"How are you?" Estimation of anxiety, sleep quality, and mood using computational voice analysis

Samuel Kim[1], Namhee Kwon[1], Henry O'Connell[1], Nathan Fisk[2], Scott Ferguson[2] and Mark Bartlett[2]

***Abstract*— We developed a method of estimating impactors of cognitive function (ICF) - such as anxiety, sleep quality, and mood - using computational voice analysis. Clinically validated questionnaires (VQs) were used to score anxiety, sleep and mood while salient voice features were extracted to train regression models with deep neural networks. Experiments with 203 subjects showed promising results with significant concordance correlation coefficients (CCC) between actual VQ scores and the predicted scores (0.46 = anxiety, 0.50 = sleep quality, 0.45 = mood).**

I. Introduction

There is a growing body of interest in using the voice as a biomarker to estimate human health conditions. Some conditions like Parkinson's disease [1], [2] and amyotrophic lateral sclerosis (ALS) [3] are directly related to speech production mechanisms. Others like schizophrenia [4], [5], [6], depression [7], [8], and bipolar disorder [9], [10] are indirectly related through neurological processes that can modulate voice.

This study focuses on known ICFs - specifically anxiety [11], sleep quality [12], [13] and mood [14], [15] - in a generally healthy population rather than a medically challenged one. Like other related applications mentioned above, we also assume that a subject's condition is embedded in the voice by modulating articulatory organs either voluntarily or involuntarily. Based on this assumption, we extract salient features from voice with respect to the target measurements and train them with machine learning algorithms to automatically estimate the measurements of interest.

In this work, we use VQs as ground truth of anxiety, sleep quality, and mood status. Dealing with subjective matters like these, however, is difficult partially because it is related to personal perspectives, feelings, and opinions which may vary across the subjects [16]. Therefore, we use a questionnaire based approach [16], [17] to consolidate responses from multiple instruments to address the target measurements.

The VQs and voice collection methodology will be discussed in Section II followed by examining the proposed voice analysis method in Section III and finally the experimental results in Section IV.

II. Data

A. Data collection procedures

219 subjects were recruited for data collection. The native language of the subject population was English, including 12 French bilingual subjects. Along with other demographic information, subjects were asked to report their medical conditions. No previous medical conditions were reported for 155 subjects while 48 reported having at least one medical condition; notably 22 with depression and 13 with migraine or headache.

The data collection consisted of one training session (not analyzed) followed by 4 evaluation sessions per subject. For each session, the subjects were asked to answer a set of questionnaires using mobile devices; four VQs and a sequence of recorded voice responses. 203 subjects completed at least three of the four evaluation sessions. See Table I for demographic distribution of the participating subjects. We conducted Wilcoxon significance tests to see if there was any potential impact of the demographic characteristics and found none had a significant influence on the primary research objective.

TABLE I: Distribution of demographic groups of the collected database.

Demographics	Group	Num of subjects	%
Age group	25-34	138	67.9
	35-44	35	17.2
	45-54	30	14.7
Gender	Male	88	43.3
	Female	112	55.2
	Refused	3	1.4
Ethnicity group	Caucasian	120	59.1
	Asian	46	22.7
	Other	18	8.8
	African American	10	4.9
	Hispanic	5	2.5
	Middle East	1	0.5
	Refused	3	1.4
Education level	Bachelors	92	45.3
	CEGEP/Professional	14	6.9
	High School	21	10.3
	Post Graduate	52	25.6
	Vocational	24	11.8
Medical condition	Stated Condition	48	23.6
	No Condition	155	76.4

TABLE II: Statistics of VQ measurements.

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

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

[2]Pharmanex Research, NSE Products Inc., Provo, UT, U.S.A. {nafisk, scottf, mrbartle}@nuskin.com

B. Questionnaires

1) Clinically Validated Questionnaires: As discussed earlier, we used VQs to collect measurements of the subjects' well-being. The VQs are designed to measure anxiety, sleep quality and mood as follows...

