The Munich Biovoice Corpus: Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production

Björn Schuller[1,2], Felix Friedmann[2], Florian Eyben[2]
[1]Department of Computing, Imperial College London, United Kingdom
[2]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
E-mail: bjoern.schuller@imperial.ac.uk

Abstract

We introduce a spoken language resource for the analysis of impact that physical exercising has on human speech production. In particular, the database provides heart rate and skin conductance measurement information alongside the audio recordings. It contains recordings from 19 subjects in a relaxed state and after exercising. The audio material includes breathing, sustained vowels, and read text. Further, we describe pre-extracted audio-features from our openSMILE feature extractor together with baseline performances for the recognition of high and low heart rate using these features. The baseline results clearly show the feasibility of automatic estimation of heart rate from the human voice, in particular from sustained vowels. Both regression - in order to predict the exact heart rate value - and a binary classification setting for high and low heart rate classes are investigated. Finally, we give tendencies on feature group relevance in the named contexts of heart rate estimation and skin conductivity estimation.

Keywords: Speech Database, Heart Rate, Skin Conductance

1. Introduction

Audio-based measurement of heart rate or skin-conductance has been addressed rather sparsely in the literature so far. This is probably due to the fact that at first it might appear that heart-rate or skin-conductance estimation from speech is unnecessary and complex, as there exist simple, reliable sensors, such as small finger clips, for measuring both.

However, in times where telepresence and telecommunication gain importance, measuring physiological parameters via physically attached sensors constitutes an additional overhead. Speech is inherently present in such systems, and if physiology parameters can be reliably estimated from a normal speech signal, the door to many, novel and innovative applications is opened.

Examples include monitoring of physiological parameters in the health and safety sector, e.g., in emergency calls, or stress level analysis from phone conversations, as well as lie detection.

Orlikoff & Baken (1989) investigated the connection between human voice and heartbeat. In their study, six male and six female participants were measured with an electroglottograph (EGG) during speech production. By signal-averaging and autocorrelation, they observed that the heartbeat leads to around 0.2% up to 19% of absolute perturbation of the fundamental frequency (jitter) measured on pronunciations of sustained vowels.

In a former study, we evaluated heart rate (HR) and skin conductance (SC) prediction as well as a simpler high pulse / low pulse (HP/LP) classification based on acoustic features (Schuller, Friedmann & Eyben 2013). The results base on recordings of breathing, pronunciation of sustained vowels and text reading before and after physical exercising and are encouraging in the sense that general feasibility was shown. To our knowledge, a similar study has so far only been attempted for HR in vowels by Skopin & Baglikov (2009) and Mesleh & al. (2012). For automatic voice-based skin conductance assessment no other study is known to us.

In this contribution we introduce the database that served for analyses in (Schuller, Friedmann & Eyben 2013) in detail and thus make it accessible for future studies to the community. We present recordings and analyses in detail (Section 2) alongside benchmarks for recognition with different machine learning algorithms (Section 4). Further, we describe a set of standard audio features extracted by our openSMILE toolkit (Eyben & al., 2013) and show initial results of feature relevance analysis in Section 3.

2. The Munich Biovoice Corpus

For the creation of the Munich Biovoice Corpus (MBC for short) speech from participants was recorded alongside with the physiological parameters heart rate and skin-conductance in a synchronised way. Participants were recorded in a "neutral", or low load state and a in a high load state after they had performed (physical) exercise.

Wild Divine Inc.'s "iom" device was used to record HR and SC data from three sensors attached to a subject's fingers. A Zoom Q3Hd camcorder equipped with an X-Y hd microphone was used to record audio ("room microphone") at a sampling rate of 92 kHz in uncompressed PCM-wave format. In addition, a Logitech Clearchat Headset to capture close-talk speech recordings was used. All devices were connected to the same recording computer to ensure synchronisation.

19 subjects (4/15 female/male, 3 Chinese, 15 German, 1 Italian) participated in the experiment and gave their consent to data recording and storage. All were free of temporary diseases, but the subjects include smokers and such with mild cardiac and neurological disorders. All completed a questionnaire about their height, weight, nationality and general health condition as well as the BFI-10 short personality test by Rammstedt & John (2007).

