







You Sound Relaxed Now – Measuring Restorative Effects from Speech Signals

Yong Ma^(✉) , Jingyi Li , Heiko Drewes , and Andreas Butz 

LMU Munich, Munich, Germany

{yong.ma, jingyi.li, heiko.drewes, andreas.butz}@ifi.lmu.de

<http://www.medien.ifi.lmu.de>

Abstract. The recently proposed *restorative environments* have the potential to restore attention and help against fatigue, but how can these effects be verified? We present a novel measurement method which can analyze participants' speech signals in a study before and after a relaxing experience. Compared to other measurements such as attention scales or response tests, speech signal analysis is both less obtrusive and more accessible. In our study, we found that certain time- and frequency-domain speech features such as short-time energy and Mel Frequency Cepstral Coefficients (MFCC) are correlated with the attentional capacity measured by traditional ratings. We thus argue that speech signal analysis can provide a valid measure for attention and its restoration. We describe a practically feasible method for such a speech signal analysis along with some preliminary results.

Keywords: Speech feature analysis · Attention measurement · Restoration

1 Why Measure Attention Restoration from Speech?

The increasing stress for humans in modern urban environments makes physical and mental recovery a vital research topic in human-centered computing. Studies indicate that the exposure to (virtual) natural environments can effectively restore attention [20] and mitigate the feeling of fatigue. Such a *restorative environment* could, for example, be used in automated driving and offer travellers a way of reconnecting with nature and mentally recharging during travel. To measure the effects of attention restoration, existing research mostly uses self-report questionnaires or response tests. However, these traditional approaches interrupt the flow of the experiment and influence the attention being measured. Other proposed attention detection methods use physiological sensors. For instance, eye movement can be used to gauge attention [7] and analyze its recovery. Signals from EEG [2, 19] and ECG [4] can also effectively detect attention restoration, but are currently still much less convenient to measure. Analyzing the human speech signal may present a very unobtrusive and effective alternative.

The idea of using speech signal processing (SSP) in the analysis of drowsiness or fatigue detection was introduced by Dhupati et al. [6]. SSP allows the automated detection and evaluation of certain mental states. More specifically, speech signals have been used for emotion detection and can also support the automatic assessment of mental recovery. Moreover, unlike physiological sensors that record data under constrained conditions, speech data can be recorded in completely natural and unpredictable situations [8]. This motivated us to combine speech feature analysis with an efficient speech segmentation algorithm [26] to detect and evaluate the attention restoration effect of restorative environments. More concretely, we investigated whether a user's speech signals, recorded in a traditional response test before and after a restorative experience, could provide an additional objective measure of the restorative effects and would align with the results of the response test and attention scale. Using traditional machine learning methods, we were also able to classify and predict different restoration levels based on speech features.

2 Background and Related Work

Our initial use case was measuring the restorative effect of an in-car restorative environment developed in another project [14]. In brief, our work there built on the paradigm of attention restoration [20], the effects of which are mainly measured by attention scales and response tests. The technical basis of our work is speech signal analysis and automatic speech segmentation. We will therefore briefly introduce related work from these two areas below.

2.1 Measuring Attention

Virtual Restorative Environments (VREs) aim to reduce stress and restore attentional capacities [25] by recreating scenes of natural beauty and peacefulness in VR. The timeline of a typical trial in attention restoration research consists of a stress-induction phase followed by a restoration phase [24]. Experimenters usually collect attention data at the beginning or between two phases as the baseline for comparison with the post-restoration data. Previous work tried to measure attention by established methods such as self-reporting scales and response tests [20]. However, these conventional measurements are inconvenient because they inevitably interrupt the flow of the experiment and also suffer from subjective factors. Furthermore, physiological signals such as heart rate [13] or other signals from mobile sensors [27] can also be used to measure attention recovery effects. As a less obtrusive measuring method, we propose to use speech signal analysis to detect and quantify attention restoration.

2.2 Speech Feature Extraction and Speaker Segmentation

Speech signal analysis means the analysis and processing of phonetic characteristics, generally including features in both the time- and frequency-domain [23].

