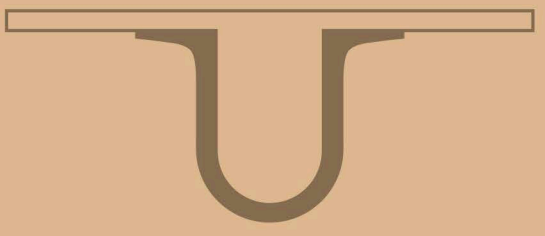
UNIVERSIDADE D
COIMBRA



Gustavo Miguel Santos Assunção

# Human Emotion Recognition Through Speech Analysis on Convolutional Neural Networks

Dissertation submitted to the Department of Electrical and Computer Engineering
of the Faculty of Science and Technology of the University of Coimbra
in partial fulfillment of the requirements for the Degree of Master of Science

June 2019


GUSTAVO MIGUEL SANTOS ASSUNÇÃO

# HUMAN EMOTION RECOGNITION THROUGH SPEECH ANALYSIS ON CONVOLUTIONAL NEURAL NETWORKS

Thesis submitted to the

University of Coimbra for the degree of

Master in Electrical and Computer Engineering

Supervisors:

Prof. Dr. Paulo Jorge Carvalho Menezes (ISR)

Prof. Dr. Fernando Manuel dos Santos Perdigão (IT)

**Coimbra, 2019**

This work was developped in collaboration with:

**University of Coimbra**

UNIVERSITY OF
COIMBRA

**Department of Electrical and Computer Engineering**

DEEC

**Institute of Systems and Robotics**

INSTITUTE OF SYSTEMS AND ROBOTICS
UNIVERSITY OF COIMBRA

*People say nothing is impossible, but I do nothing every day*

Winnie the Pooh

# Dedication

This dissertation, though seemingly short of content, is the culmination of years of work and dedication encompassing all educational and even some non-educational stages of my life. Stages filled with loving friends and family, colleagues and teachers who supported me and helped me develop the skills necessary to overcome the obstacles I had to face. And so, even though it is certainly not enough, I would like to thank and dedicate them my work. To my friends and colleagues from the AP4ISR team, for all the late nights and plenty of laughter. To my advisors Paulo Menezes and Fernando Perdigão not only for their valuable input in this work but also for mentoring me in the right direction. To my girlfriend Joana for her support and company through plenty of asian food. To my friends Bruno and Carlos for being my companions through these last five years. To my brother André and Elif for often being a much needed getaway during turmoil and stressful times. Above all, to my parents Dina and Jorge for endowing me with their own values, always nurturing my abilities and constantly making sure I was on the right track to success but also while being happy. I hope to some day be able to repay them somehow.

*Gustavo Assunção*

# Acknowledgments

The author would like to thank the respective database owners and curators for providing access to their emotional speech datasets and for allowing their use in this research. Without these, empirical results would not have been attainable.

These special thanks are also extended to the Institute of Systems and Robotics (ISR) for providing the resources necessary for the carried out experiments.

Cover image designed by macrovector / Freepik.

# Abstract

The idea of recognizing human emotion has recently received considerable attention from the research community, due to its many possible forensic applications and potential boosting of interactive systems. As such, and following the current trend of research, many machine learning models have been proposed addressing the topic of speech emotion recognition (SER), the idea of classifying a person's emotional state based on speech analysis. These models have far surpassed the performance of previous classical techniques. Nevertheless, even the most successful methods are still rather lacking in terms of adaptation to specific speakers and scenarios, which causes them to be incapable of meeting real human performance standards. In this dissertation, a large scale machine learning model for classification of emotional states is evaluated. This model has previously been trained for speaker identification but is instead used here as a front-end for extracting robust features from emotional speech. The proposed hypothesis is that adaptation to a speaker's emotional prosody can greatly improve the accuracy of a SER system. Several experiments using various state-of-the-art classifiers were carried out, using the Weka software, in order to evaluate the robustness of the extracted features. Considerable improvement was observed when comparing the obtained results with other SER state-of-the-art techniques, which demonstrates the importance of speaker adaptation in this matter.

# Resumo

A noção de reconhecer emoções humanas tem, recentemente, vindo a receber considerável atenção por parte da comunidade científica, devido às suas variadas aplicações forenses e potencial melhoramento de sistemas interactivos. Assim sendo, e seguindo a actual tendência de investigação, bastantes modelos de *machine learning* têm sido propostos com foco na questão de reconhecimento de emoções na fala (SER), o conceito de classificar o estado emocional de uma pessoa com base na análise da sua fala. Estes modelos já deveras ultrapassaram a performance de outras técnicas clássicas a eles precedentes. Não obstante, mesmo os modelos com mais sucesso incorporam um certo nível de défice em relação à adaptação a locutores e cenários específicos, fazendo com que sejam incapazes de atingir os padrões de performance real humana. Nesta dissertação, um modelo de *machine learning* de grande escala é avaliado para classificação de estados emocionais. Este modelo foi treinado para identificação de locutor mas é, ao invés, aqui usado como uma componente basilar para a extracção de características robustas de fala emocional. A hipótese aqui proposta é que a adaptação à prosódia emocional de um locutor pode seriamente melhorar a precisão de sistemas SER. Diversas experiências foram feitas usando vários classificadores de estado-da-arte, com recurso ao software Weka, de vista a avaliar a robustez das características extraídas. Foram observados melhoramentos consideráveis quando comparados os resultados obtidos com outras técnicas de SER de estado-da-arte, demonstrando então a importância de adaptação ao locutor nesta matéria.

# List of Figures

1.1 An idealization of how a situation of automatic speech emotion recognition would play out in the real world with an interactive system. . . . . 4

2.1 A Venn diagram like visualization of a Naive Bayes classifier, as similar to the common representation of Bayes theorem. Using previously observed data and considering class conditional independence, the classification space is divided. A new instance is then assigned the class which yields the maximum a posterior likelihood (MAP). . . . . 10

2.2 Example of a kNN classifier in 2-dimensional space, where $k = 3$ and $k = 5$ cases are considered. It can be understood how by considering the $k = 3$ case, the new yellow instance would be assigned to class 2, while in the $k = 5$ case, it would be assigned to class 1. . . . . 11

2.3 A decision tree diagram showing 5 splits total, based on evaluations of data instance features. As it is clear, a new instance is assigned the class of the leaf node it reaches after a certain number of feature evaluations. . 12

2.4 Example of a binary SVM classifier in 3-dimensional space. An unclassified data instance would then be assigned a class depending on which side of the hyperplane it would be placed on. . . . . 14

2.5 Visualization of how considering an extra dimension may solve the data separability issue. On the left, data in 2D space seems inseparable. But by considering a 3rd dimension (changing perspective), the data becomes clearly separable. . . . . 14

2.6 Happiness Version Spectrogram . . . . . 15

2.7 Sad Version Spectrogram . . . . . 16

2.8 Diagram of a simple artificial neural network (ANN), including one input and one output layers with threen hidden layers inbetween. . . . . 17

2.9 Overview of the system in place within a network's node/neuron. The inputs are fed to the node, which applies the weights to them as a linear combination, passing the result of this to the activation function which yields the output. . . . . 18

2.10 Examples of already existing social companion systems which would benefit from speech emotion recognition (SER). From left to right, Google Home, Amazon's Echo and Jibo ©. . . . . 24

2.11 Exemplary diagram of a 100-frame spectral representation of a 1-second audio file progressing through the layers of the *VGGVox* model. The asterisk symbol is used to identify the layers whose outputs were used as feature matrices for emotional classification. . . . . 29

2.12 Weka Explorer UI and console. On the top left, the preprocessing section of Explorer where data can be read and prepared for evaluation. Top right shows the classification section where data can be evaluated on several implemented classifiers. Bottom left shows the visualization section where some data visual representations are provided. Bottom right displays the simple Weka console. . . . . 30

2.13 High-level diagram of the overall system. The layer names in green correspond to the layers from where the features were extracted. Having 5-fold cross validation been used, each time four folds were used for training (orange) while the reamining fold was used for validation (blue). . . . . 32

4.1 Parameter pair mappings of the same randomly chosen parameters across all classes on the left, and distributions of another different but also randomly chosen parameter across all classes on the right, for the $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$ feature arrays, respectively from top to bottom. Either of the visualizations is representative of the distributions/mappings of most parameters in all the data, for the respective feature arrays. The seven colors represent the seven emotional states (classes). . . . . 37

5.1 Overview of the demo's system, encompassing audio acquisition, feature extraction and finally emotional state classification. . . . . 39

5.2 Demo application user interface, with the four sections described in the text. 40

# List of Tables

2.1 Catalog of emotional speech databases. . . . . 23
2.2 Average Pooling layer's k-th dimension adaptation to clip's n-second Duration . . . . . 28

3.1 State of the art classifier performance on full emotional corpora feature arrays. . . . . 35
3.2 State of the art classifier performance on standalone emotional database $1 \times 1 \times 4096$ feature arrays. . . . . 35

# Contents

**List of Figures** **xv**

**List of Tables** **xix**

**1 Introduction** **1**

1.1 Relevance of Study . . . . . 1

1.2 Main Objectives . . . . . 3

1.3 Related Work . . . . . 4

**2 Methodology** **8**

2.1 The Multiclass Classification Issue . . . . . 8

2.1.1 All-versus-all (AVA) . . . . . 9

2.1.2 One-versus-All (OVA) . . . . . 9

2.1.3 Extensible Classifiers . . . . . 9

2.1.3.1 Naive Bayes . . . . . 10

2.1.3.2 $k$-Nearest Neighbours ($k$NN) . . . . . 11

2.1.3.3 Decision Trees . . . . . 12

2.1.3.4 Neural Networks . . . . . 13

2.1.4 Support Vector Machines (SVM) . . . . . 13

2.2 Audio Spectral Representations . . . . . 15

2.3 Machine Learning and Neural Networks . . . . . 16

2.3.1 Learning Stage . . . . . 17

2.3.1.1 Activation Function . . . . . 18

2.3.1.2 Entropy . . . . . 19

2.3.1.3 Cross-Entropy as a Loss Function . . . . . 20
2.3.2 Testing Stage . . . . . 21
2.3.3 Convolutional Neural Networks . . . . . 22
2.4 Database Acquisition . . . . . 22
2.5 Speaker Adaptation . . . . . 23
2.5.1 Recent Work . . . . . 25
2.6 The *VGGVox* model for Speaker Recognition . . . . . 27
2.7 The Weka Software . . . . . 29
2.8 System Overview . . . . . 30

**3 Results** **33**
3.1 Data Preparation . . . . . 33
3.2 Classifier Performance . . . . . 34

**4 Discussion** **36**

**5 Demo Application** **39**

**6 Conclusion** **41**

**7 Future Work** **43**

**Bibliography** **44**

# 1

# Introduction

A study on the automatic and speaker adaptive recognition of human emotions through speech analysis using a convolutional neural network is now presented as the leading topic of research of this M.Sc. dissertation. This section aims to induct the reader into the topic, introducing its key aspects and relevance of study followed by a summary breakdown of its main objectives. By the end of this section it is hoped that the reader has acquired some interest for and a minimal understanding of the topic.

## 1.1 Relevance of Study

Emotion, in any sense of the word, is unquestionably a fundamental aspect of human interaction and everyday life. Despite lack of solid corroboration, many recent reports converge onto a notion of emotion as an evolutionary trigger for adaptive behavior of a being to specific circumstances [1]. It should, therefore, come as no surprise that any decision made or action taken by an individual would be directly or indirectly influenced by that individual's emotionality, regardless of the nature and/or nurture it is built on.

The development of one's emotionality may well be a source of disagreement among members of the research community. In fact, several distinct emotionality models have been proposed in the last few decades considering emotions as positions in activation-emotion space, some of which are presented by Cowie [1]. Plutchik's *emotional wheel* [2], where full blown emotional states are seen as angular measures, and Fox's levelled model [3], where emotions originate in two opposing tendencies being further differentiated by

levels and forming a pyramid-like structure, are two good examples of emotional models. Yet, by being two-dimensional structures, these models are rather lacking in terms of visual representation of emotionality. Thus the pleasure-arousal-dominance (PAD) model [4], a three-dimensional representation where a score is attribtued on each scale, effectively placing emotions on different positions in 3D-space, poses as a better alternative for representation of emotional space.

The effects and impact caused in human life by emotional development, are well apparent and upheld enough to spark interest on the possibilities involved with recognizing someone's emotional state. Cowie [1] lists some of these possibilities, detailing specific areas where emotion recognition would be beneficial, out of which the convergence of a machine's communicational register to conversation specific standards would be a major catalyst for the development of a system capable of performing what is described in the here presented topic of research.

Human users of an automatic emotion/sentiment recognition aid system could also, for example, greatly benefit from the fact that such a system would be capable of alerting themselves or others of emotional state swings. Additionally, such a functionality would also permit some degree of behavioral adaptive plasticity in an interactive system, not only as the aforementioned convergence of communicational registry but also as the generation of more genuine machine speech.

Speech, in a broad sense, refers to the human ability to convey some idea or state of mind as a mixture of utterances. It is, therefore, only natural that speech would contain highly relevant emotional and sentiment information convolved with its many variable features. Extraction of that information from speech remains a complicated task and mostly overlooked when compared with emotion/sentiment analysis in text [5] and facial expression. Hence, automatic recognition of human emotions through speech analysis remains an interesting and opportunity rich research topic.

This dissertation is divided in the following manner: This section describes what were the main objectives of this project and includes an overview of recent related work. Following that, section 2 outlines the methodology of the approach used, delving into the specifics of the proposed technique for emotional feature extraction. This is followed by

section 3 where detail is given about the experiments carried out and the obtained supporting results. Proceeding, an analysis and discussion of the results is made on section 4. Finally, a conclusion is available in section 5 and a developed demo is described in section 6. An overview of future work is presented in section 7.

## 1.2 Main Objectives

Despite the subjective nature of emotion recognition and the lingering disagreement among researchers as to how emotions can be classified, human beings, whether due to their nature or nurture, are still inherently able to identify different emotional states among each other, in most cases with a relatively high accuracy rate. As according to Cowie [1], we as people have scored recognition rates of 90% or above on emotion databases, something that is not surprising given the fact that we are able to extract information from a panoply of distinct modalities. Seeing that this project focused mainly on speech analysis and feature extraction from utterances, it was only natural to expect lower recognition rates than those obtained by humans. Nevertheless, acceptable results corresponded to rates of 70% or greater, which were in fact obtained on this project as expected. Such achievements prove the viability of machine learning techniques as means of emotion recognition. Plus, obtaining these high recognition rates should allow for several practical uses of an automatic emotion recognition system, as previously specified.

Further, by incorporating speaker adaptation into emotion recognition, it was hoped to obtain boosted results with regards to emotional state accuracy. Consequently this would support the notion that a person's prosody is affected by their current emotional state, which would in turn back the idea that exploring more speaker and context intrinsic speech information may be worthwhile for SER.

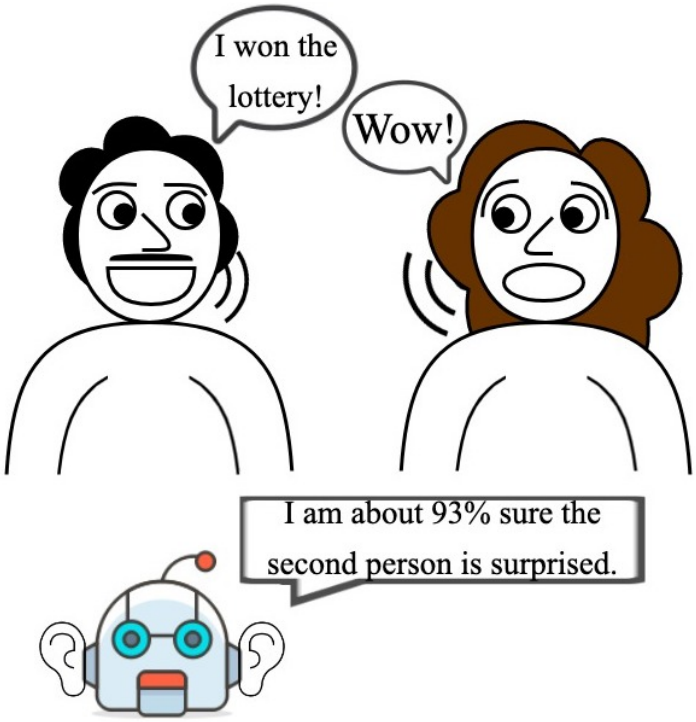



**Figure 1.1:** An idealization of how a situation of automatic speech emotion recognition would play out in the real world with an interactive system.

## 1.3 Related Work

The recognition of human emotion is not exactly a new field of research as is. This topic has long been studied and many techniques have been developed to assess a good range of modalities, such as motion fashion or facial expression, the latter of which has long been the sub-category of this topic given the most attention. Speech analysis for emotion recognition, however, is a relatively new area of interest, with some past research techniques being based on conventional classification methods using feature extraction or crafting.

Though these past methods, and other variants, have achieved significantly good results, their performance has been gradually surpassed by machine learning models and neural networks (NN), which have recently been gaining more and more advocates, due to their ability to model real neural/nervous systems and directly discern emotions. As such, the state of the art is currently accredited to machine learning techniques such as neural net-

works, and the various learning and data processing techniques it encompasses. Recently, Shahsavarani [6] developed a CNN architecture with 2 convolutional layers followed by a fully connected layer, to assess wideband spectrograms of emotional utterances, justifying these with glottal pulse being associated with one period of the vocal fold vibration. Results on both real and acted emotion databases were optimal in the 90-100% range, with the better ones being achieved by a combination of drastically higher quantities of training epochs, larger convolutional kernel in the first layer, data augmentation and average pooling.

Differently, Lotjidereshgi *et al.* [7] devised a novel method involving the use of a Liquid State Machine (LSM) for emotion recognition. In it, the speech signal was divided into two inputs for standard integrate-and-fire neuron reservoirs. One for the vocal tract, following the extraction of Linear Predictive (LP) coefficients and Equivalent Rectangular Bandwidth (ERB) scalling of those coefficients, and another for the source, obtaining the residual correspondent to said LP coefficients and further decomposing it through a gammatone filterbank and ERB scalling. The the Spiked NN, trained with Asymmetric Spike Time-Dependent Plasticity (STDP) to adapt the conductance of the synapses throughout the speech sample, then uses Principal Component Analysis (PCA) to average the activity of the neurons from each reservoir and Linear Discriminant Analysis (LDA) for final recognition. Results of this new technique faired around the 80% rates, making it highly comparable to other state of the art techniques.

An interesting approach by Lim *et al.* [8] involved feeding Short Time Fourier Transform representations of speech utterances to a CNN with two convolution and max pooling layers, followed by a fully connected layer which would forward them to a two layered Long Short-Term Memory (LSTM) recurrent neural net (RNN) architecture, in order to synthesize sequential dynamics in the speech signals and extract valuable information embedded in the temporal properties of said signals, resulting in a kind of time distributed CNN. The performance of this combo architecture was tested against simple CNN and simple RNN versions of the same technique, both having been surpassed by the combination. Results averaged at around 87%, deeming the performance of this method quite acceptable for a machine learning approach to this topic.

On a side note, Chenchah *et al.* [9] presented their take on speech emotion recognition, addressing the negative influence a noisy acoustic environment, expected in real life conditions, may have on recognition systems. An emotion recognition approach was proposed, based on emotion modelling using the common feature extraction technique of Mel Frequency Cepstral Coefficients (MFCC), followed by speech classification with a 3-state left-to-right Hidden Markov Model with Gaussian Mixture. Following their goal, an initial step was also taken employing three different noise reduction techniques and evaluating their effects on the overall recognition system. Based on experimental results using the IEMOCAP database [10], spectral subtraction proved to be the only method capable of improving the obtained recognition rate, typically raising them by around 2-3% for airport and babble noise, but showing no significant improvement for car and train noise.

