TOWARDS AN ARTIFICIAL THERAPY ASSISTANT
Measuring Excessive Stress from Speech

Frans van der Sluis, Egon L. van den Broek
Human-Media Interaction (HMI), University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Ton Dijkstra
Donders Institute for Brain, Cognition, and Behavior, Radboud University
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

Keywords: Stress, Psychiatry, Diagnosis, Speech, Agent, Model.

Abstract: The measurement of (excessive) stress is still a challenging endeavor. Most tools rely on either introspection or expert opinion and are, therefore, often less reliable or a burden on the patient. An objective method could relieve these problems and, consequently, assist diagnostics. Speech was considered an excellent candidate for an objective, unobtrusive measure of emotion. True stress was successfully induced, using two storytelling sessions performed by 25 patients suffering from a stress disorder. When reading either a happy or a sad story, different stress levels were reported using the Subjective Unit of Distress (SUD). A linear regression model consisting of the high-frequency energy, pitch, and zero crossings of the speech signal was able to explain 70% of the variance in the subjectively reported stress. The results demonstrate the feasibility of an objective measurement of stress in speech. As such, the foundation for an Artificial Therapeutic Agent is laid, capable of assisting therapists through an objective measurement of experienced stress.

1 INTRODUCTION

In 1975, Malcolm Lader stated: *"Psychiatric research has been unsuccessful in developing scientific methods of its own but has relied on a series of techniques borrowed from other disciplines. Too often the outside discipline has been chosen because of its relevance, and the high hopes at the outset of such studies have lessened as concrete advances have failed to materialise."* (Lader, 1975). Since these words were published significant progress has been made in both science and engineering. However, psychiatry is still struggling in some sense. On the one hand, it is now generally accepted that mind and body go hand in hand. Consequently, for example, psychopharmacy have gained in popularity since Lader's words. On the other hand, science is not even close to truly understanding the relation between mind and body, which is illustrated by the lack of computational models.

Computational models have been proposed in the shape of decision support systems and agents. In some branches of industry these models have been successfully employed but not in psychiatry. Various reasons for this lack of success can be opted; for example, overly complex models, unjustified simplifications, a lack of validation procedures, and simply a lack of domain knowledge. To prevent from suffering the same pitfalls, this study limits its aims. The core concept under investigation will be: stress.

The usage of the term stress has been liberal, leaving it as a poorly defined term with many definitions. However, as a common denominator, the different views build upon some form of a process model, in which *"environmental demands tax or exceed the adaptive capacity of an organism, resulting in psychological and biological changes that may place persons at risk for disease"* (Cohen and Oviatt, 2002). Several traditions have emphasized different aspects of this process: environmental, psychological, and biological aspects. These can be roughly translated to, respectively, the stressor (Kessler, 1997), the appraisal (Lazarus, 1993), and their (often physiological) responses (Cohen and Oviatt, 2002).

Table 1 gives an overview of a few prevalent

van der Sluis F., L. van den Broek E. and Dijkstra T..
TOWARDS AN ARTIFICIAL THERAPY ASSISTANT - Measuring Excessive Stress from Speech.
DOI: 10.5220/0003175203570363
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2011), pages 357-363
ISBN: 978-989-8425-34-8
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

Table 1: Some Stress-Related Psychiatric Disorders.

Post-Traumatic Stress Disorder (PTSD) is caused by a severe trauma, originating from a range of situations; e.g., warfare, natural disasters, inter-personal violence such as sexual, physical, and emotional abuse, intimate partner violence, and collective violence. Key characteristics of PTSD are a persistent reexperience of the stressor, and persistent symptoms of increased arousal (American Psychiatric Association, 2000).
Depression cannot always be related to a specific cause, though several contributing factors have been identified: e.g., genetic vulnerability and unavoidability of stress. More specific, certain stressful life events (e.g., job loss, widowhood) can lead to a state of depression. Furthermore, chronic role-related stress is significantly associated with chronically depressed mood (Kessler, 1997). Important to note is that the experience of stress is associated with the onset of depression, and not with the symptoms of depression (American Psychiatric Association, 2000).
Insomnia often has a fairly sudden onset caused by psychological, social, or medical stress. Though, in some cases, it may develop gradually and without a clear stressor. Insomnia is characterized by sleep deprivation, and associated with increased physiological, cognitive, or emotional arousal in combination with negative conditioning for sleep (American Psychiatric Association, 2000).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

