University of Tartu

Faculty of Science and Technology

Institute of Technology

Joosep Hook

Automatic Speech-based Emotion Recognition

Bachelor's thesis (12 ECTS)
Computer Engineering

Supervisors:

Prof. Gholamreza Anbarjafari
Fatemeh Noroozi, MSc

Tartu 2018

Resümee/Abstract

Automaatne kõnepõhine emotsioonituvastus

Afektiivse arvutiteaduse peamisteks eesmärkideks on inimese erinevate afektiivsete seisundite tuvastamist võimaldavate süsteemide uurimine ning väljatöötamine. Tänini pole leitud universaalseid tunnuseid, mille abil täpselt tuvastada inimese kõnes väljenduvaid emotsioone. Käesolevas töös luuakse kaks kõnepõhise emotsioonituvastuse süsteemi. Tugivektorklassifitseerijat kasutatakse esimeses süsteemis koos Surrey Audio-Visual Expressed Emotion, Berlin Database of Emotional Speech, Polish Emotional Speech database ning Serbian emotional speech database andmebaasidega. Keskmistatud emotsioonituvastusvõimed, vastavalt 80.21%, 88.6%, 75.42% ning 93.41%, saavutatakse 87 tunnust kasutades. Teine süsteem, kus kasutatakse klassifitseerijana otsustusmetsi, koosneb kahest mudelist, mis on treenitud esimese kahe andmebaasi modifitseeritud versioonide peal. Esimene mudel treenitakse vaid meessoost rääkijate, teine mudel nii mees- kui naissoost rääkijate salvestustega. Teise süsteemi peamine eesmärk on uurida pakutud tunnuste kasutamise võimalikkust päriselulises olukorras ning uurida kõneleja soo mõju emotsioonituvastusvõimele. Selle testimiseks lindistati kahe naissoost ning kahe meessoost inglise keelt teise keelena kõnelevate inimeste kõnet. Salvestusi kasutati sisendiks treenitud mudelitele. Keskmistades emotsioonituvastusvõime üle kõigi emotsioonide saavutati mõlemas mudelis 28% täpsus emotsiooni tuvastamisel, mis on parem kui juhuslik arvamine. Naissoost kõnelejate puhul oli süsteemi keskmine emotsioonituvastusvõime esimeses mudelis 19%, teises 29%. Meessoost kõnelejate puhul olid tulemused vastavalt 36% ja 28%. Saadud tulemused näitavad, kuidas muutub emotsioonituvastusvõime kummagi soo jaoks, kui treenimisetapil on kasutada rohkem või vähem kummastki soost kõnelejate salvestusi.

CERCS: P176 Tehisintellekt

Märksõnad: kõnepõhine emotsioonituvastus, masinõpe, tugivektorklassifitseerija, otsustusmets

Automatic Speech-based Emotion Recognition

The main objectives of affective computing is the study and creation of computer systems which can detect human affects. For speech-based emotion recognition, universal features offering the best performance for all languages have not yet been found. In this thesis, a speech-based emotion recognition system using a novel set of features is created. Support vector machines are used as classifiers in the offline system on Surrey Audio-Visual Expressed Emotion database, Berlin Database of Emotional Speech, Polish Emotional Speech database and Serbian emotional speech database. Average emotion recognition rates of 80.21%, 88.6%, 75.42% and 93.41% are achieved, respectively, with a total number of 87 features. The online system, which uses Random Forests as it's classifier, consists of two models trained on reduced versions of the first and second database, with the first model trained on only male samples and the second trained on both. The main purpose of the online system was to test the features' usability in real-life scenarios and to explore the effects of gender in speech-based emotion recognition. To test the online system, two female and two male non-native English speakers recorded emotionally spoken sentences and used these as inputs to the trained model. Averaging over all emotions and speakers per model, it is seen that the features offer better performance than random guessing, achieving 28% emotion recognition in both models. The average recognition rate for female speakers was 19% in the first and 29% in the second model. For male speakers, the rates were 36% and 28%, respectively. These results show how having more samples for training for a particular gender affects emotion recognition rates in a trained model.

CERCS: P176 Artificial intelligence

Keywords: support vector machines, random forests, speech emotion recognition, machine learning

Contents

List of Figures

List of Tables

Abbreviations, Constants, Definitions

ESD - Emotional Speech Database

SVM - Support Vector Machines

SER - Speech-based Emotion recognition

LLD - Low-Level Descriptors

UNPLP - Non-Uniform Perceptual Linear Predictive (features)

LPCC - Linear Predictive Cepstral Coefficients

MFCC - Mel-Frequency Cepstral Coefficients

RF - Random Forests

SAVEE - Surrey Audio-Visual Expressed Emotion database

EMO-DB - Berlin Database of Emotional Speech

PESD - Polish Emotional Speech Database

GEES - Serbian Emotional Speech Database

1 Introduction

1.1 Machine Learning

Machine learning can be broadly categorized as supervised and unsupervised learning. In supervised learning, a machine learning algorithm is provided with input data and the expected result, such as a class label or a number. The algorithm then iterates over the data, trying to make predictions, adjusting it's internal parameters along the way to better match the expected outcome (class label or target number).

Input data consists of features, which are measurements or categorical values associated with each observation, and a class label or a target value. A human being could be described by numerical (height, weight) or categorical (eye color, hair color) features.

If the features present in the input data are sufficient, then the algorithm would learn how to predict a class label or target number with great accuracy. But if the features cannot describe the relationship between input data and target output (predicting body mass index without height), then more features are needed. On the other hand, using too many features when only a handful are needed can harm performance by making the input data ”noisy”.

Feature selection is a process where a subset of the features, usually providing the best accuracy or smallest error rate, is chosen for training the machine learning classifier. Using fewer features can reduce training, testing and classification time. To further improve classification performance, the parameters associated with the specific classifier can be tweaked or a different algorithm can be used altogether.

Before training and testing can begin, the data needs to be partitioned to non-overlapping training and testing sets. The testing set will be used to evaluate the performance and generalization ability of the classifier. However, with this approach, the performance depends on which samples end up in the testing and training sets. One way to counter this is to use n-fold cross-validation, described in further detail in section 4.7.1. Achieving good accuracy on input data is desirable, but the classifier is more useful if you can predict the class label of new data not used for training. The ability to correctly make predictions on new unseen data is called generalization. n-fold cross-validation can be thought to train n different classifiers, testing each classifier with data not used in training, and averaging the result. This can be a better estimate of the performance of the classifier.

In this work, classifiers are used to predict the emotional state of the speaker. Let us look at an example of a classifier. A decision tree [16] classifier in machine learning is a tree-like structure with nodes containing tests on input data and leaves containing the final classification result. The features to test in the nodes of the tree are automatically chosen based on different metrics when the tree is grown, such as information gain or impurity [16], which results in better classification performance. Intuitively, they are a set of questions that can be asked about it's input data and the answers to these questions most accurately predict the correct class of the input data. An example decision tree is depicted in Fig. 1.1.

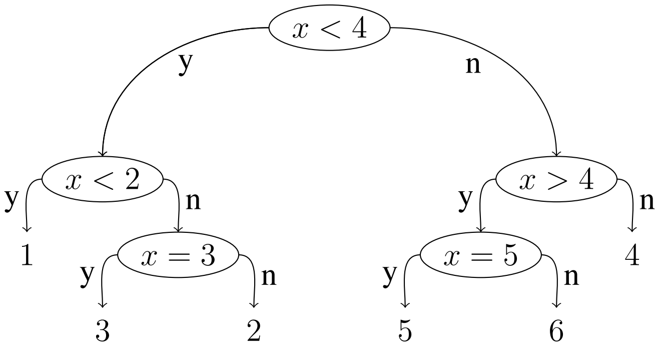


Figure 1.1: A decision tree for guessing a random number x chosen from an interval of $[1, 6]$. Each node represents a question, y - yes, n - no.

1.2 Speech-based Emotion Recognition

According to [1], Speech-based Emotion Recognition (SER) research started in the mid 1980s. During the early days, research focused on using statistical properties of certain acoustic features for emotion recognition [1]. After a decade of evolution in computer architectures, more complicated algorithms became feasible and iterative algorithms allowed for estimating acoustic features with greater accuracy [1]. Authors also said that the current efforts in research are focused on finding ways to improve classification in real-life applications. New applications are discovered rapidly due to the popularity of telecommunication services and multimedia devices [1].