Trigeorgis *et al.* [11] proposed a new method for end-to-end spontaneous emotion recognition, addressing the problem of context awareness and the idea of it having relevant feature information which should be considered. The method consisted of a recurrent NN with Long-Short Term Memory (LSTM) cells, which first performs temporal convolution on the data twice, with the results being pooled accross time inbetween convolutions. This would be followed by max pooling and feeding to the recurrent layers. The concordance correlation coefficient was employed in the objective function used for training, instead of the common mean square error technique, in order to evaluate the agreement level between the predictions of the network and the gold-standard derived from annotations. Testing of this new method with the RECOLA database [12] showed significantly better performance than other state of the art techniques such as Support Vector Regression and Bidirectional LSTM deep recurrent NN, even though it evaluated raw signals whereas these other methods used extracted features from said signals.

A highly compelling approach was presented by Deng *et al.* [13], based on a novel unsupervised domain adaption method for emotion recognition named *Universum* Autoencoder, capable of combining knowledge acquired from both labelled and unlabelled data. In it, an encoder made up of several feed-forward NN layers builds complex representations of the data it receives, passing on this information to both a decoder, also made

up of several feed-forward NN layers, and a *Universum* learning path, consisting of yet another feed forward NN. Following, the decoder rebuilds the data as closely as possible, based on the complex representations it has received from the encoder, by minimizing the reconstruction error. Parallely, the learning path minimizes a total error based on a tradeoff of minimization of classification errors and maximization of contradictions, on label and unlabelled data, respectively. As the main goal, the model's complex representations are optimized by minimization of a linear combination of the total error seen by the learning path, with the reconstruction error of the decoder. This new technique employed several well known emotional speech databases for training, and its performance was compared to that of five popular state-of-the-art recognition methods in several situations, among which some where said popular methods performed poorly. The *Universum* Autoencoder method significantly outperformed all other techniques, averaging a 7-9% increase in unweighted average recall (UAR) from mid 50% to low 60% values, which proves its efficacy and the importance of domain adaptation in speech emotion recognition scenarios.

Additionally, Peng *et al.* [14] introduced their novel auditory-inspired take on the matter, involving an end to end system made up of a two-layer recurrent neural network preceded by two 3D convolutional layers, which extract 3D features such as acoustic and modulation frequency, and model spectral-temporal (ST) representations of the data using said extracted features, respectively. This path was followed as it was previously observed that different emotions have distinct ST modulation representations. The two recurrent layers are then, in order, employed to obtain short-term and utterance-level dependencies, respectively. The recognition system's performance was tested using the IEMOCAP database, having demonstrated a 5-10% increace in accuracy rates, with respect to those of other state of the art techniques, effectively proving the importance of sequential dependencies information for recognition of emotions in speech. Further, it was observed that a 3D CNN by itself had worse performance than the proposed model due to the lack of a recurrent layer.

# 2

# Methodology

Considering the extensiveness of this work, several topics were approached during its development. As such, this section aims at providing a theoretical background to each of the topics, as well as detailing their roles in the overall system constructed.

## 2.1 The Multiclass Classification Issue

This topic refers to the matter of assigning a specific class from a closed set of classes, to a data instance based on the use of models built using previously observed instance-class pairs. The problem can be symbolically formulated as attempting to build a model $H$, using a training dataset of pairs $(\mathbf{x}_i,y_i)$ where $\mathbf{x}_i \in \mathbb{R}^n$ corresponds to the *i*th instance, and $y_i \in \{C_1,...,C_K\}$ corresponds to the *i*th label, chosen from a closed set of $K$ labels. The model should then function successfully with new unseen instances by accurately assigning them labels as in $H(\mathbf{x}_i) = y_i$. It differs from binary classification in that at least three classes exist for instance classification, instead of exactly two. Nevertheless, by considering $K = 2$ (two classes, usually represented by values $0$,$1$ or $+1, -1$), multiclass can generalize binary classification.

Were it possible to know the density $p_j(\mathbf{x})$, for each of the $K$ classes, the prediction model's loss function would simply be seen as:

$$H(\mathbf{x}) = arg \quad max_{j\in 1,...,K} p_j(\mathbf{x}) \tag{2.1}$$

However, given the considerable complexity of estimating densities in high dimensions when the available data is limited, the multiclass classification issue becomes one of finding an adequate loss function or extending binary classification techniques to work on multiclass classification problems. These binary approaches may be naturally extensible, or be based on a separating function rather than density estimation. Extension approaches include Error-correcting output code, Single Machine and the following All-versus-All (AVA) and One-versus-All (OVA) techniques.

### 2.1.1 All-versus-all (AVA)

For this approach $K(K-1)$ separate binary classifiers $H_{ij}$ are constructed, where $i$,$j \in \{C_1,...,C_K\}$ classes and $i \neq j$, in order to distinguish each class pair. Considering class $i$ corresponding to positive examples and class $j$ corresponding to negative examples, classification is performed using:

$$H(\mathbf{x}) = arg \quad max_i\Big(\sum_j H_{ij}(\mathbf{x})\Big) \tag{2.2}$$

### 2.1.2 One-versus-All (OVA)

For this approach $K$ separate binary classifiers $H_i$ are constructed, where $i \in \{C_1,...,C_K\}$ classes. Considering the $i$th classifier, the positive cases correspond to all points of class $i$ and the negative cases correspond to all points not in class $i$. Classification is the performed using:

$$H(\mathbf{x}) = arg \quad max_i H_i(\mathbf{x}) \tag{2.3}$$

### 2.1.3 Extensible Classifiers

The following binary classification methods are either naturally extensible or rely on some supplementary technique such as AVA or OVA.

#### 2.1.3.1 Naive Bayes

This Bayesian statistical classifier gets its basis from maximum *a posteriori* (MAP) estimation and is naturally extensible to multiclass applications. Considering our set of classes $\mathbf{C} = \{C_1,...,C_K\}$, with probabilities $P(C_1),...,P(C_K)$, a label $l \in \mathbf{C}$ can be assigned to some new data instance with $N$ features $\mathbf{x} = \{x_1,...,x_N\}$ by choosing the class with the MAP probability, given previously observed data. Symbolically, the model becomes:

$$H(\mathbf{x}) = arg \quad max_l P(C = l|x_1,...,x_N) \tag{2.4}$$

Evidently, the MAP probability $P(C = l|x_1,...,x_N)$ is obtained using Bayes Theorem. However, since computing the class conditional probabilites of the features given the class set would be a laborious task due to feature interdependence, class conditional independence is assumed. As such, the model can be further simplified into:

$$H(\mathbf{x}) = arg \quad max_l P(C = l)P(x_1|C = l)...P(x_N|C = l) \tag{2.5}$$

A visual representation of a Naive Bayes classifier using the traditional Venn diagram application to Bayes theorem is shown in Figure 2.1 to ease understanding of the classification process.

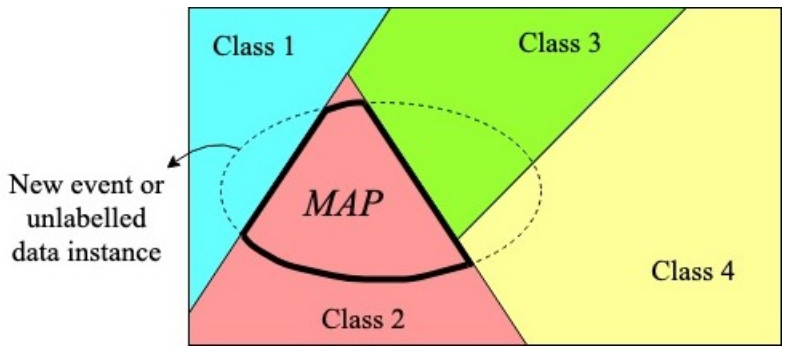



**Figure 2.1:** A Venn diagram like visualization of a Naive Bayes classifier, as similar to the common representation of Bayes theorem. Using previously observed data and considering class conditional independence, the classification space is divided. A new instance is then assigned the class which yields the maximum a posterior likelihood (MAP).

#### 2.1.3.2 *k*-Nearest Neighbours (*k*NN)

This simple classifier is also naturally extensible and in it each $N$ features vector is considered an $N$-dimensional point. In order to classify a new data instance, some metric is applied to measure the distance between its feature vector and all other feature vectors from the training data. By considering the $k$ smallest distances (neighbors), the class represented by the most neighbors is assigned to the new data instance. Let $D = \{d_1^{C_i},...,d_k^{C_j}\}$ be the set of $k$ smallest distances between the new instance and all training data, each corresponding to a class of set $\mathbf{C} = \{C_1,...,C_K\}$. Let $n_i$ be the number of times class $C_i$ is represented in $D$. The model is then:

$$H(\mathbf{x}) = arg \quad max_{C_i} n_i \tag{2.6}$$

An optimal value for $k$ is usually obtained using either cross-validation or some validation set. A visual representation of this classifier in 2-dimensional space is shown in Figure 2.2.

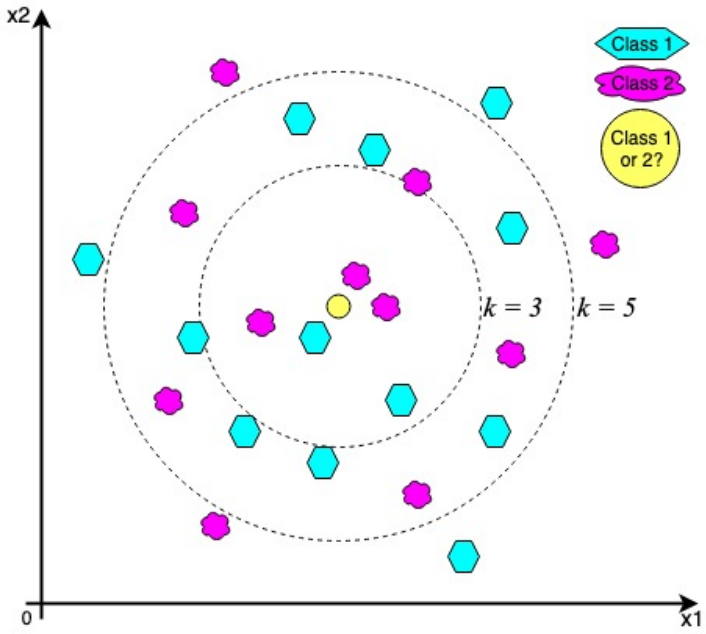



**Figure 2.2:** Example of a kNN classifier in 2-dimensional space, where $k = 3$ and $k = 5$ cases are considered. It can be understood how by considering the $k = 3$ case, the new yellow instance would be assigned to class 2, while in the $k = 5$ case, it would be assigned to class 1.

#### 2.1.3.3 Decision Trees

These high-level tree resembling structures are attempts to produce considerably good data generalizations using the feature set. The nodes that make up the tree can be either splits or leaves, leaving the remaining path inbetween two to be considered a branch. Leaf nodes correspond to classes, and an instance is assigned a class by following a path from the root of the tree to a leaf node. This path is defined by performing tests on the instance features at each node.

Depending on the amount of splits performed a varying number of leaf nodes, and consequentely classes, can be obtained. Hence these methods may handle both binary and multiclass classification problems. A visual representation of a simple decision tree is shown in Figure 2.3.

Further, during training, metrics are used so that a split node is formed based on the feature which provides the most information gain (i.e. the feature showing data separation most clearly). These metrics are implementation dependent and as such, several types of decision trees exist. Some examples include Random Forest, an algorithm which builds several decision tree classifiers with randomly formed subsets of data and makes its final decision based on a metric involving the decisions of all trees, and Logistic Model Trees (LMT), decision trees which besides running an instance through a branch path, also build logistic regression functions at the leaves, which then make the predictions, as aproximations to the target functions.

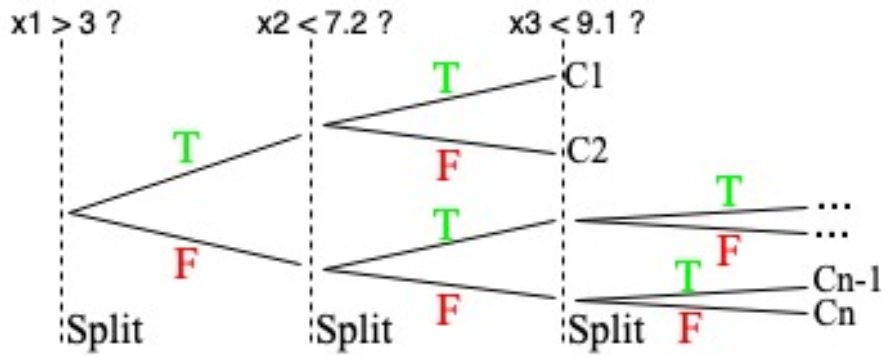



**Figure 2.3:** A decision tree diagram showing 5 splits total, based on evaluations of data instance features. As it is clear, a new instance is assigned the class of the leaf node it reaches after a certain number of feature evaluations.

#### 2.1.3.4 Neural Networks

Given the importance of Neural Networks to this work, these will be further detailed in a following section. Nonetheless, it terms of binary extensibility, these are based on using $K$ neurons in the output layer, corresponding to the $K$ classes, instead of a single binary neuron. This allows the classification, for example, to take place using the generated binary word rather than a single binary digit.

### 2.1.4 Support Vector Machines (SVM)

This binary classification algorithm is used for finding the hyperplane which best separates all data between two classes, considering the feature vectors as $N$-dimensional points. Optimization is performed by maximizing the minimum distance between the separating hyperplane and the data poitns closest to it, which are called *support vectors*. As such, the best hyperplane corresponds to whichever one produces the highest margin between the two opposing classes.

Considering the $N$-dimensional training dataset of $(\mathbf{x}_i,y_i)$ pairs, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1,1\}$, and a generalized hyperplane $h(x) = \mathbf{x}'\beta + b = 0$, with $\beta \in \mathbb{R}^n$ and $b \in \mathbb{R}$, then the idea is to compute the optimal $\beta$ and $b$ which minimize $\|\beta\|$ and guarantee $y_i h(\mathbf{x}_i) \geq 1$ for all instances. In the case of *support vectors*, $y_i h(\mathbf{x}_i) = 1$. Given the constraints and the minimization goal, the problem becomes one of quadratic programming and many implementations are, therefore, possible. Let $(\beta_o,b_o)$ be the optimal solution pair, and $\mathbf{x}$ be an unseen data instance, the model may classify it by:

$$H(\mathbf{x}) = sign(\mathbf{x}\beta_o + b_o) \tag{2.7}$$

A simple diagram exemplifying a SVM classifier in 3-dimensional space is shown in Figure 2.4.

Essentially the class decision is obtained using the sign of the distance between the unseen instance and the boundary. When it happens that the data is nonseparable, the largest

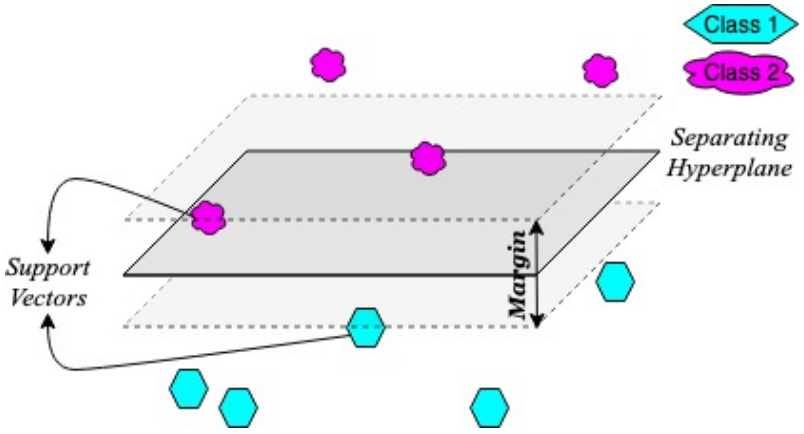



**Figure 2.4:** Example of a binary SVM classifier in 3-dimensional space. An unclassified data instance would then be assigned a class depending on which side of the hyperplane it would be placed on.

possible *soft* margin is considered, one which may contain data points within, in place of a regular margin, which cannot. This involves further use of quadratic programming concepts such as *slack* variables and a penalty parameter. Some approaches involving the extension of data into higher dimensions, much like how it is done with other classifiers, have also been proposed. A visual example of how this is achieved is shown in Figure 2.5.

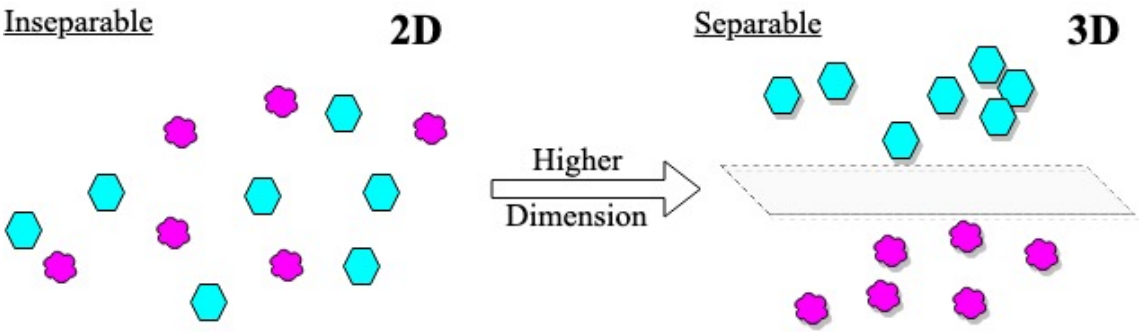



**Figure 2.5:** Visualization of how considering an extra dimension may solve the data separability issue. On the left, data in 2D space seems inseparable. But by considering a 3rd dimension (changing perspective), the data becomes clearly separable.

SVM extension to a multiclass application is done by breaking it into multiple binary classifications. Techniques such as AVA and OVA may be employed, though in different ways. In the case of an OVA approach, the constructed SVM which provides the highest output function for a new instance gets to predict its class. Alternatively, for an AVA approach, each of the SVM assigns one of their two classes to the new instance. As such,

each class will get a certain number of votes from the entire SVM set, and the final prediction is the one corresponding to the class with the most votes.

## 2.2 Audio Spectral Representations

Spectrograms, visual representations of the scope of frequencies present in a signal as it varies with time, have long been an essential tool in the field of signal processing. By depicting energy variations at different frequencies, over time, these images pose as a great option for accurate representation of speech features such as tone, pitch, volume and others. As previously explained, significant emotional information is embedded in these speech features, meaning that having representations of them will enable their analysis from speech recordings, and ultimately allow for an assessment of the labelled emotional state.

These representations can be either wideband, with lower frequency resolution but profiting from higher time resolution, or narrowband, with higher frequency resolution but at the cost of lower time resolution. Both variations can be appropriate, depending on the situation at hand and what needs to be evaluated as, for example, narrowband spectrograms can resolve individual harmonics whilst wideband spectrograms are more suitable for showing individual glottal pulses (spectral peaks due to vocal tract).

Using the SAVEE database (see Table 2.1 later on), example spectrograms were generated for the same phrase by the same speaker, depicting two distinct emotional states, in the same environment, shown in Figures 2.6 and 2.7. The differences, which can be seen when comparing both spectrograms, support feature disparity between utterances.