stress-related psychiatric disorders. Although not an exhaustive list, this overview illustrates how different aspects of stress can explain different disorders. For example, depression and insomnia have a strong appraisal component, whereas PTSD is mainly explained by a severe stressor. Moreover, Table 1 highlights the temporal course of the stressor as well as the stress response. Although the actual stressor can be both acute and chronic, there is a chronic stress response for all diseases; either at the onset (e.g., depression) or as a symptom (e.g., PTSD).

In general, the diagnosis of stress-related psychiatric disorders is, amongst other methods, performed with a careful interview (American Psychiatric Association, 2000). During this interview, the clinician has to determine if the patient suffers from excessive stress. Moreover, the clinician has to identify the possible stressor causing an excessive stress response. Hence, a key diagnostic task is to determine whether or not the patient suffers from excessive stress in relation to specific stressors.

The diagnosis of excessive stress is repeated during treatment as well, in order to indicate the progress of the treatment. Depending on the treatment type, this diagnostic repetition can even be part of the treatment itself. (Everly, Jr. and Lating, 2002) differentiate between three therapeutic genres: 1) avoid/minimize/modify stressors; 2) reduce excessive arousal and organ dysfunction; and 3) ventilate or express the stress response. The latter incorporates a repeated expression of the stress response and requires a repeated measurement of it; for example, as is done with the treatment of PTSD. Technology is beginning to play a more significant role in the treatment of stress disorders, evidenced by a new treatment method of self-help and minimal contact therapies which has proven to be succesful for certain types of patients (Newman et al., 2010).

Currently, the measurement of excessive stress is problematic. A clinician uses diagnostic criteria based on a range of questionnaires to support this aim. Introspection and the expert opinion of the clinician are at the basis of these tools. Inherently, subjective measures can be unreliable. Moreover, these questionnaires can be a burden for the patient.

The aim of the study is to lay the foundation for an Artificial Therapy Assistant (ATA), capable of assisting therapists through an objective measurement of stress. Moreover, such a system can be useful for minimal-contact or self-help interventions as well (Newman et al., 2010). The next section will identify the prerequisites that had to be taken into account for this system. After that, in Section 3, a clinical study of objective stress measurement will be introduced. This study involved the participation of patients suffering from a PTSD, which enabled the salient determination of stress characteristics and, with that, an indicator of stress. Section 4 presents the results obtained through this study and defines a model that can serve as the foundation for an ATA. Finally, in Section 5 we discuss the contribution of this work to the diagnosis and treatment of stress disorders.

2 CONSIDERATIONS AND SPECIFICATIONS

The aim of the study was an agent that is able to support psychiatrists and psychologists in their diagnosis of excessive stress. Key to this system is the measurement of signals indicative of emotions, in particular of stress, and the determination of intensity. It is the intensity that can help the therapist in determining whether or not the patient is suffering from an excessive stress response. As such, the agent aims to support the decision of the clinician, contrary to giving a decision itself.

It is known from literature that multiple physical sources can be applied as stress indicator (Lader, 1975; van den Broek et al., 2010). This research elaborates on speech, which has a number of advantages: i) In therapy sessions, speech is often already recorded. Hence, using speech requires no additional effort for the therapists; ii) Speech processing is fully unobstrusive; and iii) There is fairly little noise in the speech signal, as therapy sessions are generally held in a controlled environment.

2.1 Feature Extraction

For the following features there is a fair amount of support for their affective information: pitch, energy, high-frequency energy, and to a lesser extent zero-crossings rate (Kedem, 1986; Scherer, 2003; El Ayadi et al., xxxx). Although there is no general concensus on the best features for stress detection, there is substantial evidence for these. Hence, they extracted from the audio signal.