SER systems can be divided, by the types of features used, to two categories: linguistic and paralinguistic. Linguistic features, as the name suggests, describe the linguistic content of human speech (what words are being said). Paralinguistic features illustrate how the speech is delivered (how are the words being said). With linguistic systems, the main challenges are automatic speech recognition and linguistic analysis of the speech. Speech recognition is not a necessary step when using paralinguistic features because speech is modeled as an audio signal, which permits using different techniques in digital signal processing. In addition, SER systems based on linguistic features can face challenges when people speaking in different languages are present [23].

As mentioned before, paralinguistic properties of speech, such as pitch and intensity, can be extracted using a multitude of signal processing techniques. The design goal of a paralinguistic SER system is to create a robust system capable of language-independent emotion recognition by finding a combination of features that can recognize different emotions with ease. Additional desirable properties of features are language-, gender- and speaker-independence, resistance to noise etc. A picture illustrating paralinguistic features is shown in Fig. 1.2.

There is little difference in the basic stages of researching and creating SER systems when machine learning methods are used. The first step consists of any preprocessing of the samples,

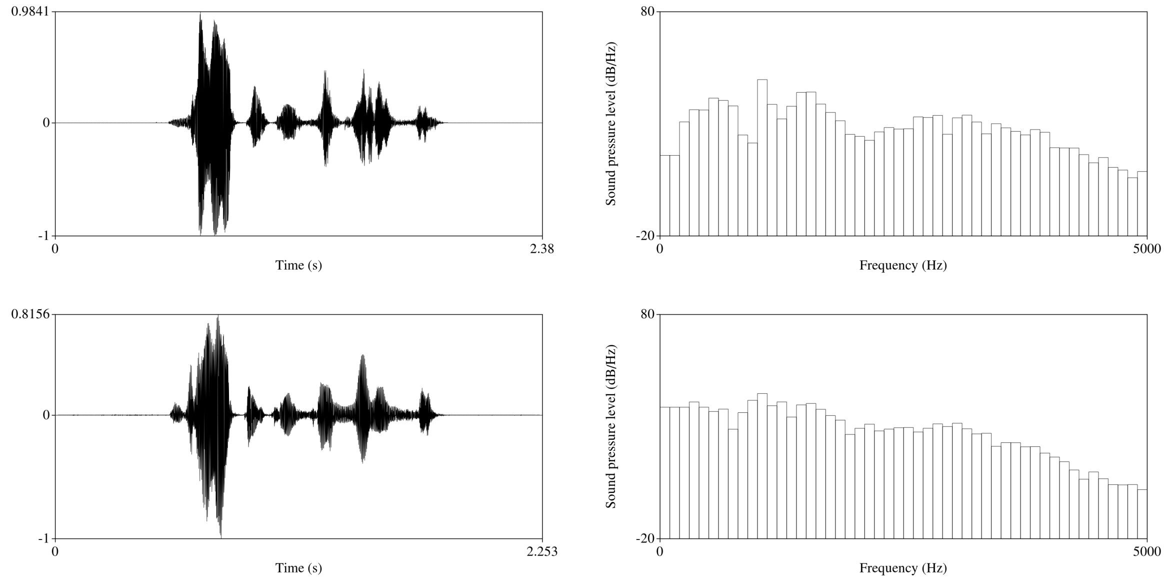


Figure 1.2: The same sentence said in anger (top) and when happy (bottom). Changes in loudness (left) and the long-term average spectrum (right) are shown. How to tell the difference?

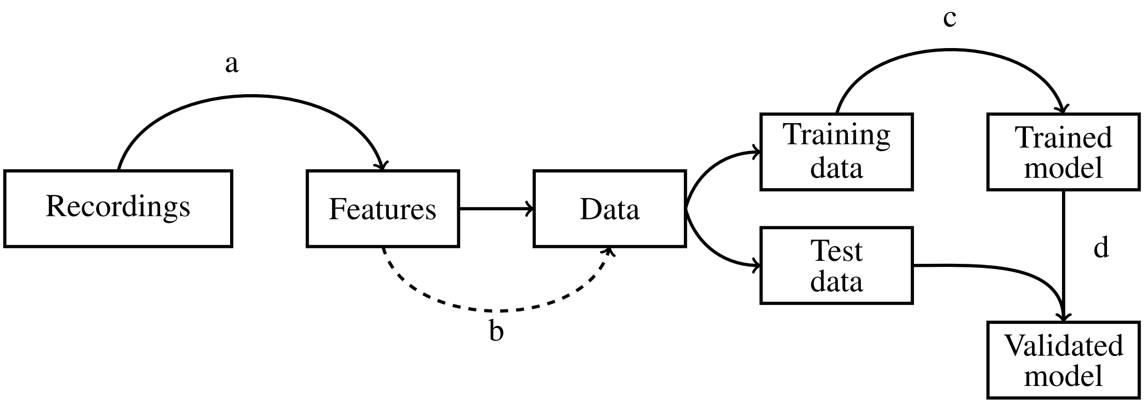


Figure 1.3: A high-level look at training a classifier for SER. a - feature extraction, b - feature selection, c - training a machine learning model, d - model validation.

such as removing noise. After preprocessing, the relevant features are usually extracted from either the whole speech signal or from each segment separately. The extracted data can be further transformed, such as replacing missing values with constants or interpolating them based on the distribution of the feature. When the data is prepared, it is split in either training and test sets or another scheme for estimating model performance, such as cross-validation, is used. The performance for average emotion recognition over the whole database or for a subset of emotions are most often reported and presented as results. This process is illustrated in Fig. 1.3.

SER has a broad range of real-life applications, such as in call centre environments [35,40,47]. It can be used to enhance driver safety in cars [24,32,53], enhance learning and motivation in online education environments [22,36,55], monitor health [27,56]. Finally, SER can be used as a subsystem in affect-aware video games [51].

The main objective of the thesis is finding features capable of good SER performance. In addition, the features must have good performance across many Emotional Speech Databases (ESD). Furthermore, testing the features will be done by using Support Vector Machines (SVM) as the classifier. The secondary objective is keeping the number of features used by the system at a minimum. Reducing computational complexity can save money, in addition to reducing training and testing time, if the system is running on cloud-based hosting services like Amazon. Finally, the power requirements needed to extract the features are lowered, making it better suited for embedded applications.

2 Related Works

The authors of [1, 48] have reviewed and summarized the history and developments of SER research. In their work, the authors explore and explain the differences between feature types, the usage of different classifiers, list and analyze existing emotional speech databases, the most common tools used in SER and more. Finally, they provide an overview of how focus in SER has changed over the course of time regarding the types of features, features and classifiers used.

In the beginning of SER research, the features most prominently focused on were intensity, duration and pitch [1, 48]. After a while, researchers started to focus Low-Level Descriptors (LLDs) describing voice quality like shimmer, spectral and cepstral measurements, jitter and harmonics-to-noise ratio [1, 48]. Following the aforementioned features, Non-Uniform Perceptual Linear Predictive (UNPLP) features, rhythm, Mel-Frequency Cepstral Coefficients (MFCCs), sentence duration and Linear Predictive Cepstral Coefficients (LPCCs) were focused on with greater intensity and interest [1,48].

The first people to study and use the Serbian emotional speech database in SER were Shaukat and Chen [49]. In their work, a strategy consisting of multiple stages, each using SVMs, was created for SER. The first stage of the multistage strategy was tasked with classifying the input signal as either passive or active. Passive emotions consisted of fear and non-fear. Non-fear contained both sad and neutral emotions. Emotions considered to be active were anger and sadness. Using a strategy similar to divide-and-conquer enabled the authors to achieve results similar to human listeners'.

The authors of [25] decided to use binary SVMs for emotion recognition. Hassan and Damper note that they were not the first authors to implement different structures consisting of binary decision trees [49]. In their work, SVMs were placed in a total of four different configurations: unbalanced decision tree, directed acyclic graph, one-versus-rest and one-versus-one. For each of the configurations, multiple ESDs were used for training, testing and comparing the models. Using this novel strategy enabled Hassan and Damper to reach state-of-the-art performance on two ESDs.

In [50], the existing multistage strategy created by Shaukat and Chen is improved upon by the authors themselves. The organization of the SVMs was changed in order to accommodate all of the basic emotions: anger, disgust, fear, happiness, sadness, surprise. This new approach, inspired mainly by psychology, can be adapted to any particular ESD that contains a subset of the aforementioned basic emotions. Enhancing the authors' previous strategy helps them to improve upon their earlier results.

