**The Open University of Israel**

**Department of Mathematics and Computer Science**

# The Fingerprints of Pain in Human Voice

Thesis submitted in partial fulfillment of the requirements

towards an M.Sc. degree in Computer Science

The Open University of Israel

Computer Science Division

By

**Yaniv Oshrat**

**Prepared under the supervision of**

**Dr. Anat Lerner, Dr. Azaria Cohen, Dr. Mireille Avigal**

**The Open University of Israel**

July 2014

# Contents

# List of Figures

# List of Tables

## Abstract

Obtaining an objective assessment of pain is an important challenge for clinicians. The purpose of this study was to examine the connections between subjective reports of pain and measureable parameters of human voice at the time of the report, as a step towards coping with this challenge.

Patients reporting pain were voice recorded on several occasions, to attain reports on different levels of pain. Recording was done in the patients' natural environment at the medical center. Voice samples were cut from the recordings and audio features were extracted, including features that were exclusively developed for this study. A machine-learning based classification process was performed in order to distinguish between samples with "no significant pain" reported and samples with "significant pain" reported. This classification process distinguished well between the two categories using a short list of features. Classification with a large number of features achieved higher rates of success. Moreover, features developed during this study improved classification results in comparison to classification based on known-features only. Differences between genders in reference to classification success rates were found.

Results indicate that there is evidence of a connection between measureable parameters of human voice and the simultaneous self-reported pain level. This finding might be useful for developing future methods to assess pain in absence of verbal communication or when objective assessment is critical[1].

The recorded raw material was used to establish The Open University of Israel corpus of speeCH in pain (OUCH-corpus) – an open-source voice samples corpus which may be used for further investigation of the relations between pain and human voice[2].

[1] A paper based on this work was written and would be submitted to a journal soon.

[2] A paper presenting the OUCH-corpus was written and would be submitted to a journal soon.

# 1. Introduction

Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage. Pain is a very significant phenomenon in medical treatment, and the assessment of patient's pain by a physician is a major challenge to the latter, especially when direct verbal communication is limited, e.g. as a result of the patient's medical situation. Another factor that complicates pain assessment is the relativity of pain between different persons [1].

Numerous researches have been conducted in order to find connections between pain and measureable physical parameters. The most prominent parameters to be investigated are heart rate, heart rate variability, skin conductance and systolic blood pressure. These studies yielded interesting finding regarding the sought connections, as we describe in section 1.1 below.

The human voice encapsulates a large amount of information about the speaker [2]. Many studies in recent years showed the ability to extract information about the physical and mental conditions of a speaker from his voice. Again, we elaborate on these studies in section 1.1.

Recognizing that pain affects physiological parameters of the human body, and that physiological characteristics of the body influence the human voice, it should be interesting to explore whether there is a connection between the subjective experience of pain and a person's voice. Since speech is the fastest and most natural means of human communication, this possible connection may have practical implications, for instance assistance in medical diagnosis, in verification of the patient's complaints or in acquisition of a pain report when verbal communication with the patient is impossible. Currently physicians base their pain assessment exclusively on the patient's subjective pain report. This approach might evolve in the future into a medical aide that will provide physicians with more objective pain assessments.

## 1.1. Related work

### 1.1.1.Pain and physiological parameters

Researchers showed connections between pain and measureable physical parameters. Loggia, Juneau and Bushnell [3] studied heart rate and skin conductance in response to pain inflicted as heat stimuli, combined with levels of pain rated by the subjects. They found that skin conductance and heart rate significantly increase during pain. This result confirms a wide range of previous studies that have observed this relationship with either heart rate, skin conductance, or both, using many types of pain.

Treister et al. [4] studied multiple parameter response to pain inflicted as heat stimuli, combined with the level of pain rated by the subjects. It was found that all of the 5 tested parameters – including heart rate and heart rate variability - successfully differentiated between no pain and all other pain categories, but none of the parameters differentiated between all 3 pain categories that they used. However, the combination of autonomic parameters demonstrated stronger associations with stress responses or noxious stimuli than each parameter separately, thus indicating a clear superiority of the multi-parametric approach in this regard.

Lindh, Wiklund and Harkansson [5] assessed pain by frequency domain analysis of heart rate variability during a routine heel lancing procedure that was performed in new-born infants as part of neonatal screening. It was shown that the squeezing of the heel is the most stressful event during the heel prick procedure.

In addition to the above, several other studies showed relations between heart rate and systolic blood pressure, and pain [6], [7]. In all cases, stimuli or exercise caused an increase in the physiologic parameters measured.

### 1.1.2.Voice and physiological parameters

Studies showed that the human voice is affected by physiological factors, and specifically by some of the parameters that were discussed above. Orlikoff and Baker [8] found that the cardiovascular system has a consistent affect on vocal fundamental frequency (F0). Orlikoff [9] showed that vowel amplitude variation is affected by the cardiac (ECG) cycle.

Human voice contains vast information, and many studies have shown that extensive knowledge about a speaker's physiology can be extracted from his voice [2]. In recent years numerous attempts have been made to extract similar information about a speaker's emotional state as well [10], [11], [12]. One of the most interesting aspects of mental condition, that attracts many researchers, is stress, due to the significant consequences that it has on human behavior. In many studies participants who were exposed to different types of stress were recorded, in order to show the effect stress has on human voice [13], [14].

Stress has been approached frequently in this kind of research. Often the way to arouse stress in participants in studies included physical manipulation, such as heat stimuli

[4], and cold stimuli [15]. Despite this, a review of literature shows no discussion on a direct connection between pain and voice.

Among the studies that investigated a connection between voice and physiological parameters, several approaches can be found: There are studies that performed special interviews with the participants , recorded the voice only, extracted vocal features and used statistics to classify the samples [10], [11]. Other studies used voice recordings of patients from interviews that were recorded for other objectives, and performed an analysis on voice samples from these interviews [13]. There are studies that intended to investigate not only the voice, but other parameters as well. The voice recordings were part of a wider set of measurements that were taken. Thus participants were fitted with pneumotachograph facemasks that recorded not only the voice of the participant, but also his oral and nasal air flow, inter-oral pressure, blood pressure and heart rate [15]. This kind of recording requires major set-up procedures and cast limitations on the consent to take part in the study and on the repeatability of the collection (if more than one sample per participant is desired).

### 1.2. The scope and goals of our study

The current study was a preliminary investigation aimed at revealing connections between subjective self-report of pain and parameters of voice at the time of the report. We could not find previous works that investigated these possible connections. If found, such connections might be very instrumental for developing future methods to assess pain in absence of verbal communication, or when objective assessment is critical.

To achieve this goal we considered a large variety of possibilities to establish such a connection, and then focused on the ones that showed potential. We employed known techniques and developed new ones as needed.

The study is based on data collection of audio material from persons who suffer pains due to injuries. The collection was designed and performed especially for the current study, but eventually was found adequate for further investigations as well. From this reason we decided to organize it as a general database of audio samples of speech in pain, and to make it available to other researchers via the internet.

In section 2 we describe the participants, the data collection and the sample processing. We portray the machine-learning analysis and the feature extraction, and elaborate on the special features that were developed during the study. In section 3 we present the results we achieved. In section 4 we discuss the results and their consequences. In section 5 we point at several issues that should be taken into special consideration regarding our results, and suggest ideas for further investigation. Section 6 presents The Open University of Israel corpus of speeCH in pain (OUCH-corpus) - the open source audio corpus which we built and made available for research usage, based on the data collected during our study.