For a domain $[0,T]$, consisting of N number of samples, the energy of the speech signal is defined as:

$$20\ \log_{10}\frac{1}{\mathrm{P}_0}\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}x^2(n)}, \tag{1}$$

where the amplitude or sound pressure of the signal is denoted in Pascals (Pa) as $x(n)$ and the auditory threshold P_0 is $2\cdot 10^{-5}$ Pa (Boersma and Weenink, 2006). The energy of the speech signal is also described as the Sound Pressure Level (SPL). It is expressed in decibels (dB) relative to the auditory threshold P_0; i.e., in dB (SPL).

To extract speech's high-frequency energy (i.e., the energy for the domain $[1000,\infty]$ in Hz), the signal first has to be transformed to the frequency domain (Banse and Scherer, 1996). This is done by a fast Fourier implementation of the discrete Fourier transform. The discrete Fourier transform (Lyons, 2004):

$$X(m)=\frac{1}{N}\sum_{n=0}^{N-1}x(n)e^{-j2\pi nm/N}, \tag{2}$$

with j representing the $\sqrt{-1}$ operator and where m relates to frequency by $f(m)=mf_s/N$. Here, f_s is the sample frequency and N is the number of bins. The number of bins typically amounts to the next power of 2 for the number of samples being analyzed; e.g., 2048 for a window of 40 msec. sampled at 44100 Hz. The energy for the domain $[M_1,M_2]$, where $f(M_1)=1000Hz$ and $f(M_2)=f_s/2$ (i.e., the Nyquist frequency), is defined by:

$$20\ \log_{10}\frac{1}{\mathrm{P}_0}\sqrt{\frac{1}{M_2-M_1}\sum_{m=M_1}^{M_2}|X|^2(m)}. \tag{3}$$

The F0 or pitch is extracted using the autocorrelation method. The autocorrelation is the cross-correlation of the signal with itself, where the cross-correlation denotes the similarity between two signals as a function of a time-lag between them. The autocorrelation R of signal x at time lag m is defined as:

$$R_x(m)=\sum_{n=0}^{N-m-1}x(n+m)\bar{x}(n) \tag{4}$$

where N is the length of the signal. The autocorrelation is then computed for each time lag m over the domain $M_1=0$ and $M_2=N-1$. The global maximum of this method is at lag 0. The local maximum beyond 0, lag m_{max}, represents the fundamental frequency, if its normalized local maximum $R_x(m_{max})/R_x(0)$ (its harmonic strength) is large enough (e.g., above .45). The fundamental frequency is derived by $1/m_{max}$. We refer to (Boersma, 1993) for a detailed description of the (implementation of) the F0 extraction.

The zero crossings rate of the speech signal is also computed. This is defined as:

$$\frac{1}{N}\sum_{n=1}^{N-1}\mathbb{I}\{x(n)x(n-1)<0\}, \tag{5}$$

where N is the number of samples of the signal amplitude x. The $\mathbb{I}\{\alpha\}$ serves as a logical function (Kedem, 1986).

3 CLINICAL STUDY

The exact relation between the identified features of speech and stress is as yet unclear. Two problems make it hard to compare most previous studies and methods on stress detection. First, many studies use mimicked emotions instead of true emotions as the

basis of their model of stress (i.e., acted vs. experienced emotions). Second, since there is often no ground truth, it is unclear if the measured vocal parameters represent an (induced) affective state. For more information on these problems, see (Scherer, 2003). To arrive at an acoustic stress indicator, this section presents a study of stress in speech, dealing with the two identified problems.

The study consisted of two phases, triggering either a happy or an anxious state in the patients. Hence, anxiety was the stressful emotion chosen to induce stress. The order of sessions was counterbalanced over the participants. 25 Female PTSD patients (mean age: 38) participated voluntarily. An informed consent was signed by all participants. Having PTSD patients as participants had several advantages. First, PTSD patients are relatively sensitive to stress and, thus, to stressors. Hence, they were expected to react more intensively to the emotion elicitation. Second, within the context of this study, the use of real patients increases the ecological validity.