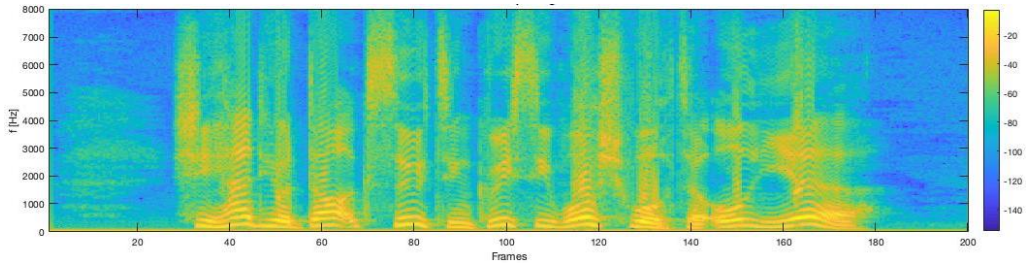



**Figure 2.6:** Happiness Version Spectrogram

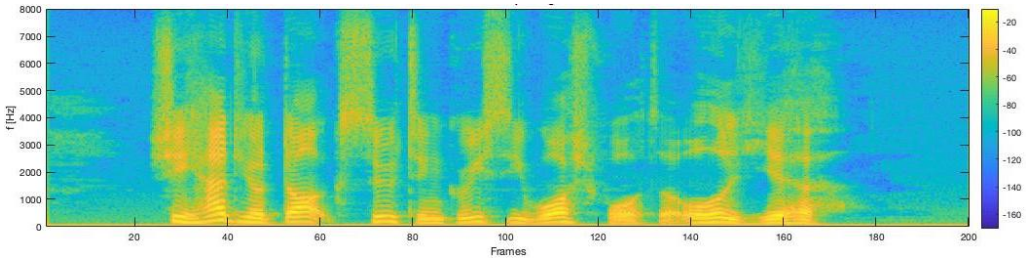



**Figure 2.7:** Sad Version Spectrogram

## 2.3 Machine Learning and Neural Networks

The topic of Machine Learning refers to the use of algorithms which are capable of, in a sense, learn to perform a certain task/skill from knowledge gained by observing sets of training data. These algorithms vary greatly and are generally quite adaptive to different situations, with some exceptions. By analyzing extensive amounts of data, machine learning systems are able to construct complex representations of that data, continuously improving them until finding an ideal middle point between underfitting and overfitting of training data. This optimal place is usually identified by other techniques also employed in the system. Post training, these systems are usually employed in evaluating the similarity of other data with respect to the data used during training.

The learning process involved in training of the machine learning framework can be divided into three categories: supervised, unsupervised and reinforcement. In the first case, the training data incorporates the desired response in the form of labels, meaning the system will learn a function based on the labels it has seen before. Contrastingly, in the second case, the data fed into the system for training contains no labels, and the framework attempts to find some patterns rather than trying to learn how to recognize and label the data. In the third case, learning does not require labelled data either. Instead, much like the name implies, the framework is reinforced to find an ideal place between usage of knowledge it already has and acquisition of new knowledge, based on some technique.

Artificial Neural Networks (ANN), a sub-topic of machine learning, have long been an interesting area of research, as they are closely modelled after the biological neural networks which make up the brain. As such, these networks are capable of, to some

degree, learn a few of the same functions that their biological counterparts perform, such as emotion and speaker recognition. The networks are organized by layers, one input, one output and a certain number of hidden layers. Each layer is made up of a certain number of neurons, and the connections between neurons of different layers depend on the network structure. An example diagram of an ANN structure is shown in Figure 2.8. A simple representation of a node's anatomy is shown in Figure 2.9.

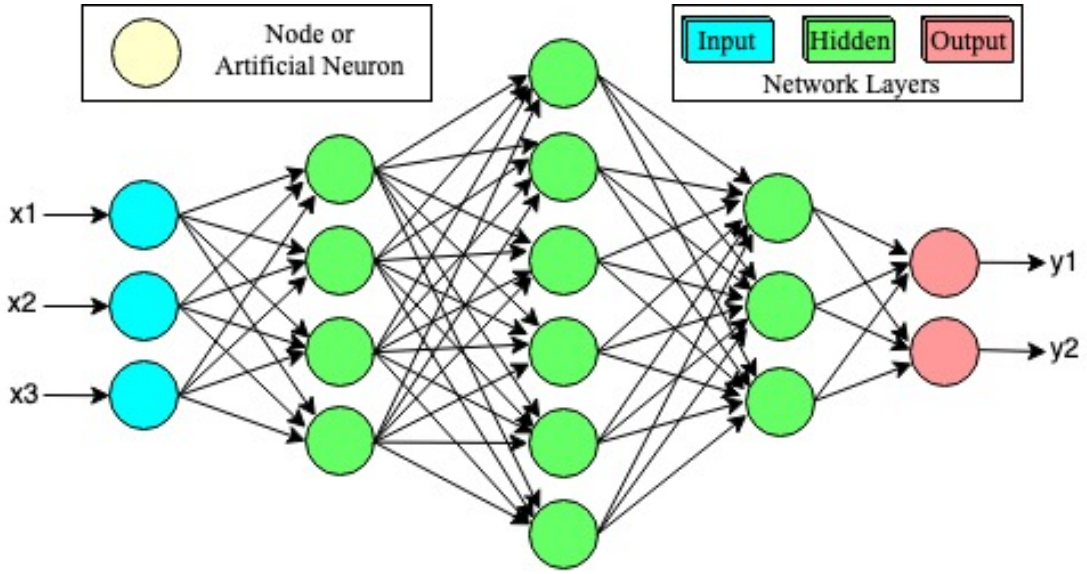



**Figure 2.8:** Diagram of a simple artificial neural network (ANN), including one input and one output layers with threen hidden layers inbetween.

### 2.3.1 Learning Stage

All neural networks encompass a stage focused solely on learning from training data. This stage is related to two types of functions, activation and cost, and happens continuously by propagation of data through the network, followed by backwards propagation through each layer of the error between the expected and calculated outputs (backpropagation). The forward propagation of data is simply done by applying a layer's weights to its input data, producing output results which are fed to the next layer as input, or as a final result in case of the network's output layer. The application of weights is done at each neuron and usually by multiplication of the weights and the data followed by a general summation (linear combination). The result of this is then passed to the neuron's activation function, which defines its output.

#### 2.3.1.1 Activation Function

These functions, which are present at each node, are mostly responsible for mapping the values they receive from weight application to a specific range of values. Considering a binary network example, one that deals solely with binary digits, each activation function will always output either 0 or 1 to the next neuron.

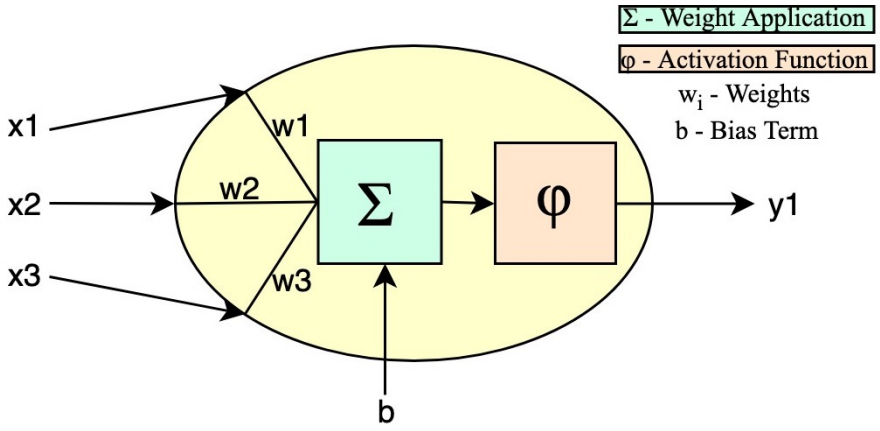



**Figure 2.9:** Overview of the system in place within a network's node/neuron. The inputs are fed to the node, which applies the weights to them as a linear combination, passing the result of this to the activation function which yields the output.

Given the characteristics of these functions, they are often used for final classification in the last layer of a network. By adapting the range of mapped values to the number of classes, these functions can be used to dictate the final result of a prediction. A common activation function used for these cases is the sigmoid function, a smooth and differentiable function bounded by 0 and 1:

$$\sigma(x) = \frac{1}{1 + e^{-x}}, x \in \mathbb{R} \tag{2.8}$$

However, when it happens that a problem is multiclass, the softmax function becomes a more suitable option. This is because it calculates the probability of each target class over all possible target classes, assigning the result to each corresponding class. It then

becomes advantageous as the sum of all probabilites will equal 1, and the class with the highest probability will be the one assigned to the new data instance. Softmax is usually implemented using:

$$h(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}, i = 1,...,K \tag{2.9}$$

Here K corresponds to the number of classes in the addressed problem.

Considering backpropagation, in order to have the computed vs. expected error propagate back in the network, it must first be calculated. As many other techniques, this is generally based on the usage of a loss or cost function, whose values the network aims to minimize. A common choice for a loss function in multiclass classification problems where an ANN undergoes supervised training is cross-entropy, a relation between two probability distributions based on the concept of entropy, which will be presented later on.

#### 2.3.1.2 Entropy

This highly useful metric is tightly coupled with the probability distribution of a class and can be interpreted as a way to understand the information's degree of randomness in some process. From the moment this measure of unpredictability is known, it also becomes possible to assess the minimum size necessary to encode all the information of that process, which is another interpretation of entropy. As such, the concept is defined as the expectation (average) of the negative log probability of some distribution $P$, or symbolically:

$$S = \mathbb{E}_{x\sim P}[-\log P] = -\sum_i P(x_i) \log P(x_i) \tag{2.10}$$

Here $x$ refers to some discrete event, which corresponds to a class in the context of machine learning, with a probability distribution $P$. However, as stated before, there is the issue of not knowing the distributions of the several classes in classification problems. As such entropy cannot be directly calculated and must instead be estimated. Considering

2.10, to estimate entropy an estimate of the corresponding probability distribution must first be provided. In any case, when using a distribution estimate $P'$, entropy becomes affected by uncertainty at two different points:

- The expectation of the negative log probability assuming $x \sim P'$;
- The negative log probability itself $-\log P'$;

Given these, the estimated entropy can be overly wrong and far from its real value, meaning it may not at all be useful as a metric. Hence the use of cross-entropy instead, in multiclass classification problems.

#### 2.3.1.3 Cross-Entropy as a Loss Function

In the context of neural networks and supervised learning, cross-entropy can be used as tunning metric of a network during its training phase. Considering a set of labelled training data, by using some technique to encode the data, such as one-hot encoding, then the real probability distribution $P$ of all classes can be estimated. Further, the larger a training dataset is, the more accurately the probability distribution of the classes will depict the real world.

Following that, considering now a machine learning model such as an ANN, without much training, it will classify data with low confidence. This naturally means that by using the results of the model's classification alone, a new and imprecise probability distribution $Q$ of all classes will be obtained. Cross-Entropy can then be calculated for some class as:

$$H(P,Q) = \mathbb{E}_{x\sim P}[-\log Q] = -\sum_{j} P(j)\log Q(j) \tag{2.11}$$

As a way to interpret this formula now, it should first be considered that as training continues, the model's prediction confidence in each class will improve. That is, the neural network estimated distribution $Q(j)$ will increase for a specific value of $j$, approaching 1, and decrease for all other values of $j$. Noting also the one-hot enconding for the real probability distributions, then only one of the $P(j)$ values will be 1 while all others are 0. Bearing this in mind, it is clear how provided the value of $j$ for which $Q(j)$ is increasing

towards 1 is the same for which $P(j) = 1$, then $H(P,Q) = -(1 \times \log Q(j) + 0 + ... + 0)$ will decrease to 0.

In the case of binary classification, cross-entropy can be greatly simplified. Considering only two opposing classes $c_1$ and $c_2$, then cross-entropy can be obtained as:

$$H(P,Q) = -\sum_{j=(c_1,c_2)} P(j) \log Q(j) = -P(c_1) \log Q(c_1) - P(c_2) \log Q(c_2) \quad (2.12)$$

Given how $P(c_1) = 1 - P(c_2)$, then the formula can be further simplified to use the values with respect to only one class. This results in a much more understandable formula:

$$H(P,Q) = -P(c_1) \log Q(c_1) - (1 - P(c_1)) \log(1 - Q(c_1)) \quad (2.13)$$

Finally, it can be concluded at a higher level how cross-entropy does go down as predictions get more and more accurate, making it a highly suitable candidate for a loss function used in the training of a neural network.

### 2.3.2 Testing Stage

At this point, a network has completed the learning phase and is ready to make predictions on new unseen data. In order to accomplish this, unlabelled instances of this new data are simply forward propagated in the network, resulting in a class being assigned to each of them by the output layer of the network. No backwards propagation of error or weight altering is performed, as no previous label information is available, and the final prediction is taken as true.

In case some validation testing is desired, labelled data not used during training may be propagated through the network. The predicted labels are compared against the real labels of the validation data instances, producing some accuracy metric of the network's performance.

### 2.3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) [15], [16] are a particular type of neural networks, possessing one or more hidden layers which, potentially among other activities, perform the mathematical operation of 2D convolution on the data they receive. By doing so, CNNs are able to accurately follow the structure of the visual cortex, the brain sector in charge of receiving and processing visual information taken in by the retinas in the eyes. Hence, these networks allow for a better response to stimuli of visually specific locations [6], meaning even slight changes in speech features during an utterance could be perceived by the neural network when analyzing the visual representations that are spectrograms. Therefore, a CNN architecture was the key aspect of the recognition system used here for development.

## 2.4 Database Acquisition

Areas such as this one naturally require some form of testing data. This data can be fed to the recognition system in order to measure its accuracy rate. Furthermore, given the fact that machine learning techniques were employed, copious amounts of training data also become a necessity. In order to satisfy this need for big data, 6 well-known and established emotional speech databases were gathered, whose basic metadata is shown below in Table 2.1, along with some other important details.

Emotion, as it can be considered a highly personal aspect of human life, is by nature a hard element to collect in the form of a database. As the majority of people is unable to properly portray a real instance of emotion at will, or when some emotional state does unfold they would rather not have that moment be recorded, researchers are mostly left with no other option but to employ the help of actors, whose job is literally to convincingly mimic any requested emotional state. Though real human emotion databases exist, these are somewhat arduous to find and are never public, making them extremely difficult to acquire. As such, the databases below were all created with the aid of actors.

Almost all available labels were used for fundamental testing, those being the archetypal

emotions *Anger*, *Joy/Happiness*, *Sadness*, *Fear*, *Disgust*, *Surprise* and *Neutral*. Other labels were used only for complementary testing. Care was also taken in gathering databases of different languages so as to keep the databases diverse and prevent the recognition system from being language-dependent.

The databases showing a sampling rate higher than $16kHz$ saw their clips suffer downsampling to this very value, as it was intended to use all databases in the same conditions. Also, it has been shown that $16kHz$ is enough to represent all the relevant information embedded in human speech witout significant losses, making it ideal for human voice analysis.

**Table 2.1:** Catalog of emotional speech databases.

| Database | Access | Language(s) | Total Duration [H] | Avgerage Clip Duration [S] | Number of Clips | Sampling Rate [kHz] | Labels | Source | References |
|---|---|---|---|---|---|---|---|---|---|
| EMODB | Public | German | 00:24:47 | 2.7796 | 535 | 16 | Anger, Joy, Sadness, Fear, Disgust, Boredom, Neutral | Professional Actors | [17] |
| EMOVO | Public | Italian | 00:30:35 | 3.1210 | 588 | 48 | Anger, Joy, Sadness, Fear, Disgust, Surprise, Neutral | Professional Actors | [18] |
| SAVEE | Public | British English | 00:30:42 | 3.8395 | 480 | 44.1 | Anger, Happiness, Sadness, Fear, Disgust, Surprise, Neutral | Volunteer Actors | [19] |
| RAVDESS | Public | American English | 01:28:48 | 3.7007 | 1440 | 48 | Anger, Happiness, Sadness, Fear, Disgust, Surprise, Calmness, Neutral | Professional Actors | [20] |
| RML | Private | English, Mandarin, Urdu, Punjabi, Persian, Italian | 01:01:13 | 5.1027 | 720 | 16 | Anger, Happiness, Sadness, Fear, Disgust, Surprise | Volunteer Actors | [21] |
| ELRA S0329 | Private | Spanish | 07:52:36 | 4.6941 | 6041 | 16 | Anger, Joy, Sadness, Fear, Disgust, Surprise, Neutral | Professional Actors | [22] |

## 2.5 Speaker Adaptation

It is self-evident that the emotional phenomena experienced by a person should tend to mold their behavior and conversational register in the social settings they engage with. Effectively, by adulthood most humans will have developed a large set of highly nature/nurture dependent, distinct behavioral responses to the multiple emotional states they experience throughout their personal lives. An emulated corroboration of this notion can be inferred by the particular articulation patterns observed in professional actors simulating emotion [23]. Further support is given by the widely accepted effects of recurrent stress in an individual's emotional state [24], [25] naturally affecting their prosody.

The role of someone's personality, and consequently their way of communicating, are often overlooked when it comes to emotion recognition from speech. In fact, even though the state of the art machine learning models are exceptionally competent at evaluating data

with an unspecified set of relevant features, most of the designs tend to focus directly and solely on overt emotional cues. As of now, this should be considered an erroneous approach given the observed higher performances of systems with deeper adaptation levels, such as domain-based [13], [26] or context-based [11], [8], [27]. Hence, more information channels should be considered when analyzing emotion in speech. One of these is the potential speaker dependency of human emotional prosody.

Empirical evidence of prosody variations for identical emotional sates in different people is also of particular relevance to interactive systems and other social robotic applications. Provided a machine is capable of identifying an intervening speaker, the ability to further adapt to the speaker themselves and their emotional conveyance mannerisms, can certainly boost the quality of the system's behavior and response suitability to the situation at hand. This concept is not unlike how pet social companions are able to perceive how their owners feel and modify their behavior to conform with the respective emotional state. Applications would be many and varied, from home assistants such as *Google Home* or Amazon's *Echo* (see Figure 2.10), to active ageing and nursing aides in the medical field. Thus, it is of great usefulness to any field of inteligent robotics that emotion recognition techniques encompassing some level of specific speaker awareness and adaptation be researched.


**Figure 2.10:** Examples of already existing social companion systems which would benefit from speech emotion recognition (SER). From left to right, Google Home, Amazon's Echo and Jibo ©.

Considering what has been observed, it becomes clear the field of speaker recognition also requires some attention in order to be successfully incorporated in a speech emotion recognition system. Through adjustment of a technique from this related area, speaker

adaptation becomes possible, given how it may be considered a subset of speaker recognition.

### 2.5.1 Recent Work

Due to the fact that speaker recognition as a field of research has been around for a reasonable amount of time, many diverse approaches have been proposed, with varying success rates. The earlier portion of these included classifier methods of unsupervised training such as Vector Quantization (VQ), where a test vector is assigned, by comparison, to a group of centroids of the data cluster from feature space the vector is closest to. Also, the well-established kNN method, where the distances between a test vector and k given training vectors are computed, being the test vector assigned the most common label among the obtained k neighbors (i.e. the ones with the smallest distances). Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) have also been employed in this sub-area of pattern recognition.

Much like the case with emotion recognition, the performance of these past methods has largely been surpassed by machine learning models and their advocates, as these models are able to more accurately mimic biological neural systems and more easily distinguish speakers. As such, the state of the art is currently accredited to deep neural networks (DNN) and their various training techniques when it comes to classifiers, while total factor analysis (i-vector [28]) dominates speaker modelling. One of the earlier examples was published by Farrel *et al.* [29], who developed a modified neural tree network (MNTN), combining the concept of decision trees with neural nets to improve the hierarchy and decision structure, ultimately resulting in lower computing time. This process learned by assigning a vector to some class and related weight vector, through minimizing of the corresponding classification error, and showed novelty at the time by using forward pruning of the decision tree to reduce overfitting. Testing on the TIMIT database [30] showed the MNTN to outperform older classifier methods, such as VQ and kNN, though these results might have been even better would the computing power of now be available at the time.

Following a different path, Wang *et al.* [31] developed a simple combo of mel-frequency cepstral coefficients (MFCC), which highlight the relevant feature information present in the frequency domain while also disregarding noise interference, with the work done by Rumelhart *et al.* [32] on back propagation (BP) neural networks (NN), which continuously adjust weights of connections in the net in order to minimize the difference between training and testing vectors. Data was gathered in different environments to improve training quality and results proved fruitful for a small number of closed-set speakers, though a larger number of hidden layers might have proved more optimal.