## 2. Methods

The study was conducted by orally interviewing patients in the Inpatient Department of Neurological Rehabilitation and the Outpatient Mobility Rehabilitation Clinic of the Chaim Sheba Medical Center at Tel-Hashomer, Israel. Permission was obtained from the Sheba Medical Center ethical committee and from the Open University of Israel ethical committee before recruiting participants. Informed consent was obtained from all the participants after receiving a full explanation of the goals and protocol of the study.

### 2.1. Participants

Participants included 27 adults (20 men, 7 women, age range: 23-65 years, mean age: 44 years). All of the participants suffered spinal and/or brain injuries, and reported injury-related pain. No pain, physical pressure or any intrusive action was inflicted upon participants as part of the study.

Since the study included a short interview in Hebrew, only patients who speak Hebrew fluently participated in the study. Most of the participants' mother-tongue is Hebrew and the rest have been speaking Hebrew for many years with good command of the language.

Exclusion criteria for all participants were: (1) communication or cognitive problems that may interfere with understanding the researchers’ explanations; and (2) psychiatric conditions that may cause inappropriate behavioral responses during the study.

As mentioned, the population of participants included males and females. All procedures were maintained the same for both males and females. In the following text we use musculine grammer only for reasons of convenience.

#### 2.1.1.Natural environment principal

Participants were approached in their natural environment at the medical center, and in their regular day-to-day course of behavior. Daily schedule, including sleep, meals, physical and psychological therapies and other elements of rehabilitation program, was maintained intact. Participants taking medicines (pain killers or other kind of drugs) continued to take them regularly during the study. The participant's posture was maintained in accordance with his regular habits, i.e., sitting on a chair, sitting in a wheelchair, sitting or lying in bed.

### 2.2. Data Collection

The data collection took place between December 2013 and June 2014. Participants were approached in the morning and afternoon (between 09:30 and 15:00). Each session included a short interview that was voice-recorded. No video recording was taken. This is due to the fact that the setting and the medical condition of the patients might lower their motivation to be photographed and as a consequence video recording might have had an effect on their willingness to participate in the study.

The interview took place in a closed location that enabled quiet and privacy in the participant's surrounding. Most of the recordings were done in the participant's own room, and when this option was not possible (because of lack of privacy or unsuitability for recording), a quiet room in proximity was used. The same person conducted all the interviews with all the participants.

In the interview, each participant was asked to state, in his voice, his name and ID number. Then he was asked to evaluate his current level of pain on a numeric scale: Degree 1 signified no pain at all; degree 10 denoted the most intensive pain the participant has ever suffered in his life. The 1-10 pain scale was chosen since it is the standard scale both in studies of this kind and in medical questioning on pain in the Sheba Medical Center. In some cases participants stated a two-level pain degree (e.g. 5-6), in which case the mid-level degree (i.e. 5.5) was recorded.

The interviews were recorded using the ZOOM Handy Recorder Type H4n (Japan), in stereophonic recording, defining a sampling rate of 96 kHz. The recording device stood on a tripod no more than 1 meter from the participant's head. The height of the recording device was adjusted to the participant's head based on his posture. Audio files were stored in the WAVE file format.

After the recording phase, interviews were listened to and checked to verify recording quality and lack of significant surrounding noises. The total number of interviews that were found to comply with the study requirements was 97.

### 2.3. Sample processing

Each interview was cut into short voice samples of either digits from the participant's ID number or words from the participant's name. A single digital-sample included one to four digits; a single word-sample included one to four words. The mean length of a sample was 0.93 seconds (min. 0.23sec, max. 2.00sec). These characteristics were chosen in order to take similar samples from each participant in all his interviews, and in order to ensure the respone was a natural answer to a question and not a response given in a reciting mode: A person's ID number and name do not change and the participant knows them very well by heart, without any external assistance.

We decided to cut the interviews into multiple short samples since we believed that significant patterns, should they exist, would be found in every voice sample. Moreover, thin-slicing may help neutralize arbitrary noises and disturbances that might have occurred during the recording, and are not inherent to the situation. Cutting each interview into several samples would reduce the significance of such random occurrences, but would keep the consistent phenomenon intact. The thin-slicing approach proved to be a better method in a previous study that aimed at diagnosing a person's mental condition from speech samples [11].

Cutting the recorded interviews into voice samples was done manually using Audacity 2.0.5 software for editing sounds. All the interviews of a specific participant were cut the same way, into samples that included the same digits or words. From a single interview, 3 to 6 samples were derived in this manner, depending on the participant's speed of speech, pronunciation and choice of words (for example, there were participants who stated their names twice in an interview: once their first name preceding their family name and once the opposite). After removing the bad samples (surrounding noises, speech mistakes, etc.) 400 samples remained from 97 interviews of the 27 participants. Figure 1 shows the number of interviews that were maintained (light bar) for every participant (on the X-axis) and the number of voice samples that were produced from the interviews (dark bar).

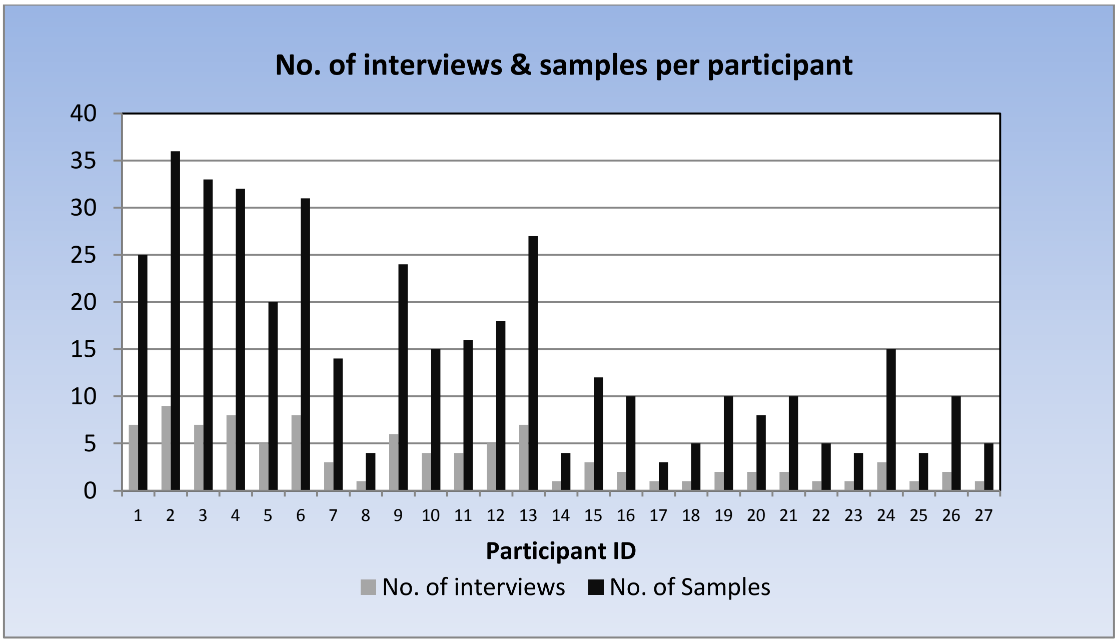

Figure 1 - Number of interviews and samples for each participant

### 2.4. Pain-level analysis

Participants reported pain of all levels from 1 to 8, and of all mid-levels excluding 3.5. This resulted in a scale of 14 levels (1, 1.5, 2, 2.5, 3, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8). None of the participants reported pain at levels above 8. Common sense indicates that a person with pain at these levels would not be able to conduct his regular course of life, and thus would not cooperate with an activity such as participating in an interview for a study. Conversations with participants revealed that some of them indeed had high-level pains during the period of the data collection, but when such levels of pain occurred they were preoccupied with getting treatment and were not available for interviews.

