching 94%), after which the accuracy stabilizes between approximately 96% and 98%.

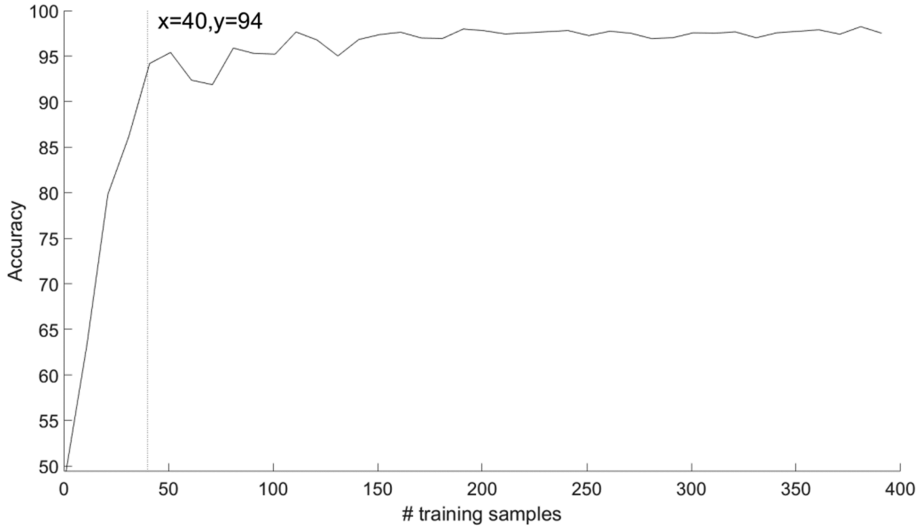

Fig. 3. Performance of the algorithm related to the number of training samples.

Overall, these results are good, which allows us to give an affirmative answer to our ***first research question*** put forward in Sect. 2.

3.4 Relation to the Arousal-Valence Space

To be able to answer our ***second research question***, we changed the SVM model to measure Arousal and Valence. According to many papers [12, 17, 19] it is possible to relate samples of human speech to emotional states, represented as points in the Arousal-Valence space. The question answered in this section is whether there is a pattern when projecting our Empathic and Dominant samples in this A-V space. Figures 4 and 5 show these results for the 8 volunteers that we used to create the dataset. As can be seen, in most cases a clear separation is visible between the Dominant and the Empathic samples, which confirms our finding that it is possible to distinguish between both stances. However, interestingly, the exact pattern differs per individual. For example, the pattern for participant B1 is completely different from the pattern for participant D2. On average, Dominance seems to be correlated with a slightly higher arousal than Empathy, whereas the correlation with Valence seems to depend on the individual.

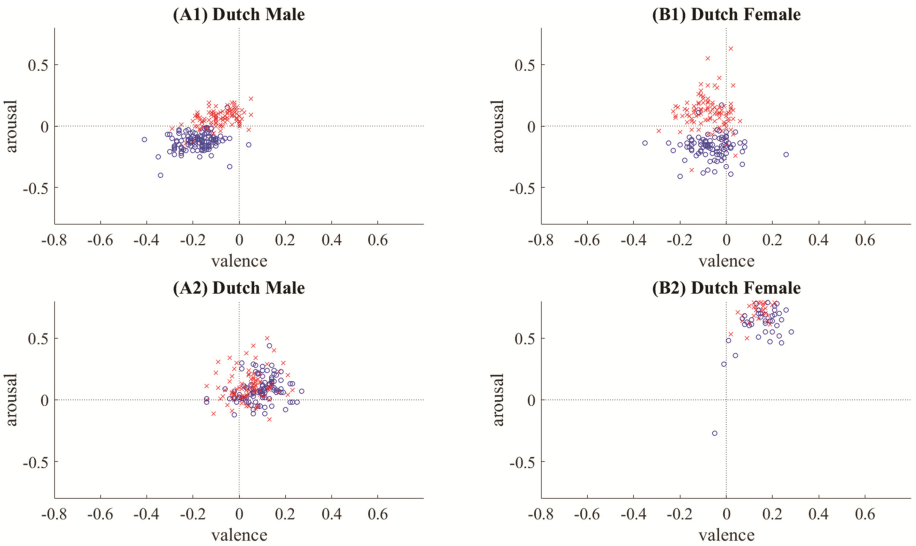

Fig. 4. Arousal and Valence distribution of Dutch samples. Dominant samples are indicated by a cross, whereas Empathic samples are indicated by a circle.

