CANARY SPEECH

PEER-REVIEWED PUBLICATION

Toward Estimating Personal Well-Being Using Voice

TOWARD ESTIMATING PERSONAL WELL-BEING USING VOICE

Samuel Kim, Namhee Kwon, Henry O'Connell

Canary Speech, LLC, Provo, Utah, U.S.A.
{sam, namhee, henry}@canaryspeech.com

ABSTRACT

Estimating personal well-being draws increasing attention particularly from healthcare and pharmaceutical industries. We propose an approach to estimate personal well-being in terms of various measurements such as anxiety, sleep quality and mood using voice. With clinically validated questionnaires to score those measurements in a self-assessed way, we extract salient features from voice and train regression models with deep neural networks. Experiments with the collected database of 219 subjects show promising results in predicting the well-being related measurements; concordance correlation coefficients (CCC) between self-assessed scores and predicted scores are 0.41 for anxiety, 0.44 for sleep quality and 0.38 for mood.

Index Terms— Affective computing, well-being, axiety, sleep quality, mood

1. INTRODUCTION

There is a growing body of interest to use voice as one of biomarkers to estimate one's health conditions. Some health conditions like Parkinson's disease [1, 2] and amyotrophic lateral sclerosis (ALS) [3] are directly related to speech production mechanism and others like schizophrenia [4, 5], depression [6, 7], and bipolar disorder [8, 9] are indirectly related through neurological processes that can modulate voice.

This work focuses on personal well-being (or, quality of life) in general population with relatively healthy conditions. Previous related researches include speech emotion recognition (SER) [10, 11] targeting to classify emotional status with a given voice segment. This work expands the scope to the personal well-being; in particular, we study various measurements such as anxiety, sleep quality and mood.

Like other related applications mentioned above, we also assume that persons' condition is somehow embedded in their voice by modulating articulatory organs either voluntarily or involuntarily. Based on this assumption, we propose an approach to extract salient features from voice with respect to the target measurements and train them with machine learning algorithms to automatically estimate using voice.

In this work, we use subjects' self-assessed measurements as ground truth of their well-being status. Dealing with subjective matters like well-being, however, is difficult partially because it is related to personal perspectives, feelings, and opinions which may vary across the subjects [12]. Therefore, we use a questionnaire based approach [12, 13] to consolidate responses from multiple questions that reflect various aspects of target measurements.

Details about the questionnaires will be discussed in the next session where we describe the data collection campaign in Section 2 followed by the proposed method in Section 3 and the experimental results in Section 4.

2. DATA COLLECTION

2.1. Procedures

We recruited 219 participants from Canada for data collection. Besides 13 participants that are not specified their demographic information, there are 114 males and 92 female and their age range is from 25 to 55 (33.6 ± 8.1) years. Their native language is English including 12 French bilingual subjects. They were also asked to report their medical conditions. Being allowed multiple choices, no previous medical condition was reported for 159 participants while some reported that they have been treated due to some medical conditions notably 22 with depression and 13 with migraine or headache.

The data collection was designed to conduct five sessions per participants with several days of intervals in between. During the data collection campaign, 202 participants completed all five sessions and the intervals between sessions per participants are from 1.2 to 15.3 (3.7 ± 1.9) days. At each session, the participants were asked to answer a set of questionnaires using mobile devices; four written surveys and a questionnaire that requires voice responses.

After removing sessions that are incomplete or have missing fields, we have total 1,048 sessions collected (approx. 4.8 sessions per subjects).

2.2. Questionnaires

2.2.1. Written surveys

As discussed earlier, we use a questionnaire based approach to collect self-assessed measurements to represent subjects' well-being status. In this regard, we adopt clinically validated questionnaires that are designed to measure anxiety, sleep quality and mood as follows.

- Anxiety: State-trait anxiety inventory (STAI) [14] and Generalized anxiety disorder (GAD7) [15] are used to measure the level of anxiety. STAI has six emotional status to score themselves based on how they feel at the moment, while GAD7 has seven emotional status to score how often they felt over the past two weeks. Both have consolidating rules to generate the level of anxiety based on the answers; higher value means higher anxiety.
- Sleep quality: We use Pittsburgh sleep quality index (PSQI) [16] which was design to measure sleep quality. The questionnaire comprises various aspects of sleep quality such as length of sleep as well as disturbing factors and their frequencies. The scoring guideline is to generate the level of sleep discomfort; higher value indicates worse sleep quality.
- Mood: We use positive affect and negative affect schedule (PANAS) [17]. There are ten different emotional status (five positive and five negative) to answer what extent the subject had felt over the past week. It gives an aggregated score to represent status of the subject; higher value indicates more negative mood and lower value indicates more positive mood.