Figure 2 shows, for each level of pain (on the X-axis), the number of interviews in which the pain-level was reported by participants (light bar), and the number of voice samples that were produced from the interviews (dark bar).

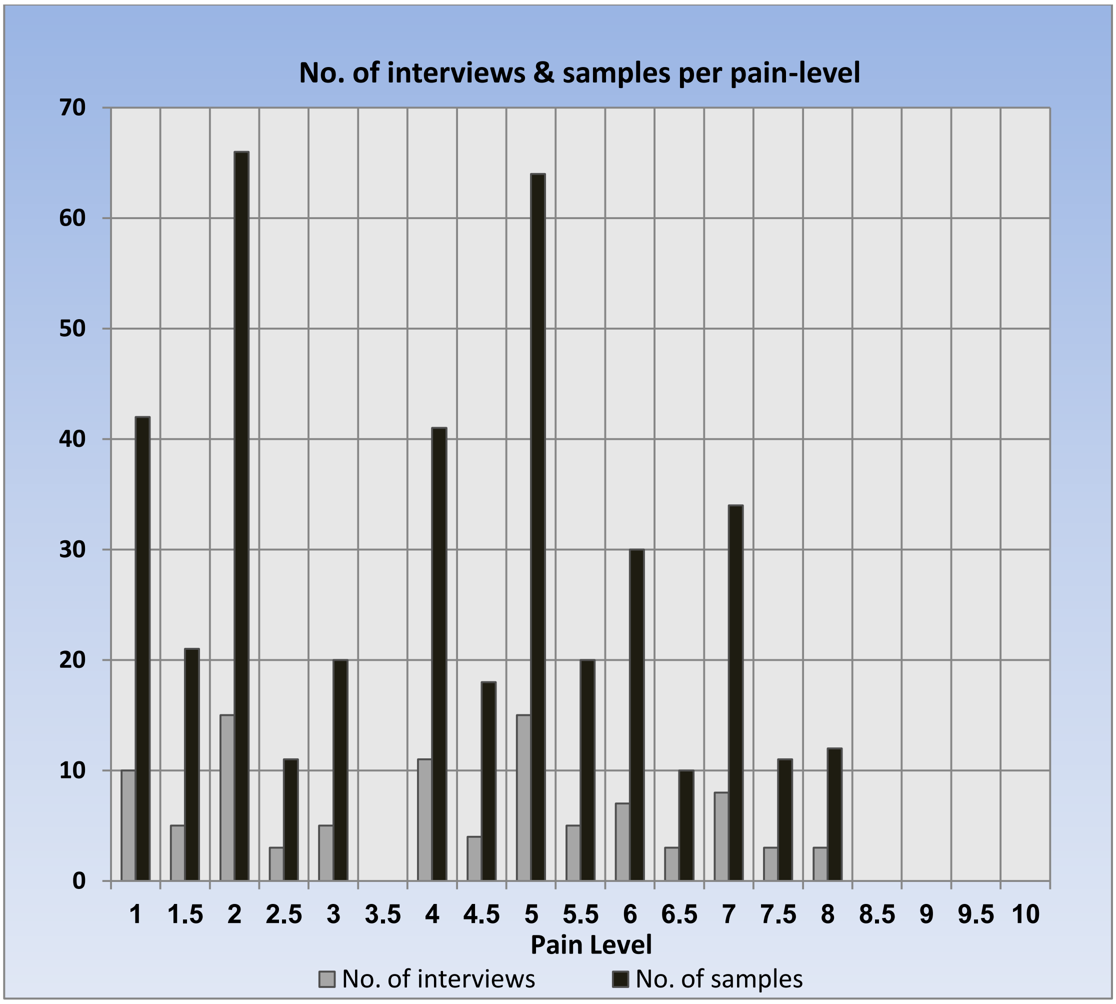

Figure 2 - number of interviews and samples for each pain level

### 2.5. Feature extraction

Audio features from the recorded samples were extracted at this stage using openSMILE (the Munich open Speech and Music Interpretation by Large Space Extraction toolkit), version 1.0.1 [16]. The openSMILE 'emobase 2010' reference set of features, which is based on the INTERSPEECH 2010 Paralinguistic Challenge feature set, was used. This set is very comprehensive and contains 1582 features, based on 34 low-level descriptors (LLD), corresponding delta coefficients and functionals, and is recommended by the developers as a reference feature set [17].

### 2.6. Machine learning analysis

Analysis of the data was done using the WEKA (Waikato Environment for Knowledge Analysis), version 3.6.10 [18]. We used the attribute selection feature of WEKA to choose the most effective attributes from the long list of features presented by 'emobase 2010' reference set, using the Correlation-based Feature Selection (CFS) [19], implemented in WEKA as CfsSubsetEval. This procedure reduced the number of attributes from the 1582 features of 'emobase 2010' to less than 50 features (the "OS short-list"). Classification was done using the Support Vector Machine (SVM) which is considered state-of the art classifier [20], assisted by the Sequential Minimal Optimization (SMO) algorithm [21].

Our first attempt was to classify the samples into the 14 different pain levels mentioned above. This attempt was not successful, and no consistent results were shown in the confusion matrices that were produced: Some of the levels had moderate Correctly Classified Instances (CCI) ratios, while other levels had low-to-very-low CCI ratios. We were not able to provide reasonable account for the varying rates of success of the different levels, or explain why a specific level was classified much better than the adjacent level in our scale. This finding was consistent with other attempts to differentiate between several categories of pain using physiological parameters [4].

We decided to try a different approach, and attempted to distinguish "significant pain" from "no significant pain". This task demanded that we define the line which separates levels of "totally no pain", "almost no pain", "no significant pain" and the similar, from levels of "significant-although-moderate pain" and higher. The task is a very delicate one, since pain is a relative-subjective experience, with different persons having different ways of defining the level of pain they experience [1].

Loyal to the nature of the study, we preferred to closely examine the preliminary results prior to setting this separation line. We ran the SMO algorithm on a target-class that included all 14 levels. From the confusion-matrix that was provided by the SMO algorithm, it was clear that the border between no-pain and pain lies in the neighborhood of levels 2 and 3 on our scale. The algorithm could distinguish quite well levels below this neighborhood from levels above it, but tended to confuse levels from the same side. In order to define the exact place to draw the line, we changed the target-class and tried three different two-level target-classes. The best results were achieved using the target-class that divides the scale between level 2 and level 2.5, or specifically as follows: No significant pain = levels 1, 1.5, 2; Significant pain = levels 2.5, 3, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8.

Using this separation-line, the data was divided into 271 samples with significant pain and 129 samples with no significant pain, i.e., a ratio of approximately 2:1.

### 2.7. Extended feature extraction

Until this stage we used only features from the openSMILE 'Emobase 2010' feature set. The results are presented in Table 1 in the results section of this article. In order to achieve better classification, we searched for more features that might comply with the kind of distinction we were seeking. Two types of features were developed: Sequence Indication and Heat Map Thresholding.

#### 2.7.1.Sequence Indication

Several studies, regarding different kinds of pain, showed that one of the symptoms of pain is making human responses less rapid [22], [23], [24]. This symptom might be expressed by alteration in the rate of change in voice parameters, which in turn may modify the length of sequences in the values of these parameters.

Sequence, in this context, is a list of consecutive measurements that are close in value to each other, or residing in a close neighborhood. The term "close" is relative to ε-value; two values $V_1$ and $V_2$ will be considered close if and only if $|V_1-V_2|<\varepsilon$.

We developed two kinds of sequence-length indications for lists of values that are typical of audio samples: In α-sequence indication, the sequence starts at the first measurement and ends when one measurement is not close to the previous one. The length of the sequence is the number of measurements in the sequence. In β-sequence indication, the sequence starts at the first measurement and ends when one measure is not close to the first measurement of the sequence. Here, again, the length of the sequence is the number of measurements in the sequence. The α-value of a list of measurements is defined as the mean length of all α-sequences in the list. Similarly, the β-value of a list of measurements is defined as the mean length of all β-sequences in the list.

Differences between α-sequence and β-sequence are most prominent in cases of moderately monotone sequences of measurements (either increasing or decreasing). In such cases, the α-sequence would have high values (i.e., long α-sequences and high α-value), since each measurement is close to the previous one, even though the values become very distant from the beginning of the sequence. The β-sequence would have lower values for the same input, since it compares the current value to the beginning of the sequence, no matter what the rate of each step. On the other hand, cases of alternate behavior sequences (for example, a sequence of the following pattern: 0, 5, (-5), 5, (-5), 5, (-5)…) would typically yield higher β values and lower α values.

Figure 3 demonstrates the computation of α-value and β-value on a list of measurements, when the ε was set to 1/20 of the range of the measurements. The sequences are marked by gates (" ┌─┐ "), and the length of each sequence is printed in the gate. At the right-hand side of the list the mean length of a sequence, i.e. the α-value and β-value, is indicated.

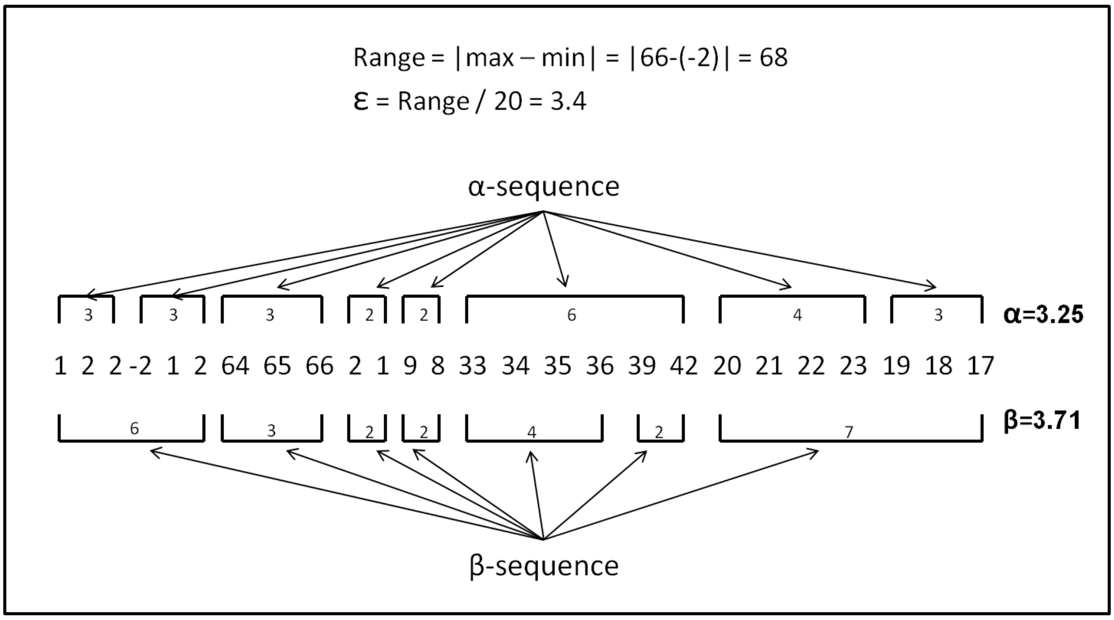

Figure 3 – Demonstration of computation of α-value and β-value for a list of measurements

We decided to deploy the sequence indications on the PLP (Perceptual Linear prediction) analysis that was performed on the voice samples [25]. PLP, like the LPC (Linear Predictive Coding) method on which it is based, is built on a model that approximates the human vocal tract and describes it as a set of filters [26]. Consequently, we presumed that vocal phenomena of the type that we were searching for might be found in the PLP analysis of speech. Moreover, since we were looking for alterations in the rate of change in voice parameters, we believed that the derivatives would be the right place to investigate.

The PLP analysis produces 13 coefficients for a single time-frame. When PLP is applied to a voice sample (that contains numerous time-frames), 13 vectors are produced: one vector per coefficient. For each of these 13 vectors we computed the first derivative vector (DEL) and the second derivative vector (DDEL). We computed the α-value and the β-value for both vectors. Thus, we computed 52 values for each voice sample (13 PLP coefficient vectors * 2 [DEL + DDEL] * 2 [α-value + β-value] = 52). In each computation, the value of ε was set to 1/20 of the range (|max-min|) of the values in the vector (i.e., ε=|max-min|*0.05). Implementation of these measures was made using MATLAB version 7.12.0.365 (R2011a).

### 2.7.2.Heat map thresholding

The second type of feature that was developed is based on Heat Map Thresholding. The motivation for this approach emerged after viewing the heat map image presentations of common feature extraction techniques, such as LPC, MFCC and PLP. In this presentation, the coefficients that were extracted from a voice sample are set in a 2-dimensional matrix: The first dimension is the number of the frame, and the second dimension is the coefficient values of that frame. The matrix is then displayed in a heat map format, using a color scheme that helps to illustrate phenomena in the image. This presentation can be applied to any feature extraction technique that produces a list of coefficients for a single frame. Figure 4 presents a heat map of a voice sample that was produced in the aforementioned way, and colored using the "jet" colormap.

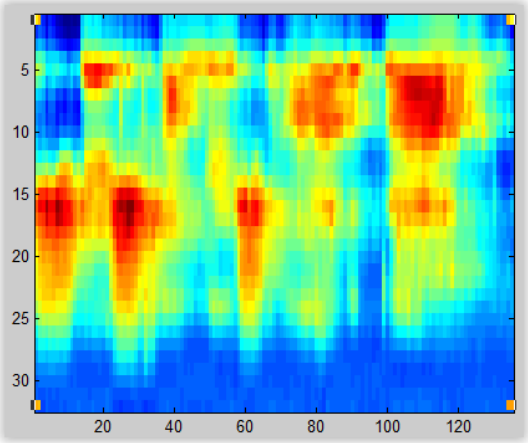

Figure 4 – Voice sample presented as a 2-dimesion matrix of RASTA-PLP coefficients and colored with the "jet" colormap. The X-axis is the number of the frame, the Y-axis presents the coefficients.

We compared pairs of such images, where each pair was produced by two voice samples that were taken from the same participant. The first sample is with a low pain-level reported and the second is with a high pain-level reported. Visually we noticed differences in the color patterns of the two images, where these differences motivated us to investigate them in a more rigorous manner than merely observing the image.

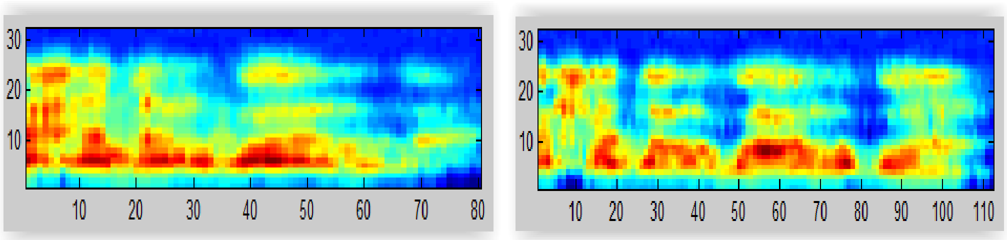

Figure 5 – Visual comparison between two images of voice samples of the same word, uttered by the same participant, once with a high pain-level reported (left) and once with a low pain-level reported (right).

We produced images using this method for the RelAtive SpecTral PLP (RASTA-PLP) coefficients. RASTA-PLP is an extension of the PLP technique that deals more effectively with problems of real-world recording and communication environment [27]. Similarly to PLP, RASTA-PLP computes the critical-band spectrum and its logarithm, but then it estimates its temporal derivative using regression and spectral values at 5 consecutive time instants, and not just a simple difference [28]. It actually filters the PLP coefficients with a band-pass filter that corresponds to the timing characters of the speech. This operation cancels the convolutional noise that is typical of recordings and communication media. Since we use recorded audio material that was recorded in field conditions, this technique seemed appropriate.

We performed image analysis in the following manner: The image underwent thresholding with five different threshold values (1/6, 2/6, 3/6, 4/6 and 5/6 in the range of values in the original image). We now had five new images for each voice sample. Each image contained a different number, pattern and size of spots ("spot" is a simply-connected space on the 2-dimensional image). Figure 6 illustrates an original image and its five derivatives.

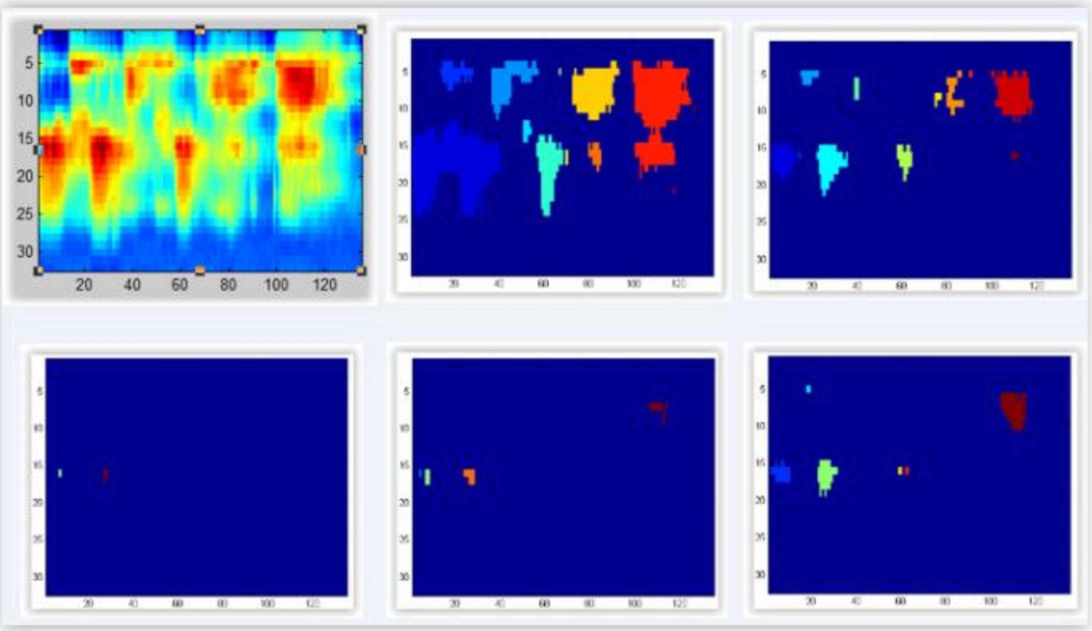

Figure 6 – Original RASTA-PLP 2 dimensional image (upper, left) and its five threshold derivative images (clockwise). The upper-right image contains 10 spots; the lower-left image contains 2 spots.

For each of the five new images, six measures were taken:

i. The number of spots in the image.

ii. The relative area of the spots of the total area of the image.

iii. The mean size of a spot in the image.

iv. The relative sum of values of the spots of the total sum of values in the image.

v. The mean of the values in the spots in the image.

vi. The standard deviation of the values in the spots in the image.

These two types of features (the α/β sequence indication features and heat map thresholding features), which together we call "original features", include altogether 260 features. As in the former case, we performed a feature selection phase to form a short list of features.

### 2.8. Feature sets

In conclusion, we had the following feature sets to test as classifiers:

i. The full openSMILE 'Emobase 2010' feature set, annotated "Emobase 2010", containing 1582 features.

ii. The full openSMILE 'Emobase 2010' feature set + all the 260 "original features", annotated "Emobase 2010 + original features", containing 1582 + 260 = 1842 features.

iii. The short-list of selected features from openSMILE 'Emobase 2010', annotated as "OS short-list", containing 43 features.

iv. The short-list of selected features from openSMILE 'Emobase 2010' + the short-list of "original features", annotated as "OS + original features short-list", containing 43 + 20 = 63 features.

The first two feature sets were very comprehensive and actually included all the features that were available. Our goal was not to use this type of giant set of features, but to find a short list of parameters which are the most significant for the classification.

## 3. Results

| **Corpus** | **No. of Instances** | **Real Instances Division** | | **Feature set** | **Classification Instances Division** | | **Correctly Classified Instances (CCI ratio)** | **Kappa** |
|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | | True positive | True negative | | |
| Whole (males + females) | 400 | 271 | 129 | OS short-list only | 240 | 59 | 74.75% | 0.3725 |
| | | | | OS + original features short-list | 243 | 61 | 76.00% | 0.4022 |
| Males only | 317 | 220 | 97 | OS short-list only | 200 | 56 | 80.75% | 0.5177 |
| | | | | OS + original features short-list | 196 | 59 | 80.44% | 0.5201 |
| Females only | 83 | 51 | 32 | OS short-list only | 44 | 25 | 83.13% | 0.6440 |
| | | | | OS + original features short-list | 45 | 24 | 83.13% | 0.6398 |

Table 1 – Classification results for small sets of features ("short-lists")

Table 1 presents the results of the analysis. Each line represents a classification process that was executed on the data. The results are presented initially as one unified corpus of 400 instances ("whole (males+females)"), and then as two sub-sets: Male participants ("males only") and female participants ("females only"), containing 317 and 83 instances respectively. The Real Instances Division column shows the way the instances were actually divided: "Positive" is "Significant Pain" reported (i.e. pain level 2.5 to 8); "Negative" is "No Significant Pain" reported (i.e. pain level 1 to 2). The Feature Set column indicates which set of features was deployed in the classification. The Classification Instances Division column demonstrates the success of the classification: The number of instances that were classified "True Positive" and the number of instances that were classified "True Negative". The Correctly Classified Instances (CCI) ratio presents the percentage of the correctly classified samples from the total number of samples in the corpus. Kappa is the kappa inter-rater agreement value (a.k.a. "Cohen's kappa", [29]) which serves as a better rater of the classification quality than the CCI [30], [31].

Examining the results in table 1 we can see that in all three cases the classification algorithm succeeded to classify the samples with relatively high ratio. Dividing the corpus into separate sub-sets according to gender made the results much better both in CCI ratio and in kappa statistic.

The addition of original features improved the results of the "whole (males+females)" corpus. In the "males only" corpus they provided a minor improvement. In the "females only" corpus they actually had no significant impact.

The results shown on table 1 relate only to classifications which include low number of features (<100), since it was our goal to find a short list of parameters that are the most significant in distinguishing between "pain" and "no pain". It should be stated that leaving the "short-list demand" out, the "original features" can obtain a large improvement to the "Emobase 2010" feature set classification. As an example, Table 2 demonstrates a significant increase both in the CCI ratio and in the Kappa statistic when the "original features" are appended to the "Emobase 2010" list.

| **Corpus** | **No. of Instances** | **Real Instances Division** | | **Feature set** | **Classification Instances Division** | | **Correctly Classified Instances (CCI ratio)** | **Kappa** |
|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | | True positive | True negative | | |
| Whole (males + females) | 400 | 271 | 129 | Emobase 2010 (1582 features) | 219 | 76 | 73.75% | 0.3981 |
| | | | | Emobase 2010 + original features (1842 features) | 230 | 79 | 77.25% | 0.4697 |

Table 2 – Example of classification results for a large set of features

## 4. Discussion

In this study, we aimed to find connection between pain experience as reported by the patient and the patient's voice at the time of the report. The results show that there is indeed a connection. We were able to classify the voice samples, distinguishing between samples that were uttered when the patient reported having no significant pain, and samples that were uttered when the patient reported having pain. This classification was done using the recorded voice samples only, with no further information about the situation or about the patient.

This connection between pain and voice might be indirect. Perhaps the pain, which the patients reported having, caused a physiological or mental effect, and that effect caused the vocal change which we found. For example, possible mental effects that might be the middle link in this type of connection are stress, anxiety or fear. Nevertheless, we indicate a chain of causes that begins with pain and ends with distinguishable vocal results.

The classification was performed using a relatively small number of audio features from the comprehensive openSMILE 'Emobase 2010' feature set (the "OS short-list"), with additional features that were developed during the study ("original features"). The most effective acoustic features for pain classification from the 'Emobase 2010' feature set were the MFCC (Mel-Frequency Cepstral Coefficients 0-14), the logMelFreqBand (logarithmic power of Mel-frequency bands 0-7) and the lspFreq (8 line spectral pair frequencies computed from 8 LPC coefficients). Features derived from these LLDs formed the "OS short-list" feature set. From the "original features", both sequence indication features and heat map thresholding features were found to be significant and were included in the short list.

The "original features" that were added to the "OS short-list" feature set improved the classification in part of the tests, i.e., in the general case and in the male-patients-only case. The "original features" had no significance in the female-patients-only case.

Our results comply with previous studies which found it possible to distinguish "no pain" from "pain" using various physiological indications, but had difficulty in differentiating between the levels of pain [32]. A possible explanation for this finding might be the subjectivity of the experience of pain, and the significant dissimilarity in the ways different people assess their pain [1]. This dissimilarity is very dominant in fine classification, but may be prevailed in two-category classification.

Our results were achieved with voice samples that were recorded in the patient's natural environment. This fact had some disadvantages (that will be further discussed below), but it indicates that future practical use of the findings might be done in medical and therapeutic areas, and is not confined to "sterile" surroundings only.

## 5. Limitation and future direction

We maintained the Natural Environment Principle, since the participants were approached in their natural environment at the medical center, and in their regular day-to-day course of behavior. This principle allowed us to include numerous participants and conduct interviews in their most authentic conditions, while compelling us to ignore several factors that may have affected the results. Medication taken regularly or irregularly by participants, diet changes, depression and brain injury effects are all factors that were overlooked in this study. Furthermore, differences of gender were found to affect the success of the classification. Since our study is a preliminary one, we believe that there is need for further study in these areas. Researchers that try to isolate and control these factors to find their separate affects on classification abilities might reveal interesting results.

All of the participants in our study suffered spinal and/or brain damage. These kinds of injuries have special characteristics that affect both the pain and the reporter, and might be different from other types of paining situations. We believe that other populations should be investigated as well.

Another factor that was influenced by the Natural Environment Principle is the voice recording set-up. The recordings were performed in quiet rooms, but these were regular quiet rooms in the medical center and not recording studio rooms with noise absorption equipment. We checked each recording post factum and removed samples that included surrounding noises. Nevertheless, the recording quality of the interviews is not that of a studio.

We tried several methods of feature extraction and found a number of effective features, while other features were found to be useless. We believe that more relevant features may be found using various feature extraction methods.

## 6. The Open University of Israel corpus of speeCH in pain (OUCH-corpus)

During our study we understood that the data we collected might enable many other directions of research. Some of these possible directions are described above, but many other possibilities exist. We felt that making the data available for other researchers would contribute to the research of voice, speech and pain.

To enable use of the data we had to process the audio in a different manner, in order to maintain participants' privacy and confidentiality. We decided to cut only single-digit samples, and no more than 5 digits per participant. The selection of digits from the recorded interviews is deliberately partial and non-sequential, in order to prevent identification.

### 6.1. Content of the Database

The Open University of Israel corpus of speeCH in pain (OUCH-corpus) contains good quality voice samples of single digits processed from the raw material. It contains 437 voice samples from 97 interviews of the 27 participants. Statistical information about the distribution of the data is presented in figures 7 and 8.

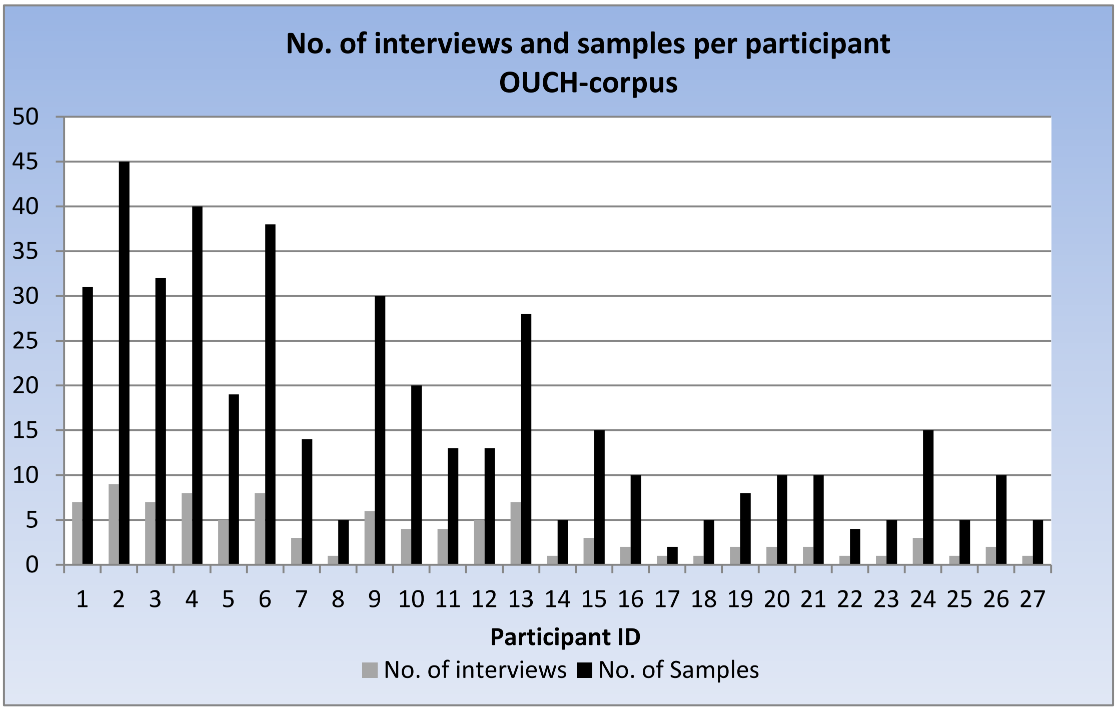

Figure 7 - Number of interviews and samples for each participant in OUCH-corpus. For each participant (on the X-axis), the number of interviews that were maintained with him (light bar) and the number of voice samples that were produced from these interviews (dark bar) are shown on the Y-axis.

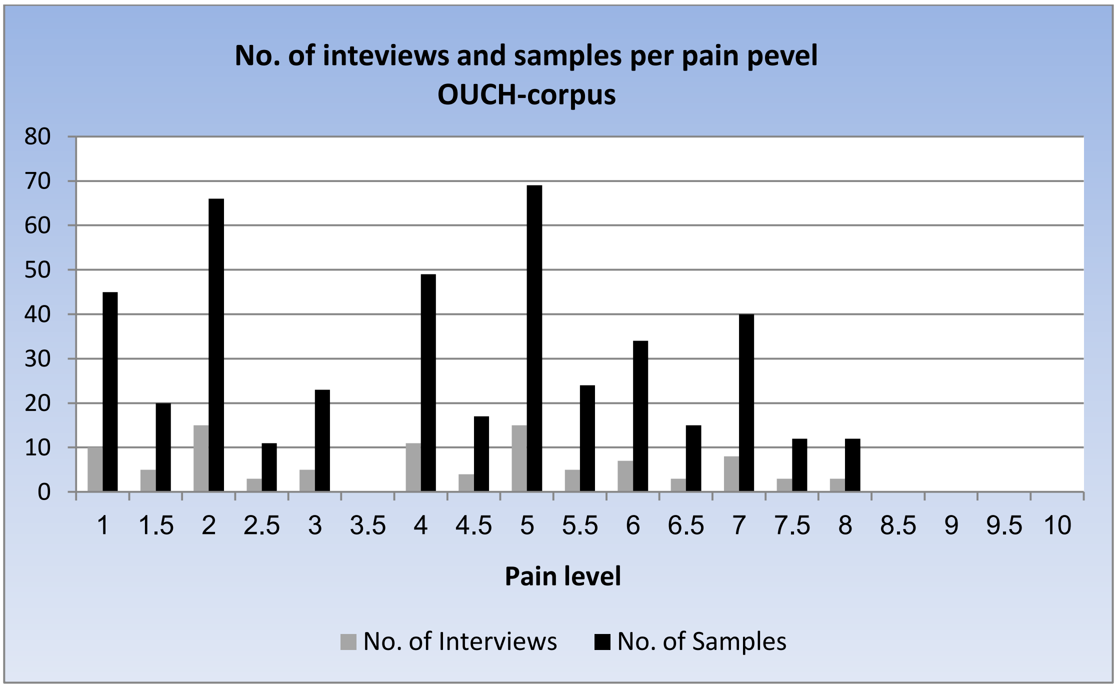

Figure 8 - number of interviews and samples for each pain level in OUCH-corpus. For each level of pain (on the X-axis), the number of interviews in which this pain-level was reported by participants (light bar), and the number of voice samples that were produced from these interviews (dark bar) are shown on the

Y-axis.

### 6.2. Technical Organization of OUCH-corpus

The samples are stored in an MS-Excel spreadsheet. Each line represents a single sample. The line contains the participant's number, age, gender, and the level of pain reported by him in the interview from which the sample was taken. The line includes also the name of the audio file of the sample, and a direct link ("attachment") to the audio file (in WAVE format). Naturally, the spreadsheet may be sorted, screened and searched in order to investigate its contents according to user needs.

In addition, all the audio files are stored in a folder. Files can be accessed via the MS-Excel file (as described above) or directly from the folder.

### 6.3. Comparison to other voice and speech databases and corpora

There are numerous voice and speech databases that were collected for various objectives. Many of the real-life speech databases were collected to assist in development of automatic speech recognition (ASR) systems. Such systems need voice material for training and calibrations, and various corpora of speech, in many languages, dialects and communication media, were collected for this purpose [33], [34], [35]. Other objective, with much resemblance to the previous one, is language identification [36].

For emotion identification, real-life collection is much harder, since it is very difficult to specify the emotion that the recorded person felt at the real-life situation in which he was involved. Moreover, often more than one specific emotion exists at a certain moment, and a single emotion cannot be isolated. Hence, acted emotions which are played by actors are recorded and stored in databases for emotion studies [37]. There are different opinions about the authenticity of such acted recordings and their applicability in research of emotions [38].

There are databases which are neither acted nor real-life, but are in reading speaking style, i.e. the participants are recorded while reading written text [34]. Again, these databases are satisfactory for ASR, but might be inadequate for emotion recognition.

As it was presented in the previous section, OUCH-corpus contains speech samples taken from interviews. The participants neither read nor acted, but answered genuinely to the personal questions they were asked. Therefore the corpus can actually be considered a real-life database.

Moreover, OUCH-corpus contains pain level reports for each sample. The pain was not a result of an artificial stimuli inflicted during research, which naturally, because of humanitarian reasons, is limited to low levels of pain [3], [4]. Part of the participants in our study experienced high levels of pain due to their injuries, and the amplitude of values in the corpus is wide.

### 6.4. Corpus availability

The OUCH-corpus can be accessed via the Open University site in the Internet at the following URL: http://www.cslab.openu.ac.il/proj/ouch. Terms of use are described at that location, too.

## 7. Conclusions

Our study established evidence to connections between human voice and pain. We achieved good classification abilities between voice sample of "no significant pain" and samples of "Significant pain", using the audio recordings only. The classification was based upon known audio features and original features that were developed during the study.

The study is a preliminary one. We pointed at several directions for future investigation. We established The Open University of Israel corpus of speeCH in pain (OUCH-corpus) and made it open-source in order to facilitate access to the data by other researchers.

Further investigations in this direction may improve the ability to distinguish "pain" from "no pain", and might evolve into a skill that can distinguish between different levels of pain. This capability might be useful in many practical uses, and especially as an aide to physicians in assessing the pain which their patients experience.

## 8. Acknowledgements

We wish to thank the participants in the study and the staff at the Inpatient Department of Neurological Rehabilitation and the Outpatient Mobility Rehabilitation Clinic of the Chaim Sheba Medical Center at Tel-Hashomer, for their time and effort.

## 9. References

[1] J. J. Bonica, "The need of a taxonomy," *Pain,* no. 6(3), pp. 247-248, 1979.

[2] C. Giddens, K. Barron, J. Byrd-Craven, K. Clark and A. Winter, "Vocal Indices of Stress: A Review," *Journal of Voice,* no. 27, pp. 390-398, 2013.

[3] M. Loggia, M. Juneau and M. Bushnell, "Autonomic responses to heat pain: heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity," *Pain,* no. 152, pp. 592-8, 2011.

[4] R. Treister, M. Kliger, G. Zuckerman, I. Goor Aryeh and E. Eisenberg, "Differentiating between heat pain intensities: the combined effect of multiple autonomic parameters.," *pain,* 2012.

[5] V. Lindh, U. Wiklund and S. Harkansson, "Heel lancing in term new-born infants: an evaluation of pain by frequency domain analysis of heart rate variability," *Pain,* no. 80, pp. 143-148, 1999.

[6] A. Moeltner, R. Hoelzl and F. Strian, "Heart rate changes as an autonomic component of the pain response," *Pain,* vol. 43, no. 1, pp. 81-89, Oct 1990.

[7] B. Robinson, "Relation of Heart Rate and Systolic Blood Pressure to the Onset of Pain in Angina Pectoris," *Circulation Research,* no. 35, pp. 1073-1083, 1967.

[8] R. Orlikoff and R. Baker, "The effect of the heartbeat on vocal fundamental frequency perturbation," *Journal of Speech and Hearing Research,* no. 32, pp. 576-582, 1989.

[9] R. Orlikoff, "Vowel amplitude variation associated with the heart cycle," *Journal of Acoustical Society of America,* no. 88, pp. 2091-2098, 1990.

[10] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear and G. Parker, "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech," *In FLAIRS Conference,* May 2012.

[11] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," *In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on,* pp. 7547-7551, May 2013.

[12] M. ElAyadi, S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* no. 44, pp. 572-587, 2011.

[13] M. McHenry, P. A. Parker, W. F. Baile and R. Lenzi, "Voice analysis during bad news discussion in oncology: reduced pitch, decreased speaking rate, and nonverbal communication of empathy," *Supportive Care in Cancer,* no. 20, pp. 1073-1078, 2012.

[14] R. Ruiz, c. Legros and A. Guell, "Voice analysis to predict the psychological or physical state of a speaker," *Aviation, Space, and Environmental Medicine,* no. 61, pp. 226-271,

1990.

[15] C. L. Giddens, K. W. Barron, K. F. Clark and W. D. Warde, "Beta-adrenergic blockade and voice: a double-blind placebo-controlled trial," *Journal of Voice,* no. 24, pp. 477-489, 2010.

[16] F. Eyben, M. Woellmer and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *In Proc. ACM Multimedia (MM),* pp. 1459-1462, Oct. 2010.

[17] F. Eyben, M. Woellmer and B. Schuller, "The openSMILE book - openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," 23 may 2010. [Online]. Available: http://opensmile.sourceforge.net/. [Accessed 23 Apr 2014].

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, no. 1, 2009.

[19] M. A. Hall, *Correlation-based Feature Selection for Machine Learning,* Hamilton: The University of Waikato, 1999.

[20] B. Schuller, A. Batliner, S. Steidl and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication,* vol. 53, no. 9-10, p. 1062–1087, 2011.

[21] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," 1998.

[22] S. Taimela, K. Osterman, H. Alaranta, A. Soukka and U. Kujala, "Long psychomotor reaction time in patients with chronic low-back pain: preliminary report," *Archives of Physical Medicine and Rehabilitation,* vol. 74, no. 11, pp. 1161-1164, 1993.

[23] A. Reinersmanna, G. S. Haarmeyer, M. Blankenburg, J. Frettlöh, E. K. Krumova, S. Ocklenburg and C. Maier, "Left is where the L is right. Significantly delayed reaction time in limb laterality recognition in both CRPS and phantom limb pain patients," *Neuroscience Letters,* vol. 486, no. 3, pp. 240-245, 2010.

[24] G. Crombez, C. Eccleston, F. Baeyens and P. Eelen, "The disruptive nature of pain: An experimental investigation," *Behaviour Research and Therapy,* vol. 34, no. 11-12, pp. 911-918, 1996.

[25] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* vol. 87, no. 4, pp. 1738-1752, 1990.

[26] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials,* vol. 7, no. 1, pp. 29-32, 1988.

[27] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "RASTA-PLP Speech Analysis Technique," *Acoustics, Speech, and Signal Processing, IEEE International Conference*

*on,* vol. 1, pp. 121-124, 1992.

[28] B. Gerazov and Z. Ivanovski, "Overview of Feature Selection for Automatic Speech Recognition," *Audio Engineering Society Convention 132,* Apr 2012.

[29] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, no. 1, pp. 37-46, 1960.

[30] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Coputational Linguistics,* vol. 22, no. 2, pp. 249-254, 1996.

[31] I. H. Witten, E. Frank and M. A. Hall, Data mining, Practical Machine Learning Tools and Techniques, 3rd ed., Burlington MA: Morgam Kaufmann, 2011.

[32] R. Treister, M. Kliger, G. Zuckerman, I. Goor Aryeh and E. Eisenberg, "Differentiating between heat pain intensities: the combined effect of multiple autonomic parameters.," *pain,* 2012.

[33] V. Zue, S. Sennef and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech database development at MIT: TIMIT and beyond,* vol. 9, no. 4, pp. 351-356, 1990.

[34] B. C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha and S. Prasanna, "Multi-Variability Speech Database for Robust Speaker Recognition," *Communications (NCC), 2011 National Conference on,* pp. 1-5, 2011.

[35] J. Hennebert, H. Melin, D. Petrovska and D. Genoud, "POLYCOST: A telephone-speech database for speaker recognition," *Speech communication,* vol. 31, no. 2, pp. 265-270, 2000.

[36] S. Maity, A. Vuppala, K. Rao and D. Nandi, "IITKGP-MLILSC speech database for language identification," *Communication (NCC), 2012 National Conference on,* pp. 1-5, 2012.

[37] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, "A database of German emotional speech," *Interspeech,* vol. 5, pp. 1517-1520, Sep. 2005.

[38] E. Douglas-Cowie, R. Cowie and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion,* 2000.

# תוכן העניינים

# תקציר

השגת חיווי אובייקטיבי של עוצמת כאב היא אתגר לרופאים ולקלינאים. מטרת מחקר זה היא לחפש ולבחון קשרים בין דיווחי-כאב סובייקטיביים לבין פרמטרים בני-מדידה של הקול האנושי בעת הדיווח, כצעד לקראת התמודדות עם האתגר הנ"ל.

מטופלים המדווחים על כאב הוקלטו מספר פעמים, עם דיווחים על רמות שונות של כאב. ההקלטות בוצעו בסביבה הטבעית של המטופלים במרכז הרפואי בו הם שוהים. דגימות קול נחתכו מתוך ההקלטות והופקו מהן מאפיינים קוליים (features), ובכלל זה גם מאפיינים שפותחו במיוחד במסגרת מחקר זה. בוצע תהליך סיווג המבוסס על למידה חישובית (machine learning) במטרה להבחין בין דיווחים מסוג "אין כאב משמעותי" לבין דיווחים מסוג "קיים כאב משמעותי". תהליך הסיווג השיג שיעורי הצלחה טובים בין שתי הקטגוריות, תוך שימוש במספר קטן, יחסית, של מאפיינים קוליים. סיווג המבוסס על מספר גדול של מאפיינים הגיע לשיעורי הצלחה טובים אף יותר. סיווג תוך שימוש גם במאפיינים הייחודיים שפותחו במהלך המחקר הביא לשיפור בשיעורי ההצלחה לעומת סיווג שהתבסס על מאפיינים מוכרים בלבד. נמצאו הבדלים בין גברים לנשים בשיעורי ההצלחה של תהליך הסיווג.

התוצאות מצביעות על קשר בין פרמטרים פיזיולוגיים בני-מדידה של הקול האנושי לבין עוצמת הכאב המדווח. ממצאים אלה עשויים לסייע בפיתוח שיטות להערכת עוצמת כאב בהיעדר תקשורת מילולית ובמקרים בהם נדרשת הערכה אובייקטיבית[1].

חומרי השמע שנאספו **במסגרת** המחקר עובדו, נערכו והונגשו. באמצעות חומרים אלה הוקם
The Open University of Israel corpus of speeCH in pain (OUCH-corpus) – קורפוס דגימות שמע, זמין דרך האינטרנט, שיוכל לשמש כבסיס למחקרי-המשך על-אודות הקשר בין כאב לבין הקול האנושי[2].

[1] מאמר המבוסס על עבודה זו נכתב ויישלח לכתב-עת.
[2] מאמר המתאר את OUCH-corpus נכתב ויישלח לכתב עת.

**האוניברסיטה הפתוחה**

**המחלקה למתמטיקה ולמדעי המחשב**

# טביעות האצבע של הכאב בקול האנושי

עבודת תזה זו הוגשה כחלק מהדרישות לקבלת תואר

"מוסמך למדעים" M.Sc. במדעי המחשב

באוניברסיטה הפתוחה

החטיבה למדעי המחשב

על-ידי

**יניב אשרת**

**העבודה הוכנה בהדרכתם של ד"ר ענת לרנר, ד"ר עזריה כהן וד"ר מיריי אביגל**

**יולי 2014**