- Anxiety: State-trait anxiety inventory (STAI) [18] and Generalized anxiety disorder (GAD7) [19] were used to measure levels of anxiety. STAI-6 has six emotional variables for users to score themselves based on how they feel at the moment, while GAD7 has seven emotional variables to score how they felt over the past two weeks. Both instruments have consolidating rules to generate a final level of anxiety based on the answers; higher values mean higher anxiety.
- Sleep quality: The Pittsburgh sleep quality index (PSQI) [20] was used as a measure of subject sleep quality. The questionnaire evaluates various aspects of sleep quality such as length of sleep as well as disturbing factors and their frequencies. The scoring guideline generates a final level of sleep discomfort; higher values indicate worse sleep quality.
- Mood: The positive affect and negative affect schedule (PANAS) [21] was used to evaluate the subject's mood. The applied PANAS contains ten different emotional variables (five positive and five negative) to evaluate the user's mood over the past week. It gives an aggregated score to represent the overall mood of the subject; a higher value is indicative of a negative mood and a lower value indicates a positive mood.

Table II summarizes the statistics of the collected scores along with possible score ranges. Note that the collected data covers the possible range of individual measurements except PSQI, which covers only the lower range.

2) Voice responses: Subject speech responses were designed to capture vocal behaviors and to use them in estimating the actual VQ scores. 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 instructions to speak freely on whatever topic they wanted for about a minute. For sentence or paragraph readings, test subjects were prompted to read aloud phonetically balanced reading materials.

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

TABLE III: 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

III. Methodology

A. Feature extraction

Types of voice features came in two categories: acoustic and linguistic features. Acoustic features capture signal-level modulations due to the speaker's status, while linguistic features capture language-level patterns which may be influenced by the condition.

Acoustic features were calculated on a per-frame basis. Frames were defined as 25 ms sliding windows that were 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 features were generated every frame [22]. MFCC and PLP provided spectral characteristics of speech signals considering human auditory characteristics. Prosody and voice quality features are related to voice tonality and its characteristics, i.e. fundamental frequency (a.k.a. pitch), shimmer/jitter and harmonic-to-noise ratio, etc. Each feature's delta and delta-delta were concatenated to capture frame-level context. To summarize the features in a response-level, we used 19 statistical functions such as mean, median, skewness, kurtosis, quartile, percentile, and slope.

Language features were 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 neural network (TDNN) architecture in Kaldi [23]. On top of common features such as part-of-voice ratio, syllable duration, filler (ah, hmm, eh, uh, etc.) ratio and word repetition ratio over the total number of spoken words, we extracted a different feature set whether the response was spontaneous or read.

For the spontaneous voice responses where no prompted text was given, the semantic features were extracted including word popularity percentiles[1], and word frequency of the 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 were computed based on the given text.

The feature dimension was 2,357 for a read response and 2,364 for a spontaneous response. For feature selection, we computed the Pearson's correlation coefficient between the extracted features and the individual VSI scores and then selected top n correlated features for the model. Note that this was done for response and measurement levels so that different features could be selected from different responses depending on the target measurement.

As a baseline, we used the OpenSmile toolkit [26] with eGeMAPS configuration [27] which is widely accepted and

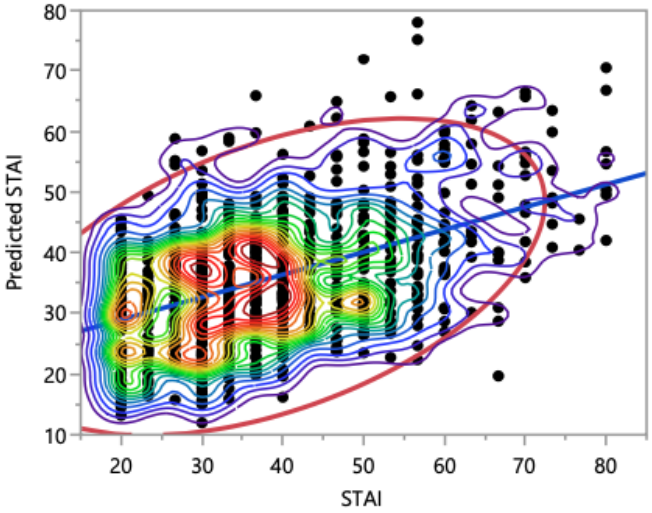
[1]The word popularity was computed from the general English language model with 130K words and the distribution of the values per response was examined and its percentile values were used.

