The sound of silence: Breathing analysis for finding traces of trauma and depression in oral history archives

Almila Akdag Salah, Albert Ali Salah, Heysem Kaya and Metehan Doyran

Information and Computing Sciences, Utrecht University, Netherlands

Evrim Kavcar

Faculty of Arts, Mardin Artuklu Universitesi, Turkey

Abstract

Many people experience a traumatic event during their lifetime. In some extraordinary situations, such as natural disasters, war, massacres, terrorism, or mass migration, the traumatic event is shared by a community and the effects go beyond those directly affected. Today, thanks to recorded interviews and testimonials, many archives and collections exist that are open to researchers of trauma studies, holocaust studies, and historians, among others. These archives act as vital testimonials for oral history, politics, and human rights. As such, they are usually either transcribed or meticulously indexed. In this work, we propose to look at the nonverbal signals emitted by victims of various traumatic events when they describe the trauma and we seek to render these for novel representations without taking into account the explicit verbal content. Our preliminary paralinguistic analysis on a manually annotated collection of testimonials from different archives, as well as on a corpus prepared for depression and post-traumatic stress disorder detection indicates a tentative connection between breathing and emotional states of speakers, which opens up new possibilities of exploring oral history archives.

Correspondence:
Almila Akdag Salah,
Information and Computing
Sciences, Utrecht University,
Netherlands.
E-mail:
a.a.akdag@uu.nl

1 Introduction

Oral history archives constitute an important branch of historical research with the aim of preserving subjective narratives about historically significant events, using structured interviews as the main methodology. With the digital turn, collecting audio-visual materials for the preservation of orally transmitted personal experiences and memories took a new turn (Kemman *et al.*, 2014). While the earlier collections, many of which were generated for a specific research project, were

doi:10.1093/llc/fqaa056 Advance Access published on 17 December 2020

fully or partially transcribed by a team of researchers in order to answer project-specific questions, today (semi-) automatic speech-to-text transcription tools are developed to prepare such collections for further use (de Jong *et al.*, 2011; Draxler *et al.*, 2020).

With the development of technologies such as paralinguistic speech analysis, audio-visual collections can be automatically processed for information that goes beyond the content of the spoken text (Ordelman *et al.*, 2008). Nonverbal cues of communication, if automatically processed, can for instance highlight emotional moments in thousands of hours of archival material (de Jong *et al.*, 2006). This opens new possibilities in curating such material, retrieving information and new ways of perusing the archives. In this article, we pursue such an approach by focusing on one paralinguistic feature, i.e. breathing, with the aim to analyze the emotional state of the narrators of traumatic events. The main hypothesis of this article is that trauma leaves traces in a person, and some of those traces can be automatically detected in the breathing patterns of people talking about their traumatic experiences. Subsequently, our main research problem is to assess the contribution (or predictive potential) of breathing-related features for automatic prediction of emotional states from speech. Such features will contribute to the exploration of oral history archives, but potentially, can also be used for providing quantitative indicators for trauma diagnosis, for tracking the alleviation of symptoms over time, or they can be transformed into a means of making the emotional intensity of the trauma accessible without revealing the content of the trauma.

2. Trauma Diagnosis, Post-Traumatic Stress Disorder, and Depression

American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) defines a traumatic event as an event that is 'generally outside the range of usual human experience' and would 'evoke significant symptoms of distress in almost everyone' (APA, 2000; APA, 2013). In the diagnosis of post-traumatic stress disorder (PTSD) and acute stress disorder (ASD), DSM symptoms are essential. If the duration of symptoms is less than 3 months, it is defined as ASD; otherwise, it is defined as PTSD. There are scarcely any studies that detect PTSD and ASD without clinical data. Some of these studies include information about heart rate, pulse, and breathing patterns of individuals (Davis *et al.*, 1996; Shalev *et al.*, 1998; Ogden and Minton, 2000; Scherer *et al.*, 2015). Although vocal parameters are used in order to detect clinical depression, no studies investigating the relationship between voice and trauma were found. Some recent work investigates PTSD symptom severity in a multimodal fashion, using questionnaires and skin conductance physiology (Mallol-Ragolta *et al.* 2018).

