Co-funded by the
Erasmus+ Programme
of the European Union

Emotion Recognition from Speech using Machine Learning algorithms

Author: Natallia Chaiko

Advisors: Prof. Eva Navas[1]
Prof. Roberto Zamparelli[2]

[1] University of the Basque Country
[2] University of Trento

European Masters Program in
Language and Communication Technologies (LCT)

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Master's Thesis

September 2019

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineering.

Abstract

Speech emotion recognition (SER) is one of the most popular research directions nowadays, however, most of the solutions suggested by the researchers are tested on different databases and different features, which makes it impossible to compare their results. In this work, we aim to compare three of the most commonly used speech emotion classifiers on the database of spontaneous emotional speech to find the most successful one. Three feature sets have also been used in attempt to find the best speech signal representation.

After a number of experiments, SVM has turned out the best of the three solutions on the classification of two and three classes of emotions.

Keywords: emotion, speech emotion recognition, SER systems, SVM, DNN, LSTM

Acknowledgments

First of all, I would like to thank my supervisors, professor Eva Navas for support and guidance throughout the project and professor Roberto Zamparelli for nonstop cheering up. I am really grateful to my colleagues at Neosound Intelligence for letting me be part of their team and constant reminding me of the deadlines.

I would also like to express my gratitude to the European Union Erasmus Mundus programme for selecting me as one of the scholarship holders of the Language and Communication Technologies programme and thus, allowing me to gain all the experience and knowledge, meet new people and extend my horizons.

Finally, I wish to thank my family and friends for their encouragement throughout my study.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

Acronym/Abbreviation	Description
ANN	Artificial neural network
ASR	Automatic speech recognition
BLSTM	Bidirectional long short-term memory network
CN	Chinese language
CNN	Convolutional neural networks
DA	Danish language
dB	Decibel
DCT	Discrete cosine transform
DE	German language
DNN	Deep neural network
EL	Greel language
ELM	Extreme learning machine
EN	English language
F0	Fundamental frequency
FFT	Fast Fourier transform
FR	French language
HMM	Hidden Markov Model
HNR	Harmonics-to-noise ratio
HU	Hungarian language
Hz	Herz
JP	Japanese language
LPC	Linear predictive coefficients
LPCC	Linear predictive cepstral coefficients
LSTM	Long short-term memory
MEDC	Mel energy spectrum dynamics coefficients
MFC	Mel-frequency cepstrum
MFCC	Mel-frequency cepstral coefficients
PCA	Principal component analysis
PDF	Probability density function
RBF	Radial basis function
RNN	Recurrent neural network
SER	Speech emotion recognition
SR	Serbian language
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TH	Thai language
UA	Unweighted accuracy
WA	Weighted acuracy
ZCR	Zero-crossing rate

Table 1. Table of acronyms and abbreviations

1 Introduction

Automatic emotion (affect) detection is currently among the most popular research directions in the field of computer science. A lot of applications are being developed to be able to recognize human emotions. These applications can serve as separate systems, for example used in call centres for distinguishing emotional calls from customers, or as parts of larger systems such as personal assistants, information providers, receptionists etc.

However, before developing an automatic emotion detection system it is necessary to look into theoretical research done in the field of human emotions if truly effective systems are to be achieved. The main reason for that is because human interaction is a complicated process, where the same words may be used as a joke, a threat, or as a genuine question looking for an answer. Since the way humans communicate is closely connected to their emotional state, a system that is unable to tell the difference will have difficulty in detecting emotions in case of emotion detection systems or producing correct responses in case of machine communicators.

Nonetheless, as Roddy Cowie together with his co-authors mentioned in (Cowie et al. 2001), in the context of automatic emotion recognition, understanding the nature of emotion is not an end in itself. It matters mainly because ideas about the nature of emotion imply that certain features and relationships are relevant to describing an emotional state, distinguishing it from others.

The study of human emotions has a long history, and many emotion theories have been proposed over time, which also means that there is no agreement among scientists on the definition of an emotion. We provide an overview of currently the most dominant ones in the following chapter.

1.1 Master’s thesis motivation and objectives

Although a lot of research has been done in the field of speech emotion recognition, different classification models have been proposed together with various speech features as a speech signal representation, the experiments are usually performed on different databases, which makes their comparison impossible. Moreover, a great deal of the experiments are executed on the databases with non-spontaneous speech, which may be useful for some applications, however, for a speech emotion recognizer trained on non-

spontaneous speech, it may be challenging to detect emotions happening in real-life conversations.

This master's thesis's purpose is to present the traditional speech emotion recognition systems as well as the systems currently used in the field and test some of the speech emotion classifiers, which have proved to be successful for the task, on the same database with the same set of features representing a speech signal to see which of them is more capable of correctly detecting emotions. Not many similar researches have been done by others (Schuller 2018). Furthermore, the database, on which the experiments are performed, contains spontaneous real-life conversations, and since it has been proved that acted emotions are more easily recognized than realistic emotions (Vogt and André 2005), the classification of the latter is more challenging. Therefore, another objective of the thesis is to check how well the classifiers are able to detect real emotions happening in everyday life.

1.2 Master's thesis outline

This document is organised as follows: Section 2 discusses main emotion theories, speech features employed to represent a speech signal and some of the commonly used speech emotion classifiers. Section 3 describes the speech emotion recognition systems utilised for the experiments, the database on which the experiments have been performed and feature sets used for a speech signal representation. Section 4 contains the results of the experiments and their analysis. In Section 5, an overview of the work done is presented and possible directions for future work are outlined.

2 Literature Review

This section presents theoretical review on emotion theories, as well as the description of the traditional approach to building a speech emotion recognition system, which includes the summary of speech features and an overview of existing speech emotion classifiers. We also explore some of the models currently used in the sphere of speech emotion recognition together with the experimental results obtained during their testing.

2.1 Emotion theories

2.1.1 Emotions as Expressions

Perhaps the most long-standing theory of emotions is described in terms of discrete categories. Darwin was the first to scientifically analyse emotions (Darwin 1872). His theory provides evidence that six emotions – happiness, sadness, anger, fear, disgust and surprise – are universal and can be recognized from facial expressions (hence, the name of the theory), therefore are considered basic. This view was also supported by Ekman (Ekman 1971), who indicated that humans perceive these emotions with respect to facial expressions in the same way, regardless of culture.

Because of the strong influence of the theory of basic emotions, a lot of automatic affect recognition systems focus on recognizing these six basic emotions. This and the fact that it is convenient and intuitive to describe observed emotions in categories make the theory even more attractive. However, basic emotions illustrate only a part of human everyday emotional display, so a deeper view on emotions needs to be developed.

2.1.2 Continuous dimension approach

The approach models emotions in several dimensions, arousal, or activation and valence being the main ones. Activation refers to the amount of energy needed to produce a certain emotion and it is known to be well accessible through acoustic features. According to the studies of Williams and Stevens (Williams and Stevens 1981) of the emotion production mechanism, the sympathetic nervous system is aroused when happiness, anger and fear are expressed, which leads to increased heart rate and blood pressure and dryness of the mouth. The corresponding speech is loud and fast, uttered in a strong high-frequency energy, a wider pitch range and a higher average pitch. Thus, happiness, fear and anger are considered high activation emotions. Contrarily, when the parasympathetic nervous

system is aroused, in case of low activation emotions such as sadness, heart rate and blood pressure decrease and salivation increases, leading to a slow, low-pitched speech with little high-frequency energy. Nonetheless, emotions cannot be distinguished through activation dimension only, as, for example, happiness and anger, being high activation emotions, are completely different in their valence, the former having a positive valence and the latter – a negative one. However, there is no agreement among researchers on how or if there is a correlation between acoustic features and a valence dimension (Liscombe 2007). Another dimension that is used to describe emotions is potency, or power, which describes how strong the emotion is (Schlosberg 1954). However, usually a simpler, two-dimensional model (activation and valence dimensions) is employed. Both approaches can be combined to locate discrete emotions (see Figure 1).