Such audio features have also helped to analyze human social behavior and assess the state of humans' mental health [8], or enable automatic mood detection [16]. When looking for suitable signal features for measuring attention, we found that the time-domain features *speech entropy*, *short-time energy* and *speech intensity* can identify stress or fatigue [17]. The frequency-domain features *Mel Frequency Cepstral Coefficients (MFCC)* can indicate happiness or stress [15]. Thus, it seemed plausible that these speech features could also be used to measure other mental properties, such as attentional capacity.

A robust and reliable speaker segmentation will substantially improve the accuracy of the extracted features. Existing segmentation methods can be split into supervised and unsupervised algorithms. Supervised segmentation algorithms, such as the GMM method [18] and artificial neural networks [32], recognize the speaker's voice after being trained on it beforehand. Unsupervised segmentation algorithms detect the speaker's voice without prior training, for example, from time-domain [12] and frequency-domain [11] features. Other traditional voice segmentation methods are based on energy estimation [22] and hidden Markov models (HMM) [26]. Since we wanted to use recordings from previously unknown study participants, any supervised method that requires prior training on a specific voice was out of question. Based on the existing literature on fully automatic segmentation, we thus decided for an unsupervised speech segmentation.

3 Establishing Speech Signal Analysis as a Measure

Our original study used a within-subject design in which each participant experienced an in-car VRE. We collected subjective and objective attention measures both before and after the restorative experience and the control condition. As a side effect, we recorded speech signals during these measures. For the purpose of this paper, we then analyzed the correlation between these other measures of attention and the features of the recorded speech signal. We hypothesized that certain speech features would be correlated with the attention measures.

3.1 Apparatus, Participants and Experimental Procedure

The VRE was implemented in Unity 3D and installed on a standard PC connected to an Oculus Rift HMD. We used a separate noise cancelling headphone for audio output. A digital audio recorder was used to record speech signals at 48 kHz and 16 bit resolution. Speaker distance to the microphone was about 15 cm. Matlab 2017a and MIRtoolbox 1.7.1 were used for speech signal analysis. Our study procedure was approved by the local ethics review board (ID: EK-MIS-2020-011). We invited 21 participants (5 male) aged between 19 and 33 years ($M = 26.7$, $SD = 4.0$) to our lab. More than half of them had experience in driving and VR.

The study consisted of 7 steps as shown in Fig. 1. After a demographic questionnaire, we asked participants to fill the 13-item Attentional Function Index (AFI) questionnaire [5] as a subjective measurement of attention. As an objective measurement, participants were then asked to complete the Digit Span Backward test



Fig. 1. Timeline of the study procedure.

(DSB) [30] on a computer and we simultaneously recorded their voice. The study was conducted in the quietest room of our lab to reduce background noise. In the third step, participants were shown a video clip of a traffic jam¹ to induce a context-specific type of stress [3] before the restorative experience. The purpose of this step was to ensure that participants were in a state of lowered attentional capacity. We then took the same measurements and recorded audio data again. In the fifth step, participants experienced the in-car restorative environment or (as a control condition) closed their eyes in VR for around 10 min [14]. After this intervention, participants were asked to complete the same tests, again recording their voice. To end the study, the experimenter conducted an interview asking participants to talk about their feelings during the restorative experience.

3.2 Methods Used for Speech Signal Analysis

We used the recorded speech data to explore the relationship between speech features and attention restoration. From the voice recordings of the DSB test, we selected the speech parts and extracted time- and frequency-domain speech features including short-time energy, zero-crossing rate, formant and MFCC. The processing pipeline is shown in Fig. 2 and explained below.

