Chapter 8
"You Can Do It!"—Crowdsourcing Motivational Speech and Text Messages

Roelof A. J. de Vries, Khiet P. Truong, Jaebok Kim and Vanessa Evers

Abstract Recent approaches for technology, that assist or encourage people to change their exercise behavior, focus on tailoring the content of motivational messages to the user. In designing these messages, the mode and style of presentation, e.g., spoken or written and tone of voice, are also thought to play an important role in the effectiveness of the message. We are interested in studying the effects of the content, mode, and style of motivational messages in the context of exercise behavior change. However, we are not aware of any accessible database on motivational messages. Moreover, collecting a large database of spoken and written messages is not a trivial task. Crowdsourcing can be an effective way to collect a large amount of data for all sorts of tasks. Traditionally, crowdsourcing tasks are relatively easy for participants (microtasks). In this work, we use crowdsourcing to collect a large amount of data for more complex tasks (macrotasks): designing motivational messages in text and recording spoken motivational messages. We present and discuss the approach, database and challenges we ran into, and report findings on unsupervised explorations of the emotional expressiveness and sound quality (signal-to-noise ratio, SNR) of the crowdsourced motivational speech.

8.1 Introduction

Recently, there is a growing interest to investigate and develop motivational technology that assists or encourages people to change their behavior (Hekler et al. 2013). This technology can be used to encourage the user, for example, to exercise more by pushing motivational messages to the user on mobile phones (Klasnja and Pratt 2012). Many studies describing the development of their technology do not explain in detail *how* they designed the motivational messages used (Latimer et al. 2010). The framing, content, and designer of motivational messages is an important and not

R. A. J. de Vries (✉)
Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands
e-mail: rajdevries@gmail.com

K. P. Truong · J. Kim · V. Evers
Human Media Interaction, University of Twente, Enschede, The Netherlands

V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,
Human–Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_8

a trivial aspect that should be considered carefully when developing motivational or behavior change technology (de Vries et al. 2017a; de Vries 2018). Personalization (e.g., tailoring to the user's personality, Arteaga et al. 2010; de Vries et al. 2016a, 2017b) could for example be a framing method with a positive influence on exercise adherence, but this needs to be investigated in more detail. Furthermore, the mode of presentation and style, e.g., spoken or written and tone of voice, could play an important role in exercise adherence.

In order to study the design and effectiveness of different types of motivational messages for motivational technology, a large database with varying motivational messages in different modes of presentation (i.e., spoken and written) was developed by the authors. For our purpose, namely developing a smartphone application to support exercise behavior change, we decided to gather motivational messages not only in text form but also in spoken form. Rather than generating a small set of messages or relying on experts, we opted for generating a large set of motivational messages by non-experts (peers) through *crowdsourcing*. In our crowdsourcing survey, participants were asked to come up with motivational messages (submitted in *written* and *spoken* form) for a hypothetical person in a given scenario about exercising. This setup allows us to collect a large number of written and spoken motivational messages to study the effectiveness of the message's modality (written vs. spoken), content (themes and topics of the messages relating to the scenarios), and vocal expressivity. In this chapter, we focus on the vocal expressivity of the spoken messages.

Crowdsourcing is usually used for small and easy tasks called microtasks (Cheng et al. 2015). Crowdsourcing written transcriptions, translations or annotations (e.g., Marge et al. 2010; Zaidan and Callison-Burch 2011; Hsueh et al. 2009) is a relatively frequent natural language processing task. However, eliciting spoken data through crowdsourcing seems to be less common and to our knowledge, this is the first effort in using crowdsourcing for the complex task of collecting text-based motivational messages as well as spoken motivational messages. A challenge for a complex task, also called a (non-decomposable) macrotask (Schmitz and Lykourentzou 2018), like this is evaluating the quality of the workers' output, because no ground truth is available (Haas et al. 2015). Crowdsourcing spoken messages brings along additional challenges: loss of control over the recorded sound quality and the speaking style of the participant are among those challenges. Participants will have different types of microphones with varying qualities and there is no knowing to what extent the spoken material actually reflects a motivational speaking style after listening to all the audio recorded. Despite these challenges, it would be useful to explore the feasibility of acquiring spoken data through crowdsourcing involving variations in speaking styles (i.e., motivational) that enables paralinguistic research, which is still a rather uncovered area in crowdsourcing.