Studying the values of Figs. 4 and 5 in detail, another interesting observation is that the Dutch samples are generally closer to each other, while the Portuguese samples are more spread over the entire spectrum. Possibly, this could be linked to cultural aspects, since all Dutch speakers are from The Netherlands and all Portuguese speakers are from Brazil. Another possible explanation would be that the Portuguese language in itself covers a wider spectrum of prosodic characteristics.

In any case, it is clear from these pictures that there is no direct mapping between the Dominant-Empathic dichotomy and the A-V space. This is interesting, because the algorithm nevertheless showed a good performance in classifying the samples as Dominant or

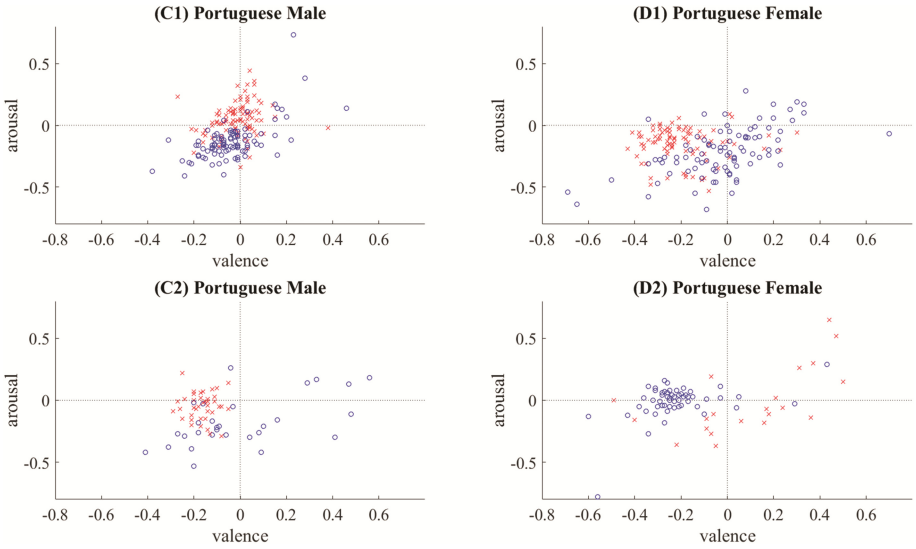

Fig. 5. Arousal and Valence distribution of Portuguese samples. Dominant samples are indicated by a cross, whereas Empathic samples are indicated by a circle.

Empathic, and this algorithm applied the same model to all 8 speakers. So, apparently, it makes use of more low level features that cannot be directly related to the A-V space.

4 Human Recognition of Interpersonal Stance

To be able to compare the performance of the algorithm with the performance of human beings, an experiment via Amazon's Mechanical Turk was set up. The experiment and the results are described in this section.

4.1 Participants and Design

Via Amazon's Mechanical Turk, 88 participants from 3 different countries were recruited: 32 Portuguese speakers that are not able to understand Dutch, 26 Dutch speakers that are not able to understand Portuguese and 30 people from the Philippines that do not understand Portuguese nor Dutch.

Each participant performed the experiment on-line, using the following procedure. They were offered 32 samples which they could hear by clicking a button, and for each of them they were asked to classify them as either Empathic or Dominant. The 32 samples were taken randomly from the database described in Table 2, with the restriction that they included 4 samples from each {language X gender X stance} combination. That is, they were offered 4 samples that were Dutch, Male, and Dominant, and so on.

4.2 Performance of the Human Participants

Table 5 shows the performance of each group, highlighting in bold their overall accuracy. As shown in this table, the overall performance of the human participants is substantially lower than the performance of the algorithm. This provides an answer to our ***third research question*** put forward in Sect. 2.

Table 5. Human performance results in Empathy and Dominance Recognition.