Table 1 summaries the statistics of collected scores of the measurements along with possible score ranges. Note that the collected data covers possible range of individual measurements except PSQI covers only lower range.

2.2.2. *Voice responses*

The questionnaire that requires voice responses is designed to capture vocal behaviors and to use them in estimating the above described measurements. We asked seven different questions to elicit three types of voice responses; one spontaneous speech, four sentence readings and two paragraph readings.

For spontaneous speech, participants were given an instruction to speak freely on whatever topic they want for about a minute. For sentence or paragraph readings, phonetically balanced reading materials are prompted to the subjects so that they can read aloud (8.5 words for sentence readings and approx. 130 words for paragraph readings on average).

Table 2 shows the statistics of collected voice responses in terms of recorded length. Overall, we collected approximately 54 hours of voice data.

Table 1: Statistics of self-assessed measurements.

	Possible range	Collected Data	
		Range	Mean ± STD
STAI	20 - 80	20 - 80	38.4 ± 13.2
GAD7	0 - 21	0 - 21	6.1 ± 4.9
PSQI	0 - 21	0 - 15	5.3 ± 2.8
PANAS	10 - 50	10 - 50	24.4 ± 6.7

Table 2: Length of voice responses.

Voice responses		Length (in secs.)
Spontaneous	Q1	64.0 ± 10.3
Sentence reading	Q2	5.4 ± 1.5
	Q3	5.1 ± 1.6
	Q4	4.8 ± 1.6
	Q5	5.1 ± 1.6
Paragraph reading	Q6	48.6 ± 8.6
	Q7	48.6 ± 12.4

3. METHODOLOGY

3.1. Feature extraction and selection

The types of features are in two categories, i.e. acoustic and linguistic features. Acoustic features are to capture signal-level modulations due to speakers' status, while linguistic features are to capture language-level patterns that may be influenced by the condition.

Acoustic features are calculated on a per-frame basis. Frames are defined as 25 ms sliding windows that are created every 10 ms. A 41-dimensional supervector of various features such as mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), prosody and voice quality related features is generated every frame, and its delta and delta-delta are concatenated to capture frame-level context. To summarize, we use 19 statistical functions such as mean, median, skewness, kurtosis, quartile, percentile, slope, etc to generate a response-level feature vector.

Language features are based on the results from automatic speech recognition (ASR). We used Canary's general English model which is trained on publicly available datasets like Tedlium and Librispeech using the time delayed neual network (TDNN) architecture in Kaldi [18]. On top of common features such as part-of-voice ratio, syllable duration, filler ratio and word repetition ratio over the total number of spoken words, we extract a different feature set whether the response is spontaneous or read.

For the spontaneous voice responses where no prompted text was given, the semantic features are extracted including word popularity percentile[1], and word frequency of the

[1] The word popularity was computed from the general English language model with 130K words and the distribution of the values per response was

Table 3: Concordance correlation coefficients (CCC) between self-assessed scores and predicted scores using voice responses. Statistical significance is denoted with stars (* for $p < 10^{-2}$ and ** for $p < 10^{-5}$).

		Individual features							Concatenated features
		Spontaneous	Sentence reading				Paragraph reading		
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	
STAI	eGeMAPS	0.05	0.03	0.07^{*}	0.02	0.04	0.03	0.07	0.04^{**}
	Proposed	0.09^{**}	0.16^{**}	0.19^{**}	0.15^{**}	0.14^{**}	0.14^{**}	0.14^{**}	0.30^{**}
GAD7	eGeMAPS	0.02	0.05^{*}	0.04	0.03	0.01	0.05^{*}	0.04^{*}	0.03^{**}
	Proposed	0.14^{**}	0.19^{**}	0.26^{**}	0.22^{**}	0.17^{**}	0.08^{**}	0.09^{**}	0.41^{**}
PSQI	eGeMAPS	0.03	0.08^{*}	0.04	0.01	0.02	0.10^{**}	0.10^{**}	0.06^{**}
	Proposed	0.14^{**}	0.17^{**}	0.21^{**}	0.20^{**}	0.21^{**}	0.12^{**}	0.09^{**}	0.44^{**}
PANAS	eGeMAPS	-0.01	0.00	0.01	0.01	0.00	0.03	0.05^{*}	0.02^{*}
	Proposed	0.12^{**}	0.18^{**}	0.20^{**}	0.16^{**}	0.17^{**}	0.12^{**}	0.15^{**}	0.38^{**}