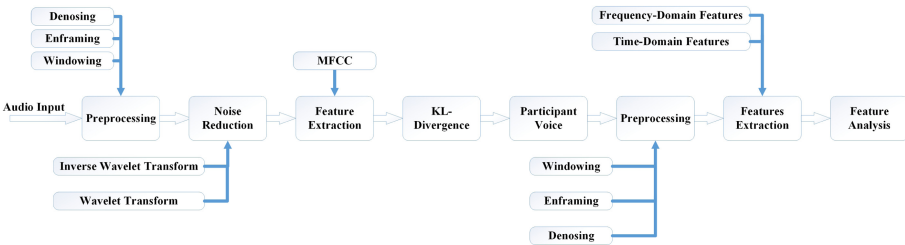


Fig. 2. Our signal analysis pipeline tightly integrates speech feature extraction and speaker segmentation.

In a preprocessing step, we denoised and enframed the audio data and defined a time window. For the frame size, we used 25 ms and a frame step of 10 ms. We included a 4-Daubechies wavelet transform [21, 28] and its inverse transform for noise reduction and speech enhancement, and then extracted the MFCC feature. Then, we also computed the Kullback-Leibler Divergence (KL-Divergence) [31]

¹ <https://www.youtube.com/watch?v=GlCazmVBUMg>.

and used it, together with the extracted features, to reliably segment the data into speech and non-speech segments. Features for attention detection were only extracted from speech segments and forwarded to further processing. In speech features analysis, we mainly explore the relationship between the speech features and attention restoration effects, which includes Pearson correlation coefficient [1], significance test, visual comparison and attention restoration level classification.

4 Comparison to Traditional Measures

To verify our method, we compared it to the traditional, objective and subjective tests of attention. We used the Pearson correlation coefficient to determine whether there was a strong correlation between these attention measures and our method before and after the restorative experience. We also compared them using traditional machine learning to verify whether speech signal analysis can actually predict the effect of attention restoration. A strong correlation effect is reported for $r \geq 0.5$ and the accuracy in our machine learning method is above 0.85 in two-class and 0.7 in three-class classifications.

4.1 Results of the Traditional Measures

For the original study, the attention score as provided by the AFI reflects the participants' subjective evaluation of their attentional state at the moment. We observed a stronger average increase in the VRE condition compared to the control condition. In the VRE, participants on average gained 2.9 (SD = 7.6) points, compared to an increase of 1.4 (SD = 8.4) points in the control condition. In the VRE, also the absolute AFI score was slightly higher after the restorative experience (Mdn = 61) than before (Mdn = 59). Across conditions, we observed a slightly stronger improvement in the DSB test in the VRE condition than in the control condition: Participants improved their short-term memory by 0.16 (SD = 0.7) digits in the VRE, but only by 0.11 (SD = 0.5) digits in the control condition. Moreover, in the test after the VRE, participants achieved their best working memory span (M = 5.11, SD = 1.0) compared to all other DSB tests.

4.2 Results Based on Speech Signal Analysis

Comparison Using Pearson Correlation. As shown in Table 1, the time domain features short-time energy, zero-crossing rate, max peak in autocorrelation and 3 formants show a strong positive correlation with the AFI score. For the formants, we chose the first three frequency peaks in the spectrum which have a high degree of energy. These six speech features thus can be effectively used to analyze and assess the attention restoration in our experiment. In addition, the short-time energy and zero-crossing rate before the VRE experience were lower than after it. In other words, these two features seem to increase when participants go from a state of fatigue to a state of relaxation.

Table 1. Pearson Correlation Coefficients between various signal features and attention restoration as measured by the AFI score

<i>Short-Time Energy</i>	0.8994
<i>Zero-crossing Rate</i>	0.9192
<i>Max peak in autocorrelation</i>	0.9209
<i>Formant 1</i>	0.9209
<i>Formant 2</i>	0.5798
<i>Formant 3</i>	0.7515

Comparison Using Significance Tests. As shown in Fig. 3, the short time energy and zero-crossing rate before the VRE experience are slightly lower than after. We used a Wilcoxon Signed Rank Test [29] and found the difference to be significant between these two states in short time energy ($p = 0.0453$) and zero-crossing rate ($p = 0.0475$). This is consistent with the results of the conventional tests and means that these features can detect attention restoration effects. For all other extracted speech features, differences were not significant.