All subjects were recorded while breathing, while pronouncing the sustained vowel /a/ repeatedly and while reading a standard text which is used frequently in phonetics. These recordings were performed in the two

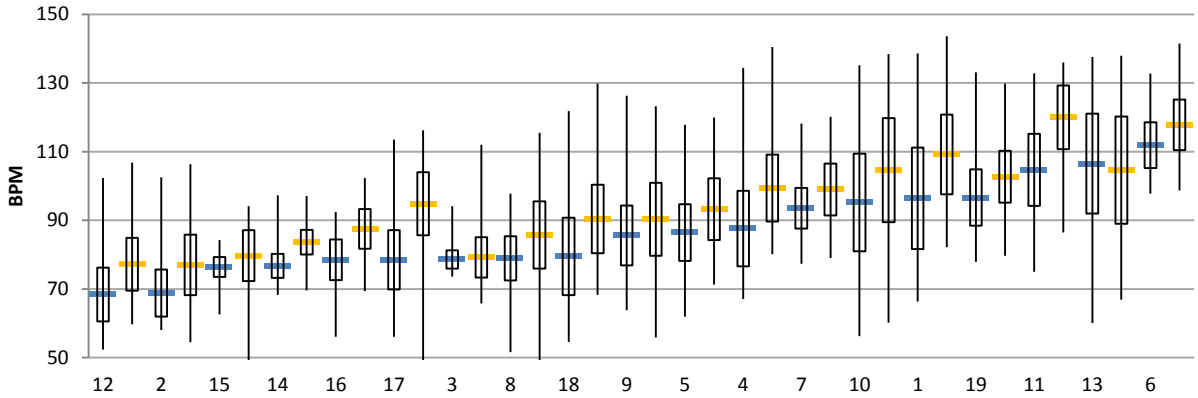

Figure 1: Range of recorded heart rate values in BPM for each subject: Subject ID (on x-axis, sorted by BPM in ascending order from left to right) and sound type: blue/left: Sustained Vowel, orange/right: Breathing.

above named physical load states with rather low pulse and rather high pulse. The states were clearly defined by the activities the participants had to perform.

Before starting the experiments, the subjects had to sit down in front of the recording equipment and they were instructed on the procedures. They then had to perform a practice recordings session to assure they are well familiar with the procedure before the actual recording takes start. The time the instructions and the practice session took should also ensure that all participants were in a low physical load state, i.e. had a resting pulse, regardless of their load before entering the experiment.

Then, the actual recording session was started. After this first session with low pulse/load, the subjects raised their physical load by exercising, i.e., running quickly up and down stairs over three stories and running along a long hallway which led to the recording room. The subjects were required to physically exercise until their pulse exceeded 90 BPM. Right after the exercise unit, the second recording session was performed.

In both recording sessions, subjects were recorded sitting on a chair in front of a desk. A Zoom Q3Hd was placed 50 centimetres from the subjects' lips; the Logitech Clearchat headset was head-worn by the subjects. The iom's heart rate sensor was connected to the left middle finger and the two skin conductivity sensors were connected to the left ring and forefinger. The subject's left hand was electrically grounded in order to minimize the influence of electromagnetic und electrostatic noise in the sensors. The iom connected to a subject's hand is shown in Figure 2.

For each subject, a comfortable frequency F_c for the pronunciation of a sustained /a/ vowel was determined during the practice phase with visual feedback from a live frequency analysis. This frequency F_c was marked on the screen, and the subject had to train repeating the /a/ vowel in the 'comfortable' frequency (/a/$_c$) as precise as possible for several times. After the subject was able to intentionally produce /a/$_c$ within a tolerance of ± 7 Hz during 5 subsequent attempts, it was assumed that /a/$_c$ could be produced reliably during the recording. The subject had to undergo the same training procedure for /a/$_l$, an /a/ vowel with a frequency F_l which is four semi tone levels below the frequency F_c.

In each of the two recording sessions they had to pronounce /a/$_l$ and /a/$_c$ four times each and then read a continuous text aloud. Native German speakers read out the text "Der Nordwind und die Sonne"– non-native subjects read the English translation of the text "The Northwind and the Sun".

Figure 1 shows details per subject on the collected heart rate (pulse) range during all recordings. The range reaches from 22 BPM up to 79 BPM deltas between minimum and maximum pulse.

Overall, the final database consists of heart rate and skin conductivity labelled audio recordings from 19 speakers. The instances are divided into 74 text periods, 644 breath periods and 630 sustained vowel expressions. They are further divided into low pulse and high pulse recordings and into headset (close-talk) and Q3Hd microphone recordings. Sustained vowels are labelled with F_0 data and divided into sustained vowels at comfortable fundamental frequency (F_c) and sustained vowels in low fundamental frequency (F_l).