		All samples	Empathic samples	Dominant samples	Dutch samples	Portuguese samples
Dutch Speakers	Accuracy	**69.71%**	73.79%	65.62%	73.31%	66.10%
	Average	18.12	19.18	17.06	19.06	17.18
	σ	3.57	3.00	3.78	2.96	3.87
Port. Speakers	Accuracy	**71.67%**	81.64%	61.71%	76.36%	66.99%
	Average	22.93	26.12	19.75	24.43	21.43
	σ	5.71	3.05	5.96	3.60	6.90
No Dutch and Port. Speakers	Accuracy	**70.52%**	73.95%	67.08%	73.12%	67.91%
	Average	21.15	22.18	20.12	21.93	20.37
	σ	4.28	3.77	4.51	3.78	4.60

Additionally, the differences in performance between the groups seems to be small, which indicates that none of the groups performed significantly better in the classification task. This was confirmed by executing paired t-tests between the groups. Only for classifying the Empathic samples, the Portuguese participants turned out to perform significantly better, with 81.64% accuracy, than the Dutch participants, with 73.97% accuracy ($p < 0.01$) and the participants from the Philippines, with 73.95% accuracy ($p < 0.01$).

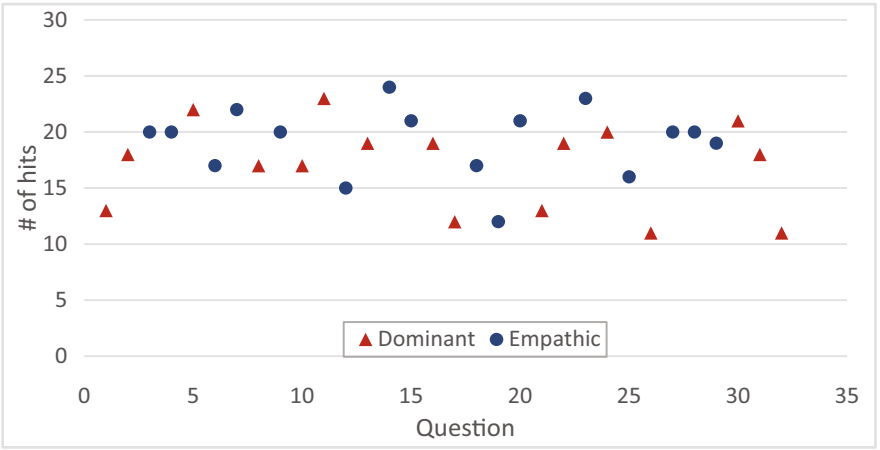

Fig. 6. Number of hits of Dutch participants for each question of the experiment.

We also tested whether the performance of the participants increased over time (i.e., during the experiment). Results about this are shown in Figs. 6, 7 and 8. The question numbers are on the horizontal axis, whereas the number of participants that gave a correct answer is on the vertical axis. These numbers are absolute numbers, which explains why they are highest for the Portuguese group (which was the largest group). The main observation is that in all of these graphs, there are no increasing (or decreasing) trends, what means that the participants did not improve throughout the experiment. This is a fundamental difference with the computer algorithm, which showed a strong improvement in performance until 40 samples and after that remained constant (Fig. 3).

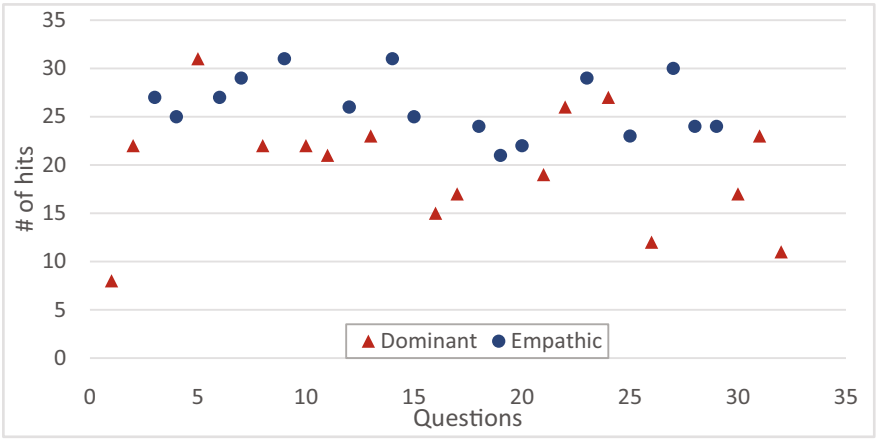

Fig. 7. Number of hits of Portuguese participants for each question of the experiment.