Kobayashi and Calag made good use of different ensemble methods such as Random Forests

(RF) and kernel factories to improve SER accuracy in their work [33]. Instead of using suprasegmental features, the authors decided to use segmental features. This choice made the authors aware of the complexity regarding syllable boundary and word boundary identification. To combat this complicated matter, the authors decided to use a simple splitting strategy which involved splitting the samples at fixed relative positions. According to Kobayashi and Calag, the approach of splitting the sample is more suited towards stream analysis and real-time processing.

In [11], the authors have the goal of finding the smallest possible set of features which provide maximum accuracy in SER. There are multiple reasons for wanting to reduce the size of the feature vector used for classification. In addition to reducing the computational complexity of extracting and processing the features, all features might not increase the accuracy of the system. Chiou and Chen had 6552 features as their starting point which yielded them an average performance of 85.2%. When the baseline features were reduced down to a total of 37, the authors report just a 5% decline in emotion recognition performance.

Yüncü *et al.* created a system capable of mimicking the human auditory canal [59]. The motivation for creating such a system stems from the authors' knowledge that the human ear performs filtering which is dependent on the frequency of the sound. The samples in ESDs were first input to the created system which then performed frequency-dependent filtering. After this step, the features were extracted from the filtered samples. A binary decision tree structure consisting of binary SVMs was then created in order to classify the extracted data.

3 Emotional Speech Databases

The data used in SER research comes from ESDs. They consist of recordings where humans speak in various affective states. These databases are created and used to study automatic emotion and speech recognition, emotional speech synthesis, evaluating human emotion perception and for various medical applications [57].

Expression of emotions can be categorized as natural, acted or elicited [9]. Although recordings of spontaneous expression of emotions in day-to-day conversations would be preferred, it is riddled with ethical issues due to the possibility of these recordings revealing intimate and personal details about the speakers [9]. According to [9], eliciting certain emotions, such as fear or panic, have similar implications. Thus, acted emotions are often chosen as the most suitable option [57]. Even if acted emotions might be seen as insincere or unnatural, they offer little ethical or legal complications [57] and allow for controlled experiments [8, 31]. Given that the emotions are expressed in an acted manner, overacting is usually not allowed [57].

Although the list of emotions to include in the database is not set in stone, ranging from shame to pride, the affective states most often present are anger, sadness, happiness, fear and disgust [57]. Of the numerous ESDs, four will be used in this thesis to assess the performance of the proposed features.

Surrey Audio-Visual Expressed Emotion database (SAVEE) [30] was created as a prerequisite for developing an automatic emotion recognition system. The database consists of audio-visual recordings from 4 male actors in 7 different emotions for a total of 480 samples. The database was validated by 10 listeners to test the recognizability of all emotions and the average emotion recognition rate was approximately 66.5% [30].

Berlin Database of Emotional Speech (EMO-DB) [8] contains recordings from 5 male and 5 female actors in 7 different emotions. Validation was done by 20 listeners and the recognition rates for emotions ranged from 96.9% for anger to 79.6% for disgust [8].

Polish Emotional Speech Database (PESD) [12] was created by the Medical Electronics Division of the Technical University of Lodz. The database contains 5 sentences from each of the 8 speakers (4 male, 4 female) in 6 different emotions each.

Serbian Emotional Speech Database (GEES) [31] consists of 6 speakers, 3 male and 3 female, expressing 5 emotions. Testing humans' and computers' SER capability was the motivation behind creating this database [31]. The database was validated by 30 listeners and the average emotion recognition rate was 95% [31].

For a more detailed overview of the databases, please refer to Table 3.1.

Table 3.1: ESDs described in detail. **M** - number of male samples, **F** - number of female samples, **T** - samples in total, **M/T** - the proportion of male samples to female samples in a database, **L** - the average duration of the samples.

Database	**Labels**	**M**	**F**	**T**	**M/T**	**L**
SAVEE	Anger	60	0	60	1	3.71
	Disgust	60	0	60	1	3.95
	Fear	60	0	60	1	3.75
	Happiness	60	0	60	1	3.8
	Neutral	120	0	120	1	3.61
	Sadness	60	0	60	1	4.48
	Surprise	60	0	60	1	3.8
Total	**7**	**480**	**0**	**480**	**1**	**3.84**
EMO-DB	Anger	60	67	127	0.472	2.64
	Boredom	35	46	81	0.432	2.78
	Disgust	11	35	46	0.239	3.35
	Fear	36	33	69	0.522	2.23
	Happiness	27	44	71	0.380	2.54
	Neutral	39	40	79	0.494	2.36
	Sadness	25	37	62	0.402	4.05
Total	**7**	**233**	**302**	**535**	**0.434**	**2.78**
PESD	Anger	20	20	40	0.5	2.06
	Boredom	20	20	40	0.5	2.86
	Fear	20	20	40	0.5	2.31
	Happiness	20	20	40	0.5	2.14
	Neutral	20	20	40	0.5	2.04
	Sadness	20	20	40	0.5	2.44
Total	**6**	**120**	**120**	**240**	**0.5**	**2.31**
GEES	Anger	276	276	552	0.5	2.61
	Fear	276	276	552	0.5	2.82
	Happiness	276	276	552	0.5	2.82
	Neutral	276	276	552	0.5	2.65
	Sadness	276	276	552	0.5	3.31
Total	**5**	**1380**	**1380**	**2760**	**0.5**	**2.84**

4 Methodology

4.1 Types of Features Used

SER features which are calculated over the whole duration of the speech signal are called suprasegmental features [1]. In contrast, segmental features are calculated over short consecutive segments of speech [1]. Both systems in this work use suprasegmental features. While this approach describes the detailed changes in features poorly, it does help by reducing the complexity of the system and keeping the number of features used low. When specifying a timing window was necessary to create a Praat object, the automatic window length selection was used.

Features can also be categorized as LLDs and functionals (applied to LLDs) [1]. Because suprasegmental features are used in both systems, all features in this work are functionals applied to LLDs. In summary, all features used in both systems are suprasegmental functionals presented in Table 4.1. The total number of features is 87.

4.2 Features

Pitch, according to [43], is the relative highness or lowness of a tone, as it is perceived by humans. It depends on the frequency of vibration of the vocal chords [43]. Pitch usually rises at the end of a sentence when a question is asked ("Did you like it?").

Intensity describes the relative effective pressure of sound [44]. The reference value is usually 20^{-6} Pa, which is often considered the threshold of human hearing [44]. Intuitively, it might be thought of as how much louder the sound is than silence.

Long-term Average Spectrum (LTAS) describes how energy contained in the speech signal is distributed across different frequencies (on average) over the duration of the whole signal [29]. An example of LTAS can be seen in Fig. 1.2.

Harmonicity, also known as Harmonics-to-Noise Ratio (HNR), describes the degree of acoustic periodicity [3]. It can be calculated with the following formula:

$$HNR = 10 \cdot \log_{10}\left(\frac{E_p}{E_n}\right), \tag{4.1}$$

where E_p is the percentage of energy in the periodic part of the signal and E_n is the percentage of energy found in noise [3]. For $E_p = 80\%$ and $E_n = 20\%$, $HNR = 6.02$ dB. Equal distribution of energy between harmonics and noise results in an HNR of 0 dB.

Table 4.1: The features used in online and offline systems. p_x - x-th percentile.

Feature	**Functionals**
Pitch	min, max, mean, p_{25}, p_{50}, p_{75}, stdev, mean absolute slope, slope without octave jumps
Intensity	min, max, mean, p_{25}, p_{50}, p_{75}, stdev
LTAS	min, fmin, max, fmax, mean, slope, stdev
Sound	min, max, mean, stdev, power, energy, RMS
Harmonicity	min, max, mean, stdev
Point process	periods, meanperiod, stdevperiod, jitterlocal, jitterppq5
MFCC(1-24)	mean, stdev

A periodic signal can be modeled as a point process. The end of each periodic vibration at time t_i would be marked by a point. To illustrate, a signal with frequency f = 5 Hz would yield 5 points during a 1 second long segment, while a signal with f = 2 Hz would yield 2 points in the same segment. Modeling sound signals this way allows us to calculate more features to analyze speech. For example, local jitter can be calculated to detect potential pathologies in speakers [54]. An algorithm for local jitter calculation is described in [4].