[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. [24]

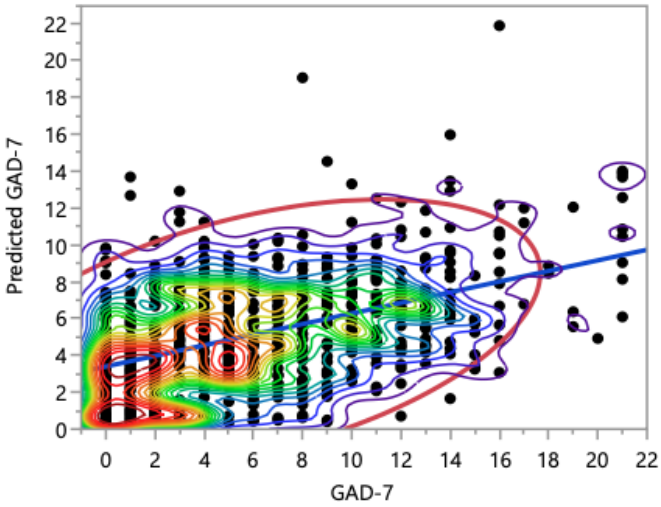
[3]The sentiment likelihood score was generated by the binary classification model trained on Stanford Sentiment Treebank using Sentiment Neuron [25].

TABLE IV: Comparison of the proposed and conventional methods in terms of concordance correlation coefficients (CCC) between VQ scores and predicted scores using voice responses. Statistical significance is denoted with stars (* for $p < 10^{-2}$ and ** for $p < 10^{-5}$).

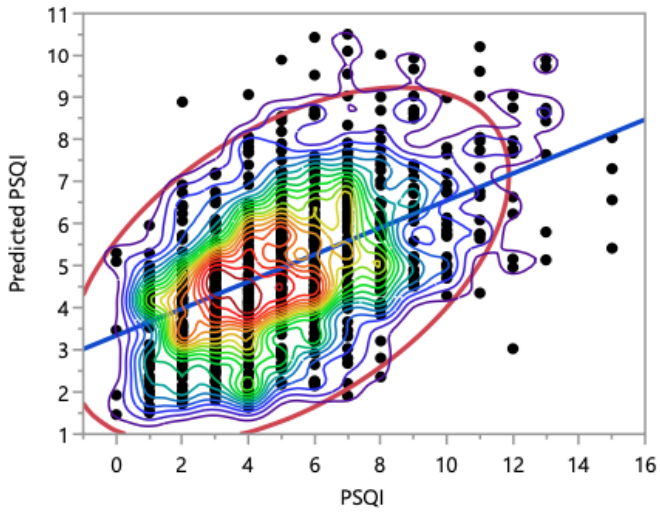
		Individual features							Concatenated features
		Spontaneous	Sentence reading				Paragraph reading		
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	
STAI	eGeMAPS	0.05	0.03	0.07*	0.02	0.04	0.03	0.07	0.04**
	Proposed	0.09**	0.16**	0.19**	0.15**	0.14**	0.14**	0.14**	0.30**
GAD7	eGeMAPS	0.02	0.05*	0.04	0.03	0.01	0.05*	0.04*	0.03**
	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**



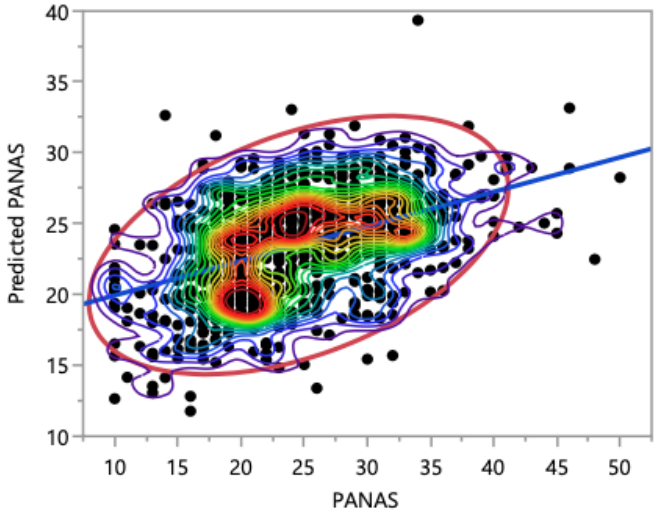
(a) Bivariate Predicted STAI and Actual STAI



(b) Bivariate Predicted GAD7 and Actual GAD7



(c) Bivariate Predicted PSQI and Actual PSQI



(d) Bivariate Predicted PANAS and Actual PANAS

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

shows promising results particularly in affective computing related fields such as speech emotion recognition. As it generated an 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 when we compared the proposed method against the conventional feature extraction.