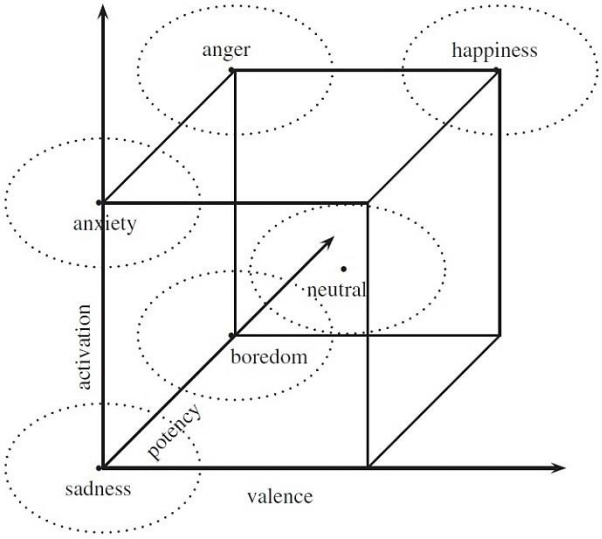


Figure 1. Mapping of discrete emotions to continuous emotion space (taken from (Yang and Lugger 2010))

2.1.3 Emotions as Embodiments

The author of the theory, James (James 1884), and later Lange (Lang 1994) described emotions as a combination of expressions and physiology, in particular, he interpreted emotions as physiological changes. This theory can help affective computing in such a way that emotion recognition systems can detect emotions by analysing the pattern of physiological changes associated with each emotion.

2.1.4 Cognitive approaches to Emotions

Cognitivists believe that in order for a particular emotion to arise, a human needs to appraise a certain event or an object as affective in some way, in other words, emotions are caused by appraisals of the situations, and not by the situations themselves. Appraisal theorists also add that two individuals can feel the same emotion when both their appraisals of the event are the same, and different emotions can arise in one individual at different times if their appraisals of the situation are different (Roseman, S. Spindel, and Jose 1990).

2.1.5 Emotions as Social Constructs

There is a group of scientists, with Averill (Averill 1980) in the lead, who claim that emotions can be explained by physiology and cognition only. They identify the appearance of emotions with the social element of human life. There are several traditions explaining the social nature of emotions. Calvo describes two of them in (Calvo and D'Mello 2010). Dramaturgical approach emphasizes the importance of cultural norms in experience of emotions, i.e. traditions and social rules manage when humans feel certain emotions and how they express them. Structural approach analyses social structure of emotions, with power and status in society being the most relevant dimensions (Kemper 1991). Gain in power and status is associated with well-being and positive emotions, while the opposite will trigger negative emotions.

2.1.6 Neuroscience

The study of emotions from neurological point of view focuses mainly on helping understand emotional processes and their neural correlates. One of the main achievements of affective neuroscience has been to provide evidence against the concept that emotions and thoughts appear in different parts of the brain, while, in fact, neural substrates of cognition and emotion overlap substantially (Calvo and D'Mello 2010). Certain cognitive processes such as memory encoding and retrieval, goal appraisal and planning permanently rely on the experience of emotions. This fact again proves that emotional aspect should be considered in any human-computer interaction.

2.2 Emotion detection

Human emotions can be detected from different channels: facial expressions, voice, body language and posture, physiology, including cognitive processes, text, brain imaging and

their combinations (Calvo and D'Mello 2010). There are numerous studies that deal with each of the sources and thus computer scientists and engineers use video, voice recordings, text, and physiology as data for building automatic affect classifiers. But here we will focus on emotion detection from speech, as it will be the main focus of the project.

Speech emotion recognition has proved to be useful in a number of spheres, where emotion detection is important. These include applications requiring natural human-machine interaction such as computer tutorial applications or web movies, where the response of the machine depends on the detected emotion of the user (El Ayadi, Kamel, and Karray 2011). It can also be used as a diagnostic tool for therapists (France et al. 2000). It may help automatic machine translation systems find correct translations, as word and sentence meanings may change due to certain emotional states of the speakers. Speech emotion recognition has also been used in call-centres for improving automatic responses to customers and automatizing data processing, particularly, highlighting the most affective calls to the managers for further analysis.

2.3 Building an emotion recognizer: the traditional approach

2.3.1 Modelling

As has been mentioned in a previous chapter, when building an emotion recognition system, it is important to acquire an adequate representation model of emotions. On the one hand, it should agree with psychological theories, while on the other hand, be convenient enough for the machine to be able to handle it. Two models are usually used for this purpose (Schuller 2018). The first model comprises discrete classes, as has been described in Section 2.1.1. The second model uses a continuous dimension approach, as has been illustrated in Section 2.1.2. Another aspect of modelling the emotion representation is the quality and making of emotions, such as acted, elicited, naturalistic, pretended, and regulated (Schuller 2018).

2.3.2 Annotation

After having decided which model is going to be used, the next step is to acquire the labelled data for training and testing that model. A peculiarity of the task is a relatively high subjectivity, uncertainty about the target labels and disagreement of the annotators on the expressed emotions (Schuller 2018). In fact, in a survey aimed to measure human

performance on emotion recognition, it was found that only 60 percent of people can correctly identify other people's emotions (Schuller, Reiter, and Rigoll 2006). There can be two ways to solve the problem. First, self-assessment techniques can be applied, i.e. people, whose voices are recorded with a certain emotion, report what emotion is expressed. However, it can be tricky, as no one can remember exactly what he or she felt at a particular moment of time. Therefore, a more appropriate solution is to use an observer rating. Usually, several external annotators label the expressed emotions, thus forming the basis of the target labels by majority vote, or average in the case of a continuous emotion representation (Gunes and Schuller 2013). Then, the elimination of the outliers as well as the weighting of annotators should be performed based on their agreement/disagreement with the majority of the raters. This way only the most relevant data and maximally objective annotation are selected for the classification.

There are a number of emotional speech databases (see Section 2.3.2.1), which allow to avoid needs of annotation, as they use professional actors to express emotions, who are well capable of the job. However, the recorded emotions may sound unnatural and forced, in comparison to the emotions happening in real conversations, so the machine trained on the former may have difficulty recognizing the latter. Therefore, it is reasonable to simply wait for the emotion to be expressed in a real-life conversation and then add it to the collected data.

2.3.2.1 Emotional speech databases

An important issue in building a speech emotion recognizer is the database used for its training and evaluation. If the database is of low quality, i.e. the annotations are not reliable, or the emotions produced are not natural enough, the results of the experiments may be unsatisfactory, or wrong conclusions may be drawn. The choice of the database for the experiments depends on the classification task: for example, the recognition of emotions in infants (Slaney and McRoberts 2003), in adults (Trigeorgis et al. 2016), the recognition of stress in speech (Vignolo et al. 2016), etc.

There are quite a few emotional speech databases known to have been used in the experiments on speech emotion recognition. However, only a part of them are available for public use. This leads to a number of problems for researchers working in this field: i) there is no coordination (or little coordination in some cases) between the researchers, so the same mistakes can be repeated during the recording of the speech databases (El

Ayadi, Kamel, and Karray 2011); ii) the experiments are mostly performed on different databases, so it is impossible to compare them and choose the best one (Schuller 2018). Some of the most commonly used databases and their characteristics are presented in Table 2.

Name of the database	Lng	Size		Source	Emotions/description of a recording set	Access
		Hrs	Spkrs			
BabyEars (Slaney and McRoberts 2003)	EN	~0.2[1]	12	Mothers and fathers	Approval, attention, prohibition	Private
Berlin emotional database (Burkhardt et al. 2005)	DE	~0.3[2]	10	Professional actors (adults)	Anger, boredom, disgust, fear, happiness, sadness, neutral	Public, free
BHUDES (Fu, Mao, and Chen 2008)	CN	~1[3]	7	Actors	Anger, joy, sadness, disgust, surprise	Partially available
CCDb (Aubrey et al. 2013)	EN	5	13	Volunteers from School of Computer Science at Cardiff University	Happiness, surprise, confusion, thoughtfulness; Natural dyadic[4] conversations	Free, available on request
CHEAVD (Li et al. 2017)	CN	2.3	238	Films/TV (aged 11 – 62)	26 non-prototypical emotional states, including the basic six	Public, free
CHEAVD 2.0 (Schuller 2018) (extension of CHEAVD)	CN	7.9	527	Films/TV (aged 11 – 62)	26 non-prototypical emotional states, including the basic six	Public, free
Danish emotional Database (Engberg and Hansen 1996)	DA	0.17	4	Nonprofessional actors	Neutral, surprise, happiness, sadness, anger	Public with license fee
FAU AEC (Steidl 2009)	DE	8.9	51	Children (aged 10 – 13)	11 emotion categories	Public, free

[1] The provided number of hours is approximate, since the authors have not provided the exact information.
[2] See Footnote 1.
[3] See Footnote 1.
[4] Dyadic – consisting of two elements or parts; double sided.