Mel-frequency Cepstral Coefficients (MFCC) [17], as the name suggests, are the coefficients of a mel-frequency cepstrum. The mel-frequency cepstrum describes how the power of a sound is distributed between bands of frequencies on the mel scale [17]. According to [18], the mel scale is believed to more accurately describe how humans perceive differences in pitches because after a threshold frequency (usually chosen as 0.625 kHz, 0.7kHz or 1kHz), the perception of pitch changes from linear to logarithmic. This means that after doubling the frequency of a tone, the pitch of the tone is not perceived to be twice as high.

Sound is speech modeled as a sound wave.

4.3 Feature Extraction

To extract the features, Praat 6.0.36 [2] was used. The program was chosen because of the built-in scripting capabilities which allowed the automation of the feature extraction process. For each class of features in Table 4.1, a Praat object with the same name exists. After creating the necessary Praat object, the object functions were used for extracting the features. In order to keep the system simple, the default parameters presented by Praat for the object functions were used. An exception to this was MFCCs: 24 coefficients were calculated instead of the default 12.

4.4 Random Forests

Random forest (RF) [7,37], as the name suggests, consist of randomly generated decision trees. A fixed amount of randomly chosen features, usually $\log_2$ of the length of the feature vector, is used to grow each decision tree [7]. After all the trees are grown, the input data is propagated down each tree and each tree gets to vote. The votes are counted and the class with the most votes is considered to be the result of the classification process [7].

4.5 Support Vector Machines

According to [13, 52], binary classification with support vector machines for two-dimensional data can be described as finding the line (hyperplane), which separates the two classes while also maximizing the distance between datapoints closest to the line. Maximizing the margins between the hyperplane and the closest datapoints, depicted in Fig. 4.1, prevents overfitting and improves generalization [52]. If the data is not separable by a hyperplane, the data can be transformed with a non-linear transformation [15]. To illustrate the transformation, consider two classes C_1 and C_2 with two features, x and y:

$$\begin{aligned} C_1 &: x^2 + y^2 < 4, \\ C_2 &: x^2 + y^2 > 6. \end{aligned} \tag{4.2}$$

Looking at the left graph in Fig. 4.2, a line cannot separate these two classes. However, after applying the non-linear transformation $\Phi((x, y)) = (\sqrt{x^2 + y^2}, \sqrt{x + y})$, the data becomes linearly separable (Fig. 4.2, right plot). The downside of finding the linear separating hyperplane in the transformed space is increased computational complexity and the need to find a suitable transformation for the data. This can be avoided by using the kernel trick, which is described in [52]. Although the examples presented here contain only two-dimensional data, the same general principles apply for data with higher dimensions.

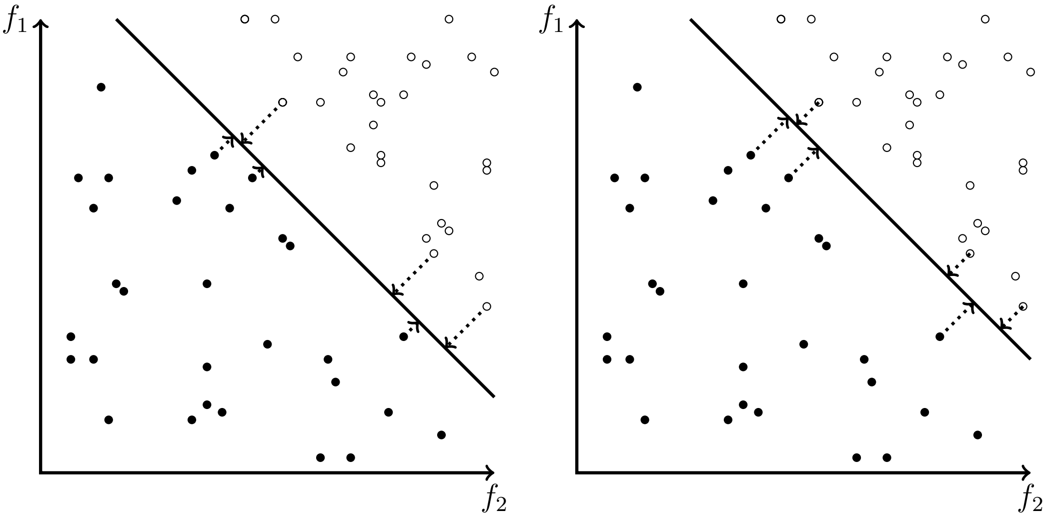


Figure 4.1: An example of a binary classification problem using features f_1 and f_2. Although both of these hyperplanes correctly classify the input data, the one on the right is more robust [52]. Notice how the lengths of the support vectors (dotted) for both classes are more similar on the right than on the left.

4.6 Data Preprocessing and Kernel Parameters

The scaling of data was performed by *svm-scale* without providing it any additional parameters. Radial Basis Function (RBF) was the kernel used for all the databases. The search for optimal

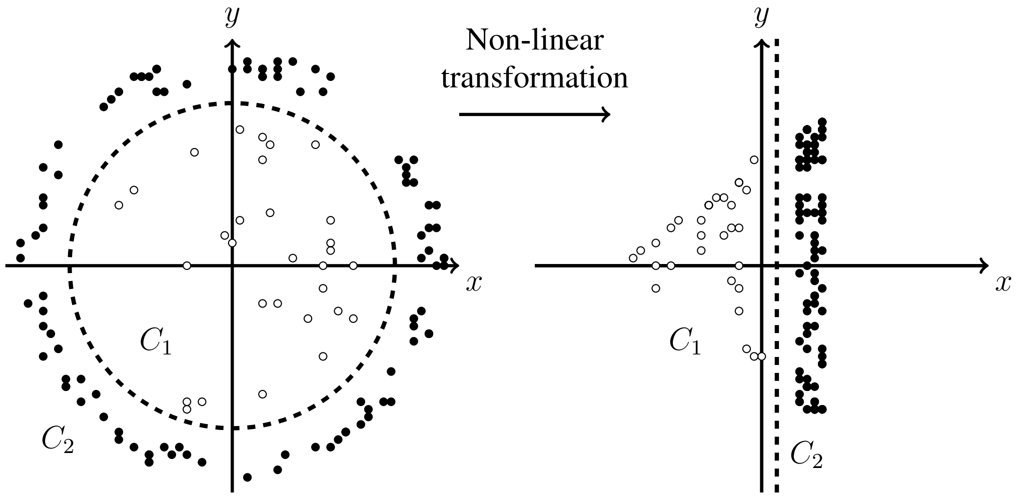


Figure 4.2: A binary classification example, white dots represent class C_1, black dots represent class C_2. After a non-linear transformation, the data becomes linearly separable [52].

hyperparameters (γ, C) was conducted in accordance to the guidelines presented in [28]. The script *grid.py* was the tool provided by LIBSVM to help finding the parameters providing maximum performance. The aforementioned program executes a search through a preset combination of different parameter values which are to be used as rough guidelines as the program does not perform an exhaustive search. The motivation behind using the provided tools was, as mentioned before, to keep the system simple and to facilitate reproducibility.

4.7 Classification

4.7.1 Offline System

A high-level description of the SER process can be described as follows: an ESD is chosen, the features are extracted, the extracted features may be further processed and finally they are used as input for the machine learning algorithms. This is the most common way of performing SER and this way is followed in the offline system. The rationale behind choosing SVM [10] as the classifier for the offline system was the abundance of comparable results that are available [11,25,26,33,49,50,59]. In addition to many similar works, the creators of LIBSVM had created a great introductory text for using SVMs and optimizing their performance [28], not to mention the plethora of useful tools bundled with the software library. To train, test and calculate the confusion matrices of the SVMs, Scikit-learn [46] was used. This Python framework uses LIBSVM as it's internal SVM implementation.

10-fold cross-validation was used to validate the model in the offline system. This is common practice in the field of SER research [14,21,26,34,38,41,42,45,58]. In n-fold cross-validation, input data is partitioned to n sets of roughly the same size. Each of the n partitions will be used for testing once while the remaining $n - 1$ partitions will be used for training the model. It can be thought of as training n different models and averaging their respective results. This approach has an advantage over a single partition strategy (divide the data to a single testing and single training set) because the constitution of the training and testing sets affects the results. n-fold cross-validation is illustrated in Fig. 4.3.