In the following, we will shift to calculating baselines for feature relevance and obtainable automatic classification and regression performances following the flow-path shown in Figure 3. and relying on broadly used open-source tools available for reproduction.

Figure 2: A subject's hand grounded and connected to the iom sensors.

3. Features

openSMILE's feature set as was designed for the INTERSPEECH 2011 Speaker State Challenge (Schuller

Low Level Descriptors (LLD)
4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
50 spectral LLD
RASTA-style filt. auditory spectrum, bands 1–26 (0–8 kHz)
MFCC 1–12
Spectral energy 25–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope
5 voice related LLD
*F*0
Probability of voicing
Jitter (local, delta)
Shimmer (local)

Functionals
33 base functionals
quartiles 1–3 and 3 inter-quartile ranges
1% percentile (≈min), 99% percentile (≈max)
percentile range 1 %–99%
arithmetic mean, standard deviation
skewness, kurtosis
mean of peak distances
standard deviation of peak distances, mean value of peaks
mean value of peaks – arithmetic mean
linear regression slope and quadratic error
quadratic regression a and b and quadratic error
contour centroid
duration signal is below 25% range / above 90% range
duration signal is rising/falling
gain of linear prediction (LP)
LP Coefficients 1–5
6 F0 functionals
percentage of non-zero frames
mean, max, min, std. dev. of segment length
input duration in seconds

Table 1: Low Level Descriptors and functionals of the INTERSPEECH 2011 Speaker State Challenge feature set.

& al. 2011) is considered in this study. The set consists of

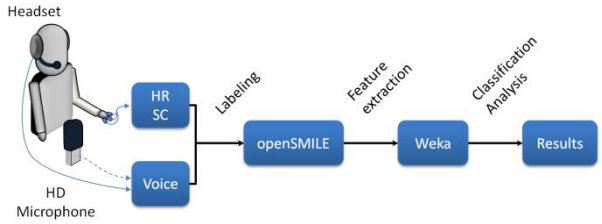

Figure 3: Steps of data processing.

4,368 features built from 4 energy-, 50 spectral- and 5 voice-related Low Level Descriptors (LLDs) to which functionals are applied (see Table 1). On the energy related and spectral LLD and their first order deltas, base functionals are applied together with min, mean, max and the standard derivation of the segment length. On the voice related LLD and their first order deltas, the base functionals are applied together with quadratic mean, rise duration and fall duration of the signal in case of voicing probability greater than 0.70. The F_0 functionals are applied on the F_0 LLD and its first order derivate.

4. Experiments

Baseline classification experiments are performed to evaluate the feasibility of automatic prediction of heart rate and skin conductivity from the voice. Further, experiments are performed to select meaningful features and feature groups that are best suited.

Support Vector Regression (SVR) is used to predict the raw, continuous values for heart rate and skin conductivity. SVR with a linear kernel function is used and sequential minimal optimisation (SMO) is applied as training algorithm as is implemented in the WEKA data-mining toolkit (Hall et. al 2009). Next to regression analysis, a binary classification task is performed for low vs. high pulse (HP/LP) with a linear Support Vector Machine (SVM). Experiments are performed on the entire sustained vowel data recorded from the headset. A randomly selected subset of 2/3 of all data (with WEKA) is used for training, the remaining 1/3 of the data are used for evaluation in the on-going.

4.1. Feature selection

Table 2 shows the results of classification and regression performed to examine how recognition accuracy is influenced by a reduction of features.

	HP/LP	HR	SC
# feat.	UA [%]	CC	CC
5	67.6	0.54	0.15
50	82.1	0.69	0.79
100	86.4	0.84	0.82
150	91.4	0.75	0.88
All	75.7	0.72	0.88

Table 2: Improvement by feature reduction – sustained vowels via headset close-talk as acoustic condition.

The group of employed features was ranked by the absolute value of the corresponding weight in the linear SVR hyperplane normal vector and reduced to the *N* highest weighted features once for HR and once for SC. Using these features, higher accuracies as for classification with respect to regression with all features

could be reached. With the 150 highest-weighted features selected, an unweighted accuracy (UA) of 91.4% could be reached for HP/LP classification and a correlation coefficient (CC) of .876 for SC recognition. For HR a CC of .838 was achieved with the top 100 features. It is to note, however, that these results are speaker dependent, i.e., that data from the same speaker is contained in the training and the test set.