depression-related terms[2], and positive and negative sentiment likelihood[3]. For the read responses, on the other hand, the ASR errors for insertion, deletion, and substitution are computed based on the given text.

The feature dimension is 2,357 for a read response and 2,364 for a spontaneous response. For feature selection, we compute the Pearson's correlation coefficient between the extracted features and individual self-assessed measurement scores and selected top *n* correlated features for the model. Note that it is done in response- and measurement- level so that different features can be selected from different responses depending on the target measurement.

As a baseline, we use OpenSmile toolkit [21] with eGeMAPS configuration [22] which is widely used and shows promising results particularly in affective computing related fields such as speech emotion recognition. As it generates a 88 dimensional feature vector per one speech signal, we set a threshold of the proposed feature selection procedure to retain only the same number of features for fair comparisons.

3.2. Modeling

We use fully connected deep neural network (FC-DNN) with 4 hidden layers of 256 neuron units. Each layer has the rectified linear unit (ReLU) as an activation function, and l2 regularizer and 50% of dropout to prevent overfitting. The output layer has only one unit as we target to build a regression model. We used an Adam optimizer with learning rate 0.0001 using mean squared error (MSE) as a loss function. The batch size is 32 and iteration stops after 100 epochs.

4. EXPERIMENTAL RESULTS

4.1. Setup

We perform a 5-fold cross validation; we randomly split the whole dataset into 5 exclusive folds and iteratively use one fold as test data and others as training data. Each fold is subject independent in a sense that different folds do not share data from the same subject. Note that we consider all the sessions independent. A longitudinal study to investigate changes over time is left for future work.

The performance is measured in terms of concordance correlation coefficients (CCC) between self-assessed scores and predicted scores [23].

4.2. Results

Table 3 shows CCC between self-assessed scores and predicted scores using voice responses with respect to the target measurement. The left part of the table shows the results with features from individual voice responses, while the right part of the table shows the results with concatenated features from individual voice responses. Fig. 1 shows the scattered plot of self-assessed measures and estimated values using concatenated features.