4.7.2 Online System

In the online system, the speaker will be asked to speak in different emotions. The speech will be recorded and Praat will automatically extract the necessary features. After the features are extracted, they are used as input to the trained models using Weka [19], which will output the class probabilities for each emotion in the models. These outputs are the basis on which the performance of the online system will be measured.

As mentioned before, two models will be trained on modified versions of SAVEE and EMO-DB. From SAVEE, surprise samples will be removed. From EMO-DB, boredom samples will be removed. This leaves us with SAVEE' (420) and EMO-DB (454), both of which contain the same exact emotions and are of very similar size. In addition, both English and German are West Germanic languages, making these two the most similar languages in our database selection. SAVEE' has samples only from male speakers, while EMO-DB' has samples from male and female speakers. This allows us to test, in addition to the usability of the proposed features, how much the absence of samples from one gender during training affects SER performance for speakers of the absent gender. Both of the models contain six different emotions: anger, disgust, fear, happiness, neutral and sadness.

RF are used as the classifier in the online system. It is a technique from the category of ensemble methods [42], introduced by Breiman in [6]. RF perform well in multi-label classification tasks [60] and often have low bias and high variance [5, 20]. In addition, it can be parallelized with ease [39].

Two male and two female non-native English speakers were included in the online experiment. After speaking and recording, the speaker was allowed to listen and re-record the sample if the sample was not deemed to be good enough by the speaker. Outputs of the models were saved for further analysis. The online system is illustrated in Fig. 4.4.

Table 4.2: Sentences used in the online system, which themselves are a subset of the ones used in SAVEE [30].

Nr	**Sentence**
1	Who authorized the unlimited expense account?
2	Please take this dirty table cloth to the cleaners for me.
3	Call an ambulance for medical assistance.
4	Those musicians harmonize marvelously.
5	The prospect of cutting back spending is an unpleasant one for any governor.
6	The best way to learn is to solve extra problems.

Figure 4.3: Offline system representation. a - Feature extraction via Praat. b - Feature scaling via *svm-scale*. c - Kernel parameter search via *grid.py*. d - Model training and 10-fold cross-validation via *svm-train*.

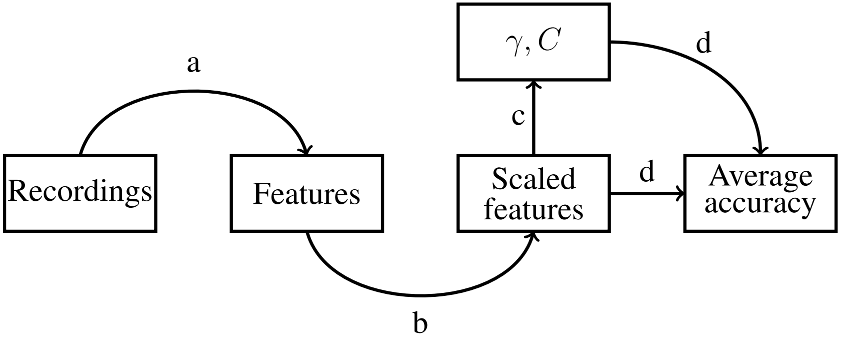
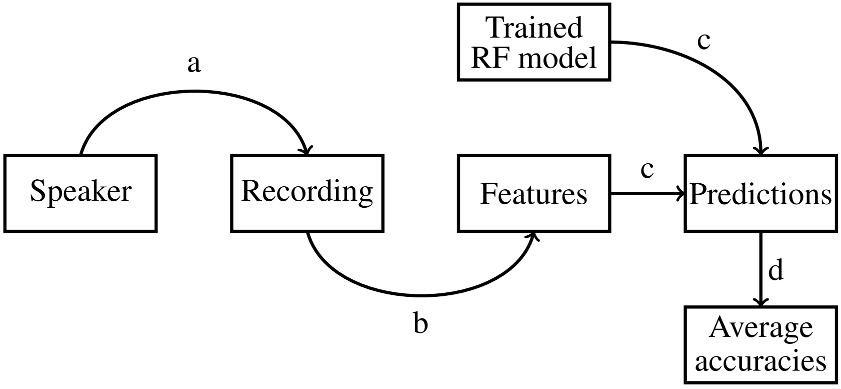


Figure 4.4: Online system representation. a - Subject speaks sentences in 6 different affective states. b - Feature extraction via Praat. c - Using Weka to get class distribution predictions with previously trained SAVEE' and EMO-DB' models. d - Averaging classification accuracy over emotions per sex and over emotions per trained model.



5 Results

5.1 Offline Results

In SAVEE, neutral was recognized the best (94.17%, Table 5.1), followed by anger (85%) and disgust (81.67%). Poorest performers were surprise (75%), fear (71.67%) and happiness (61.67%). Surprise mislabeled as fear (15%) and happiness mislabeled as anger (15%) were the biggest sources of confusion.

In EMO-DB, sadness was the best recognized emotion (98.39%, Table 5.2), followed by anger (93.7%) and neutral (92.41%). The worst performers were boredom (90.12%), disgust (82.61%) and happiness (66.2%). 21.13% of happiness samples were classified as anger, making it the worst performing emotion in all selected databases. During the validation of EMO-DB, the emotion recognition accuracy of human listeners was collected [8]. The performance of human listeners compared to the proposed system can be seen in Table 5.5. For emotions excluding happiness and sadness, the improvements range from 3.01% and 4.21%. Human listeners recognize happiness considerably more often than the system (66.2% vs 83.7%). The system outperforms humans in sadness recognition by 17.69%.

Although PESD is considerably smaller than the other databases, an average emotion recognition accuracy of 75.42% was achieved regardless. Boredom was recognized the worst (67.5%, Table 5.3), fear and sadness (both 70%) following closely. In a similar manner, boredom was one of the labels with relatively poor performance in EMO-DB. However, boredom recognition was far higher than in PESD (90.12% vs 67.5%).

GEES is the largest database used in this work in addition to it's highest average emotion recognition rate of 93.41% (Table 5.4). A large group of 30 people were involved in validating the database and their emotion recognition results are available in [31]. During the validation phase, the average emotion recognition rate was 95%, which illustrates the masterful performance of the speakers and the accurate emotion recognition of the validating audience.

Best performers in GEES are happiness (96.56%, Table 5.4), fear (95.47%) and sadness (95.29%). In contrast to other databases, neutral (87.86%) shows the worst performance here, followed by anger (91.85%). The obtained results compared to humans' is demonstrated in Table 5.6. Some minor improvements can be seen compared to human listeners except for anger, where performance is lacking (Table 5.6).

Similar works and their results are described in Table 5.7. In these works, SVM was used as the classifier and a minimum of one database also used in the offline system. For each database, the number of features and the average emotion recognition rate is presented and compared with others. The greatest gains in emotion recognition rates occur with SAVEE and PESD, resulting

Table 5.1: SAVEE confusion matrix. $\gamma = 0.0078125, C = 128$ yielded an average accuracy of **80.21%**.

	ANG	**DIS**	**FEA**	**HAP**	**NEU**	**SAD**	**SUR**
ANG	**85.00**	5.00	1.67	8.33	0.00	0.00	0.00
DIS	3.33	**81.67**	3.33	0.00	6.67	1.67	3.33
FEA	1.67	6.67	**71.67**	8.33	0.00	1.67	10.00
HAP	15.00	0.00	13.33	**61.67**	1.67	0.00	8.33
NEU	0.00	2.50	0.00	0.83	**94.17**	2.50	0.00
SAD	0.00	8.33	0.00	0.00	13.33	**78.33**	0.00
SUR	1.67	0.00	15.00	8.33	0.00	0.00	**75.00**

Table 5.2: EMO-DB confusion matrix. $\gamma = 0.0078125, C = 32$ yielded an average accuracy of **88.6%**.

	ANG	**BOR**	**DIS**	**FEA**	**HAP**	**NEU**	**SAD**
ANG	**93.70**	0.00	0.00	0.79	4.72	0.79	0.00
BOR	0.00	**90.12**	1.23	0.00	1.23	6.17	1.23
DIS	0.00	2.17	**82.61**	4.35	4.35	4.35	2.17
FEA	4.35	0.00	1.45	**91.30**	1.45	1.45	0.00
HAP	21.13	0.00	1.41	9.86	**66.20**	1.41	0.00
NEU	0.00	6.33	0.00	1.27	0.00	**92.41**	0.00
SAD	0.00	0.00	0.00	0.00	0.00	1.61	**98.39**

Table 5.3: PESD confusion matrix. $\gamma = 0.03125, C = 32$ yielded an average accuracy of **75.42%**.