Name of the database	Lng	Size		Source	Emotions/description of a recording set	Access
		Hrs	Spkrs			
IEMOCAP (Carlos Busso et al. 2008)	EN	12	10	Professional actors (adults)	Excitement, frustration, happiness, sadness, anger, surprise, neutral	Free, available on request
INTER1SP (Kerkeni et al. 2019)	SP	~3[5]	2	Professional actors (adults)	Anger, sadness, joy, fear, disgust, surprise and neutral	Partially available
Japanese emotional speech database (Mori, Tsuyoshi, and Ozawa 2006)	JP	~0.1[6]	1	1) Japanese male speaker; 2) Synthesized speech	1) anger, boredom, disgust, fear, happiness, sadness, neutral 2) 12 emotions including anger, surprise, disgust, sorrow, boredom, depression, and joy	Free, available on request
KISMET (Breazeal and Aryananda 2002)	EN	~0.4[7]	3	Nonprofessional actors	Approval, attention, prohibition, soothing, neutral	Private
RECOLA (Ringeval et al. 2013)	FR	9.5	46	French speakers	Online dyadic interactions	Partially available
SAVEE (Haq and Jackson 2009)	EN	~0.3[8]	4	Postgraduate students	Anger, disgust, fear, happiness, sadness, surprise, neutral	Public, free
SEMAINE (Mckeown, Valstar, and Cowie 2007)	EN	80	150	Undergraduate and postgraduate students	Fear, anger, happiness, sadness, disgust, contempt, amusement; Human-machine interactions	Free, available on request
SEWA (Kossaifi et al. 2019)	EN DE CN EL HU SR	44	398	Group of volunteers (aged 18 – 60+)	Watching videos – dyadic conversations	Publicly available
Audiovisual Thai Emotion Database	TH	~9[9]	6	Drama-students	Happiness, sadness, surprise, anger, fear, disgust;	Free, available on request

[5] See Footnote 1.
[6] See Footnote 1.
[7] See Footnote 1.
[8] See Footnote 1.
[9] See Footnote 1

Name of the database	Lng	Size		Source	Emotions/description of a recording set	Access
		Hrs	Spkrs			
(Stankovic, Karnjanadecha, and Delic 2011)					Reading 1,000 most common Thai words	

Table 2. Characteristics of common speech emotion databases

2.3.3 Speech features

Before feeding the labelled data into some machine learning algorithm for the classification of emotions, one usually needs characteristic audio features. This is an ongoing research in the field of speech emotion recognition: from the beginning of the research in the field up to now, the researchers have not yet found a set of speech features, which best reflects the emotional content (El Ayadi, Kamel, and Karray 2011), (Schuller 2018).

There are different classifications of speech features, some divide them into continuous, qualitative and spectral features (El Ayadi, Kamel, and Karray 2011), others as in (Johnstone and Scherer 2000), do not consider voice quality features as a separate class of features. In the following chapter, we elaborate on the features of each category, mentioning their pros and cons. However, when representing a speech signal, it is common to combine features from all the groups.

2.3.3.1 Continuous speech features

Researchers believe that prosody continuous speech features such as energy and pitch carry most of the emotional content of a phrase (Cowie et al. 2001), (Busso, Lee, and Narayanan 2009). Studies described in (Williams and Stevens 1981), (Johnstone and Scherer 2000) confirm that the arousal state of the speaker (described in detail in Section 2.3.1) influences the overall energy, frequency and duration of the pauses and energy distribution across the frequency spectrum.

According to (Johnstone and Scherer 2000) continuous speech features can be classified into the following groups: time-related features, features related to fundamental frequency (pitch-related features), intensity (energy)-related features and time-frequency-energy features. The first three groups are related to speech rate, pitch and loudness, respectively, while the fourth is linked to timbre and voice quality. As speech is a constantly changing, dynamic signal, acoustic features are usually measured on short

segments, called frames, where speech is considered more or less stable. Such features are called short-term, or local, features. However, since emotion is believed to be a lasting phenomenon, it is also common to obtain long-term, or global, features, by combining short-term features over longer time frames.

2.3.3.1.1 Time-related features

A speech signal is a temporal sequence of sounds (voiced and unvoiced) and silences, both of which can carry its affective information. The most commonly used timing features in speech emotion recognition are speech rate, ratio of duration of voiced and unvoiced regions, and duration of the longest voiced speech (El Ayadi, Kamel, and Karray 2011), which are expected to vary for different emotions. In practice, there is no clear evidence about which vocal cues are characteristic of which emotion, the same also goes for silences (Johnstone and Scherer 2000).

2.3.3.1.2 Features related to fundamental frequency

Fundamental frequency (F_0) is one of the most frequently used speech feature for describing a speech signal. F_0, measured in cycles per second, or Herz (Hz) and it is the rate at which vocal cords vibrate (open and close) when a certain sound is pronounced. It is perceived as the pitch of the voice. Some features derived from F_0 include mean and median, minimum and maximum, standard deviation, variance, range (the difference between maximum and minimum F_0), jitter (the degree of the fluctuations of the time for the vocal cords to vibrate and how they differ from one cycle to the next) (El Ayadi, Kamel, and Karray 2011), (Johnstone and Scherer 2000).

2.3.3.1.3 Intensity-related features

Intensity is related to the amount of energy used to produce a speech sound and is perceived as loudness. It is measured in decibel (dB), a logarithmic transform of the intensity. Minimum, maximum, range, mean, median, standard deviation of the energy are among regularly calculated statistics of intensity applied in the emotion recognition from speech (El Ayadi, Kamel, and Karray 2011).

2.3.3.1.4 Combined time-frequency-energy features

During the production of speech the structure of the vocal tract changes due to various placements of the articulators, such as tongue, lips and teeth. Each articulatory configuration forms a different resonant cavity, producing a specific sound of a language,

which is characterised by these amplified frequencies, which are called formants. In affect recognition systems first and second formants, their bandwidths (the range of frequencies amplified in a given formant) and formant amplitudes (the amount of resonance) are commonly used as acoustic features (El Ayadi, Kamel, and Karray 2011), (Johnstone and Scherer 2000).

Several studies have shown that there is a relation between prosodic speech features and emotions (Cowie et al. 2001), (Johnstone and Scherer 2000), (Gunes and Schuller 2013), (Murray and Arnott 1993), (Öster and Risberg 1986). However, the boundary for the distinction of emotions from prosodic features is unclear, i.e. there are similarities between characteristics of some emotions, making it difficult to differentiate between them. For example, Öster and Risberg in (Öster and Risberg 1986) report that there is an overlap of median fundamental frequency and the total range of variation between the emotions of anger, fear and 'positive', the latter may correspond to 'joy' in terms of basic emotions.

2.3.3.2 Spectral features

Another group of speech features often selected as a representation of a speech signal are spectral-based features[10], since it has been found that the emotional content of an utterance influences the spectral energy distribution across the speech range of frequency (Nwe, Foo, and De Silva 2003). For example, it is shown in (Banse and Scherer 1996) that the utterances with the emotion of sadness have low energy at high frequencies, whereas utterances with the happiness emotion have high energy levels at the same range of frequencies.

Spectral features can be obtained in a number of ways, linear predictive coefficients (LPC) model and bank-of-filters model being the most commonly used ones, as they proved to be efficient in speech processing (Rabiner and Juang 1986).