Chang *et al.* [33] presented their novel work of associating an i-vector framework for speaker recognition, a leading approach in this field, with speech separation using DNNs trained for ideal SNR mask estimation. After separation of utterances using said mask, the corresponding spectrograms were fed into both a GMM/i-vector and a DNN/i-vector systems, the latter attemping to minimize a cross-entropy loss function during training. MFCC and other types of cepstral coefficients were used as evaluation features, and the experiments proved the systems' usefulness in removing additive stationary noise but only to a limited SNR range. Nonetheless, the DNN system outperformed the recognition rate of its GMM counterpart.

An interesting and noteworthy approach to the issue at hand was done by Zhenhao *et al.* [34]. In their work, a feed-forward neural net architecture was used for classification with features extracted from MFCCs and normalized with their own mean and variance. Preprocessing of data involved voice activity detection (VAD) mechanisms, and training of the network was done through forward-BP using a dynamic regularization scheme to prevent overfitting and allow further iterations to refine the model. A cost function of binary classification using logistic regression was considered for the NN rather than cross-entropy, and other concepts such as prediction score normalization and speaker-specific thresholding were also studied. The overall system was tested across the TIMIT database for both speaker classification and verification, and results were shown to be optimal for short utterances and still reasonably good for longer duration utterances, even with a high number of imposter speakers relative to the closed-set size.

Given the importance of i-vectors in speaker recognition, and since these can be neg-

atively affected by variations caused by channel effects (e.g. noice, utterance duration), Yao *et al.* [35] set out to surpass already existing channel compensation methods such as linear discriminant analysis (LDA) and Gaussian probabilistic LDA. In their work, the team coupled together signals from both speaker classification and verification, through joining of gradients, and applied these in the learning process of a feedforward neural net which they named discriminatively learned network (DLN). Like most works, MFCCs were taken as evaluation features for the DLN, and background data used for training as well. Enlarging of inter-class differences and simultaneous reduction of intra-class variations was successful, and so the DLN significantly outperformed other compensation approaches.

In terms of actual speaker adaptation methods, not much research has been made. However, Gideon *et al.* [36] assessed the effectiveness of progressive neural networks (ProgNets) at freezing the weights of a model's initial layers, tuned for speaker recognition and gender detection from speech, and using these transitional representations as input to the posterior layers, trained for emotion recognition. Performance rates were somewhat higher than those of standard DNN or simple pre-training and fine-tuning (PT/FT) networks. On a separate note, Sidorov *et al.* [37], [38] explored the effects of adding speaker specific and gender information as features in the vectors used to train emotion recognition models, essentially further detailing the datafiles in one experiment. Parallely, the group predicted speaker and gender information with ANN-based recognizers, adding the obtained hypotheses to the feature sets fed into the used emotion recognizer. This plain method of extending the feature vector with additional speaker specific information was found to improve emotion recognition performance on both experiments. It was also found that including more specific speaker information into the feature vectors yielded better results than simply adding gender information.

## 2.6 The *VGGVox* model for Speaker Recognition

This model developed by Nagrani *et al.* [39], based on a VGG-M architecture [40] and composed of 12 layers, is quite capable of perceiving stimuli at specific locations of an

image due to its convolutional design. Such a competency makes it ideal for analyzing spectral representations of audio signals and extracting a robust set of features.

**Table 2.2:** Average Pooling layer's k-th dimension adaptation to clip's n-second Duration

| **Frames** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| **Dimension** | | | | | | | | | |
| 2 | 5 | 8 | 11 | 14 | 17 | 20 | 23 | 27 | 30 |

In order to retain as much information as possible, the raw data undergoes minimal process in the sense that narrowband magnitude spectrograms are generated using a sliding hamming window of width 25ms and step 10ms, meaning an *n*-second input will provide a 100*n* frames spectrum. Normalisation is also performed on mean and variance, at every frequency bin of the spectrum, as it was observed that such a step produced an increase of 10% in classification accuracy. Yet, no other operations are performed on the input data, and the CNN is fed essentially raw spectrograms.

Variable length inputs are also efficiently dealt with by varying the support filter dimension of the *apool6* layer (Figure 2.11). As such, the implementation is adaptable to an audio clip's duration, provided it is between 1 and 10 seconds in length, according to Table 2.2. The dimension values are conforming with the stride and padding methods used by the model, for each duration value. It should be noted that the model does handle clips longer than 10 seconds, by considering only the central 10-second segment of the clip, in spite of losing all the other potentially relevant surrounding information. A progression example of a 1-second audio input fed to the *VGGVox* model is provided in Figure 2.11, for reference.

In terms of purpose, the model was directed towards speaker classification, and trained using the *VoxCeleb1* dataset [39] also developed by Nagrani and her team. This dataset is of large scale, including over 100,000 utterances by 7000+ speakers of varied backgrounds and languages, resulting in more than 2000 hours of audio. Consequently, the model is an ideal candidate for capturing copious amounts of speaker specific cues and prosody mannerisms from any type of human speech, emotional included. Training iterations also included batch normalization [41] and used the default hyper parameter values

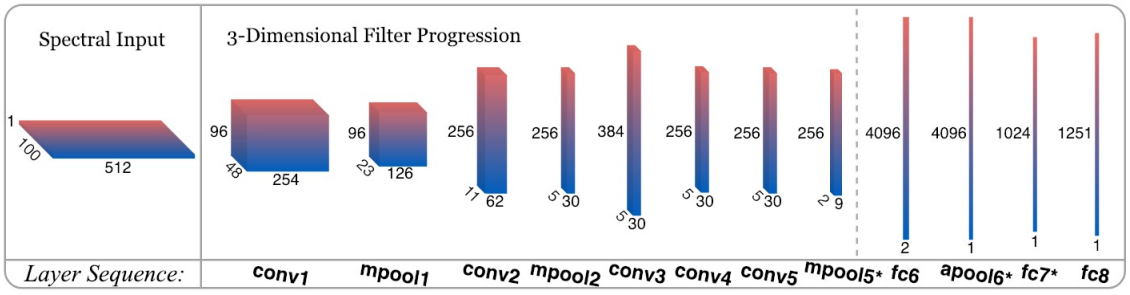



**Figure 2.11:** Exemplary diagram of a 100-frame spectral representation of a 1-second audio file progressing through the layers of the *VGGVox* model. The asterisk symbol is used to identify the layers whose outputs were used as feature matrices for emotional classification.

of the employed MatConvNet toolbox [42].

## 2.7 The Weka Software

The Waikato Environment for Knowledge Analysis, or Weka software [43], is a highly useful piece of software developed and maintained by the University of Waikato, New Zealand, which serves as an aid to machine learning related projects. The software incorporates a large set of tools for data analysis and visualization, as well as implements several state-of-the-art algorithms for regression and data classification. In a simple fashion, a user may read data from a file (in some specific format), preprocess and prepare it for evaluation and finally train and run it through a classifier. Weka also provides a console-like application which incorporates most functionalities of the GUI and some others not yet implemented in it. An overview of the Weka user interface is shown in Figure 2.12.

In this work specifically, the Weka software was used as it already contains several classifier implementations, out of which the following are included:

- Naive Bayes - Based on the algorithm described in [44].
- k-Nearest Neighbors - Based on the algorithm described in [45].
- Random Forest - Based on the algorithm described in [46].
- Logistic Model Tree - Based on the algorithm described in [47].

- Support Vector Machine - Based on the algorithm described in [48].

By using these existing and well-established implementations, available in the *Explorer* user interface of Weka, the errors related to potential mistakes made during algorithm implementation were severely reduced. Plus, considering how the classifiers available in the environment all include some (usually high) degree of adjustability, it was possible to find and choose the best suited parameters for each classifier and for the situation at hand. The data visualization portion of the specified UI was also used to analyze the degree of specialization of the evaluated features.

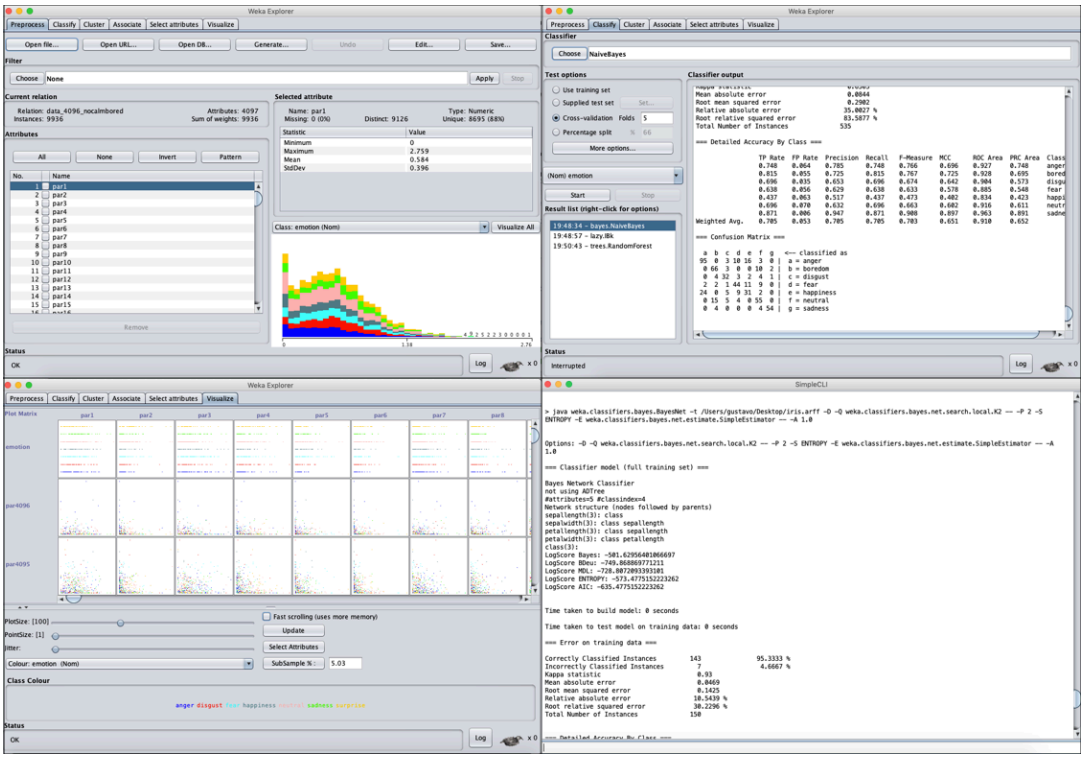


**Figure 2.12:** Weka Explorer UI and console. On the top left, the preprocessing section of Explorer where data can be read and prepared for evaluation. Top right shows the classification section where data can be evaluated on several implemented classifiers. Bottom left shows the visualization section where some data visual representations are provided. Bottom right displays the simple Weka console.

## 2.8 System Overview

The overall system for human emotion recognition through speech analysis on a neural network essentially encompassed two major stages. In the first stage, speaker adaptation,

the *VGGVox* trained model for speaker recognition was employed for feature extraction. By feeding this model with the audio obtained from the emotional speech corpora, the adaptation was performed on the audio by using the features obtained as outputs of the three different layers forming the bottleneck of the network: *mpool5*, *apool6* and *fc7*. These three layers were chosen given the fact that they provide the most summarized yet informationally rich speaker specific features. They are, therefore, the most likely to go so far as to successfully perform speaker adaptation but not so much that they would be fully specialized for speaker recognition, which was not the objective of this work. The process was achieved by simply applying the model to the audio clips without any form of processing other than the already specified. As can be observed in Figure 2.11 and inferred for inputs of any length, the obtained arrays had the following dimensions, respectively: $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$. It should be noted how the $9 \times 1 \times 256$ arrays were obtained by averaging the $9 \times k \times 256$ output arrays of the *mpool5* layer along the *k*-th dimension, which corresponds to the variable length *n*-second / 100*n*-frame input in accordance with Table 2.2. No outputs from any other layers were considered.

In the second stage of the system, emotional state classification, the previously obtained features already somewhat adapted to speaker mannerisms were used for training and testing of the previously mentioned state-of-the-art classifiers which were Naive Bayes, 5-Nearest Neighbor, 500-tree Random Forest, Logistic Model Tree and Support Vector Machine. In order to achieve this, 5-fold cross validation was used. This way, 20% of the data obtained from propagating the emotional speech corpora through the *VGGVox* model was used for testing classifiers trained with the other 80% of the same data. Accordingly with the method, this was repeated 5 times considering all the 5 different 20% data folds, each time the 20% testing fold being a different one than before.

Considering the stages previously mentioned, which are part of the overall system shown in Figure 2.13, the final outputs were produced in the form of accuracy percentages for each of the classes (emotions) considered, aside from other useful performance metrics.

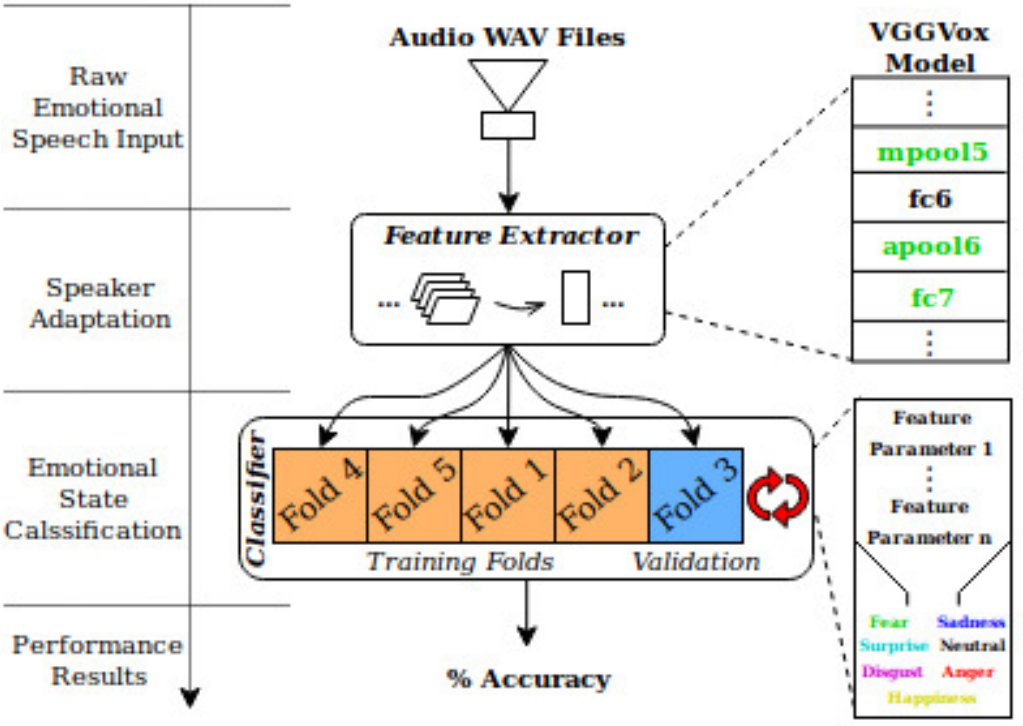



**Figure 2.13:** High-level diagram of the overall system. The layer names in green correspond to the layers from where the features were extracted. Having 5-fold cross validation been used, each time four folds were used for training (orange) while the reamining fold was used for validation (blue).

# 3

# Results

As a means to evaluate the proposed approach's performance, several experiments were carried out in order to evaluate the robustness and efficacy of the extracted feature arrays in terms of emotion recognition. The Weka software was employed so as to apply the feature arrays on the following state of the art classifiers: Naive Bayes, kNN, Random Forest, Logistic Model Tree (LMT) and Support Vector Machine (SVM). For this, the implementations considered were those previously mentioned, which are included in the Weka software. A neural network based approach was not followed during the classification stage given the fact that the available data is not enough to credibly train a machine learning model. This is because machine learning models, as opposed to statistical classifiers, by nature lack any sort of initial direction. In this section, more detail is provided on the carried out experiments and the obtained results. The following section presents an analysis and discussion of those results.

## 3.1 Data Preparation

All files from the obtained databases were converted to the WAV format, at a sampling rate of $16kHz$ as previously explained, given that this value has been proved to be more than sufficient to capture all information embedded in a speech signal. In accordance with the *VGGVox* model's implementation, and in order to take full advantage of all the provided audio, files were adapted to be between 1 and 10 seconds in length. Therefore, a small amount of clips below the 1 second mark were disregarded, as these would hardly provide any emotional information. Further, clips above the 10 second mark were divided into

equally long audio segments. As stated, $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$ feature arrays were extracted for each clip, and the $9 \times 1 \times 256$ arrays were obtained by averaging the $9 \times k \times 256$ output arrays of the *mpool5* layer along the *k*-th dimension, as explained.

## 3.2 Classifier Performance

The simple Naive Bayers classifier was used as an efficacy baseline for evaluation against the rest of the state of the art classifiers in the Weka software, when fed the provided feature arrays for emotion recognition. Performance results were obtained using 5-fold cross validation, on the entire available emotional speech corpora as one large multi-language database. These are shown in Table 3.1. Furthermore, Cohen's Kappa [49] is also provided to further support the validity of the obtained results against random chance, parallel with unweighted average recall (UAR), a favoured metric in emotion recognition systems which attributes the same significance to all possible classes [50]. As can be observed, though some performance variation exists depending on the feature array dimensions applied, the acquired results are in general highly successful. Also, they are on par or above other rates obtained using state-of-the-art techniques from recent studies, such as the ones previously cited.

Given the observed higher classifier performance, using the $1 \times 1 \times 4096$ feature arrays (highlighted row in Table 3.1), in the majority of scenarios, a second testing phase was carried out, under the same conditions, but applied to each emotional speech database individually. This was done in order to evaluate the degree to which language and cultural background have impact on a speaker's emotional prosody. The obtained results are displayed in Table 3.2.

**Table 3.1:** State of the art classifier performance on full emotional corpora feature arrays.

| | **Naive Bayes** | | | **k-Nearest Neighbor** | | | **Random Forest** | | | **Logistic Model Tree** | | | **Support Vector Machine** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* |
| 9x1x256 | 52.2% | 0.44 | 0.50 | 76.4% | 0.72 | 0.75 | 74.3% | 0.70 | 0.72 | 75.9% | 0.72 | 0.74 | 76.5% | 0.72 | 0.75 |
| 1x1x4096 | 56.7% | 0.49 | 0.54 | 80.1% | 0.77 | 0.79 | 77.3% | 0.73 | 0.76 | 81.1% | 0.78 | 0.80 | 76.8% | 0.73 | 0.75 |
| 1x1x1024 | 55.4% | 0.47 | 0.53 | 78.6% | 0.75 | 0.77 | 76.0% | 0.72 | 0.74 | 76.5% | 0.72 | 0.75 | 85.6% | 0.83 | 0.84 |

**Table 3.2:** State of the art classifier performance on standalone emotional database $1 \times 1 \times 4096$ feature arrays.

| | **Naive Bayes** | | | **k-Nearest Neighbor** | | | **Random Forest** | | | **Logistic Model Tree** | | | **Support Vector Machine** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* |
| EMODB | 72.9% | 0.67 | 0.73 | 74.9% | 0.69 | 0.73 | 76.0% | 0.70 | 0.72 | 80.4% | 0.76 | 0.80 | 74.9% | 0.68 | 0.72 |
| EMOVO | 52.8% | 0.45 | 0.53 | 66.2% | 0.61 | 0.66 | 65.3% | 0.60 | 0.65 | 68.5% | 0.63 | 0.69 | 57.9% | 0.51 | 0.58 |
| RAVDESS | 51.8% | 0.44 | 0.53 | 65.5% | 0.60 | 0.65 | 61.9% | 0.55 | 0.60 | 71.6% | 0.67 | 0.71 | 55.0% | 0.47 | 0.60 |
| RML | 70.4% | 0.65 | 0.70 | 73.5% | 0.68 | 0.73 | 75.3% | 0.70 | 0.75 | 79.9% | 0.76 | 0.80 | 71.3% | 0.66 | 0.71 |
| S0329 | 74.7% | 0.70 | 0.74 | 86.2% | 0.83 | 0.84 | 87.4% | 0.85 | 0.85 | 92.4% | 0.91 | 0.91 | 91.0% | 0.89 | 0.90 |
| SAVEE | 61.67% | 0.54 | 0.58 | 65.2% | 0.59 | 0.61 | 64.8% | 0.57 | 0.60 | 70.4% | 0.65 | 0.68 | 55.6% | 0.45 | 0.49 |

# 4

# Discussion

The results on the full emotional speech corpora, in Table 3.1, show how the $1 \times 1 \times 4096$ feature arrays appear to be, in most cases, more robust than their counterparts. This is likely due to the fact that the $1 \times 1 \times 1024$ feature arrays are already too specialized for speaker classification, having put aside a great part of emotional information, whilst the $9 \times 1 \times 256$ feature arrays are still too general, not having concentrated focus on any particular speaker or emotion representation yet. The only exception observed on Table 3.1, where $1 \times 1 \times 4096$ feature performance is surpassed, is given by the $1 \times 1 \times 1024$ features applied to an SVM. The overspecialization of the latter may justify this. Given the spatial mapping nature of the SVM algorithm, an already clear separation between the feature parameters ultimately gives SVM an advantage over other non-spatial based classifiers. In addition, kNN does give the second best performance rate using these features.