In this chapter, we present our approach to crowdsourcing spoken (and written) motivational messages and present our collected corpus. We discuss how we designed the data collection and we report on (1) the audio quality (SNR) of the crowdsourced audio material and (2) an initial, unsupervised exploration of the acoustical feature space of motivational speech. We describe related work in Sect. 8.2 and present our

data collection effort in Sect. 8.3. We report on an preliminary exploration of the quality and acoustics of motivational speech in Sect. 8.4 and discuss the conclusions and future research in Sect. 8.5.

8.2 Related Work

We explain relevant psychological concepts used in our study and discuss previous related work.

8.2.1 Motivation and Exercise Behavior Change

According to the Transtheoretical Model (TTM, Prochaska and DiClemente 1983), people, who change their exercise behavior for example, will go through certain *stages of change*. These five stages of change classify people into progressing stages of behavior change as follows: Precontemplation (not considering change), Contemplation (thinking of change), Preparation (preparing for change), Action (actively making changes), and Maintenance (maintaining the change). We expect that motivational messages attuned to the stages of change a user is in will be more effective for exercise adherence. However, in an evaluation of the text version of the spoken motivational messages described in this chapter, we found that the way people *rate* the messages on how motivating they are does not always match the expectation of what messages should be most effective for the stage of change they are in (more details are reported in de Vries et al. (2016b)).

8.2.2 Crowdsourcing Text and Speech

Over the last few years, researchers have been using crowdsourcing platforms such as Amazon Mechanical Turk (AMT) for various natural language processing (NLP) tasks. Callison-Burch and Dredze (2010) and Parent and Eskenazi (2011) summarize the kind of NLP tasks commonly addressed which include, among others, transcriptions of spoken language (Marge et al. 2010), producing and evaluating (machine) translations (Zaidan and Callison-Burch 2011; Callison-Burch 2009), and sentiment labeling (Hsueh et al. 2009). These tasks usually involve assessing text or spoken data. Crowdsourcing platforms can also be used to *acquire* spoken language data. Although challenging (for example, there is no way to control the microphone type, distance or noise level), collecting spoken language data through crowdsourcing can be a cost- and time-effective way to gather large amounts of speech data under realistic conditions. Recently, efforts to collect speech data through crowdsourcing have been undertaken involving tasks such as reading aloud street addresses (McGraw

et al. 2010), having conversations with a spoken dialogue system (McGraw et al. 2010), narrating Wikipedia articles for use by blind or illiterate users (Novotney and Callison-Burch 2010), reading aloud sentences in under-resourced languages (Lane et al. 2010), and annotating photos through spoken descriptions for a voice search system (McGraw et al. 2011). Challenges discussed in these studies include loss of (quality) control and also technical challenges since incorporating a web-based audio collection framework in crowdsourcing platforms such as AMT is not straightforward. Studies on the prosody of motivational speech, with the exception of a recent study by Skutella et al. (2014) are rare. In instructor–trainee indoor cycling sessions they found, among other things, a high frequency of prominent, accented words fulfilling a coordinative and informative function. We are aware of only one related study on collecting motivational messages, by Coley et al. (2013), where written *text* messages were crowdsourced to encourage people to quit smoking. With our effort of crowdsourcing motivational speech and text messages, we aim to address this lack of data and research and demonstrate the feasibility of crowdsourcing spoken motivational messages.

8.2.3 Defining Macrotasks and Microtasks

Macrotasking, as defined by this book, refers to complex and often creative crowd work, which may or may not be decomposable to microtask level, but which differs from microtasking in that it requires more worker time, can accept free-form worker input (i.e., not only multiple-choice standardized input), and its quality needs to be, at least partially, determined through subjective evaluation, for example peer review. Microtasks, in contrast, are small tasks that are easily performed. Microtasks are frequently used in crowdsourcing (Cheng et al. 2015).