	ANG	**BOR**	**FEA**	**HAP**	**NEU**	**SAD**
ANG	**77.50**	2.50	5.00	12.50	0.00	2.50
BOR	0.00	**67.50**	12.50	0.00	10.00	10.00
FEA	12.50	2.50	**70.00**	0.00	2.50	12.50
HAP	17.50	0.00	2.50	**80.00**	0.00	0.00
NEU	0.00	7.50	0.00	0.00	**87.50**	5.00
SAD	0.00	12.50	7.50	0.00	10.00	**70.00**

Table 5.4: GEES confusion matrix. $\gamma = 0.125, C = 32$ yielded an average accuracy of **93.41%**.

	ANG	**FEA**	**HAP**	**NEU**	**SAD**
ANG	**91.85**	0.18	7.61	0.36	0.00
FEA	1.45	**95.47**	1.27	1.09	0.72
NEU	9.78	1.45	**87.86**	0.91	0.00
HAP	0.18	0.72	0.18	**96.56**	2.36
SAD	0.00	0.54	0.00	4.17	**95.29**

Table 5.5: Offline system compared to human listeners [8], the system's improvements are highlighted in **bold**.

EMO-DB	**This work**	**Humans**	**Difference**
Anger	93.7	96.9	-3.2
Boredom	90.12	86.2	**3.92**
Disgust	82.61	79.6	**3.01**
Fear	91.3	87.3	**4**
Happiness	66.2	83.7	-17.5
Neutral	92.41	88.2	**4.21**
Sadness	98.39	80.7	**17.69**

Table 5.6: Offline system compared to human listeners [31], the system's improvements are highlighted in **bold**.

GEES	**This work**	**Humans**	**Difference**
Anger	91.85	96.06	-4.21
Fear	95.47	93.33	**2.14**
Happiness	87.86	88.95	-1.09
Neutral	96.56	94.67	**1.89**
Sadness	95.29	96.04	-0.75

Table 5.7: Offline system comparison with similar systems. The results of the offline system are highlighted in **bold**.

Reference	**Database**	**Labels**	**Features**	**Accuracy**
.	**SAVEE**	**7**	**87**	**80.21**
[59]	SAVEE	7	566	73.81
[33]	SAVEE	7	153	76.08
.	**EMO-DB**	**7**	**87**	**88.6**
[59]	EMO-DB	7	566	82.9
[25]	EMO-DB	7	6553	92.3
[11]	EMO-DB	7	4368	86.1
[11]	EMO-DB	7	180	81.1
[33]	EMO-DB	7	153	85.13
.	**PESD**	**6**	**87**	**75.42**
[59]	PESD	6	566	71.3
.	**GEES**	**5**	**87**	**93.41**
[25]	GEES	5	6553	94.6
[50]	GEES	5	318	90.63
[50]	GEES	5	162	90.96
[49]	GEES	5	318	89.7

in gains of 4.13% and 3.92%, respectively. In addition, 1.7 times less features were used (566 vs 87). With EMO-DB and GEES, the system's performance came close to state-of-the-art, resulting in decreases of 3.7% and 0.89%, respectively. However, it should be noted that this level of accuracy was achieved with roughly 75 times less features (6553 vs 87), making the proposed features more efficient in terms of computational resources while maintaining a similar level of accuracy.

5.2 Online Results

For male speakers, the average emotion recognition rate in SAVEE' was 36% and 28% in EMO-DB' (Table 5.8). In SAVEE', all emotions except anger (25% vs 33.33%) and neutral (91.67% vs 100%) were recognized better than in EMO-DB'. Fear in males was recognized much better in SAVEE' than in EMO-DB' (50% vs 8.33%).

For female speakers, the average emotion recognition rate in SAVEE' was 19% and in 29% in EMO-DB' (Table 5.8). For all emotions except fear (100% vs 8.33%) and sadness (8.33% vs 0%), EMO-DB' showed better results with female speakers. Happiness in females was recognized extremely accurately in EMO-DB' (0% vs 100%) in addition to neutral (8.33% vs 41.67%).

Overall, the system reached 28% average emotion recognition rate in both models. This is better than random guessing (16.67%).

Table 5.8: Emotion recognition rates in the online system.

	SAVEE'			**EMO-DB'**		
	Male	Female	Both	Male	Female	Both
ANG	25.00	0.00	12.50	33.33	16.67	25.00
DIS	16.67	0.00	8.33	8.33	8.33	8.33
FEA	50.00	100.00	75.00	8.33	8.33	8.33
HAP	16.67	0.00	8.33	8.33	100.00	54.17
NEU	91.67	8.33	50.00	100.00	41.67	70.83
SAD	16.67	8.33	12.50	8.33	0.00	4.17
AVG	**36.11**	**19.44**	**27.78**	**27.78**	**29.17**	**28.47**

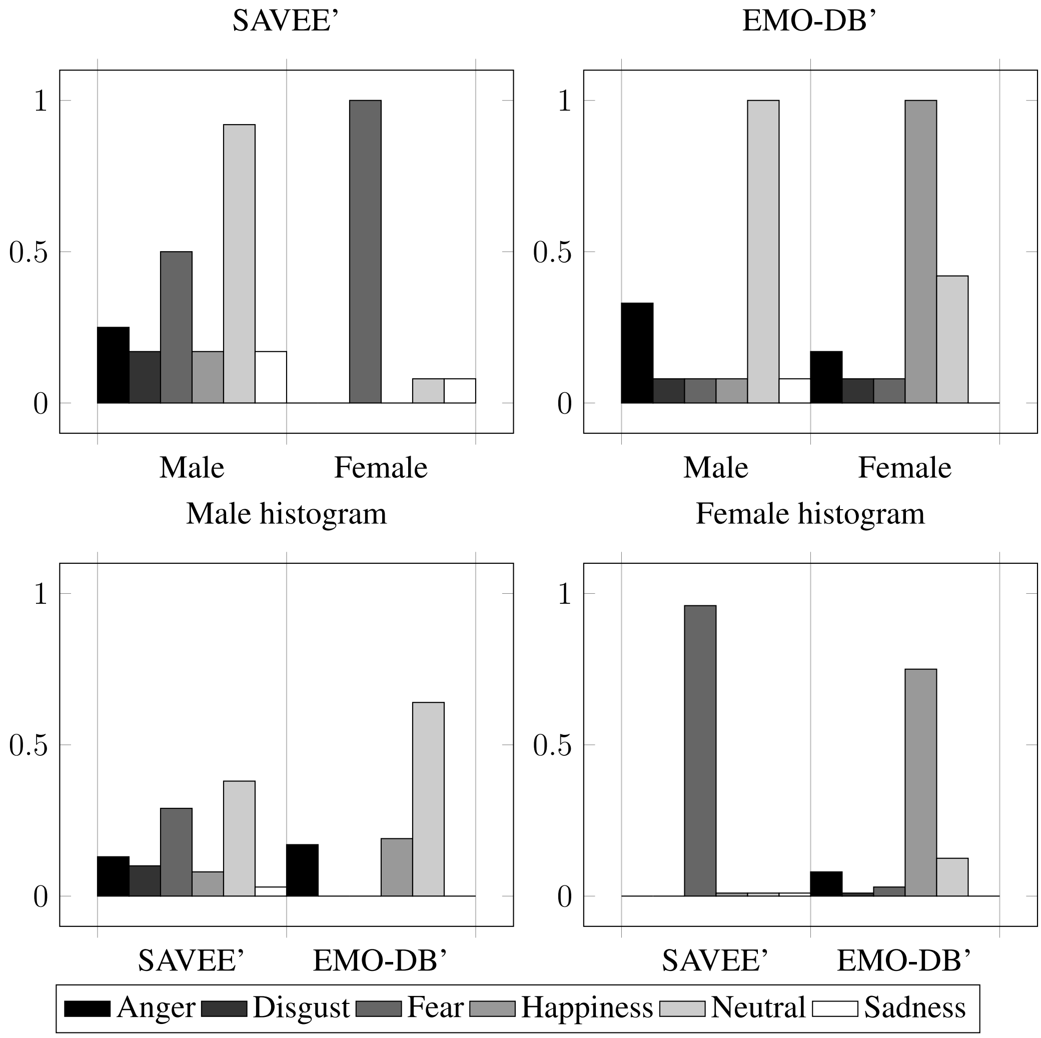


Figure 5.1: The top two charts display emotion recognition rates. The bottom charts describe average prediction distributions.