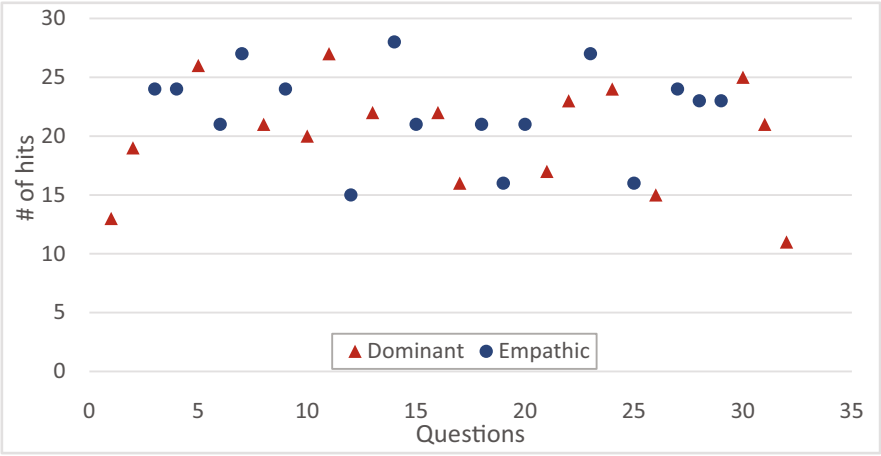

Fig. 8. Number of hits of Philippine participants for each question of the experiment.

More importantly, we tested for the total set of human participants whether they found Dominance and Empathy equally difficult to recognize. As shown in Table 5, it turned out that the Empathic samples (76.46%) were recognized more easily than the Dominant samples (64.80%). An unpaired t-test confirmed that this difference was significant ($p < 0.001$). This result is similar to the results for the automated algorithm, which also had a bit more difficulty with recognizing the Dominant samples. This gives answers to our ***fourth research question***.

Finally, our ***fifth research question*** was whether the performance of human listeners is influenced by whether or not they understand the language of the spoken fragments. To investigate this, we compared the average performance over all samples classified by a listener that understood the language (i.e., all Dutch samples classified by Dutch participants plus all Portuguese samples classified by Brazilian participants) with the average performance over all other samples. And unpaired t-test confirmed that the difference between these two performances was not significant ($p = 0.50$). Hence, in our experiment, being able to understand *what* was said did not help (or hinder) participants in recognizing *how* it was said.

5 Conclusion

This paper presented an algorithm to automatically detect interpersonal stance in vocal signals, especially related to Dominance and Empathy. In addition, an experiment was performed in which we investigated how well human beings performed in the same task. Five research questions were addressed, which yielded the following results.

First, the results of the algorithm points out good perspectives related to the capacity of computers to identify Dominance and Empathy. The average accuracy is 94.58%. Considering the results and the fact that the classification is real-time, it seems promising to apply the algorithm in serious games or other human-computer interaction applications in order to enrich the user's experience. In particular, in follow-up research we aim to use the algorithm for an aggression de-escalation training system for public transport employees.

Second, it was explored whether the concepts of Empathic and Dominant could be related to specific positions in the Arousal-Valence circumplex. In our experiments, we observed that Dominance and Empathy are generally well distinguishable when mapped into this space. However, the way in which both stances are distinguished seems to differ per individual.

Third, it was found that the algorithm substantially outperformed human listeners, which on average reached a performance of 70.64% in the recognition task. The relatively low performance of human beings may also be explained by the fact that in real life, humans use much more than only vocal signals to classify emotions. For instance, they typically combine facial expression, speech content, background knowledge about culture, context situation, and so on. On the other hand, as suggested in [20], computers can also make use of such information to improve their performance.