LPC model performs spectral analysis on speech frames with an all-pole modelling constraint. This means that the resulting spectral representation of a signal should have the form of $\sigma/A(e^{jw})$, where $A[e^{jw}]$ is a p^{th} order polynomial with z-transform (Equation 1) and σ is a sigmoid function (Equation 2):

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p} \quad (1)$$

[10] Spectrum of a time-domain signal is a representation of that signal in the frequency domain.

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad (2)$$

The order, *p*, is called the LPC order. The output of the LPC spectral analysis is a vector of coefficients that specify the spectrum of the speech frame under analysis (Rabiner and Juang 1986).

A filter is a device or process that removes some unwanted components in a speech signal. Most often, it means removing some frequencies or frequency bands. When a bank-of-filters model is used to obtain spectral information of a signal, the signal is passed through a bank of bandpass filters, i.e. a certain range of frequencies are allowed to pass in each filter, while frequencies outside that range are rejected. The output of the bandpass filter is the spectral representation of a signal, i.e. spectral features are extracted from the outputs of the bandpass filters.

As humans do not perceive pitch linearly (Rabiner and Juang 1986), the filters' bandwidths are usually evenly distributed along some nonlinear frequency scale such as the Mel frequency scale, the Bark scale (Rabiner and Juang 1986) and the ExpoLog scale (Bou-Ghazale and Hansen 2000).

Mel frequency scale is used to obtain mel-frequency cepstrum (MFC), which is a representation of the short-term power spectrum[11] of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Incorporating this scale makes speech features match more closely what humans hear. Mel-frequency cepstral coefficients (MFCCs) are coefficients that make up an MFC.

One needs to perform the following operations to compute MFCCs from a signal. First, it is necessary to take the Fourier transform of a signal frame. Fourier transform is a function representing a signal in terms of its constituent frequencies, i.e. it transforms a signal from a time domain into a frequency domain according to expression (3).

$$X(f) = F\{x(t)\} = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft}dt, \qquad (3)$$

where *t* represents time, *f* represents frequency. Then, one should map the powers of the spectrum obtained with the Fourier transform onto the mel scale, filtered by a triangular band pass filter bank. Mel scale is calculated as

[11] Power spectrum is a representation of a magnitude of frequency components of a signal. Power spectrum gives information on how much energy (power) certain frequencies contain in a given signal.

$$Mel(f) = 2595 \log_{10} 1 + \frac{f}{700}, \qquad (4)$$

where *Mel(f)* is the logarithmic scale of the normal frequency scale *f*. MFCCs are then calculated as

$$C_n = \sqrt{\frac{2}{k}} \sum_{K=1}^{K} \log S_k \cos[n(k-0.5)\pi/k], \;\; n = 1, 2, \ldots, N, \qquad (5)$$

where S_k *(k = 1, 2, ..., K)* is the output of the filter banks and *N* is the total number of signal samples (Xu et al. 2004).

The scheme of the MFCCs computation is shown in Figure 2.

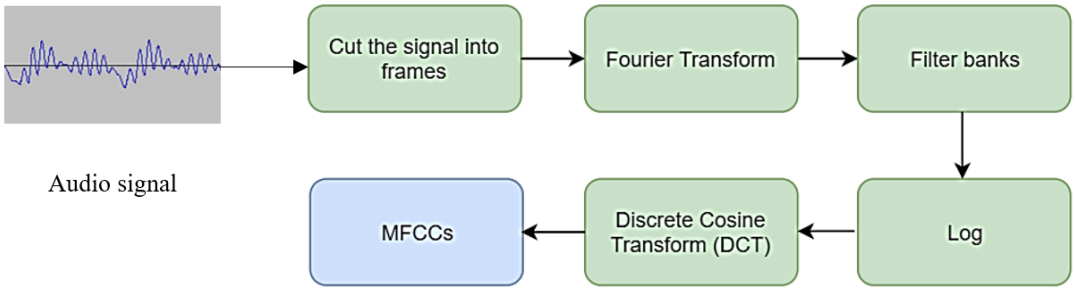


Figure 2. MFCCs computation algorithm

In addition to these spectral features used in speech emotion recognition there are also log frequency power coefficients (LFPC), used to represent energy distribution across the frequency spectrum, and linear prediction cepstral coefficients (LPCC), derived from either linear prediction (LP) analysis or a filter-bank approach (described above). The details of the computations of LFPC and LPCC can be found in (Nwe, Foo, and De Silva 2003) and (Atal 1974), respectively.

2.3.4 Emotion classification

After speech features extraction, the next step is to choose a certain classifier, which decides the underlying emotion of the utterance. Various types of classifiers have been used for the emotion classification task: Hidden Markov Models (HMM), Support Vector Machines (SVM) and neural networks among others. There are other classification models used for speech emotion classification, but in this work we will dwell only on the mentioned ones. The following chapter will elaborate on the techniques of each of the classification method, as well as provide the associated researches and their results.

2.3.4.1 Hidden Markov Model

Hidden Markov Models classifiers have been widely used in speech processing, since they are physically related to the production mechanism of speech signal (Awad and Khanna 2015). HMM are a class of probabilistic graphical model[12] which allows to predict a sequence of unknown (hidden) variables from a set of observed variables. An HMM can be represented as a dynamic Bayesian network[13] unrolled over time where the observations made at each time step are used as predictors of the best sequence of states.

2.3.4.1.1 Example of an HMM

Figure 3 demonstrates the scheme of an HMM. The scenario is having urns *X1*, *X2* and *X3* in a room, each containing a known mix of balls, labelled *y1*, *y2*, *y3* and *y4* each, which can be randomly drawn at each state. A user sees the sequence of the balls and decides which are the urns that the balls are taken from. Even if the users know the composition of the urns, which they usually do not, they cannot be sure from which urn the balls are taken, they can only figure out the likelihood of the balls being taken from the urns.

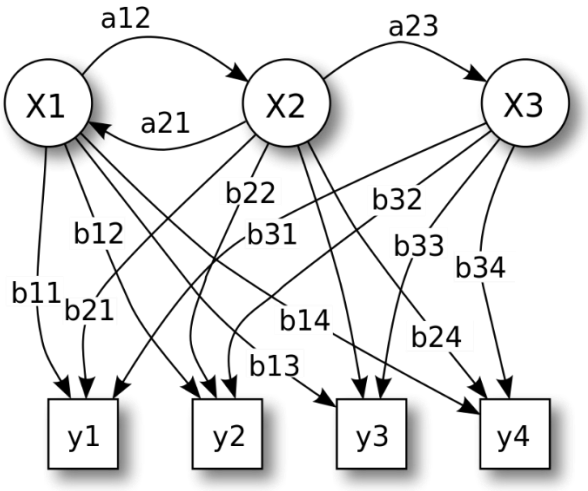


Figure 3. Hidden Markov Model scheme[14]

[12] Graphical model is a probabilistic model in which a graph is used to illustrate conditional dependence structure between random variables.

[13] Bayesian network is a probabilistic graphical model (see Footnote 12). For example, a Bayesian network can represent probabilistic relations between diseases and their symptoms: due to the presence of some symptoms the probability of having certain diseases can be computed.

[14] The image is taken from https://en.wikipedia.org/wiki/Hidden_Markov_model

2.3.4.1.2 Elements of an HMM

There are a finite number of hidden states in an HMM (*X1*, *X2* and *X3* in Figure 3), within a state a signal possesses some distinctive properties. At each time *t*, a new state is entered, or the process remains at the same state, based on a transition probability (*a11*, *a21*, *a23*), which depends on the previous state. After each transition, observed variables (*y1*, *y2*, *y3* and *y4*) are produced according to a probability distribution of the current state, which are called output probabilities (*b11*, *b21* etc.).

2.3.4.1.3 Discrete versus continuous HMM

The example in Section 2.3.4.1.1 describes a discrete HMM, where observations are defined by a set of discrete symbols from a finite alphabet, here, the balls. However, speech has continuous features, therefore observations are continuous vectors. There are two ways of handling them: either convert continuous observations into discrete ones using a codebook, which may degrade the results, or employing HMM states that have continuous observations, whose probability density function[15] (PDF) are evaluated as a combination of other distribution functions – a mixture distribution, with an associated mixture weight (Awad and Khanna 2015), Gaussian mixture model being the most common one. Gaussian distribution, or normal distribution, is a probability distribution commonly used to model different phenomena. It has a bell-curved shape as shown in Figure 4. Gaussian mixture model is a combination of several Gaussian distributions.