Automatic depression assessment via computational approaches typically focuses on language content (Althoff *et al.*, 2016), facial behavior (Cohn *et al.*, 2009), and vocal cues (Cohn *et al.*, 2009; Cummins *et al.*, 2015). State of the art methods combine multiple modalities, which increase the computational complexity of the used approach, but bring advantages in robustness and accuracy. In this work, we focus on the relationship between traumatic events, speech (recollection of traumatic events), and breathing (breathing patterns during the narration of traumatic events). Here, it is important to emphasize the differences between the everyday usage of the word 'trauma' and the clinical terminology. Many people might experience a traumatic event (as defined by DSM), but such an experience not necessarily leads to long-lasting symptoms of PTSD or ASD.

3. General Approach

Our general approach is to use speech features to automatically segment trauma survivor testimonials into 'speaking', 'breathing', and 'silence' classes. The duration of the silences, the frequency of breathing and its quality, the whole dynamics of the testimonial are investigated through unsupervised learning and visualization. We work with two different data sources. The first one is a collection of segments from various oral archive videos (nine subjects, two clips per subject, 1.5–2 min per clip), and the second one is the distress analysis interview corpus-Wizard of Oz (DAIC-WOZ) corpus (twenty-five subjects, PTSD severity ranges between 17–75 and 15–20 min per subject), recently used for the Audio/Visual Emotion Challenge and Workshop (AVEC, 2019) for depression and PTSD detection (Ringeval *et al.*, 2019). Oral archive testimonials are not collected right after the traumatic events, and do not have PTSD evaluation. Even if survivors had PTSD after the event, the time interval between the events and their recollection may have been enough for them to shake off the symptoms of PTSD. However, our previous findings indicate that simple breathing features extracted from nonverbal parts of communications provide promising cues for depression and PTSD level prediction, and could potentially complement information based on the language content (Kaya *et al.*, 2019).

4. Speech and Breath Features

The most commonly used speech processing techniques in the recognition of emotions and clinical depression in the literature are related to prosody (i.e. pitch, jitter, energy, pause time, and speaking rate), as well as the spectral features (i.e. formants) and cepstral features (i.e. Mel frequency cepstral coefficients). Prosodic, source, and acoustic features, as well as vocal tract dynamics are speech-related features affected by depression. Researchers have found that depressed subjects are prone to possess a low dynamic range of the fundamental frequency, a slow speaking rate, a slightly shorter speaking duration, and a relatively monotone delivery (Mundt *et al.*, 2007; Low *et al.*, 2010; Cummins *et al.*, 2011; Alghowinem 2013; Williamson *et al.*, 2013). The long silences and filled pause segments might also indicate a correlation with high emotional state and personal content (Heuvel and Oostdijk, 2016).

Breathing consists of two phases called inspiration and expiration. During inspiration, the diaphragm is used to increase the volume of the chest cavity, causing the air to enter by mouth or nose to fill the low-pressure lungs. During a normal expiration, the diaphragm and external muscles are relaxed, the chest volume is lowered, the pressure increases, and breath is dispelled. During speech activity, some of the air stored in the lungs is spent to produce sounds. Subsequently, long bouts of speaking also require taking a breath. Breathing is intimately connected to oxygen flow and thus is involved in all internal and external systems, including circulation, hormone systems, and the nervous system. Deep breathing in general seems to allow emotional release and processing of the energy stuck in the body from the trauma. Additionally, Ogden and Minton (2000) observed that subjects can have breathing difficulties when they are talking and extemporizing about their trauma.

Many features can be extracted from the speech signal for the automatic detection of breathing, including Mel Frequency Cepstrum Coefficient parameters, short-time energy, zero-crossing rate, spectral slope, and duration. State of the art breathing detection algorithms use convolutional and recurrent neural networks (Nallanthighal and Strik, 2019). However, breathing detection is a difficult problem under noisy recording conditions. The testimonials we are using in this work have many sources of noise to confound the automatic algorithms, including background music and other persons talking. Therefore, we have extended our experiments on a dataset collected in a lab environment (DAIC-WOZ corpus), where the PTSD scores and depression levels of the participants were measured (Gratch *et al.*, 2014; Ringeval *et al.*, 2019). In the next sections, we summarize our preliminary findings on two separate data collections with different conditions.