A visual depiction of a parameter's distribution across all classes is shown in Figure 4.1, for the three assessed feature array dimensions. This parameter was randomly chosen but still provides an accurate depiction of the common morphology of most of the parameters' distributions, for the respective feature arrays. Evidently, this figure corroborates the previously stated, as the $1 \times 1 \times 1024$ parameter distribution is seemingly too specialized, while the $9 \times 1 \times 256$ respective distribution appears yet too general, making the $1 \times 1 \times 4096$ feature arrays the ideal choice in this scenario. In addition, parameter pair mappings of randomly chosen parameters across all considered classes are also shown in Figure 4.1, for each of the feature array dimensions considered. Given that each mapping is exemplary of all mappings for the respective array dimensions, this is shown in order to

enable another exemplary visualization of the underspecialization and overspecialization issue. It can be noted how the $9 \times 1 \times 256$ mapping is indeed lacking in specialization, forming a blob, whilst the $1 \times 1 \times 1024$ mapping is too specialized, almost bound to the axes. This is a product of parameters taking a null value due to potentially valuable information being disregarded, which can decrease performance. This fact was indeed observed during testing.

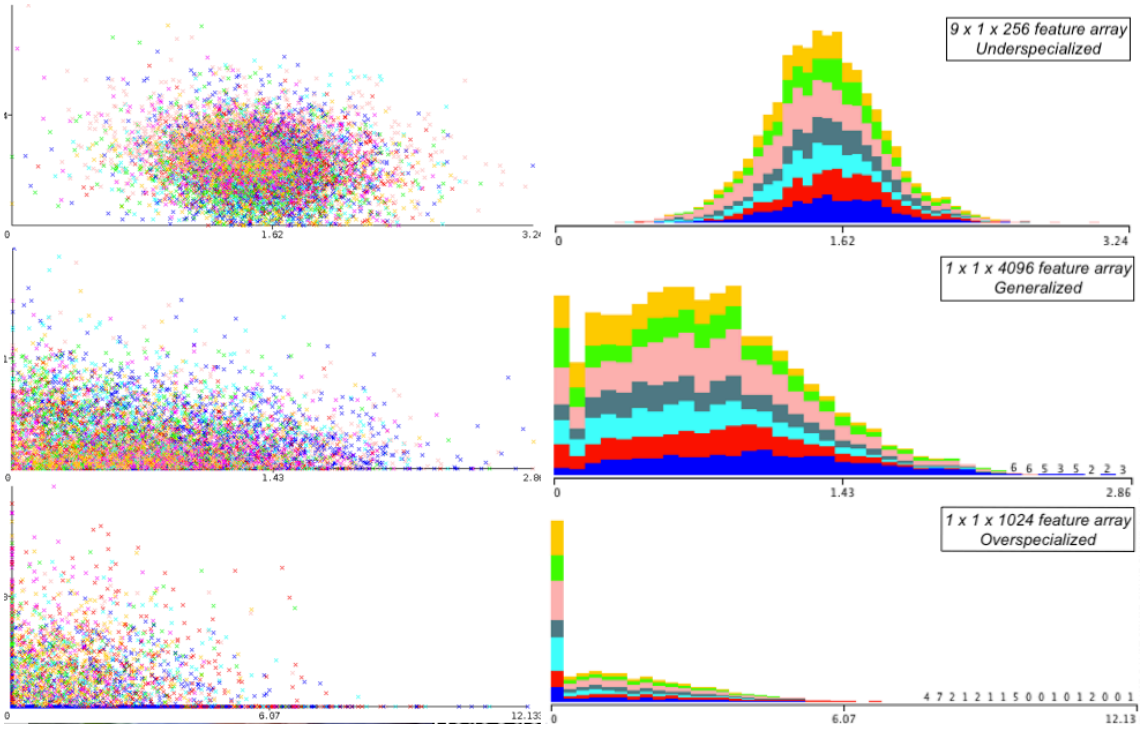



**Figure 4.1:** Parameter pair mappings of the same randomly chosen parameters across all classes on the left, and distributions of another different but also randomly chosen parameter across all classes on the right, for the $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$ feature arrays, respectively from top to bottom. Either of the visualizations is representative of the distributions/mappings of most parameters in all the data, for the respective feature arrays. The seven colors represent the seven emotional states (classes).

In terms of the database specific classification experiments, by observing Table 3.2, results are varying from database to database. Though database size must also be taken into consideration, given the language and cultural diversity between databases, the obtained results suggest that emotional prosody is affected differently in each population. As such, adaptation to cultural background, in addition to speaker, is likely an approach worthwhile researching in order to improve speech emotion recognition systems.

Altogether, the obtained results always surpassed the proposed baseline, having LMT

given the best results when considering the overall leading option of $1 \times 1 \times 4096$ features. However, if considering $1 \times 1 \times 1024$ features, spatial-based classifiers such as SVM and kNN become a better choice. In terms of the $9 \times 1 \times 256$ features, classifier performance was well balanced. Finally, all results certainly support the existence of relevant emotional information in speaker specific speech features, confirming this work's hypothesis. As such, speaker adaptation should be performed in systems attempting to perform successful human emotion recognition from speech analysis.

Undoubtedly the results also demonstrated how suitable a convolutional neural network (CNN) architecture is for analyzing and extracting features from audio, despite CNNs being image driven models. Even though spectrograms had to be obtained from the audio data in order for it to be fed into a network of this type, minimal preprocessing was performed and still the obtained features yielded excellent results. It can certainly be affirmed with confidence that CNNs are a better method for analyzing speech audio signals than other techniques, classical or machine learning based.

# 5

# Demo Application

Considering the novelty and success achieved with the approach of this work, publications were made in order to share it with the scientific community. One of these publications was a demo for the ”exp.at’19” conference[1] in the island of Madeira, Portugal. As such, after the testing phase of this work an actual implementation of the overall system was developed in order to demonstrate its effectiveness.

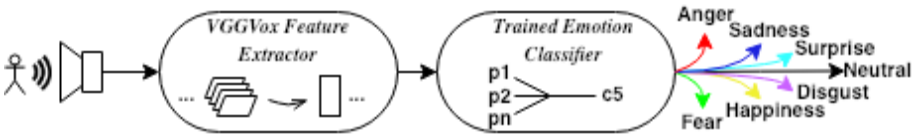



**Figure 5.1:** Overview of the demo’s system, encompassing audio acquisition, feature extraction and finally emotional state classification.

The demo was built using MatLab and made ready to function as a real time application. That being said, it is composed of three phases: audio acquisition, feature extraction and emotional state classification. The workflow carried out by the demo’s system can be observed in Figure 5.1 for easier understanding. In the first phase, a microphone turned to a speaker captures 5-second long audio clips continuously. Each clip is passed on to the second phase, whilst the first phase moves on to the next audio clip recording. In the second phase, and following what has been previously explained, the *VGGVox* model is used for feature extraction from the audio while also performing speaker adaptation. The chosen feature array is that with dimensions $1 \times 1 \times 4096$, as it was observed these produced the best results. Following this, the features are passed on to an emotion classifier trained in Weka but applied in MatLab using its Java wrapper. As expected, the final result is

[1] http://expat.org.pt/home/

a vector of 7 probability values, each corresponding to one of the considered emotional states.

Once the demo MatLab application is run, a simple user interface is presented which is made up of 4 sections and a stop button. In the bottom two sections, the user can observe the recorded audio's signal in time and its frequency spectrum so as to conclude whether or not it saturated, which would affect the validity of the classifier's final prediction. In the top right section, the user is presented with the probability results for each of the emotional states. This is visually complemented by the top left section which presents an irregular heptagon patched over a unit circle, on which each vertex corresponds to an emotional state and is as far from the center of the circle as the emotional state's respective probability value is greater (closer to 1). An example screenshot of the UI, showing the neutral state evaluated from a silent recording, is shown in Figure 5.2.

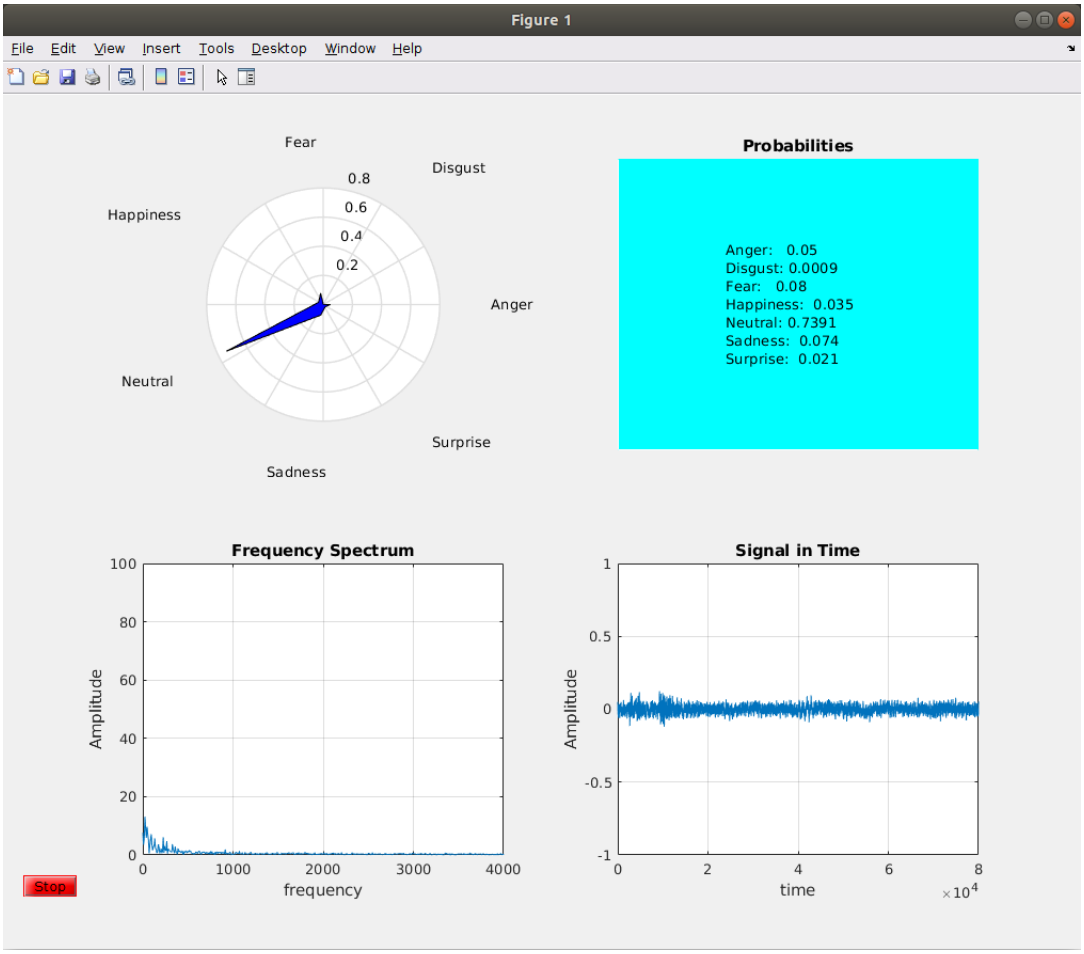



**Figure 5.2:** Demo application user interface, with the four sections described in the text.

# 6

# Conclusion

This dissertation was aimed towards the assessment of a machine learning approach to human emotion recognition by evaluation of recorded speech. Given the audio visual spectral representations to be evaluated and the success of CNNs at extracting valuable information from any kind of images, a convolutional architecture was chosen as a means for feature extraction. Taking also into account the potentially useful non-emotional information embedded in speech signals, such as speaker mannerisms, and in order to obtain valid and successful results without the need of acquiring big data, a model already trained with copious amounts of audio for speaker recognition was picked for performing speaker adaptation. This model, the *VGGVox*, was employed at extracting features from a large set of multi-language emotional speech databases. The features used came from the several layers of the *VGGVox* architecture that make up its bottleneck, and were applied on the training and testing of five different state-of-the-art classifiers using the Weka software. This software was employed in order to use tried and tested implementations of the classifiers, rather than writing new potentially error rich implementations.

Considering the testing and obtained results it was determined that, regardless of language, there is valuable emotional information embedded within speaker specific features, particularly on the ones right before the last bottleneck stage of the model. Posteriorly, acceptable but varying performance ratios were obtained on standalone databases using only the best scenario corresponding feature arrays from the first experiment. This suggests varying degrees of emotional prosody mannerism for different cultural backgrounds. Finally, and based on a general observation of the results, it was concluded that an initial step of speaker adaptation is of paramount importance and should be performed in any

speech emotion recognition systems, so as to achieve higher accuracy rates. All in all, the presented CNN-based approach excelled at the proposed task in all tests carried out.

The work here presented was also the subject of three different submissions, two articles and one demo. Out of three, two have been accepted so far and a response is expected soon on the third. Author versions of these submissinos are included as attachments to this dissertation.

# 7

# Future Work

This work focused on performing human emotion recognition through analysis of data from a single modality, speech, on a convolutional neural network (CNN) trained specifically for speaker recognition, followed by application of state-of-the-art classifiers to make the final emotional state predictions. The aim of this approach was not only to assess the adequacy of CNN application to emotion recognition, but also in great part to evaluate the usefulness of some non-emotional speech information on adaptable emotion recognition. Hence the use of a speaker recognition model to achieve some degree of speaker adaptation.

In the future, assessments will also be done on the efficacy of dimensionality reduction techniques such as PCA or LDA applied to the feature arrays, in order to reduce them to their core information. Further, it is intended to delve deeper into adaptable emotion recognition, by considering additional speaker information, such as cultural background, and other non-emotional conversational aspects such as context, setting and domain incorporation. This will be done in order to take advantage of all the information available in a real emotional speech signal, which is most times wrongly disregarded.

In addition, other modalities for emotion recognition will also be explored, such as vision. This will be done bearing in mind the goal of incorporating facial expression analysis into a multi-modal emotion recognition system. As a final step, it is planned that such a capability will be embedded in an interactive robot.

# Bibliography


[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE SIGNAL PROCESSING MAGAZINE*, vol. 18, no. 11, pp. 32–80, 2001.

[2] R. Plutchik, "Emotion: A psychoevolutionary synthesis," *New York: Harper and Row*, 1980.

[3] N. Fox, "If it's not left it's right," *Amer. Psychol*, vol. 46, pp. 863–872, 1992.

[4] A. Mehrabian, *Basic dimensions for a general psychological theory*, pp. 39–53. Oelgeschlager, Gunn and Hain, 1980.

[5] Z. Liu and N. EYK, "Application of machine learning in automatic sentiment recognition from human speech," *IRC Conference on Science Engineering Technology*, 2017.

[6] S. Shahsavarani, "Speech emotion recognition using convolutional neural networks," Master's thesis, University of Nebraska-Lincoln, bahar@huskers.unl.edu, 03 2018.

[7] R. Lotjidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," *ICASSP*, 2017.

[8] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.

[9] F. Chenchah and Z. Lachiri, "Speech emotion recognition in noisy environment," *International Conference on Advanced Technologies for Signal and Image Processing - ATSIP*, 2016.

[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, p. 335, 2008.

[11] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[12] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013.

[13] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE SIGNAL PROCESSING LETTERS*, 2017.

[14] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral-temporal representations," *IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[15] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 93–202, 1980.

[16] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*, pp. 319–, Springer-Verlag, 1999.

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," *INTERSPEECH*, 2005.

[18] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *LREC*, 2014.

[19] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," *Proc. Int. Conf. Auditory-Visual Speech Processing*, 2009.

[20] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *Plos One*, 2018.

[21] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," *IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

[22] E. L. R. Association, "Emotional speech synthesis database elra-s0329," *https://catalogue.elra.info/en-us/repository/browse/ELRA-S0329*, 2011.

[23] R. Jürgens, A. Grass, M. Drolet, and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of Nonverbal Behavior*, vol. 39, pp. 195–214, 2015.

[24] S. Paulmann, D. Furnes, A. M. Bøkenes, and P. J. Cozzolino, "How psychological stress affects emotional prosody," *PLOS ONE*, vol. 11, pp. 1–21, 11 2016.

[25] M. Spada, A. Nikčević, G. Moneta, and A. Wells, "Metacognition, perceived stress, and negative emotion," *Personality and Individual Difference*, vol. 44, pp. 1172–1181, 4 2008.

[26] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, pp. 1068–1072, 09 2014.

[27] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *INTERSPEECH*, 2015.

[28] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[29] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 2, no. 1, pp. 194–205, 1994.

[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *https://catalog.ldc.upenn.edu/LDC93S1*, 1993.

[31] Y. Wang and B. Lawlor, "Speaker recognition based on mfcc and bp neural networks," *IEEE ISSC*, 06 2017.

[32] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[33] J. Chang and D. Wang, "Robust speaker recognition based on dnn/i-vectors and speech separation," *IEEE ICASSP*, 03 2017.

[34] Z. Ge, A. N. Iyer, S. Cheluvaraja, R. Sundaram, and A. Ganapathiraju, "Neural network based speaker classification and verification systems with enhanced features," *INTELLIGENT SYSTEM CONFERENCE*, 09 2017.

[35] S. Yao, R. Zhou, P. Zhang, and Y. Yan, "Discriminatively learned network for i-vector based speaker recognition," *IEEE Electronics Letters*, vol. 54, no. 22, pp. 1302–1304, 2018.

[36] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. Mower Provost, "Progressive neural networks for transfer learning in emotion recognition," *INTERSPEECH*, pp. 1098–1102, 08 2017.

[37] M. Sidorov, S. Ultes, and A. Schmitt, "Comparison of gender- and speaker-adaptive emotion recognition," *LREC*, pp. 3476–3480, 05 2014.

[38] M. Sidorov, S. Ultes, and A. Schmitt, "Emotions are a personal thing: Towards speaker-adaptive emotion recognition," *ICASSP*, pp. 4803–4807, 05 2014.

[39] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[40] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *British Machine Vision Conference*, 2014.

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv Preprint*, 2015.

[42] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," *CoRR*, vol. abs/1412.4564, 2014.

[43] M. A. H. E. Frank and I. H. Witten, "The weka workbench. online appendix for 'data mining: Practical machine learning tools and techniques'," *Morgan Kaufmann*, 2016.

[44] P. L. G. H. John, "Estimating continuous distributions in bayesian classifiers.," *Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, 1995.

[45] D. K. D. Aha, "Instance-based learning algorithms.," *Machine Learning*, pp. 37–66, 1991.

[46] L. Breiman., "Random forests.," *Machine Learning*, pp. 5–32, 2001.

[47] E. F. N. Landwehr, M. Hall, "Logistic model trees.," *Machine Learning*, pp. 161–205, 2005.