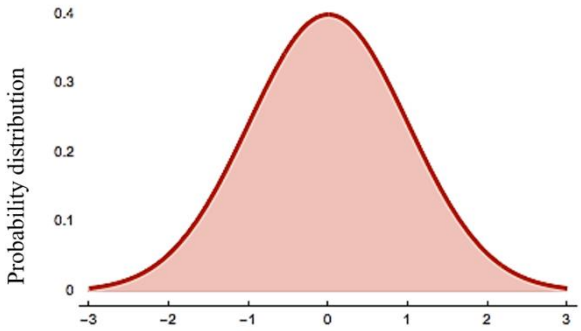


Figure 4. The graph of a Gaussian distribution with mean of 0 and standard deviation of 1[16]

[15] Probability density function defines probability distribution of a continuous random variable (for more details refer to (Knill 2009)).

[16] The image is taken from https://brilliant.org/wiki/normal-distribution/

2.3.4.1.4 HMM design

When designing an HMM classifier for emotion recognition, there are a number of issues requiring attention. First, it is necessary to choose the right topology of an HMM. HMM can have left-to-right topology, where the hidden state index either increases or stays the same, or fully connected topology, where it is possible to make transitions from any state to any state at any time (see Figure 5). Left-to-right HMMs are widely used in automatic speech recognition (ASR), as speech is a continuous signal which does not go back in time. However, for speech emotion recognition it may be better to use fully connected HMMs, since HMM states, in this case, correspond to emotional cues such as pauses. For example, when a pause is associated with the emotion of sadness, it may occur at any part of the utterance, therefore, it is necessary to be able to reach any state from any other state. Some other issues of HMM design include choosing the optimal number of states, type of observations (discrete or continuous), the optimal number of observation symbols for a discrete HMM or the optimal number of Gaussian mixtures or a continuous HMM (El Ayadi, Kamel, and Karray 2011).

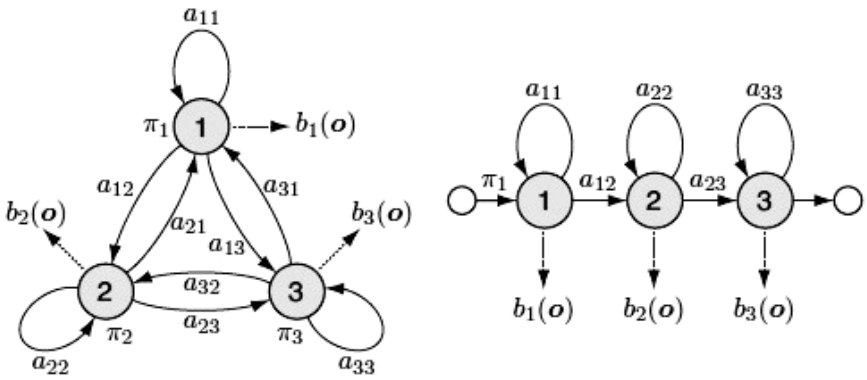


Figure 5. HMM topologies: fully-connected on the left, left-to-right on the right[17]

2.3.4.1.5 Speech emotion classification using HMM

HMM classifiers have been widely used in emotion recognition experiments. In (Atal 1974), an HMM-based classification of six basic and neutral emotions was proposed. Prosodic and energy-related features were selected as a representation of a speech signal. Left-to-right continuous HMMs were built for each emotion resulting in seven overall models. The solution was tested on a database of 5250 audio samples of acted emotions.

[17] The image is taken from http://yanfenglu.net/researchVAS_p1.htm

The overall recognition accuracy achieved was 77.8%, with fear being the most recognizable emotion: the accuracy of its detection was 95.4%.

Another HMM classification of emotions is described in (Nwe, Foo, and De Silva 2003), recognizing six archetypal emotions. The authors use LFPC, MFCC and LPCC spectral features for representing a speech signal. For each emotion a four-state fully connected discrete HMM was developed. Two databases were used for training and testing the proposed solution: Mandarin and Burmese speakers were employed to produce 720 utterances. The best average rates were 78.5% and 75.5% for the Burmese and Mandarin databases, respectively.

Speech emotion classification using HMM was also performed by (Y. L. Lin and Wei 2005). Different combinations of the following features were tried out in the experiment to represent a signal: fundamental frequency, energy, the first four formant frequencies, two MFCCs and five Mel frequency sub-band energies. Sub-band energies were computed as follows: a magnitude spectrum of each frame was calculated by a fast Fourier transform (FFT), and then input to a bank of five equidistant filters located in the Mel frequency scale between 60 Hz to 7.6 kHz. To get the five sub-band energies, the logarithm mean energies of the five filter outputs were estimated. A five-state continuous HMM was designed as an emotion classifier. Danish emotional database (see Section 2.3.2.1) was employed for training and testing the solution. 99.5% mean recognition accuracy was achieved in the experiment.

2.3.4.2 Support Vector Machine

2.3.4.2.1 Structure of SVM

Support Vector Machine (SVM) is a machine learning algorithm that is widely used for classification tasks. SVM's task is to divide the dataset into classes in the best way possible, by finding a hyperplane, a line separating two classes so that the distance (margin) between the nearest data points (support vectors) from both classes is as far as possible (Cortes and Vapnik 1995). Figure 6 shows the scheme of the SVM with two features.

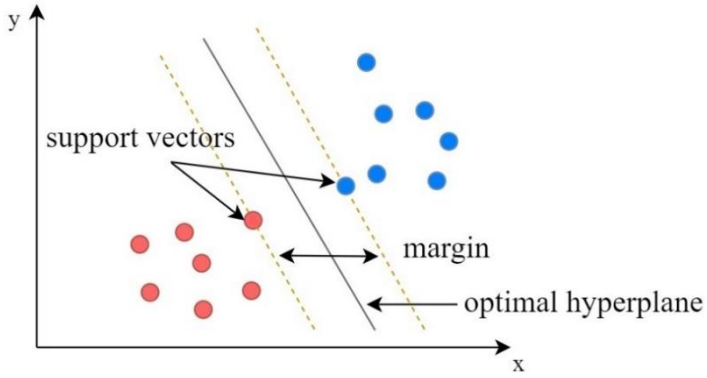


Figure 6. SVM and its components scheme

The example in Figure 6 presents an example of linearly separable data without overlapping, which in practice is rare. Usually, datasets are either linearly separable but there is overlapping between the classes, or the data are overall non-separable linearly. In the former case, SVM employs soft margins, in the latter, it uses kernel tricks.

A soft margin is a margin that allows certain data points to be misclassified, i.e. the data points within the width of the margin, whether on the correct side of the separating line or on the wrong one, may be neglected. The width of the soft margin is controlled by the parameter *C* that determines the trade-off between finding a line maximizing the margin and minimizing the misclassification (Kecman 2014).

Figure 7 demonstrates an example of a nonlinear data separation. It is clearly seen that linear separation is possible only with overlapping, whereas nonlinear separation is done without any errors. When this is the case, SVM uses kernel tricks, which by making some transformations to the existing features, create new features, with the help of which SVM is able to find the nonlinear boundary.