Considering the tasks mentioned in the related works discussed in the previous section in light of this definition of macro and microtasks, all of those tasks mentioned could be considered microtasks, although for some this is only because they are decomposed to microtask level. Narrating articles (Novotney and Callison-Burch 2010), reading aloud sentences (Lane et al. 2010), or transcribing spoken language (Marge et al. 2010) is relatively easy and straightforward and therefore fits the microtasks definition well. However, producing and evaluating (machine) translations (Callison-Burch 2009; Zaidan and Callison-Burch 2011) and sentiment labeling (Hsueh et al. 2009), depending on the difficulty of the text, can require some cognitive effort. Moreover, having conversations with a spoken dialogue system (McGraw et al. 2010) and annotating photos through spoken descriptions for a voice search system (McGraw et al. 2011) can also require quite some cognitive effort depending on the dialogue or the photo. For these tasks, it seems that what qualifies them for microtasks is that these tasks were decomposed to the simplest level, such as describing only one photo, or have one short dialogue, and in that way they require very little worker time. On the other hand, these tasks could also qualify for macrotasks because they require free-form input, the quality of the input needs to be determined through subjective evaluation, and the tasks are not easily performed.

Applying this definition to the task we designed, our crowdsourcing task can be considered a macrotask. Our crowdsourcing task required creativity and quite some worker time (participants were asked to come up with multiple motivational messages for a hypothetical person in a given scenario about exercising, see Tables 8.1 and 8.2), accepted only free-form worker input (the participants had to design all the messages from scratch), and the quality was partially determined through subjective evaluation (more details on our evaluation are reported in de Vries et al. 2016b). This is also what makes a macrotask like this challenging, because there is no ground truth available for evaluating the quality of the workers' output (Haas et al. 2015). Moreover, it is challenging because we decided to gather motivational messages not only in text form but also in spoken form. Also, crowdsourcing spoken messages brings along additional challenges (e.g., loss of control over the recorded sound quality and speaking style of the participant).

On the other hand, our macrotasks could be decomposed into smaller tasks by asking participants for only one motivational message each. In this way, the task would require less participant time and could arguably move toward a microtask. However, this task would then be non-decomposable and still require a certain creativity of the participant (to come up with a motivational message) and the quality of the message would still be determined through subjective evaluation. Moreover, for the purpose of our data collection, we were also interested in variation in the motivational messages, which is stimulated by asking participants for multiple motivational messages, in that sense the task was non-decomposable. Another facet to consider is the complexity of the tasks. According to Schmitz and Lykourentzou (2018, p. A:7): "Macrotasks are almost always complex, in that they require multiple interconnected knowledge domains …". Our task however, is not so complex or difficult that it requires worker training, in fact, the task is purposefully crowdsourced

Table 8.1 One of the five stage of change scenarios (Contemplation) and the macrotask of designing multiple motivational messages for specific time frames (with one collected example)

One of the stage of change scenarios: Contemplation
Contemplation: *"Consider a middle-aged person, with a steady personal life and solid friend foundation. This person lacks regular exercise in his/her daily life, but has been thinking about starting to exercise regularly and wonders if he/she will be able to do it. This person is opting to start in the next 6 months"*
Long: "Imagine you have to provide this person with motivational messages during a long period of time (for example, 1 year) and these messages take into account the current exercise habits as described. These messages would be provided every other week (for example, week 1 and week 3 of every month). What would be 3 messages you can think of?" **Example**: *"You have to start somewhere"*
Short: "Imagine you have to provide this person with motivational messages during a short period of time (for example, 1 month) and these messages take into account the current exercise habits as described. These messages would be provided three times a week (for example Monday, Wednesday and Friday). What would be 3 messages you can think of?"

Table 8.2 One of the three running performance scenarios (Running too fast) and the macrotask of designing multiple motivational messages for specific points of time in a run (with one collected example)

One of the running performance scenarios: Running too fast
Too fast: *"Consider a person during an actual exercise, for example running, he/she is well under way in his/her run but for the purpose of a good exercise it would be best if he/she decreases the intensity of the run"*
During: "Imagine you have to provide this person with motivational messages during this session of physical activity and these messages would be provided to encourage and motivate this person to decrease the intensity during the run. What would be 3 motivating messages you can think of?"
After: "Consider the same person after the exercise (the run), he/she has exercised and so he/she is done, but did not succeed in decreasing the intensity of the run, despite the motivational messages, and is now cooling down. Although disappointing at this moment, running regularly is what is most important. What would be 3 motivating messages you can think of?" **Example**: *"Great run, but watch your speed next time"*
Before: "Consider the same person before his/her next exercise (the run). In the last run it would have been better to have had a lower intensity. This person decides to run again, partially because of the messages during his/her cooling down the other day, and is ready to start. What would be 3 motivating messages you can think of?"

to reach people who do not have the domain knowledge to design expertise driven motivational messages (designing motivational messages is the task), but who design motivational messages from their (potentially limited) own experience. In that sense, our task does not fit the general complexity criterion of macrotasks.