The patients had to read two stories aloud, one of an anxious and one of a happy situation. To prevent any interfering factors the stories were kept similar on their syntactic structure and their complexity. Moreover, the order of the stories was counterbalanced over all participants. Before the patients started with the emotion inducing stories, they read a sample story to familiarize themselves with the task.

The stories served as emotional Stroop tasks (Williams et al., 1996), since they included words that induced either anxiety or a happy emotional state. Emotional Stroop tasks are frequently used in clinical psychology and psychiatry research and are accepted as a reliable method for eliciting emotions. Emotional Stroop tasks can be defined as the presentation of stimuli that are expected to evoke emotions, due to an attentional bias of the participants. In this research, the Stroop effect was achieved through anxiety triggering words incorporated in one of the stories.

To be able to derive stress from speech, several steps had to be performed. First, the signal was recorded. This was done using a standard PC, a microphone preamplifier, and a microphone. The recording's sample rate was 44.1 kHz, mono channel, and a resolution of 16 bits. The recordings were divided into samples of approximately one minute of speech. This resulted in a one-on-one mapping between the ground truth (explained furtheron) and the speech features. Second, other voices and speckle noise were removed from the recorded signal.

Several features were extracted from the clean signal; see also Section 2.1. From each of these features, a number of statistical parameters were derived: *mean*, *median*, standard deviation (*std*), variance (*var*), minimum value (*min*), maximum value (*max*), range ($max - min$), the quantiles at 10%($q10$), 90%($q90$), 25%($q25$), and 75%($q75$), the inter-quantile-range $10\% - 90\%$ ($iqr10, q90 - q10$), and the inter-quantile-range $25\% - 75\%$ ($iqr25, q75 - q25$). All features were computed using a time window of 40 msec. and a step length of 10 msec.; i.e., every 10 msec. over the next 40 msec. of the signal. Next, the statistical parameters were derived over time chunks of 60 sec., allowing a one-on-one comparison with the Subjective Unit of Distress (SUD) data.

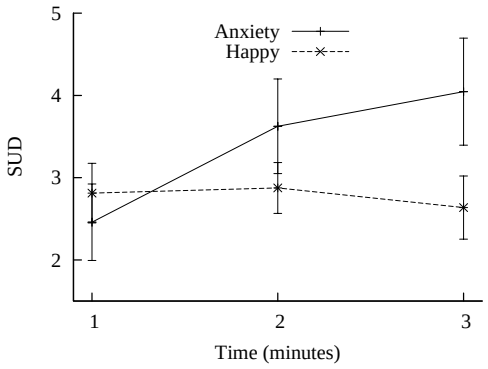

Figure 1: Reported stress per group and time.

To enable the validation of the speech parameters they were compared to a subjective measurement: the SUD. The SUD is a Likert scale indicative of the (dis)stress a participant experiences at the moment of measurement. For this study, a linear scale with range 0 to 10 was used on which the participants were instructed to place a cross or a dot. After (Wolpe, 1958) introduced the SUD, it has proven itself to be a reliable measure of one's emotional state. The participants used the SUD every minute, making it a routine task. The SUD was used as ground truth for the derived speech parameters (See Section 2.1).

4 RESULTS

If the manipulation of stress was successful, the results can be used as a stress indicator. Figure 1 illustrates the mean values of both manipulations and denotes the confidence intervals without the inter-subject variance (Cousineau, 2005) through the vertical bars. This figure shows that the manipulation has been successful. Furthermore, when isolating the anxiety condition, a trend was visible for time on SUD scores ($F(2,56) = 2.726; p = .07$).