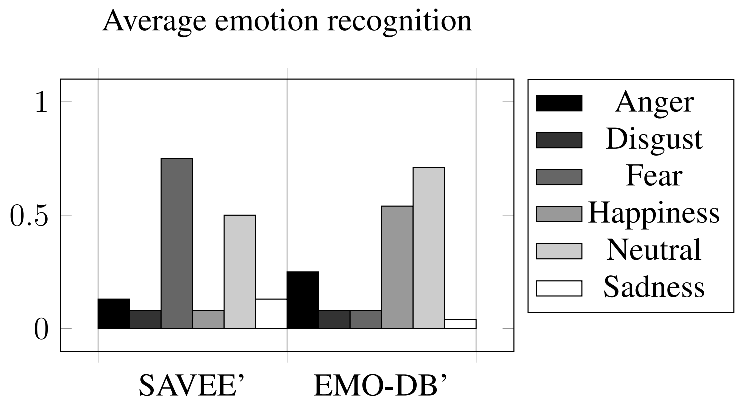


Figure 5.2: Emotion recognition rates in SAVEE' and EMO-DB'.

6 Discussion

6.1 Offline System

There is bi-directional confusion regarding anger and happiness in all databases (Tables 5.1, 5.2, 5.3, and 5.4). A similar effect is noticed in related works as well [49, 50, 59]. In addition, this confusion occurs not only in machine learning classifiers, but in human listeners as well [31].

SAVEE stands out with a few results among other databases. In SAVEE, happiness was misclassified as fear most often (13.33%) with EMO-DB a close second (9.86%). This was not observed in both PESD and GEES (Tables 5.3, 5.4). This also occurs with sadness: 13.33% of sadness samples were misclassified as neutral, compared to EMO-DB, where the rate was 1.61%. With PESD, the difference was smaller (13.33% vs 10%).

Overall, the most misclassifications occurred with happiness in EMO-DB, although it is comparable in 3 databases (Tables 5.1, 5.2 and 5.3). The recognition of disgust follows an interesting Another interesting result in EMO-DB is the misclassifications of disgust, which is almost evenly distributed between other emotions. A similar effect was observed in SAVEE.

Another less pronounced confusion can be observed with neutral, sadness and boredom. Neutral and sadness are often confused with each other, given that boredom does not exist in the database (Tables 5.1 and 5.4). Human listeners, in addition to machine learning algorithms, find these emotions to be confusing [31]. However, if boredom does exist in the database, discriminating between boredom and neutral (EMO-DB) or boredom and sadness (PESD) becomes more difficult (Tables 5.2 and 5.3).

Overall, the proposed features provide good performance on all selected databases, sometimes exceeding, sometimes almost matching state-of-the-art performance. Further research into features capable of differentiating between happiness and anger is necessary to improve emotion recognition rates in all databases as this is a common problem in SER research.

The performance of the offline system is comparable to that of human listeners' for GEES (Table 5.6). With the exception of happiness and sadness, the same stands for EMO-DB (Table 5.5). While the average emotion recognition rate is similar to humans', the extremely high recognition rate of sadness (98.39%) stands out. Human listeners are the only available means for emotion detection at this time. Because of this, all emotion recognition results that vastly exceed humans' should be met with healthy skepticism. The validation results of the database depend on the set of people selected for validation and their capability for emotion recognition in addition to the ability of the speakers to express their emotions clearly. If this is not kept in mind, then it will be easy to stop recognizing emotions and start recognizing arbitrarily assigned labels named 'sad' or 'angry'. An alternative to the arbitrary labels would be emotion distributions per sample, which could be created during the validation phase with minimal extra effort, because the database creators already collect this information while validating. To create the distributions, the people validating the database listen to the samples, after which they vote on which emotion was this sample most representative of. The emotions that can be voted for would be the emotions present in the database. This allows us to see samples that are different degrees of 'angry' or 'sad' and perhaps even mixed emotions like 'angry' and 'sad'. To increase the number of voters, the speakers and the creators themselves can be allowed to vote, resulting in more accurate descriptions of the samples. This approach is also more flexible as it provides the SER researcher an option to either use the distributions to compare their system against human listeners on a per-sample basis or to create discrete labels out of the distributions and use these instead.

6.2 Online System

The online system demonstrated the usability of the proposed features in a plausible real-world application. The models used for classification were trained on ESDs, which have been recorded in excellent acoustical conditions. Despite this, the models were able to recognize emotions from samples recorded with poorer quality recording equipment in an acoustically non-treated environment. This, in addition to the fact that EMO-DB' was able to recognize emotions from recordings in another language, suggests that the system was able to learn something general and was actually recognizing emotions. To add, the non-native speakers were speaking with accents, which gives more weight to the claim that the features are capable of language-independent SER.

In addition, the online system also illustrated how the data used for training the models affects SER results. A significant improvement in emotion recognition for male speakers was observed in SAVEE' where more male samples were present during training. A similar effect can be seen when comparing female speakers' in both models: in EMO-DB', SER performance for female speakers was considerably higher than in SAVEE', where there were no female samples present during training. These results suggest that the higher availability of training samples allow the model to be trained more thoroughly, thus resulting in higher emotion recognition accuracy. The curious case of SAVEE' predicting fear for all the female speakers can be explained by the differences of male and female voice acoustics (e.g. differences in pitch) and how the male speakers in SAVEE' expressed fear when speaking. If male speakers speak in a higher pitch and a quieter voice, then the model can get confused easily when female speakers are talking, as SAVEE' did not have any female samples available for training.

The online system also shows that SER research can be done in a different manner altogether. Although increasing emotion recognition performance is desirable up to a point, the raw performance achieved on ESDs is by no means indicative of real-world performance. In order to explore SER outside the confines of an isolated laboratory environment, different real-life testing methods could be implemented as standard research procedure. To make these results comparable and more meaningful, the whole process should be standardized as much as possible. Such an additional dimension in common research practices could help the field evolve in new, exciting directions with novel solutions for better performance in noisy environments.

7 Conclusion and Future Work

7.1 Conclusion

In this thesis a SER system was proposed. The novelty of the work has been presented as a scientific paper and published in [26]. On average, the system achieved an increase of 0.8% in SER accuracy with an 81.2% reduction in the number of features used. The system achieved recognition rates of 80.21%, 88.6%, 75.42% and 93.41% on SAVEE, EMO-DB, PESD and GEES, respectively. Comparing the system's performance with other state-of-the-art systems, the differences in performance are 4.14%, -3.7%, 4.12% and -1.19%. In addition, the total number of features used by the proposed system compared to state-of-the-art were reduced by 43.1%, 98.6%, 84.6% and 98.6%, respectively. The promising results warrant further testing on other ESDs to better assess the universal performance of the features. To better assess the accuracy of SER systems and offer the researchers more flexibility, replacing discrete labels with per-sample class distributions acquired during database validation was proposed.

The online system, in addition to illustrating the relationship between training data constitution and SER performance, showed the usability of the features in a simple real-life SER scenario. As the availability of training samples from a particular gender increased, so did the SER performance for speakers of that gender in the trained model. The inverse was also true, as the model trained on only male samples had significantly poorer performance with female speakers. In EMO-DB, the decreased amount of male training samples was also reflected in the male emotion recognition rate, which was smaller than in SAVEE.

7.2 Future Work

In future works, further testing is required to assess the language independence of the proposed feature set. In addition, an in-depth look into the happiness and anger discrimination should be taken as this in itself would improve the emotion recognition rates, benefiting the field of SER as a whole.

To further improve upon the online system, ESDs in multiple languages can be used train many models and let the different models vote like decision trees in RFs. Furthermore, a language detection component can be added and thus create a language-sensitive SER system. This would allow SER to work at maximum efficiency by using models trained on the language spoken.

Bibliography

[1] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[2] Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 6.0.36)[computer program]. retrieved january 1, 2018, 2018.

[3] Paul Boersma and David Weeninka. Harmonicity. Accessed: 03.04.2018. `http://www.fon.hum.uva.nl/praat/manual/Harmonicity.html`.