5. Experiments with Oral History Archive Data

In order to test the quality of various collections, as well as if the language, culture, and the traumatic events' nature and date have a measurable effect on breathing patterns, we have collected short clips from survivor testimonials. We sampled different languages (English, Chinese, Japanese, Spanish, Kinyarwanda), and different mass-traumas (Holocaust, Nanjing Massacre, Tsunami, Guatemalan Genocide, Tutsi Massacre). In the first phase of the project, a total of eighteen clips from nine survivors were manually annotated for 'speaking', 'silence', 'breathing', 'lip noise', 'other people speaking' classes. For each survivor, we selected a 'normal' speech and a more 'emotional' speech segment, where the subject describes events or feelings related to the traumatic experience.[1] For some survivors, the date of the events was decades ago (e.g. Nanjing Massacre and Holocaust), for some, the memories were fresh (e.g. Tsunami). Each speech segment is annotated to extract the time span of speech, silence, and breathing periods.

In Fig. 1, each line represents one survivor's speech segments, where light blue circles are speech, purple circles are breathing, and green circles are silences. The horizontal dimension shows the time (the longest segment is 1.5 min), and the circle radii are proportional to duration. On the left side, the normal speech segments are aligned, whereas the ones on the right side are emotionally charged. Typical features in the latter include long silences, pierced by deep breath—especially before telling about the most traumatic event—sometimes frequent and sharp breathing. As expected, even in such a small sample size, certain characteristics prevail. For instance, personal speech patterns are different, and set the tempo of the speech as well as breathing, but within the personal tempo, deep breathing emerges. Some of these patterns are not heard while listening to the video/audio recordings themselves but become visible only after the annotated portions of breathing patterns are visualized. Cultural approaches to trauma and disasters might dictate the way events are described, but this small sample suggests the possibility that the breathing and silence patterns that occur while telling a traumatic event are shared across cultures.

6. Experiments with the DAIC-WOZ Corpus

The DAIC is a database of recorded interactions in the format of semi-structured clinical interviews (Gratch *et al.*, 2014). It is designed to analyze indicators of PTSD and major depression. We use a part of this corpus, DAIC-WOZ (where WOZ stands for Wizard of Oz), where a human is interviewed by a virtual agent controlled by a human interviewer in another room. The average duration of these interviews was 16 min (the shortest being 530 s, and the longest 1,567 s). Each interviews' clinical measures include PTSD and depression severity levels [Patient Health Questionnaire (PHQ)]. In our analysis, we have tested automatic segmentation of the speech signal to catch nonverbal elements like silence, breathing, and other non-linguistic vocalizations, and contrasted these with features based on the Automatic Speech Recognition (ASR) transcripts. Our results showed that nonverbal parts of the signal are important for detection of depression, yet combining this with linguistic information produced the best results (Kaya *et al.*, 2019).

7. Emotional Moments and Breathing

When we look at how breathing patterns change with different PTSD levels, we notice that subjects with low

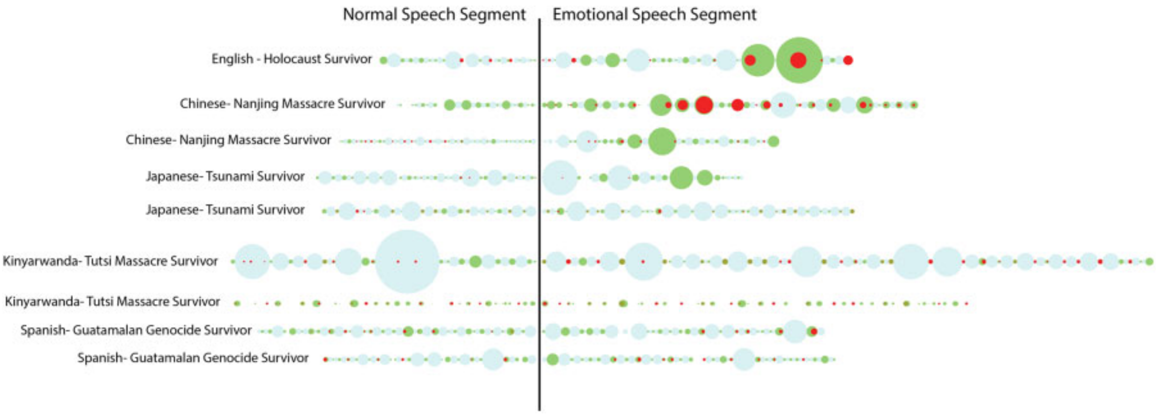

Fig. 1 Visualization of breath (red), silence (green), speech (light blue) patterns