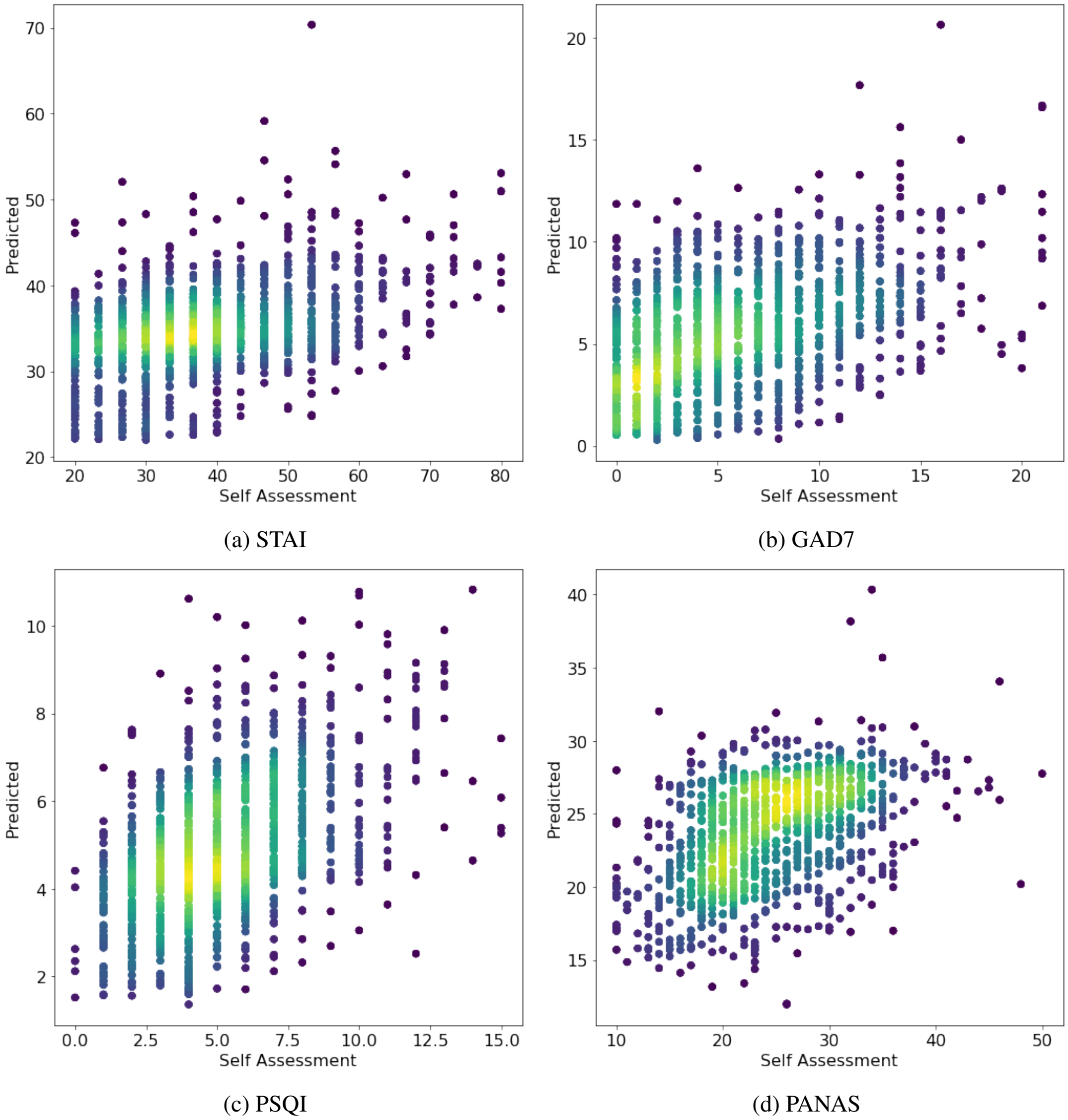
The results with individual voice responses show that the proposed method outperforms the conventional general purpose feature extraction method in estimating all measurements with all voice responses. The same applies to the results with concatenated features. This indicates that the proposed feature selection strategy that considers differences in individual responses is beneficial. The proposed linguistic features are also considered to contribute toward the improvement.

It is notable that the correlation coefficients with features from sentence reading are higher than the ones with features from spontaneous or paragraph reading responses. The results also show that concatenating features from individual voice responses boosts the performance. These suggest that con-

examined and its percentile values were used.

[2] We built a dictionary with depression-related words and negative expressions, as well as common words observed from depression patients' speech, which resulted in 486 terms [19]

[3] The sentiment likelihood score is generated by the binary classification model trained on Stanford Sentiment Treebank using Sentiment Neuron [20].



(a) STAI

(b) GAD7

(c) PSQI

(d) PANAS

Fig. 1: Scatter plots of self-assessed scores and predicted scores using the proposed method. Colors represent the density of the individual data points (yellow for high density and blue for low density).

catenating features from multiple short voice responses rather than one long voice response with a fixed set of features is beneficial to estimate self-assessed measurements. It may be partially because that salient features are averaged out over a long sentence, but in-depth study will be in the future. In the future, we will also study impact of the ASR performance in extracting linguistic features as well as longitudinal study to investigate changes over time.

5. CONCLUSION

We showed promising results in estimating well-being status with voice in terms of various measurements, i.e. anxiety, sleep quality, and mood. The proposed method utilizes acoustic and linguistic features along with response- and measurement- level feature selection strategy followed by a deep neural network based regression model. Experimental results show that sleep quality using PSQI has the strongest correlation while anxiety using STAI has the weakest correlation. Although the concordance correlation coefficients between self-assessed scores and estimated scores may not be too strong, they are statistically significant with $p < 10^{-5}$.