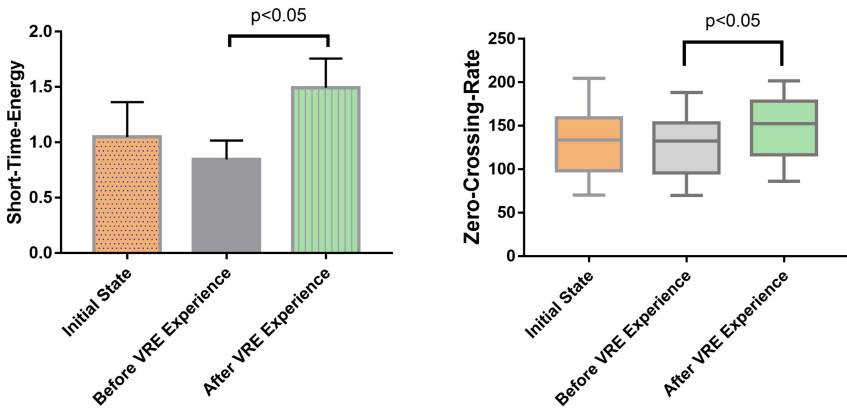


Fig. 3. Comparison of the features short-term-energy (left) and zero crossing rate (right) using a Wilcoxon Signed Rank test

Visual Comparison Using Spectrograms. From our data, we also computed spectrograms using a 128 channel Mel filter bank spanning 0 to 8 kHz (see Fig. 4). The left image shows data before the VRE and the right one after. The right spectrogram is visibly brighter than the left one, which suggests that this speech feature also is (positively) correlated to the attention restoration level and that such features can help to detect those levels.

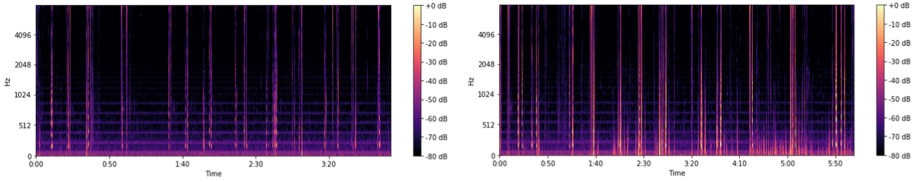


Fig. 4. Spectrograms of the MFCC feature before (left) and after (right) the VRE

Attention Restoration Level Classification. Using seven effective speech features which seemed to be correlated to attention restoration levels, we tested two traditional machine learning methods. Low attention levels represent the state before the VRE experience or control condition and high levels are the state after. The resulting accuracies for 2-class classifications were 0.92 and 0.88 respectively using the Support Vector Machines (SVM) [9] and the k-Nearest-Neighbours (KNN) [10] algorithms. Even if we define three different levels (after control condition, after VRE experience, and before), the accuracy of the SVM and KNN methods were 0.72 and 0.61, which still means that they can be used to recognize these three different states rather reliably.

5 Discussion, Limitations and Outlook

In the context of our research on attention restoration through VREs, we have found that speech signal analysis can potentially be used as an indirect measure for the effectiveness of restorative experiences. However, we are aware of certain limitations and found indications for necessary future work.

5.1 Discussion

In our original study, we had designed an evaluation procedure for the restorative effects of an in-car VRE. Through traditional attention measurements, we found that the studied VRE achieved low to moderate restorative effects in terms of improved attentional capacity and working memory. In addition to these established measures, we now also tested a novel evaluation method based on speech signal analysis: After extracting short-time energy and zero-crossing rate [12] in the time domain and MFCC [15] in the frequency domain with a preceding speaker segmentation, we conducted a correlation analysis and found these acoustic characteristics to be strongly correlated with the aforementioned traditional attention measures. This result was also confirmed by using significance tests on the change in these features and by visual comparison.

Furthermore, we found that traditional machine learning algorithms, such as KNN and SVM, can reliably detect and predict the measured attention restoration states from the recorded speech data. We therefore argue that speech signal analysis can be utilised to detect and evaluate the restorative effect in the same way as traditional measurements, such as attention scales or response tests.