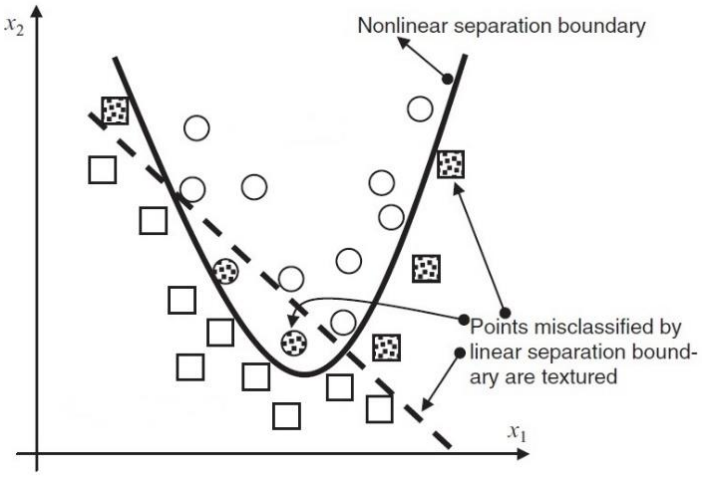
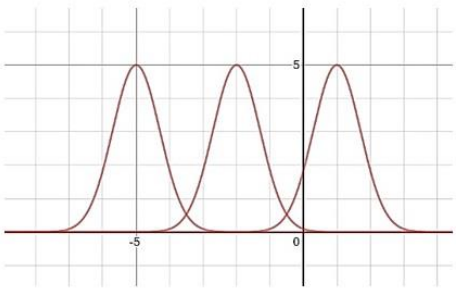
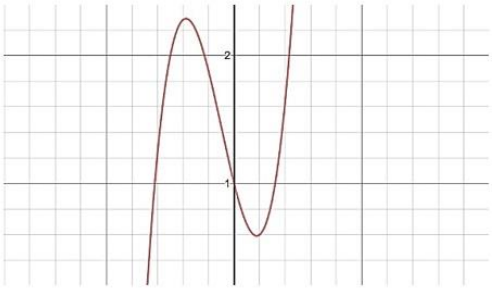


Figure 7. Example of nonlinear data separation by SVM (taken from (Kecman 2014))

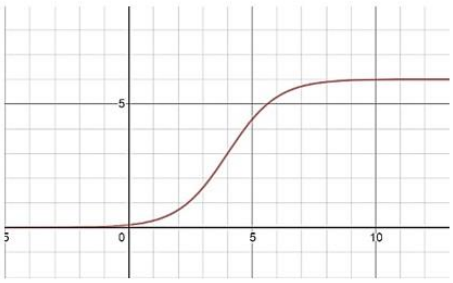
The most common kernel functions (transformations) used by SVM are radial basis function (RBF), polynomial, sigmoid and linear. The graphs of the kernel functions are shown in Figure 8. For computation details of kernel functions refer to (Kecman 2014).



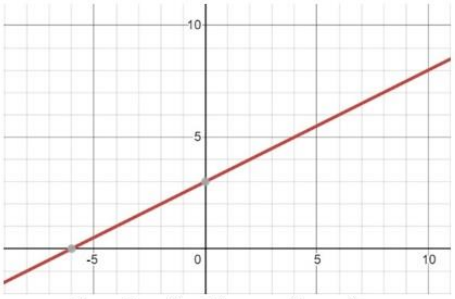
Graph of an RBF function



Graph of a polynomial function



Graph of a sigmoid function



Graph of a linear function

Figure 8. Graphs of the kernel functions

2.3.4.2.2 Emotion classification using SVM

Support vector machines are also extensively used in speech emotion recognition, however, with different success rates. In (Lee et al. 2004), an SVM classifier with

polynomial kernel functions is described. As a representation of a speech signal, prosodic features such as pitch-related features and speech rate were used. The solution was trained and tested on a dataset of 880 utterances, containing four emotions – anger, happiness, sadness and neutral – recorded by a semi-professional actress. An overall accuracy of 55.68% was achieved during the classification.

Another experiment of an SVM classifier of speech emotion recognition was performed in (Y. L. Lin and Wei 2005). Mel energy spectrum dynamics coefficients (MEDC) were proposed to distinguish five emotions of the Danish emotional database (see Section 2.3.2.1). To extract MEDC the steps below were followed. First, the magnitude spectrum of each utterance was calculated using FFT. The result was equally distributed on a mel frequency scale and was then passed through a bank of N filters. Next, the logarithm mean energies of the output of the filters were calculated together with their first and second differences. The final MEDC were obtained by combining the first and second differences. An SVM with an RBF kernel was used as a classifier and an accuracy of 88.9% was obtained during the classification.

In (Seehapoch and Wongthanavasu 2013), an SVM classifier with a linear kernel was trained and tested on three databases: Berlin, Japanese and Thai emotional databases (detailed in Section 2.3.2.1). Fundamental frequency, energy, zero-crossing rate[18] (ZCR), linear predictive coding and MFCC were selected as a representation of a signal. SVM with linear, polynomial and RBF kernels were tested. The best recognition accuracies were obtained with a linear kernel SVM and MFCC as a signal representation: 78.9%, 89.29% and 92.42% for Berlin, Japanese and Thai databases, respectively.

(Kerkeni et al. 2019) proposes an SVM classifier with a polynomial kernel for speech emotion recognition. The system is evaluated on the Berlin emotional dataset and the INTER1SP Spanish emotional database (refer to Section 2.3.2.1). MFCC and modulation spectral features (MSF) were chosen to represent a speech signal. The latter are obtained by emulating the spectro-temporal processing performed in the human auditory system and consider regular acoustic frequency jointly with modulation frequency (for computation details of MSF refer to (Kerkeni et al. 2019)). The highest classification accuracies achieved on the Berlin database is 81.10% with the combination of MFCC and

[18] Zero-crossing rate is the rate at which the signal changes from negative to zero to positive or from positive to zero to negative. It is commonly used in Voice activity detection to differentiate between silence and speech.

MSF as a speech signal representation, and 90.94% for the Spanish database with MFCC alone as a signal representation.

2.3.4.3 Neural networks

2.3.4.3.1 Basic structure of a neural network

The term artificial neural network (ANN) derives its origin from the human brain, where a large number of neurons are interconnected to simulate intelligent behaviour. In 1943, neurophysiologist Warren McCulloch and mathematician Walter Pitts in (McCulloch and Pitts 1943) modelled a simple neural network with electrical circuits. Inspired by their work, Frank Rosenblatt (Rosenblatt 1962) developed a simple artificial neuron called perceptron. A perceptron, shown in Figure 9, takes several binary inputs, *x1, x2*, …, and produces a single binary output. Rosenblatt proposed a simple rule to compute the output, having introduced weights *w1, w2,* …, real numbers, which indicate the importance of the respective inputs to the output. The neuron's output, 0 or 1, depends on whether the weighted sum of the inputs (Equation 6) is less or greater than some threshold, the value of which is decided beforehand (Nielsen 2018).

$$output = \begin{cases} 0 \; if \; \sum_j w_j x_j \leq threshold \\ 1 \; if \; \sum_j w_j x_j > threshold \end{cases} \qquad (6)$$

When dealing with neural networks, usually, the term 'threshold' is changed to the term 'bias', which is obtained by moving it to the other side of the inequality ($b \equiv -threshold$). Bias can be described as a measure of how easy it is for a perceptron to get a 1 as output, or in biological terms, how easy it is for the output neuron to fire.

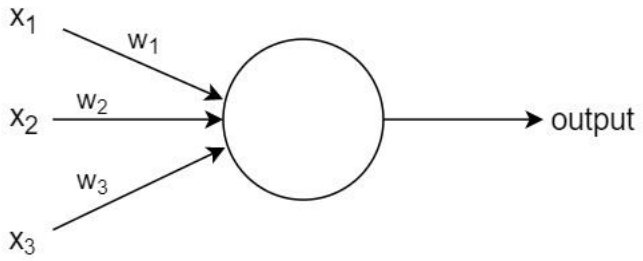


Figure 9. A perceptron

In modern artificial neural networks, though, a perceptron is rarely used, whereas slight changes in weights, which are made to let the network learn, may lead to big changes in

the output of the perceptron, e.g. 0 can become 1. Therefore, another type of neuron is used, a sigmoid neuron, which is similar to perceptron but modified so that small changes in weights and bias cause small changes in their output, which will allow the network to learn. A sigmoid neuron also has inputs *x1*, *x2*, …, but in this case they are non-binary, i.e. the input can be any number between 0 and 1. Sigmoid neurons also have weights *w1*, *w2*, …, and overall bias *b*, but its output is non-binary, instead it is defined as $\sigma(w \cdot x + b)$, where σ is a sigmoid function (see Section 2.3.3.2) and is called an activation function (other mathematical functions, like softmax or tangent, are also used as activation functions in neural networks). So, the output of a sigmoid neuron will be computed as follows:

$$\frac{1}{1+\exp\left(-\sum_j w_j x_j - b\right)} \qquad (7)$$

A neural network is then comprised of several sigmoid neurons connected with each other as shown in Figure 10. The leftmost neurons, which are the input, are called the input layer. The rightmost neuron is an output layer. Everything that is in between the input and the output layers are hidden layers. Neural networks can have one or several hidden layers.