Fourth, both for human and computer, there turned out to be a slight difference in the capability to recognize Dominance and Empathy. The algorithm performs 5.2%

better in classifying Empathy (96.8%) than Dominance (91.6%). The same trend was found for the 3 groups of participants, where the Empathic samples (76.46%) were recognized more easily than the Dominant samples (64.80%).

Finally, we found that the performance of the human participants was not influenced by whether or not the listener understands the language of the spoken fragments. Hence, being able to understand *what* was said did not help (or hinder) participants in recognizing *how* it was said.

The outcome of this research is two-fold: on the one hand, it resulted in an algorithm that can be useful for the development of human-computer interaction applications (in particular in the domain of aggression de-escalation training). On the other hand, it sheds more light on how human beings recognize interpersonal stance based on vocal signals, and how they differ in this task from computer algorithms.

Acknowledgments. This research was supported by the Brazilian scholarship program Science without Borders - CNPq {scholarship reference: 233883/2014-2}.

References

1. Hogan, K., Stubbs, R.: Can't Get Through: 8 Barriers to Communication. Pelican Publishing Company, Gretna (2003)
2. Patterson, A.E., Berg, M.: Exploring nonverbal communication through service learning. J. Civic Commitment, **21** (2014)
3. Picard, RW.: Affective computing for HCI. In: Proceedings of HCI 1999, Munich, Germany (1999)
4. Bosse, T., Gerritsen, C., de Man, J.: An intelligent system for aggression de-escalation training. In: Proceedings of the 22nd European Conference on Artificial Intelligence, ECAI 2016. IOS Press (2016)
5. Anderson, L.N., Clarke, J.T.: De-escalating verbal aggression in primary care settings. Nurse Pract. **21**(10), 95, 98, 101, 102 (1996)
6. Du Bois, J.W.: The stance triangle. In: Englebretson, R. (ed.) Stancetaking in Discourse: Subjectivity, Evaluation, Interaction, pp. 139–182. John Benjamins Publishing Company (2007)
7. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of ACM Multimedia (MM), Barcelona, Spain, pp. 835-838. ACM, October 2013. ISBN 978-1-4503-2404-5, doi: 10.1145/2502081.2502224
8. Dodge, K.A.: The structure and function of reactive and proactive aggression. In: Pepler, D., Rubin, H. (eds.) The development and treatment of childhood aggression, pp. 201–218. Erlbaum, Hillsdale (1990)
9. Berkowitz, L.: Whatever happened to the frustration-aggression hypothesis? Am. Behav. Sci. **21**, 691–708 (1978)
10. Bosse, T., Provoost, S.: Towards aggression de-escalation training with virtual agents: a computational model. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2014. LNCS, vol. 8524, pp. 375–387. Springer, Cham (2014). doi:10.1007/978-3-319-07485-6_37
11. Leary, T.: Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation. Ronald Press, New York (1957)
12. Russel, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)

13. Hirst, D., Di Cristi, A.: A survey of intonation systems. In: Hirst, D., Di Cristo, A. (eds.) Intonation Systems: A Survey of Twenty Languages, pp. 1–44. Cambridge University Press, Cambridge (1998)
14. Mark Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
15. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification (2010)
16. Rybka, J., Janicki, A.: Comparison of speaker dependent and speaker independent emotion recognition. Int. J. Appl. Math. Comput. Sci. **23**(4), 797–808 (2013). doi:10.2478/amcs-2013-0060
17. ElAyadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. J. **44**, 572–587 (2011). doi:10.1016/j.patcog.2010.09.020
18. Lin, C., Liao, W., Hsieh, W., Liao, W., Wang, J.: Emotion identification using extremely low frequency components of speech feature contours. Sci. World J. **2014** (2014). Article id. 757121, Hindawi Publishing Corporation. doi:10.1155/2014/757121
19. Yik, M., Russel, J., Steiger, J.: A 12-point circumplex structure of core affect. Emotion **11**(4), 705–731 (2011)
20. Formolo, D., Bosse, T.: Human vs. Computer performance in voice-based emotion recognition. In: Proceedings of the 19th International Conference on Human-Computer Interaction, HCI 2017. LNCS, pp 285–291. Springer, Heidelberg (2017)