B. Modeling

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

IV. Experimental Results

A. Setup

We performed a 5-fold cross validation by randomly splitting the whole dataset into 5 exclusive folds, iteratively using one fold as test data and the remaining folds as training data. Each fold was subject independent so that different folds did not share data from the same subjects. Note that we considered all the sessions independently. A longitudinal study to investigate changes over time was left for future work.

The performance was measured in terms of concordance correlation coefficients (CCC) between VQ scores and predicted scores [28].

B. Results

Table IV shows CCC between VQ 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.

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 which 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 boost the performance. These suggest that concatenating features from multiple short voice responses rather than one long voice response with a fixed set of features is beneficial to estimating self-assessed measurements. This may be partially because the salient features are averaged out over a long sentence, but further research is needed for clarification. Additional study will also examine the impact of the ASR performance in extracting linguistic features, and longitudinal research is needed to investigate changes over time.

TABLE V: Comparison of the numbers of selected features in terms of CCC between VQ scores and predicted scores using voice responses.

	STAI	GAD7	PSQI	PANAS
88 (Table IV)	0.30	0.41	0.44	0.38
100	0.46	0.46	0.50	0.45

Table V shows the performance with different sizes of selected features. Although we empirically choose this hyperparameter and it needs to be fully investigated in the future, it indicates that there is a room for improvement by how we select features as an input to machine learning algorithms. Fig. 1 depicts the scattered plots of VQ scores and estimated scores using the feature set. The red oval and blue line represent a bivariate normal ellipse that encircles 95% and the linear fit of the data points respectively. The colored zones are percentile density contours that move from red to blue as the percentage of the data in the quantiles increases. The figures illustrate, as shown in correlation coefficients, the VQ scores and estimated scores are significantly correlated across all instruments.

V. Conclusion

We show promising results using computational voice analysis in predicting three ICFs (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 regression model. Experimental results show that sleep quality using PSQI has the strongest correlation while anxiety using STAI has the weakest correlation. All assessments have statistically significant correlations with $p < 10^{-5}$. Voice as a non-intrusive measurement of ICFs is a promising research pathway that will benefit from the collection of more data to detect hidden patterns and subtle differences among larger populations.

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 Parkinson's 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] Armen C. Arevian, Daniel Bone, Nikolaos Malandrakis, Victor R. Martinez, Kenneth B. Wells, David J. Miklowitz, and Shrikanth Narayanan, "Clinical state tracking in serious mental illness through computational analysis of speech," *PloS one*, vol. 15, no. 1, pp. e0225695, 2020.

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

[8] 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.

[9] 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.

[10] 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.

[11] Bruce S. McEwen, "The neurobiology of stress: from serendipity to clinical relevance 1," *Brain Research*, vol. 886, no. 1-2, pp. 172–189, 2000.

[12] Andreana Benitez and John Gunstad, "Poor sleep quality diminishes cognitive functioning independent of depression and anxiety in healthy young adults," *The Clinical neuropsychologist*, vol. 26, pp. 214–23, 02 2012.

[13] Steven Gilbert and Cameron Weaver, "Sleep quality and academic performance in university students: A wake-up call for college psychologists," *Journal of College Student Psychotherapy*, vol. 24, pp. 295–306, 10 2010.

[14] Karuna Subramaniam, John Kounios, Todd Parrish, and Mark Beeman, "A brain mechanism for facilitation of insight by positive affect," *Journal of cognitive neuroscience*, vol. 21, pp. 415–32, 07 2008.

[15] Tabitha Payne and Michael Schnapp, "The relationship between negative affect and reported cognitive failures," *Depression research and treatment*, vol. 2014, pp. 396195, 02 2014.

[16] 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.

[24] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] Thomas Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall Press, USA, first edition, 2001.

[23] 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.

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

[26] 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.

[27] Florian Eyben, Klaus Scherer, Björn 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.

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