8.3 Data Collection

We describe how we designed our macrotasks and collected a database of spoken and written motivational messages through crowdsourcing.

8.3.1 Participants

We recruited participants via AMT. The requirements were that they needed to have completed more than a 1000 tasks on AMT, more than 98% of their tasks needed to be approved successfully, and they needed to be located in the US. These requirements ensured that we would have participants who were experienced and serious in filling in questionnaires, and that they had good proficiency in English (95% of the recruited participants reported "very good" for their self-assessed proficiency of English). The sample size consists of 500 people. Of these, 17 were excluded because their data is incomplete or have numerous outliers. Then, another 19 were excluded because they

have missing audio files (recording audio was encouraged but not strictly required to finish the survey). The final sample for spoken messages includes 464 participants (246 male). All but 4 participants were native English speakers. The minimum age was 18 and the maximum was 68. The average age was 30.93 (SD = 9.13) and the median 29.0.

8.3.2 Method

The macrotask for the participants throughout this survey was to come up with motivational messages to motivate certain people in different scenarios (in a randomized order). Since one of the features of our intended smartphone application is the use of motivational messages tailored to the stage of change, scenarios were manipulated based on the stages of change. See Table 8.1 for examples that describe a person in a situation corresponding to a certain stage of change. Participants were asked to come up with 6 different messages to motivate this person to exercise more, 3 for the **short** and 3 for the **long** term.

Another intended feature of our smartphone application is to provide motivational feedback about the quality of exercise. Hence, the second manipulation involved the running performance (running too fast, too slow or exactly right) of the person described, see Table 8.2 for example. Participants were asked to come up with 9 different motivational messages: 3 for **before**, 3 for **during**, and 3 for **after** a running session.

8.3.3 Implementation

Although we used AMT to enlist participants, the survey itself could not be embedded in AMT due to technical constraints with collecting audio. We needed to prompt the participants in the survey with the written motivational messages they had come up with earlier (and not predefined prompts) to record them on our web application outside the survey. Because we only found an option with static (predefined) prompts in AMT, we had to come up with a workaround. We used a relatively easy workaround with SurveyMonkey[1] where there is a possibility to use answer text boxes as future variables (prompts). In the web survey, this allowed us to refer to the future variable name identifier (i.e., a number in front of the to-be-instantiated variable). For the crowdsourced speech data acquisition, we set up a web application called the WAMI recorder[2] with a Google App Engine as described in McGraw (2013)[3] and from SurveyMonkey we referred the participants to this page to record their motivational

[1] https://surveymonkey.com.

[2] https://wami-recorder.googlecode.com.

[3] https://wami-gapp.googlecode.com.

messages. All audio files (∼7000) were stored in the Google App Engine in separate folders for each participant and were automatically retrieved via a script. However, the link between participant id and audio id was lost which meant that we needed to manually link each participant to the correct folder through their matching written motivational messages.

8.3.4 Measures

In addition to basic demographic information, participants were asked to fill in a 1-item stage of change measure for exercise (Norman et al. 1998), the Godin Leisure-Time Exercise Questionnaire (Godin and Shephard 1997), a 30-item processes of change measure for exercise (Nigg et al. 1999), an 18-item self-efficacy measure for exercise (Benisovich et al. 1998), a 10-item decisional balance measure for exercise (Nigg et al. 1998),[4] and the 50-item IPIP personality questionnaire.[5] These measures are not reported on in this work.