[48] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, p. 7:1–27:27, 2011.

[49] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, pp. 37–46, 1960.

[50] A. B. e. a. B. W. Schuller, S. Steidl, "The interspeech 2009 emotion challenge.," *Interspeech*, pp. 312–315, 2009.

# Appendix

- exp.at’19 - Demo
- AES Audio Forensics 2019 - Full Paper
- IROS 2019 - Full Paper

# Importance of speaker specific speech features for emotion recognition

Gustavo Assunção, ISR-UC, Paulo Menezes, ISR-UC,

Fernando Perdigão, IT-UC

**Abstract**

The recognition of emotions is an inherent ability possessed by humans, which has long intrigued many researchers. Primarily due to the possibility of its successful emulation and integration in independent systems. Further, speech, being a mixture of utterances conveying a state of mind, proves to be a suitable candidate from which emotionality can be inferred, due to its many feature variations. This is corroborated by human beings themselves using this modality for extraction of emotionality clues. Another important aspect has to do with communicational register adaptation and the skill to discern different emotions in different speakers. Sure enough, the same emotional utterance may be interpreted divergently for two different people, meaning emotionality specific information is present in a speaker's personal register. As a demo, we propose a real-time automatic emotion recognition system from speech, based on the use of the well established VGG-like convolutional neural network speaker recognition model *VGGVox*, trained with over 100,000 utterances from the *VoxCeleb1* dataset on speaker recognition, for emotional feature extraction and feeding to state-of-the-art classifiers for accurate recognition of emotional states. Positive supporting results have been captivating enough to spark interest in the technique.

# 1 Introduction

Emotion, in any sense of the word, is unquestionably a fundamental aspect of human interaction and everyday life which, despite lack of solid confirmation, many reports agree to be an evolutionary trigger for adaptive behavior of a being to specific circumstances and environmental scenarios [1]. Hence, the ability to recognise different emotional states in a user would prove highly beneficial for any interactive system aiming for the convergence of its own communicational register to conversation/situation specific standards, along with many other behavioral plasticity advantages.

Approaches to emotion recognition in speech, despite their success or lack thereof, are commonly environment and language specific, focusing on a handful of aspects known to be suitable to the techniques used, while disregarding everything seemingly unrelated and effectively losing information later proved to be valuable. Furthermore, most machine learning techniques, which currently champion the state of the art, include high levels of preprocessing on the used data, rather than employing it in its raw form. Something that not only increases the size and computational complexity of the final evaluation model but also decrease the validity of the model's success as a standalone recognition technique.

As previously stated, the state of the art is currently attributed to machine learning methods, such as neural networks. Recent work has shown how the generalization and adaptability of these structures can greatly benefit their performance in terms of classification and pattern recognition problems, easily surpassing most classical techniques. Many implementations of machine learning methods have been proposed on the subject of emotion recognition, some having produced highly positive results, including the *Universum* Autoencoder by Deng *et al.* [2], exploring domain adaptation, Trigeorgis *et al.* [3] LSTM architecture, addressing context awareness, and Lotjidereshgi *et al.* [4] novel LSM configuration, evaluating source and vocal tract components of speech signals separately. On par with these techniques are convolutional neural networks (CNN), models which have proved highly successful in the field of image analysis and object recognition. As such, a recent trend has been growing motivating the use of CNNs for analysis of visual representations of spectral information from speech signals.

Another area on which CNNs have recently been applied is speaker recognition, the idea of recognizing a certain speaker included in a set through their characteristic speech features. Additionally, considering supervised machine

learning techniques, a correlation between speaker and speech emotion recognition is apparent, due to the notable similarities between the steps taken in terms of spectral analysis of speech clips, on both topics. Given that emotion expression is tightly coupled with one's own way of acting and thus conversating, and considering the stated closeness in spectral analysis between topics, it is not farfetched to infer the existence of relevant emotionality information embedded in speaker specific speech features. Naturally, a new topic of research is formulated based on wether or not emotion recognition systems are speaker dependent and if so, should said systems first attempt to perform a speaker preclassification stage of sorts. Only after which successful emotion recognition becomes truly valid, having the evaluated utterance been situated in terms of speaker type space. Something that happens in reality with a human being and the people they interact with.

# 2 System Description

The experimental setup will consist of a simple microphone gathering speech clips from a speaker, and feeding them to a machine which in turn parses said clips into segments between 1 and 10 seconds long. These segments are then fed to the *VGGVox* model [5] for feature extraction, gathering parameters which then follow through to a state-of-the-art classifier such as Random Forest [7], resulting in one of 7 archetypal emotional states: anger, disgust, fear, happiness, sadness, surprise and the neutral state.

## 2.1 Training Data

In order to train emotion classifiers with features extracted using a speaker classification model, 6 well known and established emotional speech databases of various languages were collected. All clips of matching emotions were pooled together, forming a larger, multi-language, single database with the 7 previously stated emotional states. Clips below 1 second long were disregarded and clips above 10 seconds long were cut into smaller segments to take more advantage of the available data and fit the input needs of the *VGGVox* model.

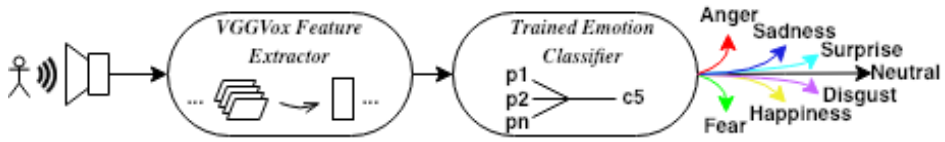



Figure 1: Overview of the demo's system.

### 2.2 *VGGVox* Feature Extraction

The *VGGVox* model for speaker classification, unlike most, has undergone extensive training with over 2000 hours of speech by 1251 different speakers, which greatly increases its validity in terms of extraction of speaker specific features. As previously stated, these also contain relevant emotionality information.

In order to extract more general features not fully specialized for speaker classification, all clips in the collected databases were evaluated through the network only up until and excluding its bottleneck stage. Feature vectors were therefore generated for each clip in the accumulated database, being given the clip's corresponding emotional label.

### 2.3 Emotion Classification

Using the feature vectors generated in the previous step, training was performed on a set of state-of-the-art classifiers, such as kNN or decision tree based methods, to successfully recognize emotion from speech, regardless of the language spoken. This was done in order to select the classifiers most suitable to our demo, based on a tradeoff between their size and corresponding success rate. Seeing that large and complex classifiers, though having better performances overall, are unlikely to fair well in a real-time setup or as integrating parts in larger automatic systems.

## 3 Supporting Results

Initial testing has been done on the proposed system, using the *VGGVox* generated feature vectors of the collected emotional speech databases to train a small set of well-known classifiers using the Weka software [6]. Performance results were obtained with 5-fold cross-validation and are comparable to other recent state-of-the-art emotional speech recognition systems, as shown in

Table 1. Given that the number of utterances in the collected database is well distributed among the seven emotional labels, the presented metrics are enough to show the potential of the technique used.

Some data reduction procedures, such as PCA and LDA, have also been considered as means of reducing the training data to its core relevant parameters and their effectiveness is currently being evaluated.

Table 1: Classifier Performance Using Extracted Features

| | *5-Nearest Neighbor* | *500-Tree Random Forest* | *Logistic Model Tree* |
|---|---|---|---|
| **Accuracy** | 80.1 % | 77.3 % | 81.1% |
| **K-Statistic** | 0.77 | 0.73 | 0.78 |

# 4 Conlusion

In this demo we propose a real-time speech emotion recognition system, based on the use of speaker specific speech features extracted from emotional speech databases, to train well-established state-of-the-art classifiers. The novel aspect and main goal of this project is to corroborate the importance of taking into account the emotional information embedded in a speaker's personal communicational register, when attempting emotion recognitin in speech. This has been partially shown by our initial supporting results. Finally, this is done so as to spark more interest in the research community on merging the two approached topics to create more successful and solid emotion recognition techniques.

# References


[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction", IEEE SIGNAL PROCESSING MAGAZINE, 18(11):32–80, 2001.

[2] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition", IEEE SIGNAL PROCESSING LETTERS, 2017.

[3] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognitionusing a deep convolutional recurrent network", IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP), 2016.

[4] R. Lotjidereshgi and P. Gournay, "Biologically inspired speech emotion recognition", ICASSP, 2017.

[5] A. Nagrani, J. S. Chung and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset", INTERSPEECH, 2017.

[6] E. Frank, M. A. Hall and I. H. Witten, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[7] L. Breiman, "Random Forests", Machine Learning, 45(1):5-32, 2001.

# Premature Overspecialization in Emotion Recognition Systems

Gustavo Assunção, ISR-UC, Fernando Perdigão, IT-UC,

Paulo Menezes, ISR-UC

**Abstract**

Emotion recognition from speech, the ability to identify expressed emotional states in vocal utterances, is an inherent ability humans apply in their daily interactions. Though a highly researched topic, it has yet to conform with real human performance levels, which may be due to the overspecialization or inability of most automatic recognition systems to adapt to non-emotional human conversational traits. Given that these traits may contain information pertinent to a speech based recognition system, generalization should be emphasized in early emotional feature extraction stages. To support this, an application of the VGGVox speaker recognition model has been evaluated for emotional feature extraction. Results on state-of-the-art classifiers were comparable to other recent speech emotion recognition techniques.

## 1 Introduction

Throughout the course of evolution, human beings have been remarkably successful in refining their ability to discern and extract pertinent information from a diverse panoply of sensorial modalities. The human brain, being the kernel of our species' accomplishments, is the perfect example of a quasi fully automatic system capable of domain adaptation, data processing and appropriate response to information rich signals, naturally driving researchers to attempt its exact emulation.

One of the modalities from which extensive amounts of information is derived is speech. This ability to communicate with one another has allowed us

not only to express what we want or need but also to inadvertently disclose emotional cues and reveal personal convictions people would rather withhold from others. Given the fact that emotion, an evolutionary trigger for appropriate response to environmental scenarios [7], governs many aspects of human behavior and conversational traits, this potential information leakage can be of great value and usefulness to any specific area in the field of forensic science, as it allows for further interpretation and possible re-apprecitiation of previously misevaluated data.

The emotion dependency of vocal acoustic waveforms is not unheard of in forensics [16]. Plus, the effects of stress on an individual's emotional prosody have received considerable empirical corroboration from the psychology research community [22], [24]. Considering the emotional pressure undergone by plausible felons and crime victims, the ability to discern an individual's emotional state based on their applied prosody could greatly benefit the analysis of statement veracity, leading to more accurate judgements and ultimately improving the forensic science field as a whole.

In spite of the observed effect of emotion in an individual's emotional prosody, most speech emotion recognition (SER) systems tend to immediately attempt emotional state classification without any regard for adaptation to other aspects such as domain, context and the speakers themselves, even though their relevancy has been thoroughly recognized by previous studies [9], [10], [25], [17], [14], [23]. The latter of those aspects, speaker adaptation, is considered here given the basis of the perceived degree of coupling between a person's own way of acting or conversing and how their emotional states should be assessed when analyzing a real human conversation. Due to this high likelihood that emotional prosody suffers from speaker dependency, most SER systems which do not take it into consideration should be considered overspecialized, experiencing premature specificity and lack of adaptation in feature extraction.

Given the broad spectre of human interaction related applications of machine learning models, we present a potential solution to the aforementioned overspecialization issue based on the use of the *VGGVox* model [21], trained with over 100,000 utterances for speaker recognition, to extract specific features from utterances of 6 standard and established emotional speech databases, with minimal preprocessing. With this effort, we hope to enable a speaker preclassification stage of sorts, yielding well adapted features and taking a better, more structured advantaged of all the information available in a speech signal.

The application, on state-of-the-art classifiers, of the features extracted using our technique confirmed the importance of speaker adaptation, and how further exploiting information embedded within other communicational traits before feature specialization can largely benefit SER systems. Such confirmation came from the obtained performance rates, which were on par with or above the ones from other state-of-the-art methods.

This paper is divided in the following manner: section 2 provides an overview of the methodology used in our approach, while section 3 details the experiments carried out and the obtained results. This is followed by section 4 where a discussion is made on the results and their supporting stance of the topic, and finally by section 5 where a summary conclusion and synopsis of future work are presented.

# 2 Methods

In this section, the emotional speech corpora are listed and their computation is detailed. Further, the applied *VGGVox* model is described and its use as an extractor of emotional speech features is outlined, so as to provide insight into our approach. Finally, the emotional classifiers trained are listed, along with a brief explanation on how the extracted features were arranged to be applied to these. A high-level diagram of the entire system is provided in Figure 1, to aid in understanding.

## 2.1 Databases of Emotional Speech

In order to test our hypothesis, 6 standard emotional speech databases were pooled together, totalling at around 9000 variable duration utterances, in 8 different languages. The pool of databases included clips portraying 9 different emotional states, but it was trimmed so as to only include the clips whose corresponding emotional state was present in all single corpora. As such, the set included utterances depicting anger, happiness, sadness, disgust, fear, surprise and the neutral state. The databases were the following:

- SAVEE [15]: 480 english utterances by 4 non-professional male actors. Includes the states of Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral;

- EMOVO [6]: 588 italian utterances by 6 professional actors (3 male, 3 female). Includes the states of Anger, Joy, Sadness, Fear, Disgust, Surprise and Neutral;
- RML [26]: 720 multi-language utterances by 8 non-professional male actors. Includes the states of Anger, Disgust, Fear, Happiness, Sadness and Surprise;
- EMODB [3]: 535 german utterances by 10 professional actors (5 male, 5 female). Includes the states of Anger, Joy, Sadness, Fear, Disgust, Boredom and Neutral;
- ELRA-S0329 [12]: 6041 spanish utterances by 2 professional actors (1 male, 1 female). Includes the states of Anger, Joy, Sadness,Fear, Disgust, Surprise and Neutral;
- RAVDESS [18]: 1440 english utterances by 24 professional actors (12 male, 12 female). Includes the states of Anger, Happiness, Sadness, Fear, Disgust, Surprise, Calmness and Neutral;

Each file from these databases was downsampled to 16 kHz, given the fact that this frequency is sufficient for capturing all essential speech information, and converted to the WAV format. Clips were also segmented in order to be between 1 and 10 seconds in length, as according with the *VGGVox* model's implementation and to take full advantage of all the audio available. Utterances below 1 second in length were disregarded as these would unlikely contain any relevant emotional information.

## 2.2 *VGGVox* model for speaker recognition

This VGG-like architecture, developed by Nagrani *et al.* [21] and made up of 12 layers, is ideal for capturing even the slightest of stimuli at specific locations due to its convolutional background. This is taken advantage of by providing the network with spectral representations of the audio signals, from which features are then extracted. These narrowband spectrograms are obtained directly from raw data, in order to retain all information, by using a sliding Hamming window of width 25ms and step 10ms. After this, normalisation is performed on mean and variance, at every frequency bin of the spectrum. Considering the former, n-second inputs provide 100n frames spectra. No further action is taken on the input data.

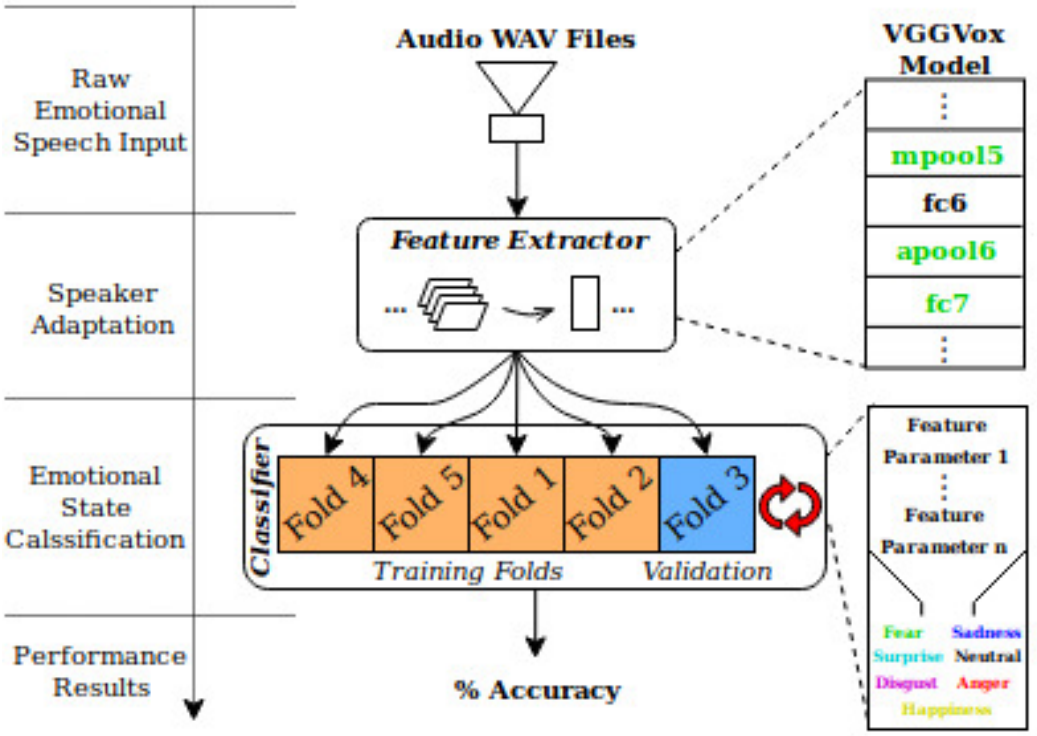



Figure 1: High-level diagram of our overall technique. The layer names in green correspond to the layers from where the features were extracted. Having 5-fold cross validation been used, each time four folds were used for training (orange) while the reamining fold was used for validation (blue).

The network deals with variable length input through the *apool6* layer (see [21]), where the filter dimension is adaptable to the corresponding clip's duration. Provided this duration is between 1 and 10 seconds long, and in accordance with the stride and padding methods used by the model, the filter dimension takes the same value as that of the input array to the *apool6* layer. The clip duration to *apool6* layer dimension correspondence is shown in Table 1. Clips which are 10 seconds in length or longer are also accepted, though the model will only consider the central 10-second window and disregarded all other encircling audio.

The model as a whole was already trained for speaker classification using the related *VoxCeleb1* dataset [20], which is made up of 100,000+ utter-

Table 1: The *apool6* adaptation to clip's duration

| ***Number of Frames*** | ***Filter Dimension*** |
|---|---|
| 100 | 2 |
| 200 | 5 |
| 300 | 8 |
| 400 | 11 |
| 500 | 14 |
| 600 | 17 |
| 700 | 20 |
| 800 | 23 |
| 900 | 27 |
| 1000 | 30 |

ances by over 7000 different speakers of different cultures and backgrounds, totalling at more than 2000 hours of audio. That being said, the model is undoubtedly capable of building complex representations of the data it receives, while also encompassing a great and varied set of speaker specific cues and prosody mannerisms. As such, it becomes an ideal candidate for performing speaker adaptation in SER systems, and preventing premature feature specialization by taking further advantage of more intertwined speech information.

## 2.3 Feature Extraction and Classification

Seeing that the ultimate goal was to perform emotion recognition and not speaker classification, features were extracted as the output arrays of the bottleneck layers of the *VGGVox* model. As such, the layers chosen for feature extraction and result comparison were the *mpool5*, *apool6* and *fc7* layers, resulting in $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$ feature arrays, respectively. These three layers were chosen given the fact that they provide the most summarized yet informationally rich speaker specific features. They are, therefore, the most likely to go so far as to successfully perform speaker adaptation but not so much that they would be fully specialized for speaker recognition, which was not the objective of our work. No outputs from any other layers were considered. Please note, the $9 \times 1 \times 256$ arrays were obtained by averaging the $9 \times k \times 256$ output arrays of the *mpool5* layers along the $k$-th dimension, which corresponds to the variable length $n$-second / $100n$-frame input in accordance with Table 1.