PTSD scores do not take long breaths when compared to subjects with high PTSD scores. The subjects with severe PTSD scores have many segments where they breathe for more than 3 s. There are no statistically significant patterns here. According to a one-sample t-test ($t = 1.30$, $P > 0.01$; Levene's test supports equal variances), the mean duration of breathing under higher and lower PTSD scores is not statistically significantly different. However, it is 2.53 s (std = 0.96) for higher PTSD, and 1.98 s (std = 0.86) for lower PTSD scores, and this indicates a tendency and suggests more data points should be investigated for a more reliable assessment. Figure 2 shows three features for high PTSD (right half) versus low PTSD (left half) subjects. We divide the clips into 40-s segments, and show the minimum speech segment, maximum silence duration, and maximum breathing duration, each as a proportion of the segment. The subjects are indicated with their PTSD scores (17 for the lowest scores, 85 for the highest). Clearly, longer silence segments are prominent in high PTSD subjects, as well as higher maximum breathing proportion. Also, these subjects have many segments with no speech at all, as indicated by zeros in minimum speech proportion.

Breathing changes depending on several factors, including physical and cognitive load, and emotional state. Obviously, breathing alone cannot reliably indicate traumatic experience, let alone PTSD. But here, we are not after a single-feature solution to automatic detection of trauma; indeed, there cannot be such a generic solution for this complex problem. Every trauma is unique and many indicators need to be considered for a clinical diagnosis.

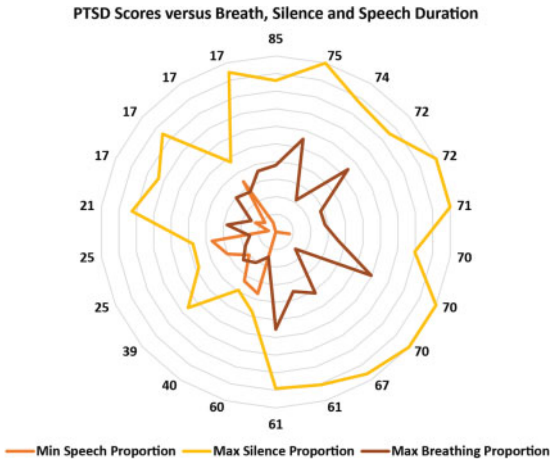

Fig. 2 Features for breath, silence, and speech presence in clips segmented into 40-s intervals. Each subject is indicated by a PTSD score (17–85)

8. Conclusions

The oral history archives could offer a rich source of data for trauma research. The traumatic events described in the archival material contain not only the political upheavals, massacres, wars, and natural disasters of the twentieth century, but also the cultural, social, and linguistic perspectives in the narrator voices and their way of narration of these traumatic events. As such, by tracing the breathing patterns of the narrators we will have a chance to see if trauma leaves somatic traces behind and beyond the cultural and linguistic barriers. We aim to enrich the semantic information contained in oral history archived by adding non-linguistic features. However, trauma cannot be 'reduced' to simple patterns and features, and purely automatic emotion mining approaches would lead to fragmentation of narratives and a loss of context, unless the limitations of these methods are made clear from the onset, and their contributions are seen as additions to existing and well-established methods.

References

Alghowinem, S. (2013). From joyous to clinically depressed: mood detection using multimodal analysis of a person's appearance and speech. *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. Geneva, pp. 648–654. doi: 10.1109/ACII.2013.113.

Althoff, T., Clark, K., and Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, **4**: 463–76.

American Psychiatric Association (APA). (2000). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn, text rev. Washington, DC: American Psychiatric Publishing.

American Psychiatric Association (APA). (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. Arlington, VA: American Psychiatric Publishing.

Cohn, J. F., Kruez, T. S., Matthews, I. *et al.* (2009). Detecting depression from facial actions and vocal prosody. 3rd Int. *Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–7.