8.3.5 Procedure

Participants signed up on AMT where they were informed of their compensation, goal of the survey and estimated time cost. They were also asked to check whether their browser and microphone worked in a test version of the WAMI recorder. Participants could then decide to proceed to the survey on SurveyMonkey where the consent form was presented. Next, participants were asked to fill in demographics and then the data collection started where they were presented with various scenarios and were asked to come up with motivational messages in written form. Subsequently, participants were asked to vocally express and record the motivational messages (on a different webpage) that they had just written. They were shown the text they had just entered and were asked to repeat the message orally as they intended it. Finally, participants were asked to fill in the questionnaires as described in Sect. 8.3.4. After completion, participants were debriefed about the detailed goals of this survey and given a completion code to fill in on AMT to receive payment. On average, the survey took about 45 min to complete. Participants were paid 3 US dollars for their participation (Table 8.3).

[4] All TTM measures adopted from http://www.uri.edu/research/cprc/measures.htm.

[5] Adopted from http://ipip.ori.org/.

Table 8.3 Descriptive statistics of the messages collected

N = 6909	Mean	Std	Median	Min	Max
Duration (s)	4.5	2.0	4.2	0.9	32.7
Number of words	9.0	5.6	8	1	97

8.4 Results

One of the main goals of this study was to collect motivational speech, but also to explore its acoustical characteristics in terms of sound quality and emotional expressiveness. We collected a total of 6960 (464 × 15) motivational messages. Using simple voice activity detection, we discarded 51 messages which did not seem to contain voice at all. First, we explore the sound quality through an analysis of Signal-to-noise ratio (SNR). Second, we analyze how feature vectors of the motivational speech are distributed in a feature vector space of emotional speech: what kind of emotion does motivational speech resemble acoustically? To the best of our knowledge, there is no other motivational speech corpus that we can use as a reference in order to validate our findings. Additionally, we do not assume whether the motivational speech collection contains spontaneous or acted emotional speech data. Therefore, we use both spontaneous and acted available emotional speech corpora as training data for our analyses. We selected the SEMAINE corpus (natural emotional speech) (McKeown et al. 2012) and the LDC Emotional Prosody Speech corpus (acted emotional speech) because of their relatively large size and variety of emotional categories.

8.4.1 SNR of the Motivational Speech Corpus

We estimate the SNR of each spoken message following a method by Hirsch (1993) using voice activity detection (VAD) and assume that each speech sample already contains some noise. Figure 8.1 illustrates the distribution of SNR for three different corpora. Our motivational speech corpus shows a median of 16.97 and mean of 16.69 ± 11.68, which are considered not optimal for automatic speech recognition (ASR) (Gong 1995; Benzeghiba et al. 2007). The SNR of the motivational speech corpus is lower than that of the other corpora considered (Kruskal–Wallis test: $\chi_2(2)$, $p < 0.0001$, followed by Nemenyi pairwise comparison $p < 0.0001$).

8.4.2 Emotional Feature Vector Space Using LDC and SEMAINE Corpus

Since we used crowdsourcing to collect a large amount of motivational speech data, we could not control for the speaking style of the participants. Moreover, there have been no studies yet (to the best of our knowledge) into prosodic characteristics of

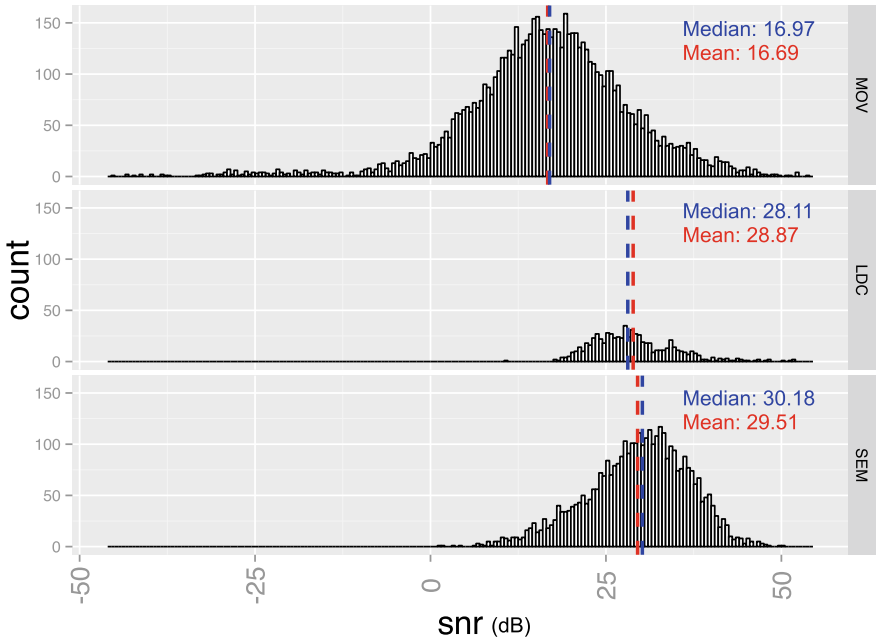

Fig. 8.1 Histogram of SNR of the speech corpora (MOV: the motivational speech corpus, LDC: the LDC Emotional Prosody Speech corpus, and SEM: the SEMAINE corpus)