In order to test the robustness of the obtained features, several state-of-the-art classifiers were trained and tested using the Weka Software [11]. Those classifiers were the following:

- Naive Bayes [13];
- k-Nearest Neighbors (kNN) [8];
- Random Forest (RF) [2];
- Logistic Model Tree (LMT) [19];
- Support Vector Machine (SVM) [4];

# 3 Results

The classifier performances were evaluated using 5-fold cross validation applied to the entire pooled emotional speech corpora. Results are displayed in Table 2, for each of the feature array dimensions considered. Results are also provided in the form of Cohen's Kappa [5], supporting the validity of the displayed accuracy rates in terms of random luck, and of unweighted average recall (UAR), a preferred metric in the field of emotion recognition systems which considers all classes equally significant [1].

As can be observed, though some performance variation exists depending on the feature array dimensions applied, the acquired results are in general highly successful. Also being on par or above other rates obtained using state-of-the-art techniques from recent studies, such as the ones previously cited.

# 4 Discussion

The analysis of Table 2 clearly shows the success of the proposed method. Even though the lowest performances, corresponding to a simple Naive Bayes classifier, were around the 50-60% mark, the rest of the accuracy rates were considerably high, even reaching scores above 80%, which are exceptional results in emotion recognition systems. Even more so given the fact that the pooled emotional speech corpora is multi-language, which usually decreases performance in other methods. Considering how the classifier performances

Table 2: State of the art classifier performance on full emotional corpora feature arrays.

| | **Naive Bayes** | | | **k-Nearest Neighbor** | | | **Random Forest** | | | **Logistic Model Tree** | | | **Support Vector Machine** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* |
| 9x1x256 | 52.2% | 0.44 | 0.50 | 76.4% | 0.72 | 0.75 | 74.3% | 0.70 | 0.72 | 75.9% | 0.72 | 0.74 | 76.5% | 0.72 | 0.75 |
| 1x1x4096 | 56.7% | 0.49 | 0.54 | 80.1% | 0.77 | 0.79 | 77.3% | 0.73 | 0.76 | 81.1% | 0.78 | 0.80 | 76.8% | 0.73 | 0.75 |
| 1x1x1024 | 55.4% | 0.47 | 0.53 | 78.6% | 0.75 | 0.77 | 76.0% | 0.72 | 0.74 | 76.5% | 0.72 | 0.75 | 85.6% | 0.83 | 0.84 |

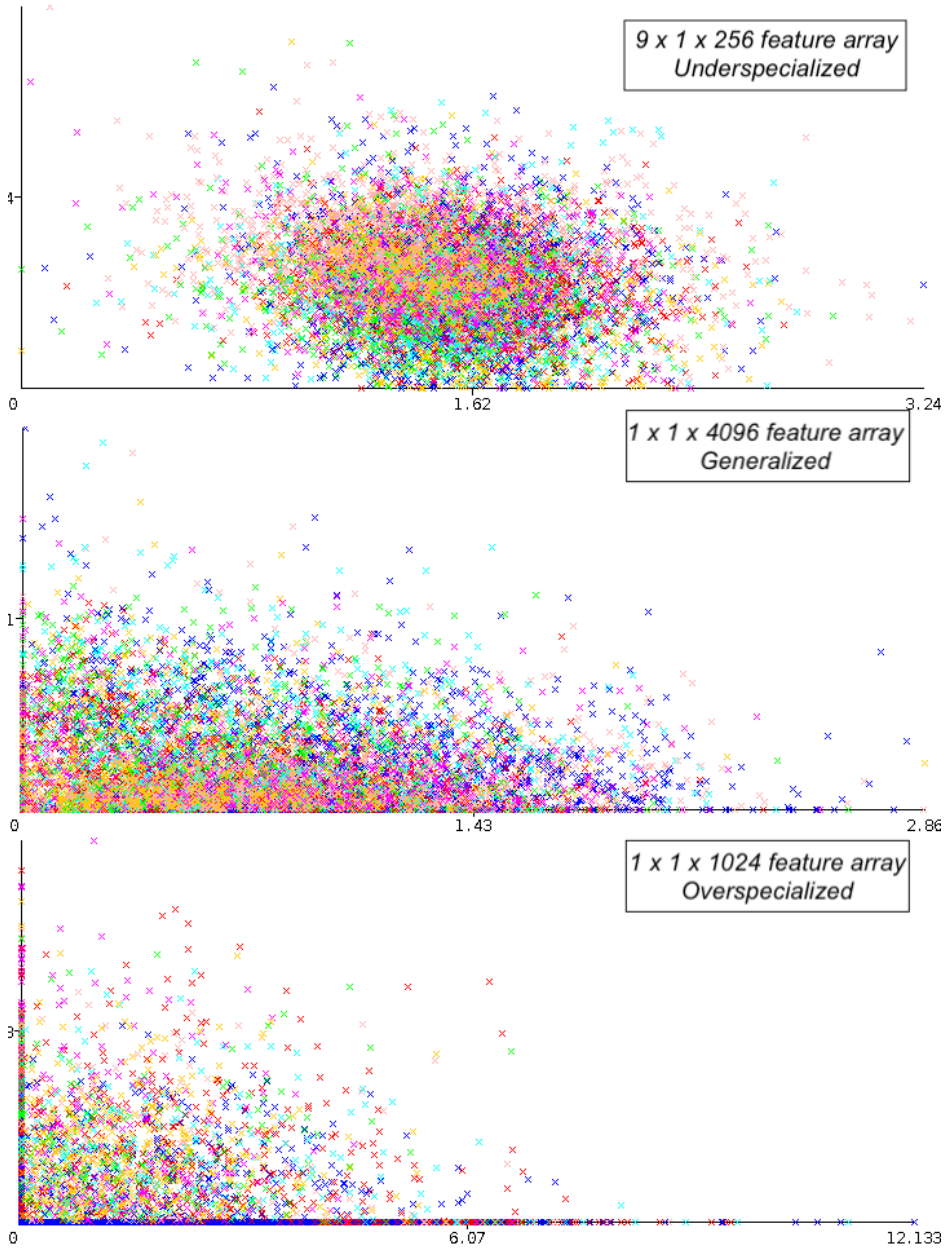



Figure 2: Parameter pair mappings of the same randomly chosen parameters across all classes, for the $9\times1\times256$, $1\times1\times4096$ and $1\times1\times1024$ feature arrays, respectively. Each is representative of all mappings, for the respective feature arrays. The seven colors represent the seven emotional states (classes).

of our technique also succesffully surpassed other recent approaches to the topic, then these results support our initial hypothesis and deem premature feature specificity a noticeable issue in SER systems. Something that can no longer be disregarded as insignificant, as it clearly decreases accuracy rates and overall system performance.

On a deeper level, by noticing how the performance obtained using the $1\times1\times4096$ feature arrays (highlighted row in Table 2) is generally better than the performances obtained using the $9 \times 1 \times 256$ or the $1 \times 1 \times 1024$ feature arrays, it can be inferred how also the underspecializing or overspecializing of obtained features on one specific communicational trait, such as speaker type, can decrease the accuracy rate. As such, a specific balance must be found on how much information can be extracted from a speech signal when considering each trait individually, in order to later build better structured data representations. Sequential adaptation to communicational attributes is therefore a suitable solution to premature specificity of extracted features.

Parameter pair mappings of randomly chosen parameters across all considered classes are shown in Figure 2, for each of the feature array dimensions considered. Given that each mapping is exemplary of all mappings for the respective array dimensions, this is shown in order to enable an exemplary visualization of the underspecialization and overspecialization issue. It can be noted how the $9 \times 1 \times 256$ mapping is indeed lacking in specialization, forming a blob, whilst the $1 \times 1 \times 1024$ mapping is too specialized, almost bound to the axes. This is a product of parameters taking a null value due to potentially valuable information being disregarded, which can decrease performance as it was observed.

# 5 Summary

In this paper, we explored the notion of using more of the information available in a speech signal to improve the complex data representations created by a machine learning model. Premature feature specificity was reduced by using a previously trained neural network for speaker classification, to obtain speaker features from emotional speech clips. These features were extracted from the three layers closer to the model's bottleneck, and were later employed for training and performance testing of state-of-the-art classifiers, using the Weka software. Given how the results surpassed those of other state-of-the-art speech emotion recognition (SER) techniques which perform

too much data preprocessing and overspecialize their features too soon, it was concluded how this premature feature specificity is undesirable and should be discouraged by improving the appropriate use of available data. Initial steps for adaptation to conversational traits are therefore extremely important for the success of SER methods.

From this point forward, we intend to explore and merge even more non-emotional human conversational traits, such as domain, context or setting, into our technique so as to take further advantage of all that is embedded in human speech. Ultimately we hope to improve SER systems and enable their use in the forensic and other human sciences with increased accuracy.

# Acknowledgments

The authors would like to thank the respective database curators for providing access to their emotional speech sets and allowing their use in our research. This work has been partially supported by OE - national funds of FCT/MCTES (PIDDAC) under project UID/EEA/00048/2019.

# References

[1] A. Batliner et al. B. W. Schuller, S. Steidl. The interspeech 2009 emotion challenge. *Interspeech*, pages 312–315, 2009.

[2] L. Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. *INTERSPEECH*, 2005.

[4] CC Chang and CJ Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, page 7:1–27:27, 2011.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pages 37–46, 1960.

[6] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. Emovo corpus: an italian emotional speech database. In *LREC*, 2014.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE SIGNAL PROCESSING MAGAZINE*, 18(11):32–80, 2001.

[8] D. Kibler D. Aha. Instance-based learning algorithms. *Machine Learning*, pages 37–66, 1991.

[9] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller. Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 24(4):500–504, April 2017.

[10] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, Sep. 2014.

[11] M. A. Hall E. Frank and I. H. Witten. The weka workbench. online appendix for 'data mining: Practical machine learning tools and techniques'. *Morgan Kaufmann*, 2016.

[12] ELRA. Emotional speech synthesis database s0329. catalogue.elra.info/en-us/repository/browse/ELRA-S0329/, 2012.

[13] P. Langley G. H. John. Estimating continuous distributions in bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[14] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. Progressive neural networks for transfer learning in emotion recognition. *INTERSPEECH*, pages 1098–1102, 08 2017.

[15] S. Haq and P.J.B. Jackson. Speaker-dependent audio-visual emotion recognition. *Proc. Int. Conf. Auditory-Visual Speech Processing*, 2009.

[16] T. Johnstone and K. Scherer. The effects of emotions on voice quality. *Proceedings of the XIVth International Congress of Phonetic Sciences*, 01 1999.

[17] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH*, 2015.

[18] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *Plos One*, 2018.

[19] E. Frank. N. Landwehr, M. Hall. Logistic model trees. *Machine Learning*, pages 161–205, 2005.

[20] A. Nagrani. The voxceleb1 dataset. `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html`, 2017. Accessed: 2019-03-14.

[21] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[22] Silke Paulmann, Desire Furnes, Anne Ming Bøkenes, and Philip J. Cozzolino. How psychological stress affects emotional prosody. *PLOS ONE*, 11(11):1–21, 11 2016.

[23] Maxim Sidorov, Stefan Ultes, and Alexander Schmitt. Comparison of gender- and speaker-adaptive emotion recognition. *LREC*, pages 3476–3480, 05 2014.

[24] Marcantonio M. Spada, Ana V. Nikčević, Giovanni B. Moneta, and Adrian Wells. Metacognition, perceived stress, and negative emotion. *Personality and Individual Difference*, 44(5):1172–1181, 4 2008.

[25] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, 03 2016.

[26] Zhibing Xie and Ling Guan. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. *IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

# Large Scale Speaker Adaptive Speech Emotion Recognition

Gustavo Assunção, ISR-UC, Paulo Menezes, ISR-UC,

Fernando Perdigão, IT-UC

**Abstract**

The idea of recognizing human emotion has recently received considerable attention from the research community, due to its many possible forensic applications and potential boosting of interactive systems. As such, and following the current trend of research, many machine learning models have been proposed addressing the topic of speech emotion recognition (SER), the idea of classifying a person's emotional state based on speech analysis. These models have far surpassed the performance of previous classical techniques. Nevertheless, even the most successfull methods are still rather lacking in terms of adaptation to specific speakers and scenarios, causing them to be incapable of meeting real human performance standards. In this paper, we evaluate a large scale machine learning model for classification of emotional states. This model has been trained for speaker identification but is instead used here as a front-end for extracting robust features from emotional speech. We theorize that adaptation to a speaker's emotional prosody can greatly improve the accuracy of a SER system. Several experiments using various state of the art classifiers are carried out, using the Weka software, in order to evaluate the robustness of the extracted features. Considerable improvement is observed when comparing our obtained results with other SER state of the art techniques, demonstrating the importance of speaker adaptation in this matter.

# 1 INTRODUCTION

Emotion and its expression undoubtedly governs many aspects of human interaction. Given its status as an evolutionary trigger for appropriate response to environmental scenarios [1], it is self-evident that the emotional phenomena experienced by a person should tend to mold their behavior and conversational register in the social settings they engage with. Effectively, by adulthood most humans will have developed a large set of highly nature/nurture dependent, distinct behavioral responses to the multiple emotional states they experience throughout their personal lives. An emulated corroboration of this notion can be inferred by the particular articulation patterns observed in professional actors simulating emotion [2]. Further support is given by the widely accepted effects of recurrent stress in an individual's emotional state [3], [4] naturally affecting their prosody.

The role of someone's personality, and consequently their way of communicating, are often overlooked when it comes to emotion recognition from speech. In fact, even though the state of the art machine learning models are exceptionally competent at evaluating data with an unspecified set of relevant features, most of the designs tend to focus directly and solely on overt emotional cues. As of now, this should be considered an erroneous approach given the observed higher performances of systems with deeper adaptation levels, such as domain-based [5], [6] or context-based [7], [8], [9]. Hence, more information channels should be considered when analyzing emotion in speech, one of which is the potential speaker dependency of human emotional prosody.

Empirical evidence of prosody variations for identical emotional sates in different people is of particular relevance to interactive systems and other social robotic applications. Provided a machine is capable of identifying an intervening speaker, the ability to further adapt itself to not only said speaker but also to their emotional conveyance mannerisms, can certainly boost the quality of the system's behavior and response suitability to the situation at hand. This concept is not unlike how pet social companions are able to perceive how their owners feel and modify their behavior to conform with the respective emotional state. Applications would be many and varied, from home assistants such as *Google Home* or Amazon's *Echo*, to active ageing and nursing aides in the medical field. Thus, it is of great usefulness to any field of inteligent robotics that emotion recognition techniques encompassing some level of specific speaker awareness and adaptation, be researched as the

one here proposed.

We introduce our approach to speech emotion recognition (SER), based on the use of the CNN model *VGGVox*, trained with over 100,000 utterances of 1,251 different speakers from the well-known *VoxCeleb1* dataset [10], for feature extraction from 6 standard and established emotional speech databases, with minimal preprocessing. Moreover, given that the set of databases used includes recordings in 8 different languages, the technique can be considered language independent.

The application of the features extracted using our technique, in state of the art classifiers, confirmed that speaker specific features extracted from speech do contain significant emotional information, robust enough to allow for clear classification of emotional states. Further, lower performance rates were observed for both under and over specialization of the speaker features extracted at different bottleneck levels of the applied CNN model, which highlights the importance of accurately leveling the balance between speaker and emotion recognition from speech.

This paper is divided in the following manner: section II provides an overview of recent related work while section III outlines the methodology of our approach, delving into the specifics of the proposed technique for emotional feature extraction using the cited speaker recognition model. This is followed by section IV where detail is given about the experiments carried out and the obtained supporting results are discussed, and finally section V where a conclusion and overview of future work are presented.

# 2 Related work

Given the necessity of evaluating a panoply of informational cues embedded in speech, added to the already complex task of considering as many vocal features as possible, the majority of classical recognition systems based on speech has seen their performance greatly surpassed by machine leaning models and more precisely, deep neural networks (DNN) [11]. As such, these architectures have been used as baselines in the performance evaluation of new emotion recognition techniques which use representations learned from other paralinguistic tasks.

Gideon *et al.* [12] assessed the effectiveness of progressive neural networks (ProgNets) at freezing the weights of a model's initial layers, tuned for speaker recognition and gender detection from speech, and using these

transitional representations as input to the posterior layers, trained for emotion recognition. Performance rates were somewhat higher than those of standard DNN or simple pre-training and fine-tuning (PT/FT) networks. On a separate note, Sidorov *et al.* [13], [14] explored the effects of adding speaker specific and gender information as features in the vectors used to train emotion recognition models, essentially further detailing the datafiles in one experiment. Parallely, the group predicted speaker and gender information with ANN-based recognizers, adding the obtained hypotheses to the feature sets fed into the used emotion recognizer. This plain method of extending the feature vector with additional speaker specific information was found to improve emotion recognition performance on both experiments. It was also found that including more specific speaker information into the feature vectors yielded better results than simply adding gender information.

Our work is novel in the sense that it does very minimal preprocessing on the raw data fed to the network and, instead of merely relying on the participation of actors, the transferred learning related to speaker recognition comes in the form of feature matrices generated directly by a large scale model trained with utterances from hundreds of persons with different ethnicities, accents, professions and ages. Plus, the referred past techniques included in their testing experiments the same utterances from speakers used for training, among other undesirable aspects. As such, high accuracy rates were already expected for in set participants whilst nothing can be concluded about the models' performance at evaluating emotions from out of set speakers. This path was not followed in our approach.

# 3 Methodology

In this section we provide an outline of the speech corpora used. Following that, detail is given on the applied *VGGVox* model, and on how feature matrices were extracted from different levels of its architecture, when fed the data from the given set of emotional databases.

## 3.1 Emotional Speech Corpora

For this work, a set of 6 emotional speech databases was gathered, made up of over 9000 utterances of varying duration in 8 different languages, portraying a total of 9 different emotional states, to be applied in a speaker

recognition model for feature extraction. The databases were pooled, forming a larger multi-language database, but the set reduced so as to only include the clips corresponding to anger, disgust, fear, happiness, sadness, surprise and the neutral state, common to all databases, which were the following:

### 3.1.1 EMODB

The Berlin Database of Emotional Speech [15] was recorded in the anechoic chamber of the Technical University of Berlin, GErmany, and it consists of 535 utterances in German of 10 different sentences, spoken by 10 professional actors (5 male, 5 female). The utterances are divided with the following labels: Anger, Joy, Sadness, Fear, Disgust, Boredom, Neutral.

### 3.1.2 EMOVO

The Emovo speech corpus [16] was recorded in the laboratories of the Ugo Bordoni Foundation in Rome, Italy, and it consists of 588 utterances in Italian of 14 different sentences, spoken by 6 professional actors (3 male, 3 female). The utterances are divided with the following labels: Anger, Joy, Sadness, Fear, Disgust, Surprise, Neutral.

### 3.1.3 SAVEE

The Surrey Audio-Visual Expressed Emotion database [17] was recorded in the University of Surrey, England, and it consists of 480 utterances in British English of 15 different sentences transcribed from the TIMIT database [18]. These sentences were spoken by 4 non-professional actors (all male), and the utterances were divided using the following labels: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral.

### 3.1.4 RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song [19] was recorded in a professional recording studio of the SMART lab of Ryerson University, Canada, and it consists of 1440 audio-video recordings of utterances in American English of 2 different sentences, spoken by 24 professional actors (12 male, 12 female). Only the audio element of this database was used, and the utterances are divided with the following labels: Anger, Happiness, Sadness, Fear, Disgust, Surprise, Calmness, Neutral.

### 3.1.5 RML

The Ryerson Multimedia Research Lab database [20] was recorded in the lab of the same name, at Ryerson University, Canada, and it consists of 720 video recordings of utterances in 6 different languages: English, Mandarin, Urdu, Punjabi, Persian, Italian. This corpus encompasses clips of 8 non-professional actors (all male), and ten different sentences for each emotional class. It should be noted that sentences were adapted to the language spoken and to the corresponding speaker's cultural background. Only the audio element of this database was extracted and used, being the utterances divided with the following labels: Anger, Disgust, Fear, Happiness, Sadness, Surprise.