[4] Paul Boersma and David Weeninka. Pointprocess: Get jitter (local)... Accessed: 03.04.2018. `http://www.fon.hum.uva.nl/praat/manual/PointProcess__Get_jitter__local_____.html`.

[5] Leo Breiman. Bias, variance, and arcing classifiers. 1996.

[6] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] Leo Breiman and Adele Cutler. Random forests. Accessed: 03.04.2018. `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`.

[8] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[9] Nick Campbell. Databases of emotional speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] Bo-Chang Chiou and Chia-Ping Chen. Feature space dimension reduction in speech emotion recognition using support vector machine. In *Signal and information processing association annual summit and conference (APSIPA), 2013 Asia-Pacific*, pages 1–6. IEEE, 2013.

[12] J Cichosz. Database of polish emotional speech. *Online: http://www.eletel.p.lodz.pl/med/eng*.

[13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[14] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, Didier Grandjean, and Björn Schuller. Fisher kernels on phase-based features for speech emotion recognition. In *Dialogues with Social Robots*, pages 195–203. Springer, 2017.

[15] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[16] The editors of Wikipedia. Decision tree learning. Accessed: 03.04.2018. `https://en.wikipedia.org/wiki/Decision_tree_learning`.

[17] The editors of Wikipedia. Mel-frequency cepstrum. Accessed: 03.04.2018. `https://en.wikipedia.org/wiki/Mel-frequency_cepstrum`.

[18] The editors of Wikipedia. Mel-frequency cepstrum. Accessed: 03.04.2018. `https://en.wikipedia.org/wiki/Mel_scale`.

[19] F Eibe, MA Hall, IH Witten, and JC Pal. The weka workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*, 4, 2016.

[20] João Gama, Ricardo Rocha, and Pedro Medas. Accurate decision trees for mining high-speed data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528. ACM, 2003.

[21] Davood Gharavian, Mehdi Bejani, and Mansour Sheikhan. Audio-visual emotion recognition using fcbf feature selection method and particle swarm optimization for fuzzy artmap neural networks. *Multimedia Tools and Applications*, 76(2):2331–2352, 2017.

[22] Mingmin Gong and Qi Luo. Speech emotion recognition in web based education. In *Grey Systems and Intelligent Services, 2007. GSIS 2007. IEEE International Conference on*, pages 1082–1086. IEEE, 2007.

[23] Jelena Gorbova, Iiris Lüsi, Andre Litvin, and Gholamreza Anbarjafari. Automated screening of job candidate based on multimodal video processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 29–35, 2017.

[24] Michael Grimm, Kristian Kroschel, Helen Harris, Clifford Nass, Björn Schuller, Gerhard Rigoll, and Tobias Moosmayr. On the necessity and feasibility of detecting a driver's emotional state while driving. In *International Conference on Affective Computing and Intelligent Interaction*, pages 126–138. Springer, 2007.

[25] Ali Hassan and Robert I Damper. Multi-class and hierarchical svms for emotion recognition. 2010.

[26] Joosep Hook, Fatemeh Noroozi, Onsen Toygar, and Gholamreza Anbarjafari. Automatic speech based emotion recognition using paralinguistics features. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 2018.

[27] M Shamim Hossain and Ghulam Muhammad. Cloud-assisted speech and face recognition framework for health monitoring. *Mobile Networks and Applications*, 20(3):391–399, 2015.

[28] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. 2003.

[29] Krzysztof Izdebski. *Emotions in the Human Voice, Volume 3: Culture and Perception*, volume 3. Plural Publishing, 2008.

[30] P Jackson and S Haq. Surrey audio-visual expressed emotion (SAVEE) database. *University of Surrey: Guildford, UK*, 2014.

[31] Slobodan T Jovicic, Zorka Kasic, Miodrag Dordevic, and Mirjana Rajkovic. Serbian emotional speech database: design, processing and evaluation. In *9th Conference Speech and Computer*, 2004.

[32] Norhaslinda Kamaruddin and Abdul Wahab. Heterogeneous driver behavior state recognition using speech signal. In *Proceedings of the 10th WSEAS international conference on System science and simulation in engineering*, pages 207–212, 2011.

[33] VB Kobayashi and VB Calag. Detection of affective states from speech signals using ensembles of classifiers. 2013.

[34] Sreenivasa Rao Krothapalli and Shashidhar G Koolagudi. Speech emotion recognition: a review. In *Emotion Recognition using Speech Features*, pages 15–34. Springer, 2013.

[35] Fu-Ming Lee, Li-Hua Li, and Ru-Yi Huang. Recognizing low/high anger in speech for call centers. In *International Conference on Signal Processing, Robotics and utomation*, pages 171–176, 2008.

[36] Wu Li, Yanhui Zhang, and Yingzi Fu. Speech emotion recognition in e-learning system based on affective computing. In *Natural Computation, 2007. ICNC 2007. Third International Conference on*, volume 5, pages 809–813. IEEE, 2007.

[37] Andy Liaw and Matthew Wiener. Classification and regression by random forest. *R news*, 2(3):18–22, 2002.

[38] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 2017.

[39] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[40] Donn Morrison, Ruili Wang, and Liyanage C De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98–112, 2007.

[41] Fatemeh Noroozi, Neda Akrami, and Gholamreza Anbarjafari. Speech-based emotion recognition and next reaction prediction. In *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, pages 1–4. IEEE, 2017.

[42] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, and Gholamreza Anbarjafari. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246, 2017.

[43] The Editors of Encyclopaedia Britannica. Pitch, britannica.com. Accessed: 03.04.2018. `https://www.britannica.com/topic/pitch-speech`.

[44] The Editors of Wikipedia. Sound pressure - wikipedia. Accessed: 03.04.2018. `https://en.wikipedia.org/wiki/Sound_pressure#Sound_pressure_level`.

[45] Pavitra Patel, Anand Chaudhari, Ruchita Kale, and M Pund. Emotion recognition from speech with gaussian mixture models & via boosted gmm. *International Journal of Research In Science & Engineering*, 3, 2017.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] Valery Petrushin. Emotion in speech: Recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering*, volume 710, 1999.

[48] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[49] Arslan Shaukat and Ke Chen. Towards automatic emotional state categorization from speech signals. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[50] Arslan Shaukat and Ke Chen. Emotional state categorization from speech: machine vs. human. *arXiv preprint arXiv:1009.0108*, 2010.

[51] Mariusz Szwoch and Wioleta Szwoch. Emotion recognition for affect aware video games. In *Image Processing & Communications Challenges 6*, pages 227–236. Springer, 2015.

[52] Tan, Pang-Ning and Steinbach, Michael and Kumar, Vipin. *Introduction to Data Mining, (First Edition)*, chapter 5. Addison-Wesley Longman Publishing Co., Inc., 2005.

[53] Ashish Tawari and Mohan Trivedi. Speech based emotion classification framework for driver assistance system. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 174–178. IEEE, 2010.

[54] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112–1122, 2013.

[55] A Tickle, S Raghu, and M Elshaw. Emotional recognition from the speech signal for a virtual education agent. In *Journal of Physics: Conference Series*, volume 450, page 012053. IOP Publishing, 2013.

[56] John Torous, Rohn Friedman, and Matcheri Keshavan. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth and uHealth*, 2(1):e2, 2014.

[57] Dimitrios Ververidis and Constantine Kotropoulos. A state of the art review on emotional speech databases. In *Proceedings of 1st Richmedia Conference*, pages 109–119. Citeseer, 2003.

[58] Na Yang, Jianbo Yuan, Yun Zhou, Ilker Demirkol, Zhiyao Duan, Wendi Heinzelman, and Melissa Sturge-Apple. Enhanced multiclass svm with thresholding fusion for speech-based emotion classification. *International Journal of Speech Technology*, 20(1):27–41, 2017.

[59] Enes Yüncü, Hüseyin Hacihabiboglu, and Cem Bozsahin. Automatic speech emotion recognition using auditory models with binary decision tree and svm. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 773–778. IEEE, 2014.

[60] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

Common License

I, Joosep Hook, (date of birth: 12.10.1992),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   (a) reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   (b) make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   ”Automatic Speech-based Emotion Recognition”, supervised by Prof. Gholamreza Anbarjafari and Fatemeh Noroozi,

2. am aware of the fact that the author retains these rights;

3. certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **20.05.2018**