motivational speech. This makes evaluation difficult. Although we can speculate that motivating people can be done by signaling positive and aroused emotions (Skutella et al. 2014), this is not verified yet. Hence, because of the relatively large amount of speech data and lack of knowledge into prosodic characteristics of motivational speech, we carried out an unsupervised cluster analysis that is exploratory of nature.

Clustering We built clusters (K-means) of each available emotional category in the feature space and investigated how close the feature vectors are to the centers of the clusters. We selected 5 representative emotional categories available in both corpora selected: neutral, happiness, anger, sadness, and boredom (Kwon et al. 2003; Huang and Ma 2006). For the SEMAINE corpus (McKeown et al. 2012), we extracted only speech segments from the users interacting with a human operator (who is playing an emotional character) that is thought to be more spontaneous. The SEMAINE corpus provides only continuous affective ratings, not discrete emotional categories. In order to map these continuous valence and arousal ratings to discrete emotional categories, we used the landmarks of the valence and the arousal dimensions as provided in FEELTRACE (Cowie et al. 2000). We calculated the Euclidean Distance between the landmarks and the values of the valence and arousal dimensions of each segment and assigned the emotional category with the smallest distance to the valence and arousal values. Lastly, we extracted segments by using VAD and time-alignment labels. Table 8.4 summarizes the emotional speech data used to build the emotional feature space.

Table 8.4 Data used to build an emotional feature vector space (No.: number of segments, F: female, M: male, A: arousal, V: valence)

Categories	No. LDC		No. SEMAINE		Landmarks	
	F	M	F	M	A	V
Neutral	34	46	1380	1314	0.00	0.00
Happiness	111	69	253	310	0.74	0.52
Anger	78	61	111	224	−0.77	0.75
Sadness	97	64	32	9	−0.7	−0.48
Boredom	88	90	219	290	−0.43	−0.48

Table 8.5 Normalized mean (standard deviations) of distances between motivational speech and emotional models

Categories	Neutral	Anger	Sadness	Happiness	Boredom
Female	0.24 (0.11)	0.25 (0.12)	0.24 (0.11)	0.24 (0.12)	0.19 (0.11)
Male	0.30 (0.11)	0.33 (0.11)	0.28 (0.12)	0.33 (0.11)	0.25 (0.10)

Feature space To build the emotional feature vector space, we extracted low-level features including energy (RMS), 12 Mel-Frequency Cepstrum Coefficients (MFCCs), prosody (F0, voice probability, zero-crossing rate), and voice quality related features (jitter, shimmer, harmonics-to-noise ratio) from only the voiced parts obtained with VAD. Feature vectors were extracted within frames of 20 ms with a Hamming window by using `openSMILE` (Eyben et al. 2010). We used only mean values of each features to construct clusters in the feature space. Since we do not know which features are dominantly related to motivational speech, we normalized all feature values by the use of the maximum and minimum values on the feature to scale them in a range of [0.0, 1.0] (de Souto et al. 2008). We found a center for each emotional category by calculating the minimum of total Euclidean distances between the center and other vectors. We normalized the distances between the motivational speech vectors and the centers of the emotional models in the same way we did for the features.

Acoustic similarity Table 8.5 presents the means of normalized distances between motivational speech feature vectors and the centers of emotional categories. For both female and male models, we can observe that the motivational speech feature vectors seem to show more acoustic similarity with boredom models than with any other models (Kruskal–Wallis test: $\chi_2(4)$, $p < 0.0001$, followed by Nemenyi pairwise comparison $p < 0.0001$). Especially, in male models, all categories show differences with significance of $p < 0.0001$ between each other except for the pair of happiness and anger.