4.2. Feature Group Relevance

Feature groups that have been compared are shown in Figure 4: Two energy-related feature groups based on the zero-crossing rate and the root mean square energy (PCM features), the sum of auditory spectrum band energies resembling loudness (AudSpec features), spectral-related features based on Mel Frequency Ceptral Coefficients (MFCCs), as well as voicing related feature groups based on jitter, shimmer, fundamental frequency (F0 features, cf. also (Johannes & al. 2007), and probability of voicing (Voicing features). For comparison, the single weights of a group's features among the top 150 features were summed and divided by the sum of the top 150 features' weights to provide a measure for relevance assigned to the features by the classifier. In total, voice-quality related features account for 4% to 18% of the top 150 features when employing a random 2/3 to 1/3 split between training and test data, respectively. In the same setting, the MFCC group accounts for 31% to 70%, auditory spectrum features for 14% to 38% and signal-based features for 11% to 23% of the top 150 features.

Recognition method	CC	MAE
ANN	.768	12.3
Simple linear regression	.423	15.2
LMSLR	.776	10.1
Linear regression	.781	10.9
SVR, RBF kernel	.748	10.7
SVR, quadratic	.809	10.0
SVR, linear	.838	9.1

Table 3: Comparison of recognition methods for HR on sustained vowels from the headset close-talk microphone based on the top 100 features. ANN: Artificial Neural Network, LMSLR: Least Median Squared Linear Regression, MAE: Mean Absolute Error, SVR: Support Vector Regression.

4.3. Choice of the Machine Learning Algorithm

Besides the linear kernel SVR and SVM, other machine learning methods have been investigated for the heart-rate task using the full feature set. Table 3 shows how the accuracy depends on the machine learning algorithm in this case. Different regressors were compared to linear SVR, which had achieved particularly good HR recognition rates when the number of features was reduced to the top 150 features as described in Section 3.1.

While a linear kernel was optimal in that case, a Radial Basis Function (RBF) kernel achieved a higher accuracy for the full feature set. Overall, one can see that the best results were obtained with SVR as compared to the considered alternatives.

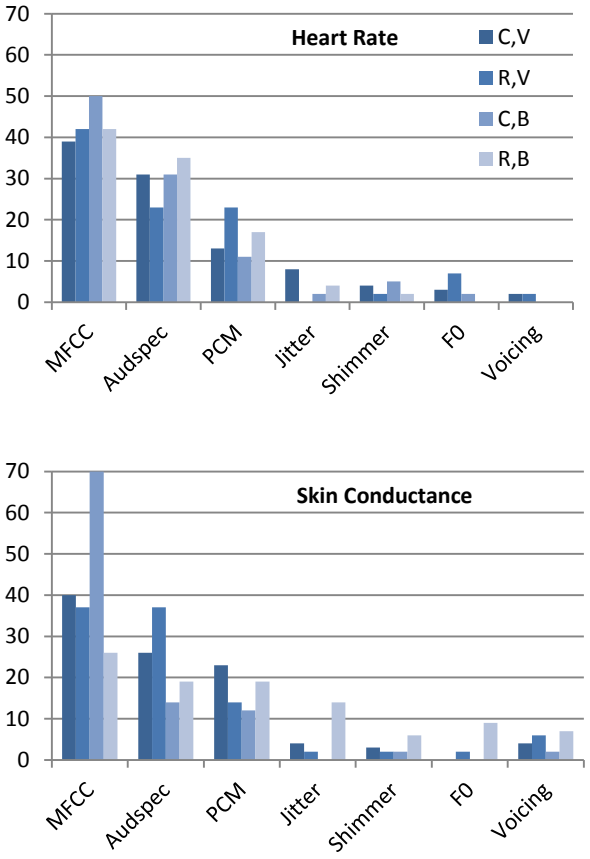

Figure 4: Feature group relevance. Top: heart rate in continuous beats per minute. Bottom: continuous skin conductance level as target. Shown are results for close-talk (C) or room microphone (R) for sustained vowels (V) or breathing periods (B).

5. Conclusion

We have introduced the Munich Biovoice Corpus (MBC) in this paper. Thereby the recording conditions have been described in detail, as well as the audio data and the labels contained in the corpus. Baseline results were shown, which – for speaker dependent settings – show very good performance when using 150 features automatically selected from a large, standard acoustic feature set of more than 4k features. As for feature group relevance, we found MFCC-type features and auditory spectrum-based ones particularly relevant. In a comparison of different classifiers, Support Vector-type algorithms prevailed.