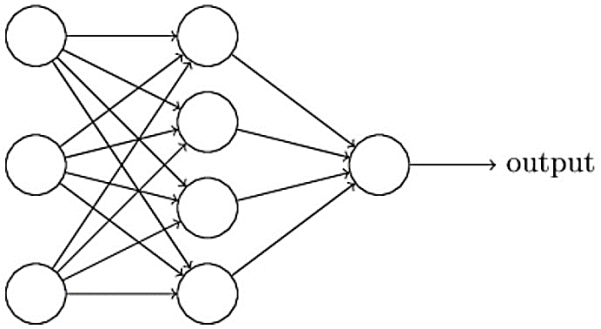


Figure 10. Neural network with one hidden layer (taken from (Nielsen 2018))

2.3.4.3.2 Recurrent neural networks

Neural networks where the output of the previous layer is the input to the next layer are called feedforward neural networks. However, there are other architectures of neural networks, which proved to be efficient in different classification tasks, including speech emotion recognition (see Section 2.3.4.3.6). One example of such a model is a recurrent neural network (RNN), where feedback loops are possible. The idea of RNN is having neurons that fire for a limited amount of time before becoming quiescent. These neurons

may cause other neurons to fire later in time, again, for a limited amount of time and so on. Therefore, over time there is a cascade of firing neurons. Loops do not cause problems in this architecture, as the output of the neuron affects its input at some later time, not at once. Figure 11 shows the scheme of an RNN architecture.

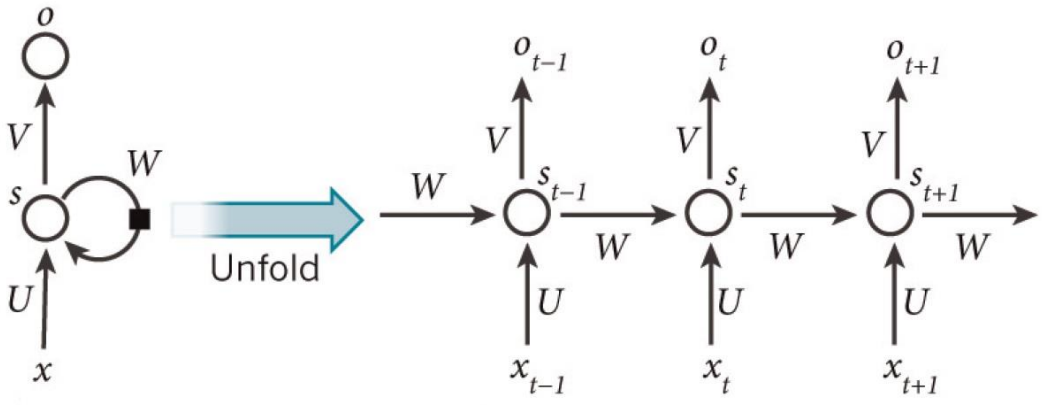


Figure 11. A basic concept of an RNN (taken from (Kerkeni et al. 2019))

2.3.4.3.3 Vanishing gradient problem

Thanks to the architecture of RNN, they are effective at learning temporal correlations, so these models work well with speech processing, thus with speech emotion recognition (Kerkeni et al. 2019). However, because of its structure, RNN suffers from a vanishing gradient problem. The goal of the network training is to find such weights and biases that the output of the network is as similar to the desired output as possible (depending on the task, the desired output can be a class, a vector etc.; in case of emotion classification, the desired output is the correctly predicted class of the emotion). Cost (loss or error) function defines how close the predicted output is to the desired output: the less is the value of the error function, the closer is the predicted output of the network to the desired one. Gradient descent algorithm is used to find the global minimum of the cost function that is going to be an optimal setup for the network. During the training, the information travels from the input neurons to the output neurons, while the error is calculated and propagated back to the neurons to update their weights. In case of RNN, every single neuron takes part in calculating the error function, not only the neurons from the previous layer, so it is necessary to propagate all the way back through time to these neurons. The problem here is connected with updating the weights, connecting the hidden layers to themselves in the unrolled temporal loops, which are called recurring weights. For example, to get from x_{t-2} to x_{t-1}, x_{t-2} is multiplied by W_{rec} (see Figure 12), to get further

from x_{t-1} to x_t, x_{t-1} is multiplied by the same W_{rec}. So there is a multiplication by the same weight multiple times, which leads to a quick decrease of the value: at the start of the training weights are assigned randomly with values close to zero, and when these small numbers are multiplied by the incoming inputs, the value of the gradient will become less and less, which is called a vanishing gradient. The lower the gradient is, the harder it is for the network to update the weights, so the network is not trained properly, and the more time it takes to for the training itself.

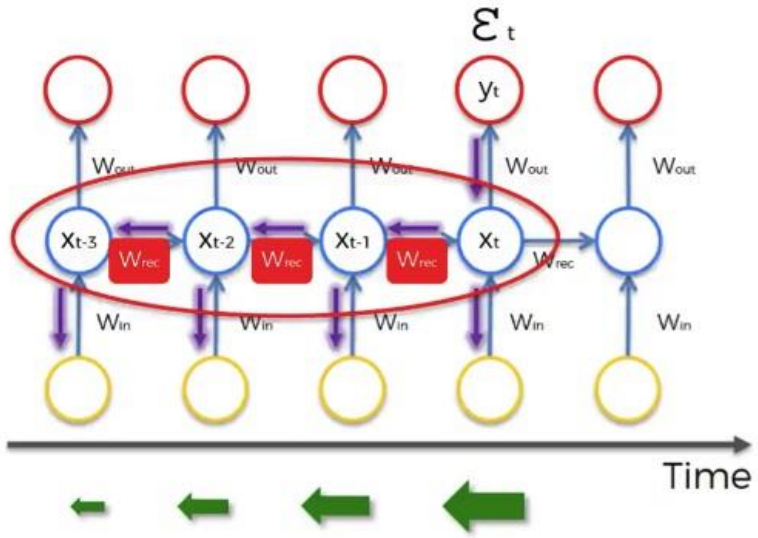


Figure 12. Vanishing gradient problem[19]

2.3.4.3.4 Long short-term memory networks

To resolve the vanishing gradient problem of RNN, long short-term memory (LSTM) RNNs were proposed by Hochreiter et al. (Hochreiter and Schmidhuber 1997), which use memory cells to store information so that it can exploit long-range dependencies in the data and overcome error back-flow problems. In particular, LSTM contain information in a gated cell, where it can be stored, written to and read from. The cell makes decisions, which information to store and pass on to the next layer and which to erase via gates that open and close. The gates are implemented with element-wise multiplication by sigmoids, which are all in the range of 0 to 1. LSTM networks are widely used in the field of speech emotion recognition (see Section 2.3.4.3.6).

[19] The image is taken from https://www.superdatascience.com/blogs/recurrent-neural-networks-rnn-the-vanishing-gradient-problem/

2.3.4.3.5 Convolutional neural networks

Convolutional neural networks (CNN) are primarily used in the field of computer vision. CNN network AlexNet (Krizhevsky, Sutskever, and Hinton 2012) won the 2012 ImageNet competition, improving the image recognition by 10.8% from the baseline, which was a breakthrough at the time. However, CNN are not limited to image recognition, they are used in other spheres, speech emotion recognition included (see Section 2.3.4.3.6).

In mathematical terms, convolution is the integral measuring how much two functions overlap as one passes over the other. In image or sound analysis, one of these functions is an input, in case of a sound analysis, it is a speech feature, and the second function, which passes through the first function through time, is a filter, which picks up this feature. The two functions relate through multiplication. At the time of training, a number of filters pass through the features creating a feature map, indicating the strength of a detected feature in an input (Brownlee n.d.). Figure 13 shows an example of applying a filter to an input.