8.5 Discussion and Conclusion

In this chapter, we presented our text and speech dataset of motivational messages collected through a crowdsourcing macrotask survey. With this data collection effort, we aimed to address the gap in both motivational technology, where datasets of motivational messages are mostly expert-written, not personalized, and relatively small, as well as in speech science, where corpora of motivational speech do not exist yet. Macrotasks, as defined by this book, refers to complex and often creative crowd work, requires more worker time, can accept free-form worker input, and its quality needs to be, at least partially, determined through subjective evaluation. Evaluating macrotasks is a challenge, because there is no ground truth available to evaluate the quality of the workers' output. We used crowdsourcing for a relatively new type of macrotask: eliciting motivational text and speech messages. This task required creative work, a long amount of worker time, free-form input, and subjective evaluation. However, the task was not necessarily very complex in that it required a lot of knowledge domains. A first unsupervised exploration of the acoustic feature space of the acquired motivational speech data was carried out which showed acoustic similarity to mostly low aroused and neutral emotional feature spaces. An SNR analysis showed relatively low SNR values by ASR standards, but we still believe that a large amount of our speech data can be used for paralinguistic research. Our study serves as a good example of how macrotasks in crowdsourcing can be used to for creative elicitation tasks, such as collecting a difficult but context-relevant text and speech dataset of crowd-designed motivational messages for cross-disciplinary use.

Although crowdsourcing seems to be a relatively easy and quick way to acquire a large amount of text and speech data, there are some limitations that one should take into account when using crowdsourcing macrotasks, in particular for a complex macrotask like speech data acquisition, see also McGraw et al. (2010), Parent and Eskenazi (2011) who discuss these limitations as well. From a practical point of view, acquiring speech through well-known crowdsourcing platforms is rather cumbersome and requires some workarounds: browser restrictions, the need to prompt the participants to read aloud what they had previously entered in text, and access to the audio files recorded lead to some cumbersome workarounds which deserve some more elegant solutions in the future. Content-wise for this specific macrotask, the unpredictability of the quality of the acquired audio is still a challenge, both the sound quality and the quality of the desired task, i.e., generating (high-quality) motivational speech. Although a comparison to existing acoustic models might give one a first insight into what the acquired speech might sound like, subsequent analyses such as perceptual rating studies are still needed for confirmation. This need for further evaluation is a general problem for macrotasks.

For future research, we will evaluate the effectiveness of the motivational text and speech messages collected through several user studies. We intend to analyze the messages for linguistic and acoustical patterns in relation to effectiveness and personalized variables such as personality and stages of change. Furthermore, the

dataset might be of interest to researchers working on speech synthesis and natural language generation: imagine an application that automatically generates motivational text and speech messages tailored to the user. Despite some limitations, we believe that our data collection effort also creates many cross-disciplinary and fruitful research opportunities.

References

Arteaga, S. M., Kudeki, M., Woodworth, A., & Kurniawan, S. (2010). Mobile system to motivate teenagers' physical activity. In *Proceedings of the 9th International Conference on Interaction Design and Children* (pp. 1–10). ACM .

Benisovich, S., Rossi, J., Norman, G., & Nigg, C. (1998). Development of a multidimensional measure of exercise self-efficacy. *Annals of Behavioral Medicine, 20*(suppl).

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786.

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, ser. EMNLP '09* (pp. 286–295).

Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 1–12).

Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064). ACM.

Coley, H. L., Sadasivam, R. S., Williams, J. H., Volkman, J. E., Schoenberger, Y.-M., Kohler, C. L., Sobko, H., Ray, M. N. , Allison, J. J., Ford, D. E., Gilbert, G. H., & Houston, T. K. (2013). Crowdsourced peer- versus expert-written smoking-cessation messages. *American Journal of Preventive Medicine, 45*(5), 543–550. http://www.ncbi.nlm.nih.gov/pubmed/24139766.