The long term vision of this approach is to establish a fully automatic processing chain for measuring attention restoration levels based on audio data from interviews, which are conducted in a study anyway, or even using audio from voice interactions.

5.2 Limitations and Future Work

The relatively small number of participants in our study and the early stage of our speech signal analysis call for future studies confirming the precise relationship between an even more comprehensive set of speech features and attentional capacity or other mental properties. With regard to the voice recordings in the Digit Span Backward test, a chat-bot based on speech signal analysis could further improve the evaluation process by eliminating the human experimenter and eliciting better audio from participants who feel under less surveillance.

For now, all our analyses were done manually and after the study. In the future, we intend to iterate on our signal processing chain and eventually provide a fully automatic assessment system based on acoustic characteristics, which can, for example, effectively assess the restorative effects of VREs in real automated driving, but also provide attention measurements in other study setups. We expect that such a system will be generally applicable in a wide variety of contexts when measuring attention restoration levels.

6 Summary

While there is a growing emphasis on human wellbeing in designing interactive technologies, the corresponding evaluation methods have been less explored so far. In this paper, we explored an evaluation method for the attention restoration effects of an in-car VRE based on speech signal analysis. We compared this novel method to conventional measures in the form of subjective ratings of attentional capacity and objective performance in response tests. We developed and presented an initial version of a complete processing chain, including feature extraction and unsupervised speech segmentation. The results show that speech signal analysis can measure the restorative effects on attention and provide results that are consistent with the traditional measurements. We thus advocate the use of speech signal analysis as a novel HCI evaluation method, especially in measuring attention, but potentially for a wider range of mental parameters.

Acknowledgements. We thank all study participants for their time and effort, as well as our anonymous reviewers for their valuable feedback. Y.M.'s contributions were funded by the China Scholarship Council (CSC), grant number 201706070119.

References

1. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*, pp. 1–4. Springer, Vienna (2009). https://doi.org/10.1007/978-3-211-89836-9_1025

2. Biesmans, W., Das, N., Francart, T., Bertrand, A.: Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**(5), 402–412 (2016)
3. Braun, M., Weiser, S., Pflöging, B., Alt, F.: A comparison of emotion elicitation methods for affective driving studies. Presented at the (2018)
4. Carreiras, C., Lourenço, A., Aidos, H., da Silva, H.P., Fred, A.L.N.: Unsupervised analysis of morphological ECG features for attention detection. In: Madani, K., Dourado, A., Rosa, A., Filipe, J., Kacprzyk, J. (eds.) *Computational Intelligence. SCI*, vol. 613, pp. 437–453. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-23392-5_24
5. Cimprich, B., Visovatti, M., Ronis, D.L.: The attentional function index—a self-report cognitive measure. *Psychooncology* **20**(2), 194–202 (2011)
6. Dhupati, L.S., Kar, S., Rajaguru, A., Routray, A.: A novel drowsiness detection scheme based on speech analysis with validation using simultaneous EEG recordings, pp. 917–921. *IEEE* (2010)
7. Franěk, M., Šefara, D., Petružálek, J., Cabal, J., Myška, K.: Differences in eye movements while viewing images with various levels of restorativeness. *J. Environ. Psychol.* **57**, 10–16 (2018)
8. Gao, B., Woo, W.L.: Wearable audio monitoring: content-based processing methodology and implementation. *IEEE Trans. Hum. Mach. Syst.* **44**(2), 222–233 (2014)
9. Gunn, S.R., et al.: Support vector machines for classification and regression. *ISIS Technical Report* **14**(1), 5–16 (1998)
10. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: KNN model-based approach in classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) *OTM 2003. LNCS*, vol. 2888, pp. 986–996. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39964-3_62
11. Hogg, A.O., Evers, C., Naylor, P.A.: Speaker change detection using fundamental frequency with application to multi-talker segmentation, pp. 5826–5830. *IEEE* (2019)
12. Jalil, M., Butt, F.A., Malik, A.: Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals, pp. 208–212. *IEEE* (2013)
13. Jiang, D., Hu, B., Chen, Y., Xue, Y., Li, W., Liang, Z.: Recognizing the human attention state using cardiac pulse from the noncontact and automatic-based measurements. *Soft. Comput.* **22**(12), 3937–3949 (2018)
14. Jingyi, L., Yong, M., Puzhen, L., Andreas, B.: A journey through nature: exploring virtual restorative environments as a means to relax in confined spaces. Association for Computing Machinery, New York, NY, USA (2021)
15. Joshi, D.D., Zalte, M.: Speech emotion recognition: a review. *IOSR J. Electron. Commun. Eng. (IOSR-JECE)* **4**(4) (2013)
16. Lam, K.Y., et al.: Smartmood: toward pervasive mood tracking and analysis for manic episode detection. *IEEE Trans. Hum. Mach. Syst.* **45**(1), 126–131 (2014)
17. Li, X., Tan, N., Wang, T., Su, S.: Detecting driver fatigue based on nonlinear speech processing and fuzzy SVM, pp. 510–515. *IEEE* (2014)
18. Maurya, A., Kumar, D., Agarwal, R.: Speaker recognition for Hindi speech signal using MFCC-GMM approach. *Procedia Comput. Sci.* **125**, 880–887 (2018)
19. Narayanan, A.M., Bertrand, A.: Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection. *IEEE Trans. Biomed. Eng.* **67**(1), 234–244 (2019)