The corpus is publicly available upon request from the authors. In future work we aim to further exploit different segmentations of the free text parts of the recordings. We also consider voice-based analysis of further bio-signals such as blood pressure (Broadwater 2002). Next, it will be interesting to analyse voice-induced heart rate (Kisilevsky & Hains 2011) and interdependence with pathologies (cf. (Baumgartner & Brutten 1982) or (Olsen & Strohl, 1987)). In particular, we aim to invite the research community to compare their results on the data in a well-formalised evaluation campaign (Schuller & al., 2014). Finally, comparison with other data and studies of speech and physical stress will be of interest (cf., e.g., the UT-Scope database (Godin & Hansen 2011) or (Meckel, Rotstein & Inbar 2002)).

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu) and No. 289021 (STREP ASC-Inclusion).

The responsibility lies with the authors.

We would like to thank all 19 participants of the study for their time and willingness to share the recordings made for scientific purposes.

References

Baumgartner, J.M. and Brutten, G.J. (1982). Expectancy and Heart Rate as Predictors of the Speech Performance of Stutterers, *Journal of Speech, Language, and Hearing Research*, 26, pp. 383-388.

Broadwater, K.J. (2002). *The Effects of Singing on Blood Pressure in Classicaly Trained Singers*, PhD Thesis, Louisiana State University.

Eyben, F., Weninger, F., Groß, F. and Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, *Proc. 21st ACM Multimedia*, ACM, Barcelona, Spain.

Godin, K.W. and Hansen, J.H.L. (2011). Analysis of the effects of physical task stress on the speech signal, *Journal of the Acoustical Society of America*, 130(6), pp. 3992-3398, ASA.

Hall, M. and Frank, E. and Holmes, G. and Pfahringer, B. and Reutemann, P. and Witten, I. H. (2009). The WEKA Data Mining Software: An Update, *SIGKDD Explorations,* 11(1), pp. 10-18. ACM.

Johannes, B., Wittels, P., Enne, R., Eisinger, G., Castro, C. A., Thomas, J. L., Adler, A. B. and Gerzer, R. (2007). Non-linear function model of voice pitch dependency on physical and mental load, *European Journal of Applied Physiology*, 101, 267–276.

Kisilevsky, B.S. and Hains, S.M. (2011). Onset and maturation of fetal heart rate response to the mother's voice over late gestation, *Developmental Science*, 14(2), pp. 214-23, Wiley.

Meckel, Y., Rotstein, A., and Inbar, O. (2002). The effects of speech production on physologic responses during submaximal exercise," *Medicine & Science in Sports & Exercise*, 34, 1337–1343.

Mesleh, A., Skopin, D., Baglikov, S. and Quteishat, A. (2012). Heart rate extraction from vowel speech signals, *Journal of Computer Science and Technology*, 27(6), pp. 1243-1251.

Orlikoff, R.-F. and Baken, R.J (1989). The Effect of the Heartbeat on Vocal Fundamental Frequency Perturbation, *Journal of Speech and Hearing Research*, 32(3), pp. 576–582.

Olson, L. G. and Strohl, K. P. (1987). The response of the nasal airway to exercise, *American Review of Respiratory Disease*, 135, 356–359.

Rammstedt, B. and John, O.P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German, *Journal of Research in Personality*, 41(1), pp. 203–212.

Schuller, B., Batliner, A., Steidl, S., Schiel, F. and Krajewski, J. (2011). The INTERSPEECH 2011 Speaker State Challenge, *Proceedings INTERSPEECH 2011*, 12th Annual Conference of the International Speech Communication Association, pp. 3201-3204, ISCA, Florence, Italy.

Schuller, B., Friedmann, F. and Eyben, F. (2013). Automatic Recognition of Physiological Parameters in the Human Voice: Heart Rate and Skin Conductance, *Proceedings ICASSP*, pp. 7219–7223, IEEE, Vancouver, Canada.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F. Ringeval, F., Marchi, E. and Zhang, Y. (2014). The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load, *Proceedings INTERSPEECH 2014*, 15th Annual Conference of the International Speech Communication Association, 5 pages, ISCA, Singapore, Singapore.

Skopin, D. and Baglikov, S. (2009). Heartbeat feature extraction from vowel speech signal using 2D spectrum representation, *Proc. 4th International Conference on Information Technology (ICIT)*, 6 pages, Amman, Jordan.