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (pp. 19–24).

de Souto, M. C. P., de Araujo, D. S., Costa, I. G., Soares, R. G., Ludermir, T. B., & Schliep, A. (2008). Comparative study on normalization procedures for cluster analysis of gene expression datasets. In *IEEE International Joint Conference on Neural Networks. IJCNN 2008 (IEEE World Congress on Computational Intelligence)* (pp. 2792–2798). IEEE.

de Vries, R. (2018). *Theory-based and tailor-made: Motivational messages for behavior change technology*. PhD dissertation, Human Media Interaction, Netherlands.

de Vries, R. A. J., Truong, K. P., & Evers, V. (2016a). Crowd-designed motivation: Combining personality and the transtheoretical model. In *Persuasive technology* (pp. 41–52). Berlin: Springer.

de Vries, R. A. J., Truong, K. P., Kwint, S., Drossaert, C. H. C., & Evers, V. (2016b). Crowd-designed motivation: Motivational messages for exercise adherence based on behavior change theory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 297–308). ACM.

de Vries, R. A., Zaga, C., Bayer, F., Drossaert, C. H., Truong, K. P., & Evers, V. (2017a). Experts get me started, peers keep me going: Comparing crowd-versus expert-designed motivational text messages for exercise behavior change. In *11th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth*. ACM.

de Vries, R. A., Truong, K. P., Zaga, C., Li, J., & Evers, V. (2017b). A word of advice: How to tailor motivational text messages based on behavior change theory to personality and gender. *Personal and Ubiquitous Computing*, *21*(4), 675–687.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia* (pp. 1459–1462). ACM.

Godin, G., & Shephard, R. (1997). Godin leisure-time exercise questionnaire. *Medicine and Science in Sports and Exercise*, *29*(6s), S36.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, *16*(3), 261–291.

Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, *8*(12), 1642–1653.

Hekler, E. B., Klasnja, P., Froehlich, J. E., & Buman, M. P. (2013). Mind the theoretical gap: Interpreting, using, and developing behavioral theory in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3307–3316).

Hirsch, H. G. (1993). *Estimation of noise spectrum and its application to SNR-estimation and speech enhancement*. International Computer Science Institute.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, ser. HLT '09* (pp. 27–35).

Huang, R., & Ma, C. (2006). Toward a speaker-independent real-time affect detection system. In *Proceedings of the International Conference on Pattern Recognition (ICPR), 1*, 1204–1207.

Klasnja, P., & Pratt, W. (2012). Healthcare in the pocket: Mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*, *45*(1), 184–198.

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. In *Proceedings of Interspeech* (pp. 125–128).

Lane, I., Waibel, A., Eck, M., & Rottmann, K. (2010). Tools for collecting speech corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 184–187).

Latimer, A. E., Brawley, L. R., & Bassett, R. L. (2010). A systematic review of three approaches for constructing physical activity messages: What messages work and what improvements are needed? *The International Journal of Behavioral Nutrition and Physical Activity*, *7*, 36.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5270–5273).

McGraw, I. (2013). Collecting speech from crowds. In *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment* (pp. 37–71).

McGraw, I., Glass, J., & Seneff, S. (2011). Growing a spoken language interface on Amazon Mechanical Turk. In *Proceedings of Interspeech* (pp. 3057–3060).

McGraw, I., Lee, C., Hetherington, L., Seneff, S., & Glass, J. (2010). Collecting voices from the cloud. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1576–1583).

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

Nigg, C., Rossi, J., Norman, G., & Benisovich, S. (1998). Structure of decisional balance for exercise adoption. *Annals of Behavioral Medicine*, *20*, S211.

Nigg, C., Norman, G., Rossi, J., & Benisovich, S. (1999). Processes of exercise behavior change: Redeveloping the scale. *Annals of Behavioral Medicine*, *21*, S79.

Norman, G., Benisovich, S., Nigg, C., & Rossi, J. (1998). Examining three exercise staging algorithms in two samples. In *19th Annual Meeting of the Society of Behavioral Medicine*.

Novotney, S., & Callison-Burch, C. (2010). Crowdsourced accessibility: Elicitation of Wikipedia articles. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 41–44).

Parent, G., & Eskenazi, M. (2011). Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Proceedings of Interspeech* (pp. 3037–3040).

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology*, *51*(3), 390.

Schmitz, H., & Lykourentzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, *1*(1), 1.

Skutella, L. V., Sssenbach, L., Pitsch, K., & Wagner, P. (2014). The prosody of motivation: First results. In *Proceedings of ESSV (Elektronische Sprachsignalverarbeitung)*.

Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ser. HLT '11, 1*, 1220–1229.