In order to create a generic stress indicator, the most relevant of all features and accompanying pa-

rameters were selected. This selection process was done with a linear regression model ($\mathcal{M}$). A $\mathcal{M}$ explains how p independent variables (predictors, x) predict dependent variable y. In order to do so, p optimal weighting factors Beta (B) over each of the $i = 1,\ldots,n$ observations are determined:

$$y_i = B_0 + B_1 x_{i1} + \cdots + B_p x_{ip} + \varepsilon_i, \qquad (6)$$

where ε_i represents unobserved random noise. The method used to determine the average optimal weighting factors over all n observations is the ordinal least squares method. To reduce the number of predictors, a backward selection algorithm was applied. Through an iterative process, this algorithm removes the non-significant predictors ($p > .10$) for subjective stress. As the backward method uses the relative contribution to the *model* as selection criterium, the interdependency of the features is taken into account as well (Harrell, Jr., 2001). This makes it a robust method for feature and parameter selection.

The model was created using the SUD scores of the anxiety and happy conditions (See Figure 1). Here, a $\mathcal{M}$ containing all features and all parameters (i.e., in total 40 predictors), explained 69.72% of the variance: $R^2 = .697$ and $\overline{R}^2 = .575$, $F(40,99) = 5.70, p < .001$. Applying the backward selection method with 22 iterations, leaving 18 predictors, the model still explained 67.37% of the variance: $R^2 = .674$ and $\overline{R}^2 = .625$, $F(18,121) = 13.88, p < .001$. The model and the used features are described in more detail in (van der Sluis et al., 2010).

5 DISCUSSION

25 Patients reported stress that had successfully been caused by reading/telling a carefully created story. By comparing speech features to a subjective report of stress, this study defined and evaluated an acoustic profile of stress characteristics in speech. The acoustic profile was shown to explain nearly 70% of variance in the subjectively reported stress. Hence, demonstrating the feasability of speech as an objective measure of experienced stress and, with that, as an ATA.

Although it is only one of many ways to induce emotions, storytelling was shown to be particularly useful in creating an emotion-induced speech signal. In particular, it is likely to create true emotions, this contrary to many other commonly used methods. The triangulation of the SUD and various speech characteristics suggests that indeed true emotions were triggered through the storytelling.

A potential problem with the acoustic stress indicator, as introduced, is described by the existing theoretical distinction of emotional and emotive communication (Caffi and Janney, 1994). Emotional communication is a type of spontaneous, unintentional leakage or bursting out of emotion in speech, while emotive communication has no automatic or necessary relation to "real" inner affective states. Emotive communication is a strategic signaling of affective information in speaking to interaction partners that is widespread in interactions; see also (Caffi and Janney, 1994). It uses signal patterns that differ strongly from spontaneous, emotional expressions and can be both intentionally and unintentionally accessed (Banse and Scherer, 1996).

Another issue is the distinction between cognitive and emotional stress, which is known as the problem of emotion specificity (Zeelenberg et al., 2008). Emotion specificity distinguishes cognitive stress, the information processing load placed on the human operator while performing a particular task, and emotional stress, the psychological and physiological tension due to emotions triggered before or during the task.

In general, the subjectively reported stress was somewhat dispersed. This is likely to be partly due to inter-personal differences and, consequently, indicates that the stories did have an influence. Moreover, a trend was shown for the anxiety inducing story to create stress over time, supporting this influence. These results suggest the value as well as the drawbacks of storytelling. Two problems can be identified:

- inducing an affective state with stories is strongly dependent on the temporal course; i.e., a story needs a build-up; and
- there were substantial inter-personal differences in the experience of the stories.

The latter problem may be useful for diagnostic goals. Inter-personal differences are likely to be caused by differences in appraisal. Hence, this can be used to assess a patient's appraisal patterns, which have been identified as a major component for certain psychiatric illnesses; for example, depression (Kessler, 1997).

The explained variance of 70% can be considered as high, especially considering the number of participants. Moreover, since the model is not personalized, some generic chacteristics of stress in speech seem to be uncovered. However, some restrictions also apply:

- only PTSD patients participated, while other patient groups might show different stress responses;

- many stressful emotions have been identified, these may be different kinds of stress; and
- restrictions applying to storytelling for emotion elicitation may have influenced the results.

These three restrictions can be seen as future research challenges. Namely, to use other patient groups, emotions, and emotion elicitation techniques.