### 3.1.6 ELRA-S0329

ELRA's Emotional speech synthesis database [21], or INTER1SP, was recorded in a noise-reduced room by the Polytechnic University of Catalonia, Spain, and it consists of 6041 utterances in Spanish. The text material is composed of 184 items, which were spoken by 2 professional actors (1 male, 1 female). The utterances are divided with the following labels: Anger, Joy, Sadness, Fear, Disgust, Surprise, Neutral.

## 3.2 The *VGGVox* Model

This model developed by Nagrani *et al.* [10], based on a VGG-M architecture [22] and composed of 12 layers, is quite capable of perceiving stimuli at specific locations of an image due to its convolutional design. Such a competency makes it ideal for analyzing spectral representations of audio signals and extracting a robust set of features.

In order to retain as much information as possible, the raw data undergoes minimal process in the sense that narrowband magnitude spectrograms are generated using a sliding hamming window of width 25ms and step 10ms, meaning an $n$-second input will provide a $100n$ frames spectrum. Normalisation is also performed on mean and variance, at every frequency bin of the spectrum, as it was observed that such a step produced an increase of 10% in classification accuracy. Yet, no other operations are performed on the input data, and the CNN is fed essentially raw spectrograms.

Variable length inputs are also efficiently dealt with by varying the support filter dimension of the *apool6* layer. As such, the implementation is

adaptable to an audio clip's duration, provided it is between 1 and 10 seconds in length, according to Table 1. The dimension values are conforming with the stride and padding methods used by the model, for each duration value. It should be noted that the model does handle clips longer than 10 seconds, by considering only the central 10-second segment of the clip, in spite of losing all the other potentially relevant surrounding information. A progression example of a 1-second audio input fed to the *VGGVox* model is provided in Figure 1, for reference.

Table 1: Average Pooling layer's k-th dimension adaptation to clip's n-second Duration

| **Frames** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| **Dimension** | | | | | | | | | |
| 2 | 5 | 8 | 11 | 14 | 17 | 20 | 23 | 27 | 30 |

In terms of purpose, the model was directed towards speaker classification, and trained using the *VoxCeleb1* dataset [10] also developed by Nagrani and her team. This dataset is of large scale, including over 100,000 utterances by 7000+ speakers of varied backgrounds, resulting in more than 2000 hours of audio. Consequently, the model is an ideal candidate for capturing copious amounts of speaker specific cues and prosody mannerisms from any type of human speech, emotional included. Training iterations also included batch normalization [23] and used the default hyper parameter values of the used MatConvNet toolbox [24].

## 3.3 Feature Extraction

In order to investigate the necessary level of speaker adaptation for greater accuracy in emotion recognition, using the gathered emotional speech corpora, feature arrays were obtained from the outputs of the three layers of the *VGGVox* model where bottleneck is apparent: *mpool5*, *apool6* and *fc7*. This was achieved by simply applying the model to the audio clips without any form of processing other than the already specified. As can be observed in Figure 1 and inferred for inputs of any length, the obtained arrays had the following dimensions, respectively: $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$. The first set of which was obtained by averaging the $9 \times k \times 256$ array along

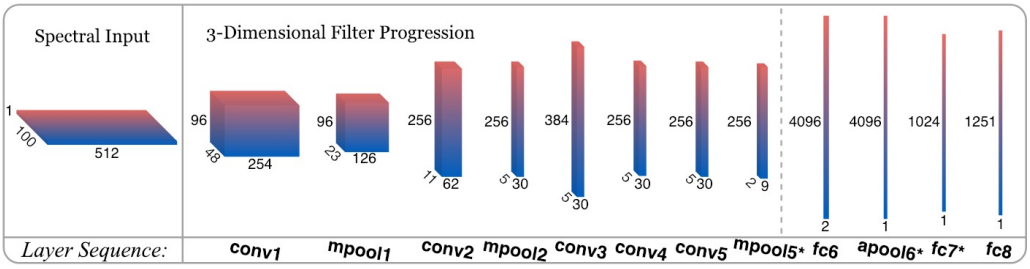



Figure 1: Exemplary diagram of a 100-frame spectral representation of a 1-second audio file progressing through the layers of the *VGGVox* model. The asterisk symbol is used to identify the layers whose outputs were used as feature matrices for emotional classification.

the $k$-th dimension, corresponding to the variable length $n$-second input (see Table 1).

# 4 EXPERIMENTAL RESULTS

Several experiments were carried out in order to evaluate the robustness and efficacy of the extracted feature arrays in terms of emotion recognition. The Weka software was employed so as to apply the feature arrays on the following state of the art classifiers: Naive Bayes [26], kNN [27], Random Forest [28], Logistic Model Tree (LMT) [29] and Support Vector Machine (SVM) [30]. A neural network based approach was not followed during the classification stage given the fact that the available data is not enough to credibly train a machine learning model, which by nature lacks any sort of initial direction, as opposed to statistical classifiers. In this section, we provide more detail on the carried out experiments and the obtained results, as well as a discussion of these.

## 4.1 Data Preparation

All files from the obtained databases were converted to the WAV format, at a sampling rate of $16kHz$, as it this value has been proved to be more than sufficient to capture all information embedded in a speech signal. In accordance with the *VGGVox* model's implementation, and in order to take full advantage of all the provided audio, files were adapted to be between 1

and 10 seconds in length. Therefore, a small amount of clips below the 1 second mark were disregarded, as these would hardly provide any emotional information, and clips above the 10 second mark were divided into equally long audio segments. As previously stated, $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$ feature arrays were extracted for each clip.

## 4.2 Classifier Performance

A simple Naive Bayers classifier was used as an efficacy baseline for evaluation against other state of the art classifiers in the Weka software, when fed the provided feature arrays for emotion recognition. Performance results were obtained using 5-fold cross validation, on the entire available emotional speech corpora as one large multi-language database, and are shown in Table 2. Furthermore, Cohen's Kappa [31] is also provided to further support the validity of the obtained results against random chance, parallel with unweighted average recall (UAR), a favoured metric in emotion recognition systems which attributes the same significance to all possible classes [32]. Given the observed higher classifier performance, using the $1 \times 1 \times 4096$ feature arrays, in the majority of scenarios, a second testing phase was carried out, under the same conditions, but applied to each emotional speech database individually. This was done in order to evaluate the degree to which language and cultural background have impact on a speaker's emotional prosody. The obtained results are displayed in Table 3.

## 4.3 Discussion

The results on the full emotional speech corpora, in Table 2, show how the $1 \times 1 \times 4096$ feature arrays appear to be, in most cases, more robust than their counterparts. This is likely due to the fact that $1 \times 1 \times 1024$ feature arrays are already too specialized for speaker classification, having put aside a great part of emotional information, whilst $9 \times 1 \times 256$ feature arrays are still too general, not having concentrated focus on any particular speaker or emotion representation yet. A visual depiction of a parameter's distribution across all classes is shown in Figure 2, for the three assessed feature array dimensions. This parameter was chosen based on its accurate depiction of the common morphology of most of the parameters' distributions, for the respective feature arrays. Evidently, this figure corroborates the previously stated, as the $1 \times 1 \times 1024$ parameter distribution is seemingly too specialized,

Table 2: State of the art classifier performance on full emotional corpora feature arrays.

| | **Naive Bayes** | | | **k-Nearest Neighbor** | | | **Random Forest** | | | **Logistic Model Tree** | | | **Support Vector Machine** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* |
| 9x1x256 | 52.2% | 0.44 | 0.50 | 76.4% | 0.72 | 0.75 | 74.3% | 0.70 | 0.72 | 75.9% | 0.72 | 0.74 | 76.5% | 0.72 | 0.75 |
| 1x1x4096 | 56.7% | 0.49 | 0.54 | 80.1% | 0.77 | 0.79 | 77.3% | 0.73 | 0.76 | 81.1% | 0.78 | 0.80 | 76.8% | 0.73 | 0.75 |
| 1x1x1024 | 55.4% | 0.47 | 0.53 | 78.6% | 0.75 | 0.77 | 76.0% | 0.72 | 0.74 | 76.5% | 0.72 | 0.75 | 85.6% | 0.83 | 0.84 |

Table 3: State of the art classifier performance on standalone emotional database $1 \times 1 \times 4096$ feature arrays.

| | **Naive Bayes** | | | **k-Nearest Neighbor** | | | **Random Forest** | | | **Logistic Model Tree** | | | **Support Vector Machine** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* | *Accuracy* | *k-Statistic* | *UAR* |
| EMODB | 72.9% | 0.67 | 0.73 | 74.9% | 0.69 | 0.73 | 76.0% | 0.70 | 0.72 | 80.4% | 0.76 | 0.80 | 74.9% | 0.68 | 0.72 |
| EMOVO | 52.8% | 0.45 | 0.53 | 66.2% | 0.61 | 0.66 | 65.3% | 0.60 | 0.65 | 68.5% | 0.63 | 0.69 | 57.9% | 0.51 | 0.58 |
| RAVDESS | 51.8% | 0.44 | 0.53 | 65.5% | 0.60 | 0.65 | 61.9% | 0.55 | 0.60 | 71.6% | 0.67 | 0.71 | 55.0% | 0.47 | 0.60 |
| RML | 70.4% | 0.65 | 0.70 | 73.5% | 0.68 | 0.73 | 75.3% | 0.70 | 0.75 | 79.9% | 0.76 | 0.80 | 71.3% | 0.66 | 0.71 |
| S0329 | 74.7% | 0.70 | 0.74 | 86.2% | 0.83 | 0.84 | 87.4% | 0.85 | 0.85 | 92.4% | 0.91 | 0.91 | 91.0% | 0.89 | 0.90 |
| SAVEE | 61.67% | 0.54 | 0.58 | 65.2% | 0.59 | 0.61 | 64.8% | 0.57 | 0.60 | 70.4% | 0.65 | 0.68 | 55.6% | 0.45 | 0.49 |

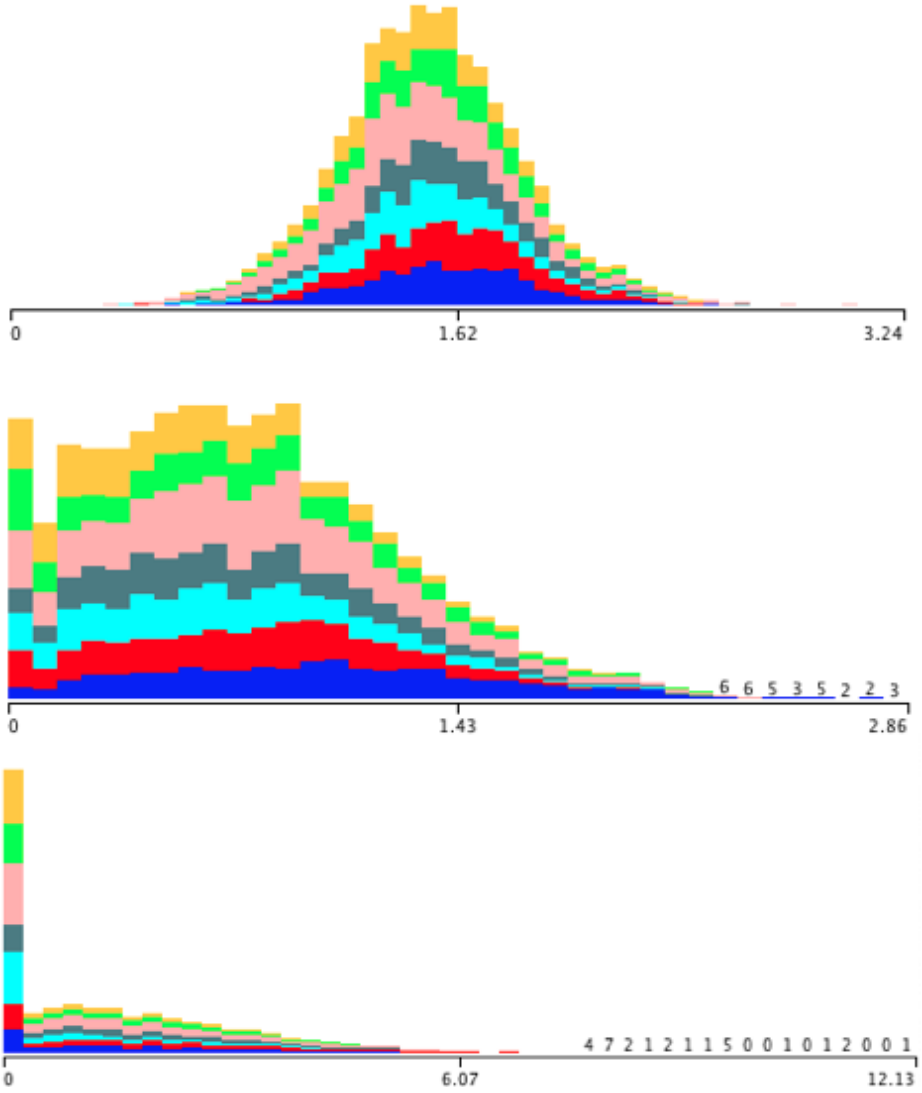



Figure 2: Distributions of the same parameter across all classes, for the $9 \times 1 \times 256$, $1 \times 1 \times 4096$ and $1 \times 1 \times 1024$ feature arrays, respectively. Each is representative of the distributions of most parameters, for the respective feature arrays.

while the $9 \times 1 \times 256$ respective distribution appears yet too general, making

the $1 \times 1 \times 4096$ feature array the ideal choice in this scenario. The only exception observed on Table 2, where $1 \times 1 \times 4096$ feature performance is surpassed, is given by the $1 \times 1 \times 1024$ features applied to an SVM. The overspecialization of the latter may justify this. Given the spatial mapping nature of the SVM algorithm, an already clear separation between the feature parameters ultimately gives SVM an advantage over other non-spatial based classifiers. In fact, kNN does give the second best performance rate using these features.

In terms of the database specific classification experiments, by observing Table 3, results are varying from database to database. Though database size must also be taken into consideration, given the language and cultural diversity between databases, the obtained results suggest that emotional prosody is affected differently in each population. As such, adaptation to cultural background, in addition to speaker, is likely an approach worthwhile researching in order to improve SER systems.

Altogether, the obtained results always surpassed the proposed baseline having LMT given the best results when considering the overall leading option of $1 \times 1 \times 4096$ features. However, if considering $1 \times 1 \times 1024$ features, spatial-based classifiers such as SVM and kNN become a better choice. In terms of the $9 \times 1 \times 256$ features, classifier performance was well balanced. Finally, all results certainly support the existence of relevant emotional information in speaker specific speech features, confirming our hypothesis. As such, speaker adaptation should be performed in systems attempting to perform successful SER.

# 5 CONCLUSIONS

In this paper, we examined the robustness of speech features extracted using a large scale speaker recognition model, for emotion recognition. We determined that, regardless of language, there is valuable emotional information embedded within speaker specific features, particularly on the ones right before the last bottleneck stage of the model. Posteriorly, acceptable but varying performance ratios were obtained on standalone databases using only the best scenario corresponding feature arrays from the previous experiment. This suggests varying degrees of emotional prosody mannerism for different cultural backgrounds. Finally, and based on a general observation of the results, we are able to conclude that an initial step of speaker adaptation

is of paramount importance and should be performed in any SER systems, so as to achieve higher accuracy rates.

In the future, we intend to assess the efficacy of dimension reduction techniques such as PCA or LDA applied to the feature arrays, in order to reduce them to their core information. Further, we intend to delve deeper into adaptable emotion recognition, by considering additional speaker information, such as cultural background, and also incorporating facial expression analysis into a multi-modal emotion recognition system, in order to ultimately embed such a capability in an interactive robot.

# ACKNOWLEDGMENTS

The authors would like to thank the respective database owners and curators for providing access to their emotional speech datasets and for allowing their use in our research. Without these, empirical results would not have been attainable. This work has been partially supported by OE - national funds of FCT/MCTES (PIDDAC) under project UID/EEA/00048/2019.

# References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias,W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computerinteraction", IEEE SIGNAL PROCESSING MAGAZINE, 18(11):3280,2001.

[2] Jürgens, Rebecca et al. "Effect of Acting Experience on Emotion Expression and Recognition in Voice: Non-Actors Provide Better Stimuli than Expected" Journal of nonverbal behavior vol. 39,3 (2015): 195-214.

[3] Paulmann S, Furnes D, Bøkenes AM, Cozzolino PJ. "How Psychological Stress Affects Emotional Prosody", (2016). PLOS ONE 11(11): e0165022.

[4] M. Spada, A. Nikčevic, G. Moneta and A. Wells. "Metacognition, perceived stress, and negative emotion". Science Direct. Personality and Individual Differences 44 (2008) 1172–1181.

[5] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition", IEEE SIGNAL PROCESSING LETTERS, 2017.

[6] J. Deng, Z. Zhang, F. Eyben and B. Schuller, "Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition," in IEEE SIGNAL PROCESSING LETTERS, vol. 21, no. 9, pp. 1068-1072, Sept. 2014.

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognitionusing a deep convolutional recurrent network", IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP), 2016.

[8] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutionaland recurrent neural networks.Asia-Pacific Signal and Information Processing AssociationAnnual Summit and Conference (APSIPA), 2016.

[9] Lee, Jinkyu, and Ivan Tashev. "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[10] A. Nagrani, J. S. Chung and A. Zisserman, "VoxCeleb: A Large-ScaleSpeaker Identification Dataset", INTERSPEECH, 2017.

[11] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, 8 2017.

[12] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis and E. M. Provost, "Progressive Neural Networks for Transfer Learning in Emotion Recognition", INTERSPEECH, 1098-1102, 2017.

[13] M. Sidorov, S. Ultes and A. Schmitt. "Comparison of Gender- and Speaker-adaptive Emotion Recognition.", LREC (2014).

[14] M. Sidorov, S. Ultes and A. Schmitt. "Emotions Are A Personal Thing: Towards Speaker-Adaptive Emotion Recognition", ICASSP, (2014).

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. "A database of german emotional speech". INTERSPEECH, 2005.

[16] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco. "Emovo corpus: an italian emotional speech database". LREC (2014).

[17] S. Haq, P.J.B. Jackson, and J.D. Edge. "Audio-Visual Feature Selection and Reduction for Emotion Classification". In Proc. Int'l Conf. on Auditory-Visual Speech Processing, pages 185-190, 2008.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue. "TIMIT Acoustic-Phonetic Continuous Speech Corpus". LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[19] S. R. Livingstone and F. A. Russo. "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english". Plos One, 2018.

[20] Z. Xie and L. Guan. "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools". IEEE International Conference on Multimedia and Expo (ICME), 2013.

[21] European Language Resources Association (ELRA). Emotional speech synthesis database Elra-S0329.

[22] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets". In Proceedings of the British Machine Vision Conference, 2014.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[24] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for MATLAB," CoRR, vol. abs/1412.4564, 2014.

[25] E. Frank, M. A. Hall and I. H. Witten, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[26] G. H. John, P. Langley. "Estimating Continuous Distributions in Bayesian Classifiers.", In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.

[27] D. Aha, D. Kibler. "Instance-based learning algorithms.", Machine Learning (1991). 6:37-66.

[28] L. Breiman. "Random Forests." Machine Learning (2001). 45(1):5-32.

[29] N. Landwehr, M. Hall, E. Frank. "Logistic Model Trees." Machine Learning (2005). 95(1-2):161-205.

[30] CC Chang, and CJ Lin. "LIBSVM: A library for support vector machines". ACM Transactions on Intelligent Systems and Technology, 2011, Vol. 2(3), pp 27:1–27:27.

[31] J. Cohen, "A coefficient of agreement for nominal scales". Educational and Psychological Measurement (1960). 20 (1): 37–46.

[32] B. W. Schuller, S. Steidl, A. Batliner et al., "The interspeech 2009 emotion challenge." in Interspeech, vol. 2009, 2009, pp. 312–315.