20. Ohly, H., et al.: Attention restoration theory: a systematic review of the attention restoration potential of exposure to natural environments. *J. Toxicol. Environ. Health, Part B* **19**(7), 305–343 (2016)
21. Popov, D., Gapochkin, A., Nekrasov, A.: An algorithm of Daubechies wavelet transform in the final field when processing speech signals. *Electronics* **7**(7), 120 (2018)
22. Rocha, R.B., Freire, V.V., Alencar, M.S.: Voice segmentation system based on energy estimation, pp. 860–864. *IEEE* (2014)
23. Schuller, B.W.: *Intelligent Audio Analysis*. Signals and Communication Technology, Springer, Heidelberg (2013). <https://doi.org/10.1007/978-3-642-36806-6>
24. Stevenson, M.P., Schilhab, T., Bentsen, P.: Attention restoration theory ii: a systematic review to clarify attention processes affected by exposure to natural environments. *J. Toxicol. Environ. Health Part B* **21**(4), 227–268 (2018)
25. Stone, R., Small, C., Knight, J., Qian, C., Shingari, V.: Virtual natural environments for restoration and rehabilitation in healthcare. *Virtual Augment. Real. Ser. Games Healthc.* **1**, 497–521 (2014)
26. Sun, Y.X., Ma, Y., Shi, K.B., Hu, J.P., Zhao, Y.Y., Zhang, Y.P.: Unsupervised speaker segmentation framework based on sparse correlation feature, pp. 3058–3063. *IEEE* (2017)
27. Visuri, A., van Berkel, N.: Attention computing: overview of mobile sensing applied to measuring attention. Presented at the (2019)
28. Wieland, B., Urban, K., Funken, S.: *Speech signal noise reduction with wavelets*. Verlag nicht ermittelbar, Ph.D. thesis (2009)
29. Wilcoxon, F.: Individual comparisons by ranking methods. In: Kotz, S., Johnson, N.L. (eds.) *Breakthroughs in Statistics*, pp. 196–202. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-4380-9_16
30. Woods, D.L., et al.: Improving digit span assessment of short-term verbal memory. *J. Clin. Exp. Neuropsychol.* **33**(1), 101–111 (2011)
31. Yang, Y., et al.: Kullback-Leibler divergence frequency warping scale for acoustic scene classification using convolutional neural network, pp. 840–844. *IEEE* (2019)
32. Yella, S.H., Stolcke, A., Slaney, M.: Artificial neural network features for speaker diarization, pp. 402–406. *IEEE* (2014)