Using the acoustic profile, one can arrive at an ATA for the diagnosis and treatment of stress-related psychiatric disorders. An ATA can help the clinical setting in several ways, to:

1. support the measurement of stress responses;
2. give decision support on whether a patient suffers from excessive stress;
3. aid the treatment of stress disorders; and
4. improve self-help and minimal-contact therapy methods (Newman et al., 2010).

Through making the measurement objective, the measurement of stress becomes more reliable; i.e., no longer solely relying on introspection. Objective measurement also increases inter- and intra-expert reliability. Moreover, diagnosis, decision-making in general, and treatment could become more fine-grained.

Concluding, an important and significant step towards an ATA for stress-related psychiatric disorders has been made. This study has shown that an objective measurement of stress through speech is feasible. Par excellence, the feasibility of objective stress measurement illustrates the possibility of more objective measures for the generally subjective fields of psychology and psychiatry.

ACKNOWLEDGEMENTS

We gratefully acknowledge the PTSD patients for voluntarily participating in this research. We thank Lynn Packwood for proof reading this article.

REFERENCES

American Psychiatric Association (2000). *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC, USA: American Psychiatric Publishing, Inc., 4 (Text Revision) edition.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to- noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, volume 17, pages 97–110. University of Amsterdam.

Boersma, P. P. G. and Weenink, D. J. M. (2006). Praat 4.0.4. http://www.praat.org (Last accessed on October 22, 2010).

Caffi, C. and Janney, R. W. (1994). Toward a pragmatics of emotive communication. *Journal of Pragmatics*, 22(3–4):325–373.

Cohen, P. R. and Oviatt, S. L. (2002). The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences (PNAS)*, 92(22):9921–9927.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, 1(1):42–46.

El Ayadi, M., Kamel, M. S., and Karray, F. (xxxx). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, [in press].

Everly, Jr., G. S. and Lating, J. M. (2002). *A clinical guide to the treatment of the human stress response*. The Plenum series on stress and coping. New York, NY, USA: Kluwer Academic / Plenum Publishers, 2nd edition.

Harrell, Jr., F. E. (2001). *Regression modeling strategies – with applications to linear models, logistic regression, and survival analysis*. Springer Series in Statistics. New York, NY, USA: Springer-Verlag New York, Inc., 1st; 6th printing edition.

Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493.

Kessler, R. C. (1997). The effects of stressful life events on depression. *Annual Review of Psychology*, 48(1):191–214.

Lader, M. (1975). *The psychophysiology of mental illness*. London, Great Britain: Routledge & Kegan Paul Ltd.

Lazarus, R. S. (1993). From psychological stress to the emotions: A history of changing outlooks. *Annual Review of Psychology*, 44(1):1–22.

Lyons, R. G. (2004). *Understanding Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2nd edition.

Newman, M. G., Szkodny, L. E., Llera, S. J., and Przeworski, A. (2010). A review of technology-assisted self-help and minimal contact therapies for anxiety and depression: Is human contact necessary for therapeutic efficacy? *Clinical Psychology Review*, [in press].

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2):227–256.

Van den Broek et al., E. L. (2009/2010). Prerequisites for Affective Signal Processing (ASP) – Parts I–IV. In Fred, A., Filipe, J., and Gamboa, H., editors, *BioSTEC 2009/2010: Proceedings of the International Joint Conference on Biomedical Engineering*

Systems and Technologies, pages –, Porto, Portugal / Valencia, Spain. INSTICC Press.

Van der Sluis, F., van den Broek, E. L., and Dijkstra, T. (2010). Towards semi-automated assistance for the treatment of stress disorders. In *HealthInf 2010: Proceedings of the Third International Conference on Health Informatics*, pages 446–449, Valencia, Spain. INSTICC Press.

Williams, J. M. G., Mathews, A., and MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological bulletin*, 120(1):3–24.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford, CA, USA: Stanford University Press.

Zeelenberg, M., Nelissen, R. M. A., Breugelmans, S. M., and Pieters, R. (2008). On emotion specificity in decision making: Why feeling is for doing. *Judgment and Decision Making*, 3(1):18–27.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS