**The Open University of Israel**

**Department of Mathematics and Computer Science**

**Do we accumulate emotions when we speak?**

Thesis submitted as partial fulfillment of the requirements towards an M.Sc. degree in Computer Science

The Open University of Israel

Computer Science Division

By

**Liat Madmoni**

**Prepared under the supervision of**

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

January 2016

# Contents

[Abstract 4](#_Toc437726914)

[1. Introduction 4](#_Toc437726915)

[1.1. Related work 4](#_Toc437726916)

[1.2. The goals and scope of our study 8](#_Toc437726917)

[1.3. Main Contribution 9](#_Toc437726918)

[1.4. Overview 9](#_Toc437726919)

[2. Recordings 9](#_Toc437726920)

[2.1. Emotions Induction Method 9](#_Toc437726921)

[2.2. Participants 10](#_Toc437726922)

[2.3. Environment & Recording Equipment 11](#_Toc437726923)

[2.4. Recording Process 11](#_Toc437726924)

[3. Speech Corpus 12](#_Toc437726925)

[3.1. Processing and Segmentation 13](#_Toc437726926)

[3.2. Annotation 13](#_Toc437726927)

[3.3. Corpus Description 13](#_Toc437726928)

[4. Acoustic Features extraction 15](#_Toc437726929)

[4.1. Features' Contribution for Emotion Recognition 16](#_Toc437726930)

[4.2. Feature Extraction 17](#_Toc437726931)

[5. Classification 18](#_Toc437726932)

[6. Results 21](#_Toc437726933)

[7. Discussion and future work 23](#_Toc437726934)

[8. References 25](#_Toc437726935)

[9. Appendix 1: The selected Acoustic Parameters 28](#_Toc437726936)

[10. Appendix 2: Sad Song Lyrics' Slide 29](#_Toc437726937)

[11. Appendix 3: Happy Song Lyrics' Slide 29](#_Toc437726938)

[12. Appendix 4: Classification results for individual speakers. 30](#_Toc437726939)

[13. תקציר 32](#_Toc437726940)

[14. תוכן עניינים 33](#_Toc437726941)

**List of Tables**

[Table 1: The flow of the recording sessions 12](#_Toc437721953)

[Table 2: The distribution of the utterances by gender and label 14](#_Toc437721954)

[Table 3: Summary of the corpus specifications 14](#_Toc437721955)

[Table 4: Classification algorithms comparison for the Females sad-happy utterances 19](#_Toc437721956)

[Table 5: Classification algorithms comparison for the Males sad-happy utterances 19](#_Toc437721957)

[Table 6: The number of utterances for each set of participants by labels 20](#_Toc437721958)

[Table 7: Phonetic match in Neutral vs Mixed 20](#_Toc437721959)

[Table 8: Two- discrete emotions classification results (Type 1 tests) 21](#_Toc437721960)

[Table 9: Three emotions classification results (Type 2 tests) 21](#_Toc437721961)

[Table 10: Two-classes groups classification results (Type 3 tests). 22](#_Toc437721962)

[Table 11: neutral vs. mixed classification results (Type 4 tests) 22](#_Toc437721963)

[Table 12: Classification results summary 22](#_Toc437721964)

[Table 13: MATLAB Acoustic Features 28](#_Toc437721965)

[Table 14: Classification results for individual speakers 30](#_Toc437721966)

# List of Figures

[Figure 1: Emotions in valence vs. arousal plane (taken from [‎‎19]) 7](#_Toc437695086)

[Figure 2: Sad Song Lyrics' Slide 30](#_Toc437695087)

[Figure 3: Happy Song Lyrics' Slide 30](#_Toc437695088)

# Abstract

Speech Emotion Recognition is a growing research field. Many researches study the recognition of a discrete emotion. In this paper we focus on the influence of mixture of emotions, located on the opposite sides of the axis of the valence and arousal dimensions. We performed 4 stages of recording in a controlled environment: Neutral stage; Sadness stage – in which we induced sadness; Mixed stage – right after the sadness stage, with no pause we induced happiness; and Happiness stage – after a long pause, we induced happiness.

We extracted 3 sets of acoustic features, using MATLAB and the open–source openSMILE program. We then compared the mixed stage with the sadness, happiness and neutral stages to study the influence of the previous induced emotions on the present state. For classification, we used WEKA (Waikato Environment for Knowledge Analysis) software with Support Vector Machine (SVM) and SMO algorithm.

Our results show that blending sadness with happiness is significantly different from neutral, sadness and happiness which suggests that human accumulate emotions.

# Introduction

In Speech Emotion Recognition (SER) the emotional state of the speaker is extracted from a given speech signal. There are many implementations for SER, like: lie detection software based on a voice stress analyzer [‎50]; medical implementations, like evaluating a patient's mental state in terms of depressions and suicidal risks [‎47]; human-machine interaction applications, such as spoken dialogue systems in call centers or smart home applications [‎20] [‎38]; personalizing E-learning experience [‎49]; personification of an interface agent for various information centers [‎48], and many more.

SER researches differ in various aspects like: recording methods staging, that is using acted versus non-acted (real-life) speech; the number and type of the tested emotions; the analysis tools that are used to identify the emotions (e.g. computational learning tools, transcript and other lexical tools using lexical cues, etc.).

Some researches use multimodal corpora, combining speech with lexical cues (bag of words), facial expressions and body gestures [‎37] [‎39] [‎48]. In [‎21], three physiological variables were measured: the electromyogram of the curragator, heart rate and galvanic skin resistance (i.e., sweat) measured on a finger. In this paper we focus on corpora that are solely based on vocal input.

In the following sub-section 1.1, we review related work.

## Related work

The challenges in constructing an emotional speech corpus (a voice database) are: planning the setup and taking the recordings; and labeling the recorded utterances. The setup of the recording can affect the authenticity of the emotions as well as the nature and number of the emotions portrayed in the speech. In the labeling process, labels taken from the space of emotions examined, are assigned to each utterance.

There are two main types of speech databases: prompted (acted) and non-prompted (also called: real-world/natural/spontaneous). Some researches use more subtle classification of the speech databases and describe three types of corpora: Acted, real-life, and elicited (also referred as induced, Wizard of Oz (WoZ) or semi-natural). Each database type has its own recording methods;

Acted recordings are usually performed in a recording studio, played by actors playing situations with predefined emotions [‎22] [‎31]. Sometimes, the utterances' content is chosen to fit the emotion (to achieve more realism). In such cases different sentences are used for different emotions [‎26] [‎36]. In other times the same sentence is repeated with different emotions [‎14] [‎22] [‎26], to ease the differentiation between the acoustic features of the same sentence based only on the different emotions.

One of the main drawbacks of acted speech is the lack of authenticity. To minimize the lack of authenticity, in some researches the recording process was skipped and the corpus was built from short clips taken from movies [‎18] [‎36]. The builders of the MURCO [‎52] database, for example, who used clips from movies, argued that by using experienced actors, context related text and relevant scenery, the actors can more easily get into the requested emotion.

For non-acted and elicited database recording there are several methods, for example: descriptive approach, WoZ approach, performing task methods, and more.

In the *descriptive speech approach*, attendees describe a picture, emotional event or a movie. In [‎21], attendees were asked to recall an emotional event and revive the feelings that they had felt at the original event.

In the *WoZ approach*, the behavior of an application or a system (e.g. a computer-based spoken language system) is simulated in such a way that the examinee believes he or she is interacting with a real interactive system. In fact the system behavior is controlled by one or more so called human `wizards', and is not really responding to the examinee's reaction. Examples for the WoZ approach can be found in Aibo experiment and in NAO-Children corpus that was based on this experiment [‎‎33]. In [‎‎33‎] children were recorded playing with a robot. In [‎48] the system named "Smartkom" asked the examinee to solve certain tasks (like planning a trip to the cinema) with its help.

In *performing task methods*, attendees are instructed to perform certain tasks. The assumption is that while being preoccupied with the tasks, the attendees are less aware of being recorded. Examples for performing task methods are: recordings during minimal invasive surgery (SIMIS Database [‎53]), and recorded attendees interacting with home center applications [‎20].

Other common non-acted methods are recording of incoming telephone calls in call centers [‎34] [‎38], and collecting speech from interviews in TV talk shows [‎35] [‎37].

The next task, after the recording, is labeling (annotating) the speech utterances with the expressed emotion.

The two most common approaches to emotional labeling are: categorical (aka discrete) and dimensional.

The *Categorical/discrete* emotion approach applies some basic emotions e.g. happiness, anger, sadness and so on (often referred to as 'the big-n') [‎21], [‎22] [‎26] [‎27] and [‎31]. For example, in [‎21] five basic emotions were used: anger, fear, joy, sadness, disgust in addition to neutral, while in [‎26] six basic emotions were used: anger, fear, surprise, disgust, joy, sadness and neutral.

The *dimensional* emotion approach is based on a psychological model, representing an emotional state using 2 or more dimensions scales. Examples for such scales are:

**Valence** (sometimes referred as: appraisal, evaluation or pleasure) - how positive or negative is the emotion.

**Arousal** (or: activation) - the degree to which that emotion inspires a person to act.

**Dominance** (sometimes referred as: power, potency or control) - the degree to which the emotion is dominant (e.g. anger is a dominant emotion, while fear is a submissive emotion).

Most commonly used dimensions are valence and arousal. Each emotional state can be represented as a combination of these dimensions. 'Sadness', for example, is an unpleasant emotional state (negative valence) and low intensity (arousal), while 'happiness' has positive valence and high arousal. An example for a research that uses valence and arousal dimensional labels can be found in [‎18].

Different models use different set of labels for each dimension. For example, in [‎38] and [‎20] the valence dimension is labeled by two possible labels: 'negative' and 'non-negative'.

According to the *circumplex model* *of emotion*s, developed by Russel [‎6], affective states (emotions) are represented as a circle in two-dimensional bipolar space. Figure 1, taken from [‎19], shows various emotions in a two dimensional space of valence and arousal. For example, happy has a higher valence value (more positive) than excited, while excited has a higher arousal value (more active) than happy.

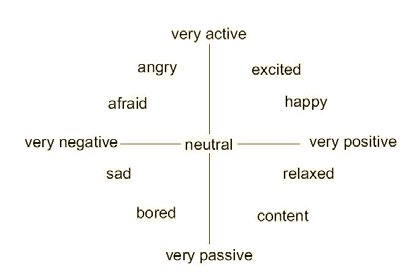


Figure : Emotions in valence vs. arousal plane (taken from [‎‎19])

In acted databases labeling is relatively an easy task, as the emotions are dictated and controlled. The same holds for semi-natural induced emotion databases [‎21].

One of the major disadvantages of using acted speech database is the lower classification validity with respect to real life speech. In [‎32] emotion recognition in acted and non-acted databases is compared, yielding the lowest classification rates when trained on the acted set and tested on the natural set. Training on the natural set and testing on the acted set yields the highest accuracy rates.

In real life databases, classification and labeling can be very challenging. The main challenge is to identify the recorded emotion.

In some cases the expressed emotion can be concluded from the context-related vocabulary used, or from non-verbal cues like yawns (for boredom, tiredness), laughs (for joy, sarcasm), and cries (for joy, anger, surprise etc.). Nevertheless, in many cases the decision is more complicated, due to multiple simultaneous (i.e. blended) or consequential emotions, and due to personal characteristics like introversion versus extroversion. More specifically, some people (especially in public) tend to restrain their verbal utterances and expressive behavior, disguise or falsely exhibit their emotions (mask).

A common strategy for identifying the emotions in speech is to use a group of professional labelers to tag the emotions. The utterances might be ambiguous and different labelers might annotate differently. Usually the final tags are chosen upon majority; another approach is to discard from the corpus all the utterances without a full agreement [‎33] [‎34] [‎38] [‎40].

When dealing with real-life corpora in SER studies, there is a need to cope with the complexity of real emotions. Real life emotions are usually not flat: people are not just 'happy' or just 'sad'; but one might experience simultaneous manifestation of more than one emotion, whether it's blended, consequential or masked.

There are relatively few studies concerning non-acted complex emotions. Most of the studies involving spontaneous speech ignored complex emotions completely and dealt only with basic emotions, even if by doing that, a part of the corpus was neglected. In [‎34] for example, only a sub-corpus of a spontaneous data from medical call center was used. That is, only utterances in which non-complex emotions were found were included. Five basic emotions were annotated: Fear, Anger, Sadness, Neutral and Relief, neglecting all complex emotions. In [‎38], the call-center data were independently tagged by 2 labelers. Only those data that had complete agreement between the labelers were chosen for the experiments, assuming that disagreement between labelers can result from complex emotions.

There are several methods to deal with labeling of complex emotions.

One approach is to use multiple labels per instance (utterance). For example, in [‎36], each instance is labeled by a major label and then by a minor label. Each label is drawn out of 16 categories of emotional labels. Only instances with full annotators' agreement on both major and minor labels were used. All other segments were discarded from the data-base. Another example that uses major and minor labeling for blended emotions can be found in [‎40] - the corpus of Stock Exchange Customer Service Center.

In NAO-Children corpus [‎33], to overcome the sequence of emotions phenomena, segment boundaries were defined. A segment is defined such that it is homogenous in terms of the emotion and its intensity.

Another way to limit the labeling conflict of complex emotions is to perform the recording in controlled environments, and hence achieving a small set of targeted specific emotions.

In some papers, the main annotation is limited to only two classes. For example, in the call center application [‎38], and in the Greek corpus from smart-home dialogue system [‎20], the 6 detected emotions are divided into two emotional categories: negative versus non-negative emotions. Utterances labeled *confused*, *anger* or *hot anger* are classified as negative, while utterances labeled delighted, pleased or neutralare classified as non-negative. In [‎40] - the corpus of Stock Exchange Customer Service Center - only the emotions: Fear, Anger and Neutral are considered, and classified into two classes: 'negative'(anger and fear) and 'neutral'. In [‎32], the annotators were given two options: 'anger and frustration' versus 'other'.

The next step after labeling is the extraction of the acoustic features. Researches differ in the number and type of acoustic features as well as in the methods for feature extraction. Recent researches use automatic tools like OpenSmile [‎10] and openEAR [‎11] (ex. [‎18] [‎28] [‎32] [‎35] [‎36]) or Praat [‎51] (ex. [‎14] [‎34] [‎40]), for feature extraction, producing a large number of features.

For the analyzing and classification step, machine learning tools or statistical models are applied. Classification algorithms (such as SVM, GMM, Bayesian networks, etc.) are often used to determine the emotion expressed in each utterance, based on the acoustic features. Many articles use WEKA data mining toolkit [‎13] (ex. [‎28] [‎30] [‎32] [‎35]), or write exclusive classification algorithms [‎22] [‎47]. In some cases feature selection algorithms (like , Correlation Feature Selection, Forward 3-Backward 1 wrapper, sequential forward floating search, Sequential Forward Selection, CFSSubsetEval supplied by WEKA) are applied on a test set to identify the most relevant acoustic features [‎15] [‎18] [‎26] [‎32].

## The goals and scope of our study

In the current research we study complex-emotions in semi-natural situations.

More specifically, we study the effect of a given discrete emotion on an opposite emotion (opposite in terms of valence and arousal levels), that immediately follows. We focus on happiness and sadness as opposite emotions that are relatively easy to induce.

The research question that we pose is:

When inducing two opposite emotions one immediately after the other: sadness and then happiness - will the successive emotion be differentiated from:

* A neutral state (do we sum our emotions?);
* The predecessor sadness;
* Independent happiness;
* All the above, i.e., a new blended emotion that is different from the previous emotion, the present emotion, and the neutral state (do we accumulate our emotions?).

To control the rapid transition of the opposite emotions, we constructed a speech database in a controlled environment.

## Main Contribution

This paper is a proof of concept. The main contribution is threefold: Firstly, we constructed a new Hebrew database for non-acted/semi-natural complex emotions (blended). As far as we know there is no other Hebrew corpus for non-acted blended emotions; Secondly, we compared different sets of acoustic features, with one of the sets exclusively picked for this purpose; and Thirdly, we compared the mixed emotion and showed that the result is a new state, different from the basic two emotions that were mixed, yet also different from the neutral state.

## Overview

The rest of this paper is organized as follows. In section 2 we describe the recording process: the emotion induction method; the participants; recording equipment and recording process. In section 3 we describe the speech corpus: wav files segmentation and annotation. In section 4 we discuss the features selection and extraction. In section 5 we deal with classification. The classification results are presented in section 6. Finally, in section 7 we suggest some directions for future work.

# Recordings

In the current research, we study mixed emotions that result from rapid change in the induced emotion form sadness to happiness.

We chose not to use existing speech databases, in order to be able to control the environment. More specifically, we found essential for this research to be as close as possible to natural speaking; we wanted to induce sadness immediately followed by happiness; and we found important the similarity of the wording of the different examinees.

## Emotions Induction Method

To construct the speech corpus, we recorded 12 participants, each participant four times (for a total of 48 recordings). The participants were instructed to read two songs. The songs were chosen such that one of them is associated with happiness and the other one is culturally associated with sadness.

To induce the appropriate atmosphere we used two steps: we first played a song and added visualizing correlated pictures on the screen; and then the participants were instructed to read out loud the lyric of this song.

Each participant was recorded four times: first with no intentionally induced atmosphere ("neutral"); then with a sad induced atmosphere ("sad"); immediately after that reading the song associated with happiness ("mixed"); and finally, after a long pause, with a happy induced atmosphere ("happy")[[1]](#footnote-1).

Psychological approaches to music listening, in general, and the induction of real-emotions, in particular, have been the subject of a wide range of research [‎1] [‎2] [‎3]. We followed [‎4] [‎5] guidelines for emotion induction with music by picking songs that elicit emotional contagion, sad or happy memory and empathy. We also attached photos to the songs' lyrics for visual imagery.

To induce sadness, we used a song that was written by a family member of a soldier that was killed in the 1997 Israeli helicopter disaster accident. This is a very well-known sad accident that people remember and can relate to. The song is sung by the family members. It describes the feeling of loss and emptiness, and thus we assumed that listening to the song will probably induce sadness and raise empathy for the family's mourning. Photos of the memorial at the crash site were used to conjure up visual images while listening to the music, or reading the lyrics.

To induce happiness, we used a well-known song. All participants reported that they knew the song (read it to their children or their parents read it to them when they were children). They also reported that the song was associated with 'happy' memories. We assumed that the song's lyrics, which describe a chorus of birds, along with the music and funny photos of birds can assist in inducing a happy atmosphere.

Although reading text is different than spontaneous (real-life) speech, we assumed that the combination of strongly emotional charged content with keeping the participants busy by the reading task should eliminate the recorder presence effect, and elicit true emotions.

## Participants

The experiment included 12 unprofessional speakers (6 females and 6 males). All of the participants are adults at ages of 21 to 60.

To ensure the participants' privacy, an ID was given to each participant.

All participants are Hebrew speakers with Hebrew as the first language.

Before the recordings each participant gave his or her consent, by signing an agreement, to store and to publish the recordings on a media and utilize it anonymously for research purposes. After each recording session, each participant was asked to fill in a short questionnaire describing his or her own recording's experience and feelings.

## Environment & Recording Equipment

The recordings were taken place in a quiet studio with no disturbances. Some very low background noises were present, yet not filtered, e.g., from the lighting or the laptop that was used to show the songs' lyrics (laptop operation was inaudible).

Recordings were done using Zoom H4n Handy recorder, equipped with high quality Stereo condenser microphones arranged in an XY pickup pattern, a built-in speaker and mounted SD card.

Most of the energy of speech signal resides in the frequency range [50-6000] Hz; therefore most speech corpora [‎18] [‎‎20] [‎24] [‎29] [‎30] [‎36] have sampling rate of at least 16 kHz. We used the Zoom H4n recorder's on-board microphones, in STEREO mode, with high sampling rate of 96 kHz and 24 bits audio bit depth. We used much higher sample rate than necessary, based on our belief that there are hints to emotion detection at high frequency band. This effort might be fruitful for future studies. The audio signals were saved on the SD-memory card mounted on the recorder as PCM (wav) uncompressed files. The recorder was placed on a table, in a distance of about 70 cm from the speaker and was directed at the speaker to have a relatively high level of speech compared to the background noises.

## Recording Process

The participants were sitting in front of a laptop computer. We made sure, that the participant felt comfortable, the lighting was sufficient and that the font size was big enough.

Like in [‎‎20], the participants knew they were being recorded, but to ensure authenticity they were not informed that the purpose of the experiment was to capture emotional reactions. We tried to reduce the Hawthorne effect [‎‎42], i.e., the phenomena where individuals change their behavior due to the attention they are receiving from researchers rather than because of any manipulation of independent variables. This was done in two ways: a) the participants were instructed to look at the computer screen. b) We ignored the participant during the recording (by avoiding eye contact or pretending to do paper work).

We recorded each session as a whole, without brakes, including the parts in which the participant did not speak at all, such as: the songs' playing and the part of feeling in the questioners. In this way we intended to draw less attention to the recording equipment (making the participants feel more comfortable).

There were four recording phases:

**Neutral phase** – the participants counted from 1 to 10. This recording was used as a reference.

**Sad phase** - Each participant was shown photos of the memorial of the 'helicopter crush' event. After telling the participants about the author of the song (a brother of one of the soldiers that was killed in the accident) and the singers of the song (the deceased's family), the song was played. During the 5:42 minutes of song's playing, the participants sat in front of the laptop, read the song's lyrics and watched the photos of the memorial at the crush site (The slide that was screened is shown in appendix 2).

In order to achieve more natural / spontaneous speech, the participants were instructed to read out loud the song as they find suitable, in terms of speed and intonation. We also told them they can avoid reading the repeating chorus (to avoid tedious or monotonic reading).

**Mixed Phase** - Immediately after the sad phase, the participants were asked to read a well-known happy song. As before, the participants were instructed to read the song as they pleased, with or without the repeating chorus (The slide that was screened is shown in appendix 3).

**Happy Phase** - After at least one hour break the participants were called to repeat the experiment of the happy song. Unlike the mixed phase, in the happy phase we used the manipulations to induce happiness, i.e., we played the 2:46 minutes long happy song before the reading, and we showed the birds' photos on the screen while reading the lyrics. Then, the participants were recorded reading the song (screening the same slide as in the Mixed phase).

In the following table we summarize the flow of the recording sessions.

Table : The flow of the recording sessions

|  |
| --- |
| **Before the break**   1. Explanation + signing the agreement. 2. Playing the sad song while the participant is silently reading the lyrics and watching the photos of the memorial for the fallen. 3. Recording the participant reading the sad song. 4. Recording the participant reading a well-known happy song, with no explanation or break. 5. The participant fills in a short questioner about his/her recording experience and feelings.   **After the break**   1. Playing the happy song while the participant is silently reading it and watching the birds' photos. 2. Recording the participant reading a well-known happy song. 3. The participant fills in a short questioner about his/her recording experience and feelings. |

# Speech Corpus

Since we recorded the whole session continuously, the raw recording materials contained a lot of surplus data (e.g. conversations with the participants during the recording sessions).

Speech utterances for the current research contain only the songs reading and the 1-10 counting. All other parts of the recordings were deleted, to protect the participants' privacy.

## Processing and Segmentation

We used the Wavosaur 1.0.8.0 software for editing sounds [‎7], to split the two long wav files (before and after the break) into 4 shorter wav files, a file per each phase, maintaining the original audio format (i.e. Stereo, 96 kHz sampling frequency and 24 bit resolution).

The 4 files were then converted into monophonic (R-channel) Windows PCM format with 96 kHz sampling frequency and 24 bit resolution, by Wavosaur. No other process, i.e., DE-noising, down sample, filtering, etc. was performed on the files.

Each participant read the song differently, in terms of speed of reading, intonation, duration of pauses (silence) between words and sentences, clarity of reading, merging words, and so on. In order to have a common ground for the different participants, we divided each file into segments (utterances) that usually ended by a pause. We then numbered the segments sequentially in each phase's file. The content of two utterances in each phase's file with the same serial number is identical (and also in the *mix* and *happy* phases). Pauses at the beginning and at the end of the sentences were not included in the corpus.

A segment in the *neutral* counting phase was chosen to be one number. The number of utterances extracted from each file is therefore: 10 utterances from the *neutral* file, and 37 or 38 utterances in the *happy* and the *sad* files respectively. These segmented files were the input for feature extraction using MATLAB [‎8] and Opensmile [‎10].

## Annotation

Utterances were manually assigned one of the following emotion labels: *sad*, *happy*, *mixed* (blend of *sadness* and *happiness*) and *neutral*. These Labels are task-dependent; they were assigned to suit the song content and the specific recording's phase (one of the four phases).

## Corpus Description

The corpus contains the following 5 tables (saved in MS-Excel spreadsheet format):

***Songs (****song-id*, *song-name*, *song-poet****)*** – This table describes the songs used for emotion elicitation, read by the participants.

***speech\_content*** (*song-id*, *segment-number*, *segment-lyric,* *segment-Hebrew-content*, segment-phonetic-transcript) – This table contains the segmentation of each of the songs listed in *songs* table.

***speaker\_info*** (speaker-id, gender, additional-information) – This table contains participants' information like: speaker-id, gender and additional information supplied by the speaker that can influence the speech features (e.g. tiredness, sour throat, vocal training, etc.)

***Speech\_Corpus*** (File-name, speaker-id, Mono/Stereo, utterance-annotation, song-id, section-id, recording-place) – This table contains description of each utterance in the corpus: File name, speaker-id (as given in the *speaker\_info* table), method of recording, i.e., mono/stereo, utterance's annotation, song-id (as listed in *songs* table), section-id (utterance content as describe in table *speech\_content*) and the recording place.

***extracted\_features*** – This table contains a small set of acoustic features, extracted for each utterance. The features' list (referred to as the **MATLAB-set**) and method of extraction is described in section 4 of this paper.

The corpus contains a total of 1553 utterances in .wav format:

* 10 neutral utterances for each speaker for a total of 120 *neutral* utterances.
* 1337 emotional utterances almost equally distributed between the various emotional states and speaker gender: about 37 utterances, in each of the *sad*, *happy*, and *mixed* phases, for each of the 12 speakers. One of the female's sad utterances was deleted due to mumbling, 6 extra utterances were picked for some of the males' happy song readings (not all speakers chose to read the repeating chorus).
* 96 full recordings of each of the 4 phases, in mono and stereo modes for each of the 12 participants (that is 4 x 2 x 12 additional utterances), were included for the benefit of future research.

Table 2 shows the distribution of the utterances by gender and label.

Table : The distribution of the utterances by gender and label

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **sad** | **happy** | **mixed** | **neutral** | **Total** |
| **Females** | 221 | 222 | 222 | 60 | 725 |
| **Males** | 222 | 226 | 224 | 60 | 732 |
| **All Speakers** | 443 | 448 | 446 | 120 | 1457 |

In Table 3 we summarize the corpus specifications.

Table : Summary of the corpus specifications

|  |  |
| --- | --- |
| Database | **Do we accumulate our emotions? – Semi-Natural complex emotions corpus** |
| Language | Hebrew |
| Emotions | sadness, happiness, neutral, mixed (blend of sadness & happiness) |
| Subjects | 12 unprofessional speakers: 6 females, 6 males |
| .corpus | Total of 1553 utterances in various lengths: 10 neutral utterances (each speaker counts from 1 to 10); 38 sad utterances, 37 happy utterances, and 37 mixed utterances – produced by reading two songs in the Hebrew language; and 96 mono and stereo utterances of the un-segmented files. |
| Recording info | **Environment**: the Open University recording studio.  **Equipment**: Zoom H4n Handy recorder using the on-board microphones.  **Sampling**: Sampling frequency of 96 kHz with 24 bits per sample.  **Files format**: .wav files. |

# Acoustic Features extraction

It is common knowledge that human voice contains clues to its emotional state [‎15]. In this work we assume that the speaker's voice carries information referring its emotional state. The human voice is not stationary. Moreover, a major part of the emotional clues are given by its dynamics, i.e., volume changes, frequency changes etc. Therefore it is common to divide the voice signal into smaller segments (usually 10-30sec [‎18] [‎19] [‎21] [‎24]). For each segment a set of parameters (features) is computed. The features cover several aspect of speech producing. For example: the variance represents the speech power which is controlled by the lung muscles and the larynx. The fundamental frequency controlled by the vocal cords and the spectral (MFCC) mostly represents the vocal tract. Above the fundamental set of voice features a second and third layer of features are constructed. For example: a series of statistics and derivatives that computes for the difference series: median, minimum and maximum, average fundamental frequency, the difference between MFCC coefficients etc. This set is arranged in a vector structure to give the feature vector. Each feature yields a sequence in time. The number of elements in this sequence is given by dividing the utterance duration and the feature computing segment duration or the hoping step size. A statistical quantity such as mean, standard deviation, etc. of the sequence elements yields a single number representing the feature value during the utterance time duration. These numbers are unified into a single high dimension array which we name the features vector.

The process that transforms the audio segment into features vector is called features extraction. Automatic tools for feature extraction (like openSMILE and PRAAT) contribute to the use of large sets of features. However, using a large feature vectors leads to low performance, a phenomena known as the curse of dimensionality, results in considerable increase in computation time and in some cases decrease in recognition rate [‎26] [‎15] [‎30].

Over the years the question of which and how many features to use occupied many of the researches, since large number of feature doesn’t guarantee better classification results [‎16] [‎26]. In [‎17] for example, one site, using ‘only’ 32 features, produced a classification performance in the same range as other sites, using more than 1000 features. In [‎26], 86 features obtained an accuracy of 84.79%, while using only 6 features improved the performance up to 86.71%. A common strategy to answer this question is to start with a large group of features, using as many features as possible, then reducing the number of features with feature selection algorithm (this strategy does not always converge to the best solution [‎17]). The most famous examples are [‎15] and [‎17], collecting 4024 features extracted in various sites. [‎16] started with almost 1300 features and [‎30] extracted 3809 features, both using WEKA with Correlation-based Feature Selection (CFS) to find the best subsets of features. [‎18] started with 988 features, while [‎28] concluded that the extremely reduced set of acoustic features obtained by the greedy forward search selection algorithm improved the results provided by the full 384 features set.

In the widespread literature of emotion detection by voice, it is common knowledge to classify the features into three types:

* Lexical features (like bag of words) - refer to the content of the speech: of textual elements and enunciation parameters like vocabulary, sighs and pauses.
* Prosodic features refer to the features: fundamental frequency, duration and loudness [‎14] [‎23] [‎24] [‎38].
* Segmental features: MFCC [‎23] [‎24].

Although there is still no common agreement on a top list of features, researchers found that feature type is also relevant for automatic classification of emotions; and combining different feature types helps to improve emotion recognition [‎22] [‎24] [‎31]. In [‎15], [‎18] and [‎28] the feature vector might be expanded by combining lexical (linguistic) features like Bag of Words (words that imply the speaker's emotions) that improves the emotion recognition.

In our work we chose to concentrate on voice features, and followed the advice of many researches [‎23] [‎24] [‎31], that stated the importance of combining prosodic (F0, energy, speech duration, etc.) features with segmental features (MFCC) for better emotion recognition. Chapter 4.1 describes the acoustic parameters we extracted. Chapter 4.2 describes the parameters' extraction and computing process.

## Features' Contribution for Emotion Recognition

In the field of emotion recognition from speech, we aim to identify the emotion of the speaker from his voice, using various acoustic features. We extracted three sets of acoustic features for emotion recognition: Two feature sets were automatically extracted using openSMILE. The third feature set was computed using MATLAB, concentrating on a small set of the most important features for emotion recognition.

In this section we describe the contribution of the third set.

**F0**, the fundamental frequency, is defined as the lowest frequency of a periodic waveform. In terms of a superposition of sinusoids (e.g. Fourier series), usually, the fundamental frequency is the lowest frequency sinusoidal in the sum [‎43]. Others define the F0 as the lowest resonant frequency [‎44], In this case, it is the largest common factor of the signal harmonics. For example: assume that an audio signal contains sinusoids at frequencies: 100, 150, 200, 400 Hz. Its F0 is 50 Hz. F0, for voiced speech, is the vibration rate of the vocal folds, which is a function of its dimension and its muscle tension.

Emotion affects the nerves system, which control the muscles and tension in the vocal cords. Since F0 measure some tension in the vocal cord muscles (As the tension goes higher the F0 goes higher), it is one of the important properties of speech that is affected by emotional modulation. F0 per se does not carry information referring emotion, but its progression might sunshades some clues. Therefore, normally statistics such as mean value of F0, variance, variation range and the contour are meaningful for emotion detection [‎31]. According to [‎14] the F0 mean and F0 range account for the most important variations observed between emotion categories.

**MFCC** (Mel Frequency Cepstral Coefficients) is widely used for phoneme detection. The sounds generated by the vocal cord are filtered and determined by the shape of the vocal tract, nose and lips. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum. The Mel scale is used to group energy originated in a different frequency bands while the logarithm is used to separate the vocal cord vibration and the vocal acoustic system. Thus, the vocal tract, nose and lips are represented by Mel Frequency Cepstral Coefficients (MFCCs) [‎25]. Since emotion affects the nerves and muscles system, which in turn affect the vocal tract system, MFCC is also used by emotion detection algorithms (as it measures the acoustic features of the vocal tract).

Many studies have shown the importance of **MFCC** features to emotion recognition [‎24] [‎16]. These studies show that combining MFCCs features improves the recognition rate [‎31].

**Power** or intensity (commonly known as energy) is the mean square amplitude. As in the F0 case, the power values per se are not meaningful for emotions, but its statistics, extreme values and time behavior. Since different emotions differ from each other in the level of intensity (arousal), intensity related functions are important features for emotion recognition [‎16].

Another important factor that should be taken in consideration is the speaker **gender**. Researchers found that gender difference in acoustic features also influence emotion classification, and when gender information is included, the classifier achieves better recognition rates [‎27] [‎41].

## Feature Extraction

In our work, we extracted three sets of acoustic features for emotion recognition:

* **The MATLAB set -** a small set of 49 commonly used acoustic features was extracted using MATLAB [‎8] with the VOICEBOX [‎9] toolbox. For this set we chose the following combination of prosodic and segmental features:

Prosodic features:

* Duration – the length of the utterance (in samples). Duration-Tempo: The samples number implies the speech rate as the number of words was conserved. Therefore the speech rate is approximated as the duration reciprocal.
* Average power/intensity – Using ITU-T P.56 recommendation the power of the speech is computed [‎45]. The active speech level measures the average power - square magnitude during speech. The outcome is also represented in decibel scale (dB).
* Derivatives of the fundamental frequency (F0) – mean, max, range, standard deviation, variance, skewness, kurtosis.

Segmental (Cepstral) features:

* + 12 MFCCs coefficients and their first and second derivatives and their statistics – min, max, mean.

We used the YIN [‎46] algorithm for **F0** estimation, as it is a simple and robust method for F0 estimation that produces fewer errors than other well-known methods. We tuned the F0 detection algorithm to use 60msec window with 5msec hop step. This window length was chosen to contain several periods of a masculine voice to gain better approximation accuracy of the YIN algorithm. We used 12 bands of the **MFCC** transform with 20 mSec segment length and 10 mSec step. We calculated the mean, minimum and maximum for each of the 12 coefficients over the entire utterance. As we mentioned before, Single F0 and intensity values are not meaningful for emotions detection, but rather their behavior over time. In our voice analysis, each 2-5 words utterance was segmented into small segments. For each segment a set of parameters and statistics was computed. This set is arranged in a single vector structure (with the 49 parameters mentioned above), to give the features vector, which represents the changes over time.

Table 13, Appendix 1 describes how the features in the MATLAB set were calculated.

As a reference to the 49 calculated acoustic features, i.e. appendix 1. We used two other feature sets of openSMILE (the Munich open Speech and Music Interpretation by Large Space Extraction toolkit), version 1.0.1 [‎10].

Since openSMILE is used by the openEAR project [‎11] for emotion recognition, various standard feature sets for emotion recognition are available. It is closable as openSMILE configuration files. Limited by the corpus size (1337 utterances), we stayed close to the 1:10 rule of thumb ratio between number of utterances in the learning set and number of features [‎26] by choosing the following 2 sets [‎12]:

* **The INTERSPEECH 2009 Emotion Challenge feature set,** (config/emo IS09.conf) - 384 features, based on the following 16 low-level descriptors (LLD):
  + pcm\_RMSenergy - Root-mean-square signal frame energy
  + mfcc - Mel-Frequency cepstral coefficients 1-12
  + pcm\_zcr - Zero-crossing rate of time signal (frame-based)
  + voiceProb - The voicing probability computed from the ACF.
  + F0 - The fundamental frequency computed from the Cepstrum.
* **The openSMILE/openEAR `emobase' set**, (config/emobase.conf) - 988 acoustic features for emotion recognition, based on the following 26 low-level descriptors (LLD):
  + Intensity
  + Loudness
  + 12 MFCC
  + Pitch (F0)
  + Probability of voicing
  + F0 envelope
  + 8 LSF (Line Spectral Frequencies)
  + Zero-Crossing Rate

For all 3 sets, 2 additional columns/features were added manually during the feature extraction: speaker gender label (male/female) and annotation (for classification purpose).

# Classification

For classification, we used WEKA (Waikato Environment for Knowledge Analysis) software [‎13], with the Support Vector Machine (SVM) and Sequential Minimal Optimization (SMO) algorithm.

Many researches used SMO algorithm for classification, showing that under the conditions of limited training data, the best performance is achieved using a Support Vector Machine (SVM) trained with the (SMO) [‎30] [‎32] compared to other algorithms, such as J48, RandomForest, NaiveBayes, SimpleLogistic, etc. . In [‎35], the most widely used classification techniques for SER: k-nearest-neighbour methods (k-NN), C4.5 decision trees, support vector machines (SVMs), artificial neural networks (ANNs), i.e., multilayered perceptrons (MLPs), and Naıve Bayes (NB) classifiers, were compared, on natural and an elicited corpora, labeled with discrete emotions or emotional dimension. SVM outperformed the other techniques in terms of classification accuracy.

We chose the classification algorithm as follows: We defined a test set that contains only two files: females happy and females sad; we then ran four different classification algorithms: SMO, SimpleLogistic (build linear logistic regression models), J48 (C4.5 decision tree learner) and RandomForest (construct random forests), trying to differentiate (by WEKA) between samples from these two files. We used the default settings for all the algorithms. In Table 4 we summarize the results of the *Correct Classification Rates* (CCRs) and the kappa statistic values for these four algorithms. The Kappa statistic measures the agreement between predicted and observed classes, on a scale of 0 to 1. Kappa of 1 indicates perfect agreement, whereas kappa of 0 indicates agreement equivalent to chance [‎54] [‎55]. SMO provided the best or near-best results in the shortest time (compared to SimpleLogistic). We repeated the process for the males' happy and sad samples and received similar results (in Table 5 ), therefor we used this algorithm for the whole corpus for all the tests.

Table : Classification algorithms comparison for the Females sad-happy utterances

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (51)** | | **emo\_IS09 (385)** | | **Emobase (989)** | |  |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **females sad-happy** |
| 83.07 | 0.6613 | 72.4605 | 0.4494 | 73.8149 | 0.4765 | RandomForest |
| 80.5869 | 0.6117 | 73.1377 | 0.4627 | 76.298 | 0.5259 | J48 |
| 83.9729 | 0.6795 | 83.2957 | 0.6659 | 82.8442 | 0.657 | SimpleLogistic |
| 83.07 | 0.6614 | 82.3928 | 0.6479 | 87.1332 | 0.7427 | SMO |

Table : Classification algorithms comparison for the Males sad-happy utterances

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (51)** | | **emo\_IS09 (385)** | | **Emobase (989)** | |  |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **Males sad-happy** |
| 77.4554 | 0.5487 | 68.75 | 0.3757 | 68.5268 | 0.3715 | RandomForest |
| 71.875 | 0.4374 | 67.8571 | 0.3577 | 66.5179 | 0.3301 | J48 |
| 79.9107 | 0.5983 | 77.2321 | 0.5447 | 76.1161 | 0.5222 | SimpleLogistic |
| 78.3482 | 0.5673 | 70.3125 | 0.4065 | 81.0268 | 0.6205 | SMO |

To better understand the newly defined mixed emotion, we conducted 4 types of tests:

* **Type 1**- classifies all pairs of emotions (happy, sad and mixed), that is, the following three tests: happy versus mixed (notice that happy and mixed used the same song, same wording); happy versus sad; and sad versus mixed. The results of these tests are shown in Table 8 . The number of utterances for these tests is listed in Table 2.
* **Type 2** – Classifies the three emotions. The results of these tests are shown in Table 9. The number of utterances for these tests is listed in Table 2.
* **Type 3** – Classifies each emotion compared with the other two. That is, the following three tests: happy versus not-happy (sad or mixed); sad versus not-sad (happy or mixed); and mixed versus not-mixed (sad or happy). The results of these tests are shown in Table 10. The number of utterances for these tests is listed in Table 6.
* **Type 4** – Classifies neutral speaking versus mixed emotion. The results of these tests are shown in Table 11.

We ran each of the above four tests three times, once per each set of parameters.

In addition we ran the above four tests three times, once per each set of participants (female, male and all the participants together).

It is important to notice that SVM was originally designed for binary classification. To improve classification results for three classes (three emotions), we took the "one-against-all" approach in test type 3. We balanced the number of utterances of the tested emotion versus the other two emotions by choosing the first half of utterances from each of the emotions that were not tested. For example, when testing sad versus not-sad, the not-sad set was created by the first half of utterances from the happy and from the mixed files. Table 6 describes the number of utterances for each class.

Table : The number of utterances for each set of participants by labels

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **sad / not-sad** | | | **happy / not-happy** | | | **mix / not-mix** | | |
| **sad** | **Not-sad** | **total** | **happy** | **Not-happy** | **total** | **mix** | **Not-mix** | **total** |
| **females** | 221 | 221 | 442 | 222 | 222 | 444 | 222 | 222 | 444 |
| **males** | 222 | 222 | 444 | 226 | 228 | 454 | 224 | 224 | 448 |
| **all Speakers** | 443 | 443 | 886 | 448 | 450 | 898 | 446 | 446 | 892 |

The results of these four types of tests shed light on the research question that compares the mixed emotion to all the other emotions: sadness happiness and neutral. To complete the comparison, we need to compare mixed versus neutral.

The 'neutral' utterances of each speaker are much shorter than the other utterances, since they contain words rather than sentences. Apart from the utterances' duration, there are also other differences that could bias the classification algorithm, like: number of utterances, silence between words, existence of certain consonants, etc.

To compare between 'neutral' and 'mixed', we needed to balance the biased elements. To achieve this balance we chose 10 words from the happy song (recorded in the 'mixed' phase), which resembled in length and sound having the same number of 'ssshh' and 'rrrr' consonants. Table 7 presents the words (and words parts) we phonetically matched from the happy song to the neutral words.

Table : Phonetic match in Neutral vs Mixed

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Neutral** | Achat  אחת | Shtaim  שתיים | Shalosh  שלוש | Arbah  ארבע | Chamesh  חמש | Shesh  שש | Sheva  שבע | Shmone  שמונה | Teshah  תשע | Eser  עשר |
| **Mixed** | Echad  אחד | Shilva  שילבה | Pashosh  פשוש | Amar  אמר | Rosh  ראש | Sheshama  ששמע | Shebah~~tzer~~  שבחצר | Simcha  שימחה | Sham  שם | Lashir  לשיר |

# Results

At first we performed the tests for discrete emotions per participant, just to get the 'feel' of things. In Table 14, Appendix 4 we summarize the results of these tests per participant, for each of the three sets of parameters. The *Correct Classification Rates* (CCRs), suggest that for each participant, the 3 emotions are distinguishable, e.g. speakers convey each emotion differently.

Table 8 shows the results for the Type 1 tests - i.e., discrete two-emotions classifications, with the best CCR highlighted. For mix vs. happy classification, the small set of parameters achieved the highest CCR rates, while for sad vs. mix and sad vs. happy classifications highest CCR were achieved for the largest (Emobase) parameters set. For each classification test, the females' utterances had the best classification results. As we expected, sad vs. happy, being opposite/contradicting emotions yielded the highest CCR.

Table : Two- discrete emotions classification results (Type 1 tests)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (48-51)** | | **emo\_IS09 (385-386)** | | **Emobase (989-990)** | | **Parameters set** |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** |  |
| 62.486 | 0.2498 | 58.1655 | 0.1634 | 62.1924 | 0.2439 | all mix-happy |
| 71.1712 | 0.4234 | 63.5135 | 0.2703 | 67.3423 | 0.3468 | females mix-happy |
| 67.3333 | 0.3466 | 60.6667 | 0.2132 | 62.2222 | 0.2443 | males mix-happy |
|  | | | | | | |
| 70.1912 | 0.4038 | 74.4657 | 0.4893 | 80.09 | 0.6018 | all sad-mix |
| 72.912 | 0.4582 | 76.9752 | 0.5395 | 80.5869 | 0.6117 | females sad-mix |
| 70.1794 | 0.4038 | 67.2646 | 0.3454 | 69.5067 | 0.3901 | males sad-mix |
|  | | | | | | |
| 75.4209 | 0.5086 | 75.6453 | 0.5129 | 79.1246 | 0.5824 | all sad-happy |
| 82.6185 | 0.6524 | 82.3928 | 0.6479 | 87.1332 | 0.7427 | females sad-happy |
| 78.7946 | 0.5761 | 70.3125 | 0.4065 | 81.0268 | 0.6205 | males sad-happy |

Table 9 shows the results for the Type 2 tests, i.e., 3 emotions classifications, with the best results highlighted. The females' utterances yielded the best CCR. It is interesting to see that for the males group, the best results were achieved using the small set of acoustic parameters.

Table : Three emotions classification results (Type 2 tests)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (48-51)** | | **emo\_IS09 (385-386)** | | **Emobase (989-990)** | |  |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** |
| 51.982 | 0.2798 | 55.1982 | 0.328 | 61.0322 | 0.4155 | all sad-mix-happy |
| 60.6015 | 0.409 | 61.5038 | 0.4226 | 64.812 | 0.4722 | females sad-mix-happy |
| 58.3333 | 0.375 | 54.7619 | 0.3216 | 56.3988 | 0.346 | males sad-mix-happy |

Table 10 sums up the results for classifications of Type 3 tests, i.e., one emotion versus the other two, with the best results highlighted. CCR in descending order are: 'sad', happy, and then 'mix'. Mix being the less significantly differentiated from the other two emotions can be explained by the fact that 'mix' might be affected by both 'happy' (happy song) and 'sad' (leftover feelings from the previous session). In most cases, it seems that the differences between one emotion to the other two are more significant for the female utterances. It is worth noting that in most cases the best results were achieved using the bigger set of features (990), although in two cases 48-51 features yielded better results. Comparing the results for emo\_IS09 and MATLAB sets shows that in half of the tests the smaller set, yielded better than the 385-386 features' set.

Table : Two-classes groups classification results (Type 3 tests).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (48-51)** | | **emo\_IS09 (385-386)** | | **Emobase (989-990)** | |  |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** |
| 71.4447 | 0.4289 | 80.2483 | 0.605 | 83.6343 | 0.6727 | all sad/no-sad |
| 74.4344 | 0.4887 | 83.9367 | 0.6787 | 87.3303 | 0.7466 | females sad/no-sad |
| 71.3964 | 0.4279 | 75.2252 | 0.5045 | 80.6306 | 0.6126 | males sad/no-sad |
|  | | | | | | |
| 67.5947 | 0.3518 | 64.922 | 0.2985 | 68.0401 | 0.3608 | all happy/no-happy |
| 72.2973 | 0.4459 | 72.2973 | 0.4459 | 72.973 | 0.4595 | females happy/no-happy |
| 69.6035 | 0.3921 | 63.6564 | 0.273 | 66.2996 | 0.326 | males happy/no-happy |
|  | | | | | | |
| 58.8565 | 0.1771 | 62.6682 | 0.2534 | 65.583 | 0.3117 | all mix/no-mix |
| 63.5135 | 0.2703 | 59.9099 | 0.1982 | 63.2883 | 0.2658 | females mix/no-mix |
| 59.8214 | 0.1964 | 58.7054 | 0.1741 | 62.2768 | 0.2455 | males mix/no-mix |

Table 11 describes the results of Type 4 tests, i.e., neutral vs. mixed classification. High CCR, for all acoustic features sets, suggests significant differences between utterances of neutral and utterances of mixed emotions.

Table : neutral vs. mixed classification results (Type 4 tests)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (48-51)** | | **emo\_IS09 (385-386)** | | **Emobase (989-990)** | |  |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **Kappa** |
| 76.6667 | 0.5333 | 81.6667 | 0.6333 | 84.5833 | 0.6917 | All |
| 74.1667 | 0.4833 | 81.6667 | 0.6333 | 85 | 0.7 | Females |
| 79.1667 | 0.5833 | 84.1667 | 0.6833 | 81.6667 | 0.6333 | Males |

In the following table, Table 12, we summarize the results of all the four types of tests.

Table : Classification results summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (48-51)** | | **emo\_IS09 (385-386)** | | **Emobase (989-990)** | |  |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** |  |
| 51.982 | 0.2798 | 55.1982 | 0.328 | 61.0322 | 0.4155 | All s-m-h |
| 62.486 | 0.2498 | 58.1655 | 0.1634 | 62.1924 | 0.2439 | All mix-happy |
| 70.1912 | 0.4038 | 74.4657 | 0.4893 | 80.09 | 0.6018 | All sad-mix |
| 75.4209 | 0.5086 | 75.6453 | 0.5129 | 79.1246 | 0.5824 | All sad-happy |
| 76.6667 | 0.5333 | 81.6667 | 0.6333 | 84.5833 | 0.6917 | All neutral mix |
|  | | | | | | |
| 60.6015 | 0.409 | 61.5038 | 0.4226 | 64.812 | 0.4722 | Females s-m-h |
| 71.1712 | 0.4234 | 63.5135 | 0.2703 | 67.3423 | 0.3468 | Females mix-happy |
| 72.912 | 0.4582 | 76.9752 | 0.5395 | 80.5869 | 0.6117 | Females sad-mix |
| 82.6185 | 0.6524 | 82.3928 | 0.6479 | 87.1332 | 0.7427 | Females sad-happy |
| 74.1667 | 0.4833 | 81.6667 | 0.6333 | 85 | 0.7 | Females neutral mix |
|  | | | | | | |
| 58.3333 | 0.375 | 54.7619 | 0.3216 | 56.3988 | 0.346 | Males s-m-h |
| 67.3333 | 0.3466 | 60.6667 | 0.2132 | 62.2222 | 0.2443 | Males mix-happy |
| 70.1794 | 0.4038 | 67.2646 | 0.3454 | 69.5067 | 0.3901 | Males sad-mix |
| 78.7946 | 0.5761 | 70.3125 | 0.4065 | 81.0268 | 0.6205 | Males sad-happy |
| 79.1667 | 0.5833 | 84.1667 | 0.6833 | 81.6667 | 0.6333 | Males neutral mix |

# Discussion and future work

This research focuses on rapid changes of opposite emotions and compares the resulted mixed emption with the two opposite emotions (sadness and happiness) as well as with the neutral state.

The corpus created is semi-natural as the participants were not actors and they were not pre-instructed to perform any emotion.

For all the classifications presented in the paper, we used 3 different sets of acoustic features that were extracted from the recording corpora. One of the sets of features ("MATLAB") was exclusively selected for this research, using MATLAB and published algorithms. We performed four types of tests using SVM classifier trained with the SMO algorithm.

Table 8 summarizes tests of type 1 – classification of all pairs of emotions (happy versus sad; happy versus mixed and sad versus mixed). The results show that the mixed emotion is differentiable from each of the two basic emotions from which it is combined of.

It is interesting to note that there is a big difference in the CCR of males versus females classification of sadness versus mixed emotions. It seems like females have more leftover feelings from the previous induced emotion than males. Another phenomenon worth noting is that CCRs in descending order are: most differentiate emotions are sadness versus happiness (as we could expect), then sadness versus mixed and finally (less differentiable) are happiness versus mixed. It is expected as happiness and mixed are based on the same song reading.

Table 9 summarizes tests of type 2 – classification of all three emotions (happy, sad and mixed). The results show low classification rates, due to limitations of classifying to three categories. We therefore continued to test type 3 in order to classify the given three emotions in three steps.

Table 10 summarizes tests of type 3 – classification of one emotion versus the other two. The results show that the CCRs in descending order are: most differentiate emotion is sadness versus non-sadness, then happiness versus non-happiness and finally (less differentiable) are mixed versus non-mixed. It is expected that sad is the most differentiable from non-sad as the songs read in the other two emotions-induction were the same. It is interesting to note that mixed was differentiated from non-mixed, but with the highest error rates. That is, mixed was more frequently misclassified as happiness or as sadness.

Table 11 summarizes tests of type 4 – classification of neutral state versus the mixed emotion. The results show that there is a high CCR differentiating neutral state versus mixed emotion for both males and females.

The conclusion from all the tests performed in this paper is therefore that people accumulate emotions: the emotion resulted from rapid emotion change is neither like the first emotion nor like the second one.

In most of the tests, the classification was more significant for the females than for the males. This interesting result requires additional study in future research.

In addition, based on Table 12, we conclude, as commonly suggested, that large number of features does not guarantee better results.

This is a proof of concept paper with a relatively small number of participants and utterances. In future work we intend to enrich the corpus to strengthen the conclusions.

In addition, we intend to alter the recording order, i.e., to record the happy induced emotion first. The motivation is to study the impact of the previous emotion. Does the impact depend on the sequential order? That is, is the impact of sadness on rapid change from sadness to happiness different from the impact of happiness on rapid change from happiness to sadness?

# References

1. G. Kreutz, U. Ott, D. Teichmann, P. Osawa and D. Vaitl, "Using music to induce emotions: Influences of musical preference and absorption," Psychology of Music, vol. 36, no. 1, pp. 101-126, 2008.
2. M. M. Marin and J. Bhattacharya, "Music Induced Emotions: Some Current Issues and Cross-Modal Comparisons," In Music Education, Hauppauge, NY: Nova Science, pp. 1-38, 2010.
3. M. Suda, K. Morimoto, A. Obata, H. Koizumi and A. Maki, "Emotional responses to music: towards scientific perspectives on music therapy," Neuroreport, vol. 19, no. 1, pp. 75-78, 2008.
4. P.N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," Behavioral and brain sciences 31, no. 05, pp. 559-575, 2008.
5. K.R. Scherer, "Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them?," Journal of new music research 33, no. 3, pp. 239-251, 2004.
6. J.A. Russell, "A circumplex model of affect." Journal of personality and social psychology 39, no. 6, pp.1161, 1980.
7. <http://www.wavosaur.com>
8. <http://www.mathworks.com>
9. <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>
10. F. Eyben, M. Woellmer, B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," In Proc. ACM Multimedia (MM), pp. 1459-1462, Oct. 2010.
11. F. Eyben, M. Wöllmer and B. Schuller, "openEAR - Introducing the Munich open-source emotion and affect recognition toolkit." In Proc. of ACII 2009, volume I, pages 576–581. IEEE, 2009.
12. 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/>.
13. 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.
14. T. Bänziger and K.R. Scherer, "The role of intonation in emotional expressions," In: Speech *communication*, Vol. 46, Issue 3–4, pp. 252–267, Jul. 2005.
15. A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous and N. Amir, “Whodunnit- searching for the most important feature types signalling emotion-related user states in speech,” Computer Speech & Language, vol. 25, no. 1, pp. 4–28, 2011.
16. T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," In IEEE Int. Conference on Multimedia and Expo, 2005 (ICME 2005), pp. 474-477, IEEE, 2005.
17. A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous and V. Aharonson, “Combining efforts for improving automatic classification of emotional user states,” Proc. IS-LTC, pp. 240–245, 2006.
18. S.G. Karadogan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," In IEEE 3rd Int. Workshop on Cognitive Information Processing (CIP), pp. 1-6, 2012.
19. Z. Xiao, E. Dellandrea, W. Dou and L. Chen, "A dimensional emotion model driven multi-stage classification of emotional speech," Res. Rep. RR-LIRIS-2007-033. 2007.
20. T. Kostoulas, T. Ganchev, I. Mporas, N. Fakotakis, "A Real-World Emotional Speech Corpus for Modern Greek," In: LREC 2008, Morocco, pp. 2676-2680, May 2008.
21. N. Amir, S. Ron and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, pp. 29–33, Sep. 2000.
22. F. N. Julia and K. M. Iftekharuddin, "Detection of Emotional Expressions in Speech," In Proc. of IEEE SoutheastCon, pp. 307-312, IEEE, 2006.
23. R. Barra, J. M. Montero, J. Macías-Guarasa, L. F. D’Haro, R. San-Segundo and R. Córdoba, "Prosodic and segmental rubrics in emotion identification," In: ICASSP 2006, pp. 1085-1088, 2006.
24. S. Nobuo and O. Yasunari, "Emotion Recognition using Mel-Frequency Cepstral Coefficients," In: Information and Media Technologies [Vol. 2](https://www.jstage.jst.go.jp/AF06S010SryTopHyj?sryCd=imt&noVol=2&noIssue=), pp. 835-848, 2007.
25. <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/> [Accessed Feb. 2015]
26. I. Luengo, E. Navas, I. Hernáez and J. Sánchez, "Automatic Emotion Recognition using Prosodic Parameters," In: Interspeech 2005, pp. 493-496, 2005.
27. D. Ververidis, C. Kotropoulos "Automatic Speech Classification to Five Emotional States Based on Gender Information," In Proc. EUSIPCO, pp. 341-344, 2004.
28. S. Planet and I. Iriondo, "Comparative study on feature selection and fusion schemes for emotion recognition from speech," International Journal of Interactive Multimedia and Artificial Intelligence 1, no. 6, 2012.
29. B. Yu, H. Li and C. Fang, "Speech Emotion Recognition based on Optimized Support Vector Machine," Journal of Software, vol. 7, no. 12, pp. 2726-2733, Dec. 2012.
30. S. Casale, A. Russo, G. Scebba and S. Serrano, "Speech emotion classification using machine learning algorithms," In Proc. of IEEE Int. Conf. on Semantic Computing, pp. 158-165, IEEE, 2008.
31. Y. Pan, P. Shen and L. Shen, "Speech Emotion Recognition Using Support Vector Machine," International Journal of Smart Home, vol. 6, No 2, pp. 101-108, April, 2012.
32. N. Visser, "Recognizing Natural Emotions in Speech, Having Two Classes," 2011.
33. A. Delaborde, M. Tahon and L. Devillers, "Affective Links in a Child-Robot Interaction," In *The Workshop Programme*, p. 1, 2010.
34. L. Vidrascu and L. Devillers, "Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features," In Proc. Inter. workshop on Paralinguistic Speech between models and data, ParaLing. 2007.
35. A. Tarasov and S.J. Delany, "Benchmarking classification models for emotion recognition in natural speech: A multi-corporal study," In Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 841–846, IEEE, 2011.
36. B. Schuller, R. Zaccarelli, N. Rollet and L. Devillers, "CINEMO-A French Spoken Language Resource for Complex Emotions: Facts and Baselines," In LREC, 2010.
37. S. Buisine, S. Abrilian, R. Niewiadomski, J.C. Martin, L. Devillers and C. Pelachaud, "Perception of blended emotions: From video corpus to expressive agent," In Intelligent virtual agents, Springer Berlin Heidelberg, pp. 93-106, 2006.
38. C.M. Lee, S.S. Narayanan and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," In IEEE International Conference on Multimedia and Expo, 2002 (ICME 2002), vol. 1, pp. 737-740. IEEE, 2002.
39. L. Devillers, S. Abrilian and J.C. Martin, "Representing real-life emotions in audiovisual data with non basic emotional patterns and context features," In Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, pp. 519-526, 2005.
40. L. Vidrascu and L. Devillers, "Annotation and detection of blended emotions in real Human-Human dialogs recorded in a Call center," In IEEE International Conference on Multimedia and Expo, 2005 (ICME 2005), pp. 4-pp. IEEE, 2005.
41. Z. Xiao, E. Dellandrea, W. Dou and L. Chen, "Hierarchical classification of emotional speech," *IEEE Transactions on Multimedia*, p.37, 2007.
42. <http://en.wikipedia.org/wiki/Hawthorne_effect> [Accessed Dec. 2014]
43. <http://en.wikipedia.org/wiki/Fundamental_frequency> [Accessed Dec. 2014]

1. <http://hyperphysics.phy-astr.gsu.edu/hbase/waves/funhar.html> [Accessed Jan. 2015]
2. <https://www.itu.int/rec/T-REC-P.56-199303-S/en>
3. A. De Cheveign´e and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” The Journal of the Acoustical Society of America, vol. 111, p. 1917, 2002.
4. M.J. Landau, "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk," Vanderbilt Undergraduate Research Journal 4, 2008.
5. W. Wahlster, "Smartkom: Fusion and fission of speech, gestures, and facial expressions," In Proceedings of the 1st International Workshop on Man-Machine Symbiotic Systems, pp. 213-225, 2002.
6. K. Chen, G. Yue, F. Yu, Y. Shen and A. Zhu, "Research on speech emotion recognition system in e-learning," In Computational Science–ICCS 2007, Springer Berlin Heidelberg, pp. 555-558, 2007.
7. <http://www.lie-detection.com/>
8. [www.praat.org](http://www.praat.org/)
9. E. Grishina, "Multimodal Russian Corpus (MURCO): Studying Emotions," In: 7th Conference on Language Resources and Evaluation LREC’2010, Valletta, Malta, pp. 32-36, 2010.
10. B. Schuller, F. Eyben, S. Can and H. Feussner, "Speech in Minimal Invasive Surgery - Towards an Affective Language Resource of Real-life Medical Operations," In: 3rd ELRA International Workshop on EMOTION, Valletta, pp.5-9, 2010.
11. J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," Computational Linguistics, vol. 22, no. 2, pp. 249-254, 1996.
12. A. J. Viera and J.M. Garrett, “Understanding interobserver agreement: the kappa statistic,” Family Medicine, vol. 37, no. 5, pp. 360-363, 2005.

# Appendix 1: The selected Acoustic Parameters

Table : MATLAB Acoustic Features

|  |  |  |  |
| --- | --- | --- | --- |
| Reference for motivation | How it was calculated | Parameter name |  |
|  | End – Start +1 | Duration | 1 |
| [‎16], [‎20], [‎21], [‎22], [‎26], [‎29], [‎‎31] | Statistics (mean, minimum and maximum) for each of the 12 MFCCs.  Extracting the 12 12 MFCCs:  mfccMatrix = mfcc(12, lineData, fs);  Calculating the MFCC statistics:  MFCC[i]-max= max(mfccMatrix(:,i));  MFCC[i]-mean= mean(mfccMatrix(:,i));  MFCC was calculated using frame size of 20msec with 10msec overlap. | MFCC[i]-max MFCC[i]-min  MFCC[i]-mean  1≤i≤12 | 2-37 |
| [‎16], [‎26], [‎29], [‎‎31] | activlev\_power=activlev(lineData,fs);  activlev\_db=activlev(lineData,fs,'d') | activlev\_power  activlev\_db | 38  39 |
| [‎16], [‎22], [‎26], [‎‎‎31], [‎38] | F0 statistics:  F0Array=F0DetectionYinModel(lineData,fs);  F0min= min(F0Array);  F0max= max(F0Array);  F0mean= mean(F0Array);  F0range= range(F0Array);  F0std= std(F0Array); %Standard deviation  F0var= var(F0Array);  F0skew= skewness(F0Array);  F0kurtosis= kurtosis(F0Array);  quantileRes= quantile(F0Array,3); | F0min, F0max,  F0mean, F0range,  F0std (Standard deviation), F0var,  F0 skewness,  F0kurtosis  F0 quantiles:  quantile1,  F0median (q2),  quantile3. | 40-50 |
| [‎27], [‎41] | Was added manually according to the speaker's gender | gender | 51 |

# Appendix 2: Sad Song Lyrics' Slide



Figure : Sad Song Lyrics' Slide

# Appendix 3: Happy Song Lyrics' Slide

Figure : Happy Song Lyrics' Slide

# Appendix 4: Classification results for individual speakers.

Table : Classification results for individual speakers

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MATLAB (48-51)** | | **emo\_IS09 (385-386)** | | **Emobase (989-990)** | | **Parameter Set** |
| **CCR (%)** | **kappa** | **CCR (%)** | **kappa** | **CCR (%)** | **kappa** |  |
| 55.7522 | 0.3361 | 72.5664 | 0.5884 | 69.9115 | 0.5485 | 237 s-m-h |
| 75 | 0.5 | 88.1579 | 0.7632 | 96.0526 | 0.9211 | 237 mix-happy |
| 58.6667 | 0.1735 | 76 | 0.5196 | 64 | 0.2796 | 237 sad-mix |
| 88 | 0.7604 | 97.2973 | 0.9459 | 98.6486 | 0.973 | 237 sad-happy |
|  | | | | | | |
| 75.2212 | 0.6283 | 71.6814 | 0.5753 | 75.2212 | 0.6282 | 240 s-m-h |
| 92.1053 | 0.8421 | 85.5263 | 0.7105 | 86.8421 | 0.7368 | 240 mix-happy |
| 78.666 | 0.5733 | 66.6667 | 0.333 | 72 | 0.4393 | 240 sad-mix |
| 88 | 0.7599 | 92 | 0.8399 | 89.3333 | 0.7863 | 240 sad-happy |
|  | | | | | | |
| 54.7826 | 0.3208 | 53.913 | 0.3088 | 60.8696 | 0.4127 | 245 s-m-h |
| 70.5128 | 0.4103 | 56.4103 | 0.1282 | 58.9744 | 0.1795 | 245 mix-happy |
| 72.3684 | 0.4443 | 75 | 0.4993 | 75 | 0.5 | 245 sad-mix |
| 80.2632 | 0.6047 | 90.7895 | 0.8158 | 85.5263 | 0.7101 | 245 sad-happy |
|  | | | | | | |
| 68.4685 | 0.527 | 68.4685 | 0.527 | 73.8739 | 0.6081 | 247 s-m-h |
| 90.5405 | 0.8108 | 71.6216 | 0.4324 | 85.1351 | 0.7027 | 247 mix-happy |
| 77.027 | 0.5405 | 72.973 | 0.4595 | 78.3784 | 0.5676 | 247 sad-mix |
| 85.1351 | 0.7027 | 94.5946 | 0.8919 | 94.5946 | 0.8919 | 247 sad-happy |
|  | | | | | | |
| 76.1062 | 0.6416 | 85.8407 | 0.7875 | 84.9558 | 0.7744 | 252 s-m-h |
| 92.1053 | 0.8418 | 97.3684 | 0.9473 | 97.3684 | 0.9473 | 252 mix-happy |
| 72.973 | 0.4595 | 77.027 | 0.5405 | 77.027 | 0.5405 | 252 sad-mix |
| 94.7368 | 0.895 | 94.7368 | 0.8947 | 96.0526 | 0.9209 | 252 sad-happy |
|  | | | | | | |
| 77.5701 | 0.663 | 72.8972 | 0.5937 | 81.3084 | 0.7195 | 255 s-m-h |
| 95.7143 | 0.9143 | 98.5714 | 0.9714 | 94.2857 | 0.8857 | 255 mix-happy |
| 72.2222 | 0.4432 | 66.6667 | 0.3318 | 80.5556 | 0.6108 | 255 sad-mix |
| 87.5 | 0.7492 | 100 | 1 | 98.6111 | 0.9722 | 255 sad-happy |
|  | | | | | | |
| 68.4685 | 0.527 | 78.3784 | 0.6757 | 80.1802 | 0.7027 | 238 s-m-h |
| 72.973 | 0.4595 | 93.2432 | 0.8649 | 97.2973 | 0.9459 | 238 mix-happy |
| 81.0811 | 0.6216 | 81.0811 | 0.6216 | 78.3784 | 0.5676 | 238 sad-mix |
| 82.4324 | 0.6486 | 91.8919 | 0.8378 | 87.8378 | 0.7568 | 238 sad-happy |
|  | | | | | | |
| 71.1712 | 0.5676 | 82.8829 | 0.7432 | 84.6847 | 0.7703 | 241 s-m-h |
| 90.5405 | 0.8108 | 82.4324 | 0.6486 | 89.1892 | 0.7838 | 241 mix-happy |
| 81.0811 | 0.6216 | 86.4865 | 0.7297 | 79.7297 | 0.5946 | 241 sad-mix |
| 94.5946 | 0.8919 | 97.2973 | 0.9459 | 94.5946 | 0.8919 | 241 sad-happy |
|  | | | | | | |
| 91.8919 | 0.8784 | 92.7928 | 0.8919 | 96.3964 | 0.9459 | 242 s-m-h |
| 93.2432 | 0.8649 | 93.2432 | 0.8649 | 97.2973 | 0.9459 | 242 mix-happy |
| 97.2973 | 0.9459 | 100 | 1 | 100 | 1 | 242 sad-mix |
| 97.2973 | 0.9459 | 94.5946 | 0.8919 | 98.6486 | 0.973 | 242 sad-happy |
|  | | | | | | |
| 76.5766 | 0.6486 | 66.6667 | 0.5 | 68.4685 | 0.527 | 248 s-m-h |
| 83.7838 | 0.6757 | 60.8108 | 0.2162 | 75.6757 | 0.5135 | 248 mix-happy |
| 79.7297 | 0.5946 | 78.3784 | 0.5676 | 75.6757 | 0.5135 | 248 sad-mix |
| 91.8919 | 0.8378 | 98.6486 | 0.973 | 95.9459 | 0.9189 | 248 sad-happy |
|  | | | | | | |
| 63.0631 | 0.4459 | 61.2613 | 0.4189 | 62.1622 | 0.4324 | 258 s-m-h |
| 87.8378 | 0.7568 | 86.4865 | 0.7297 | 87.8378 | 0.7568 | 258 mix-happy |
| 64.8649 | 0.2973 | 55.4054 | 0.1081 | 52.7027 | 0.0541 | 258 sad-mix |
| 86.4865 | 0.7297 | 93.2432 | 0.8649 | 93.2432 | 0.8649 | 258 sad-happy |
|  | | | | | | |
| 70 | 0.55 | 70.9091 | 0.5635 | 74.5455 | 0.6178 | 260 s-m-h |
| 89.1892 | 0.7838 | 89.1892 | 0.7838 | 89.1892 | 0.7838 | 260 mix-happy |
| 68.4932 | 0.3695 | 71.2329 | 0.4243 | 75.3425 | 0.5056 | 260 sad-mix |
| 84.9315 | 0.6985 | 82.1918 | 0.6439 | 89.0411 | 0.7806 | 260 sad-happy |

# תקציר

זיהוי רגשות מתוך דגימות קול הוא תחום מחקרי מתפתח. מחקרים רבים חקרו זיהוי של רגשות בדידים. בעבודה זו התמקדנו בהשפעה של עירוב רגשות המצויים בצדדים מנוגדים של הצירים במימדים valence ו- arousal.

לשם כך ביצענו 4 שלבי הקלטה בסביבה מבוקרת: מצב נטראלי, מצב עצב, בו השרנו עצבות. מצב מעורב – בו מייד אחרי מצב עצב, השרינו שמחה. ומצב שמחה - אחרי הפסקה ארוכה, השרינו מצב שמחה.

נתחנו את דגימות הקול בעזרת הפקת מאפיינים אקוסטיים של הקול ומדידת רמת המובהקות ביכולת זיהוי הרגש באמצעים של למידה חישובית, באופן הבא:

הפקנו 3 סטים של מאפיינים אקוסטיים, תוך שימוש ב-openSMILE ו-MATLAB, כדי להשוות את המצב המעורב עם מצבים שמחה, עצב ונטראלי על מנת לבחון את השפעת הרגש המושרה הקודם על הרגש הנוכחי.

לסיווג/הבחנה בין הרגשות, השתמשנו בתוכנת WEKA, עם אלגוריתם למידה חישובית SVM + SMO.

התוצאות שקיבלנו מראות שמצב מעורב של שמחה ועצב, אינו זהה למצב נטראלי (וכן אינו זהה למצב שמחה, או מצב עצב), דבר המצביע על כך שאנו "צוברים" רגשות.

# תוכן עניינים

1. הקדמה. .........................................................................................4
   1. סקר ספרות..............................................................................4
   2. היקף המחקר ומטרתו.................................................................8
   3. תרומה עיקרית..........................................................................9
   4. סקירת העבודה.........................................................................9
2. הקלטות..........................................................................................9
   1. שיטת השראת הרגשות ............................................................9
   2. משתתפים...............................................................................10
   3. ציוד וסביבת הקלטה..................................................................11
   4. תהליך ההקלטה.......................................................................11
3. מאגר קבצי שמע..............................................................................12
   1. עיבוד וחיתוך הקבצים................................................................13
   2. תיוג........................................................................................13
   3. תאור מאגר קבצי השמע.............................................................13
4. הפקת מאפיינים אקוסטיים.................................................................15
   1. תרומת המאפיינים לזיהוי הרגש...................................................16
   2. הפקת מאפיינים........................................................................17
5. סיווג...............................................................................................18
6. תוצאות...........................................................................................21
7. דיון והצעות למחקר עתידי..................................................................23
8. רשימת מקורות................................................................................25
9. נספח 1: הפרמטרים האקוסטים שנבחרו..............................................28
10. נספח 2: שקופית מילות השיר העצוב...................................................29
11. נספח 3: שקופית מילות השיר השמח...................................................29
12. נספח 4: תוצאות סיווג עבור כל דובר....................................................30

**האוניברסיטה הפתוחה**

**המחלקה למתמטיקה ולמדעי המחשב**

**האם אנו סוכמים רגשות כשאנו מדברים?**

עבודת תזה זו הוגשה כחלק מהדרישות לקבלת תואר

"מוסמך למדעים" M.Sc. במדעי המחשב

באוניברסיטה הפתוחה

החטיבה למדעי המחשב

על-ידי

**מדמוני ליאת**

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

אפריל 2015

1. The experiments and methods were approved by the Ethical Committee of the Open University. [↑](#footnote-ref-1)