6. ACKNOWLEDGEMENT

The authors like to thank Nathan Fisk, Scott Ferguson and Mark Bartlett for valuable discussions.

7. REFERENCES

[1] Kara M. Smith, James R. Williamson, and Thomas F. Quatieri, "Vocal markers of motor, cognitive, and depressive symptoms in Parkinson's disease," *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, vol. 2018-Janua, pp. 71–78, 2018.

[2] Yermiyahu Hauptman, Ruth Aloni-Lavi, Itshak Lapidot, Tanya Gurevich, Yael Manor, Stav Naor, Noa Diamant, and Irit Opher, "Identifying Distinctive Acoustic and Spectral Features in Parkinsons Disease," in *Proc. Interspeech 2019*, 2019, pp. 2498–2502.

[3] Hannah P. Rowe and Jordan R. Green, "Profiling Speech Motor Impairments in Persons with Amyotrophic Lateral Sclerosis: An Acoustic-Based Approach," in *Proc. Interspeech 2019*, 2019, pp. 4509–4513.

[4] Francisco Martínez-Sánchez, José Antonio Muela-Martínez, Pedro Cortés-Soto, Juan J.osé García Meilán, Juan A.ntonio Vera Ferrándiz, Amaro Egea Caparrós, and Isabel M.aría Pujante Valverde, "Can the Acoustic Analysis of Expressive Prosody Discriminate Schizophrenia?," *The Spanish journal of psychology*, vol. 18, no. March 2016, pp. E86, 2015.

[5] Rohit Voleti, Stephanie Woolridge, Julie M. Liss, Melissa Milanovic, Christopher R. Bowie, and Visar Berisha, "Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder," in *Proc. Interspeech 2019*, 2019, pp. 1433–1437.

[6] Denis Dresvyanskiy, Danila Mamontov, Maxim Markitantov, and Albert Ali Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *AVEC 2019*, 2019.

[7] Carol Espy-Wilson, Adam C. Lammert, Nadee Seneviratne, and Thomas F. Quatieri, "Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables," in *Proc. Interspeech 2019*, 2019, pp. 1448–1452.

[8] Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin G. McInnis, and Emily Mower Provost, "Identifying Mood Episodes Using Dialogue Features from Clinical Interviews," in *Proc. Interspeech 2019*, 2019, pp. 1926–1930.

[9] Katie Matton, Melvin G. McInnis, and Emily Mower Provost, "Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder," in *Proc. Interspeech 2019*, 2019, pp. 1438–1442.

[10] Carlos Busso, Mutaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, 2008.

[11] B. Schuller, E. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, 2013.

[12] Francis Guillemin, Claire Bombardier, and Dorcas Beaton, "Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines," *Journal of Clinical Epidemiology*, vol. 46, no. 12, pp. 1417–1432, Dec. 1993.

[13] Theodore Pincus, Christopher Swearingen, and Frederick Wolfe, "Toward a multidimensional health assessment questionnaire (MDHAQ): Assessment of advanced activities of daily living and psychological status in the patientfriendly health assessment questionnaire format," *Arthritis & Rheumatism*, vol. 42, pp. 2220–2230, 2001.

[14] Audrey Tluczek, Jeffrey B. Henriques, and Roger L. Brown, "Support for the reliability and validity of a six-item state anxiety scale derived from the state-trait anxiety inventory," *Journal of Nursing Measurement*, vol. 17, no. 1, pp. 19–28, 2009.

[15] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Lwe, "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *JAMA Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, 05 2006.

[16] Daniel J. Buysse, Charles F. Reynolds, Timothy H. Monk, Susan R. Berman, and David J. Kupfer, "The pittsburgh sleep quality index: A new instrument for psychiatric practice and research," *Psychiatry Research*, vol. 28, no. 2, pp. 193 – 213, 1989.

[17] David Watson, Lee Anna Clark, and Auke Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales.," 1988.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.

[19] Mohammed Al-Mosaiwi and Tom Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018, PMID: 30886766.

[20] Alec Radford, Rafal Józefowicz, and Ilya Sutskever, "Learning to generate reviews and discovering sentiment," *CoRR*, vol. abs/1704.01444, 2017.

[21] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM '13, pp. 835–838, ACM.

[22] Florian Eyben, Klaus Scherer, Bjrn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.

[23] Lawrence I-Kuei Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.