Cummins, N., Epps, J., Breakspear, M., and Goecke, R. (2011) An investigation of depressed speech detection: features and normalization. *Twelfth Annual Conference of the International Speech Communication Association*, pp. 1–4.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, **71**: 10–49.

Davis, J. M., Adams, H. E., Uddo, M., Vasterling, J. J. and Sutker, P. B. (1996). Physiological arousal and attention in veterans with posttraumatic stress disorder. *Journal of Psychopathology and Behavioral Assessment*, **18**(1): 1–20.

de Jong, F. M. G, Ordelman, R. J. F., and Scagliola, S. I. (2011)**.** Audio-visual collections and the user needs of scholars in the humanities: a case for co-development. In *Proceedings of the 2nd Conference on Supporting Digital Humanities.* Retrieved from http://hdl.handle.net/1765/77364

de Jong, F., Ordelman, R., and van Hessen, A. (2006). The role of automated speech and audio analysis in semantic multimedia annotation. In *Proceedings of the IET International Conference on Visual Information Engineering (VIE 2006)*, pp. 249–54.

Draxler, C., Van den Heuvel, H., Van Hessen, A., Calamai, S., Corti, L., Scagliola, S. (2020) A CLARIN Transcription Portal for Interview Data. In *Proceedings of the 12th International Conference on Language Resources and Evaluation* (LREC2020), pp. 3346–52.

Gratch, J., Artstein, R., Lucas, G. M. et al. (2014). The distress analysis interview corpus of human and computer interviews. In *Proceedings of LREC*, pp. 3123–8.

Heuvel, H. and Oostdijk, N. H. J. (2016). Falling silent, lost for words. . . Tracing personal involvement in interviews with Dutch war veterans. In *Proceedings of LREC*, pp. 998–1001.

Kaya, H., Fedotov, D., Dresvyanskiy, D. et al. (2019). Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 27–35.

Kemman, M., Scagliola, S., de Jong, F., and Ordelman, R. (2014). Talking with scholars: developing a research environment for oral history collections. In Bolikowski, Ł., Casarosa, V., Goodale, P., Houssos, N., Manghi, P., and Schirrwagen, J. (eds), *Theory and Practice of Digital Libraries—TPDL 2013 Selected Workshops*, vol. **416.** Springer International Publishing, pp. 197–201. DOI : 10.1007/978-3-319-08425-1_22.

Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. and Allen, N. (2010). Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *Proceedings of IEEE ICASSP*, pp. 5154–7.

Mallol-Ragolta, A., Dhamija, S., and Boult, T. E. (2018). A multimodal approach for predicting changes in PTSD symptom severity. In *Proceedings of ACM ICMI*, pp. 324–33.

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., and Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, **20**(1): 50–64.

Nallanthighal, V. S. and Strik, H. (2019). Deep sensing of breathing signal during conversational speech. In *Proceedings of Interspeech*, Graz, Austria, pp. 4110–14. DOI: 10.21437/Interspeech. 2019–1796.

Ogden, P. and Minton, K. (2000). Sensorimotor psychotherapy: One method for processing traumatic memory. *Traumatology*, **6**(3): 149–73.

Ordelman, R., Heeren, W., van Hessen, A. *et al.* (2008). Browsing and searching the spoken words of Buchenwald survivors. In *BNAIC 2008 Belgian-Dutch Conference on Artificial Intelligence*, pp. 403–4.

Rees B., and Smith J., 2008. Breaking the silence: The traumatic circle of policing. *International Journal of Police Science & Management,* **10**(3): 267–79.

Ringeval, F., Schuller, B., Valstar, M. *et al.* (2019). *AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition*, arXiv preprint arXiv : 1907.11510.

Scherer, S., Lucas, G. M., Gratch, J., Rizzo, A. S., and Morency, L. P. (2015*)*. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, **7**(1): 59–73.

Shalev, A. Y., Sahar, T., Freedman, S. et al. (1998). A rospective study of heart rate response following trauma and the subsequent development of posttraumatic stress disorder. *Archives of General Psychiatry*, **55**(6): 553–9.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., and Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge*, pp. 41–8.

Notes

[1] Usually, recollections of traumatic events are very emotional. However, it is also not uncommon where the survivor relates these events in a manner devoid of emotional content (Rees and Smith, 2008). How the breathing patterns change in such cases is a future research question.