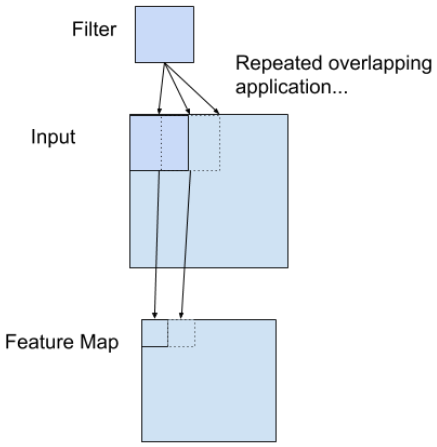


Figure 13. Example of a filter applied to an input to create a feature map (taken from (Brownlee n.d.))

CNN usually consist of a variety of layers in sequence (Badshah et al. 2017). A typical model includes several convolutional layers (described above), which are in charge of generating a feature map from the input. Pooling layers may be added after convolutional layers to reduce the spatial size of the feature maps and reduce the number of parameters and computation in the network. CNN may also contain fully connected (dense) layers, where each neuron of the input is connected to each neuron of the layer. A sequence of

convolutional, pooling and fully connected layers forms a feature extraction pipeline. Finally, a softmax layer may be introduced to do the classification task.

2.3.4.3.6 Speech emotion classification using neural networks

Numerous researches have been done on the task of emotion detection using neural networks with quite different results. The following chapter will outline some of them. It should be noted that not all the experiments use the same evaluation metrics, so it is not always possible to compare them with the others.

In (Lim, Jang, and Lee 2016), a duo-combination of CNN–RNN emotions classification is described. The authors use some publicly available database for their experiment, where six basic emotions and a neutral emotion are present. The speech signal is transformed to 2D representation using Short Time Fourier Transform (STFT) after pre-processing. The 2D representation is then analysed through CNNs and Long Short-Term Memory architectures. The authors do not report the accuracy achieved during the tests, but the precision and recall measures. The average precision on the test set is 88.01%, while the average recall is 86.86%.

In (Badshah et al. 2017), a speech emotion recognition system using CNN is presented. Speech signal is represented as spectrograms[20], which serves as an input to a CNN. The CNN model consists of three convolutional, three fully connected layers and a softmax layer to extract features from the spectrograms and perform the classification of emotions. For training and testing the model the Berlin emotions dataset (see Section 2.3.2.1) has been used. Satisfactory results (accuracy of more than 50%) is achieved for all emotions except 'fear' (25.33% accuracy). Average accuracy across all the emotions achieved is 56.22%.

Another CNN-based speech emotion classifier is described in (Zheng, Yu, and Zou 2015). The authors use IEMOCAP database (refer to Section 2.3.2.1) to evaluate their system. Log-spectrogram is computed for each utterance and the principal component analysis (PCA) technique is used to reduce the dimensionality. A CNN model consisting of two convolutional and two pooling layers is constructed to learn the representation of the emotion from the segments with labelled training speech data. The proposed model achieves 40% classification accuracy.

[20] Spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time.

(Kerkeni et al. 2019) present an LSTM speech emotion classifier. The proposed system is evaluated on the Berlin emotional dataset and the INTER1SP Spanish emotional database (see Section 2.3.2.1). MFCC and modulation spectral features (MSF) were chosen to represent a speech signal. The neural network consists of two consecutive LSTM layers with hyperbolic tangent activation, followed by two classification dense layers. Several combinations of a speech signal representation are tested (MFCC alone, MSF alone, MFCC + MSF). The best classification accuracies are achieved in both Berlin and Spanish databases when MFCC and MSF are combined, 76.98% and 90.05%, respectively.

In (Lalitha, Tripathi, and Gupta 2018), a Deep neural network (DNN)-based speech emotion classifier is discussed. It uses the Berlin emotional database (detailed in Section 2.3.2.1) for training and testing the proposed model. The speech signal is represented by several features, such as MFCC, pitch, LPC, amplitude, IMFCC (Inverted MFCC) and others. In total, 22 types of speech features are selected as a representation of a speech signal. The classifier model is a feed forward back propagation network with three hidden layers. The average classification accuracy of seven emotion classes is 88.9%.

(Parthasarathy and Tashev 2018) describe several CNN-based speech emotion classifiers. The authors use 17048 sentences in Mandarin from Microsoft spoken dialogue system XiaoIce (Z. Wang and Tashev 2017) to train and evaluate their models. The utterances in the database contain four emotions: happiness, anger, sadness and a neutral emotion. To represent a speech signal, 29 features are extracted including 26 log Mel-spectrum, fundamental frequency, energy and speech presence probability and their deltas, resulting in 58 features in total. Several classification models, such as CNN with one or several convolutional layers and different pooling techniques, and feature combinations are tested by the authors. The evaluation metric used in the experiment is a sum of weighted and unweighted accuracies[21], with the maximal value of 200. The best result is shown by a CNN with three convolution layers and pyramidal pooling[22] with 26 log Mel-spectrum as a speech signal representation, giving a total of 121.15 WA+UA.

[21] The weighted accuracy (WA) corresponds to the correctly detected samples divided by the total number of samples by classes, i.e. the number of correctly predicted instances in a class, divided by the total number of instances in that class. The unweighted accuracy (UA) is the fraction of instances predicted correctly, i.e. total correct predictions, divided by total instances. The distinction between these two measures is useful if there are classes that are under-represented in the dataset.

[22] In pyramidal pooling, the input signal is partitioned into smaller regions and the pooling operation is repeated on each region.

In (Niu et al. 2017), a CNN-based speech emotion classifier is proposed. IEMOCAP and Berlin emotional databases (see Section 2.3.2.1) are used for training and testing the model. Spectrograms are obtained from the speech signal to represent it. Deep retinal convolutional neural network (DRCNN) is proposed as a classification model, which consists of two parts. First, the data are augmented using the algorithm based on retinal imaging principle (DAARIP), which by applying the principle of retina and convex lens imaging allows to get more training data by changing the size of the spectrogram (for more details on DAARIP refer to (Niu et al. 2017)). Then, a CNN with five convolution, three pooling and three fully connected layers is trained and tested on the updated dataset. The classification accuracy achieved on IEMOCAP and Berlin databases are 99.25% and 99.79%, respectively.

2.3.4.4 Hybrid speech emotion recognizers

Apart from employing single machine learning algorithms to build speech emotion recognizers, hybrid systems have also been used in a number of experiments. In (Mao, Chen, and Zhang 2007), an HMM-ANN system has been built to recognize emotions. An earlier version of BHUDES database (refer to Section 2.3.2.1) is used for training and evaluating the system. This version contains twenty texts read by five actors with five emotions: anger, joy, sadness, disgust and surprise. HMMs is used to model speech feature sequences, which contain MFCC, LPCC, pitch, amplitude energy, log energy and the first formant. A one-hidden layer ANN is used for the emotion classification. The average recognition rate has reached 81.6%.

Another hybrid system for speech emotion recognition is proposed in (Fu, Mao, and Chen 2008). The system is trained and tested on BHUDES database (see Section 2.3.2.1). MFCC, LPCC, pitch, amplitude energy, log energy and the first formant are used as a representation of a speech signal. The classification is made in two steps. First, an SVM classifier with a polynomial kernel is employed to distinguish emotions into two classes: the first class contains ‘anger’, ‘joy’ and ‘surprise’ emotions, whereas, ‘sadness’ and ‘disgust’ belong to the second class. Respectively, 98.9% and 95.8% classification accuracies are achieved for each class. Then an HMM classifies the two classes into five, one emotion per class. The average accuracy obtained is 76.1%.

In (Huang et al. 2014), a CNN-SVM speech emotion recognition model is discussed. The solution is trained and tested on four speech emotional databases: SAVEE, Berlin

emotional database (EmoDB), Danish emotional speech database (DES) and Mandarin emotional speech database (BHUDES) (see details in Section 2.3.2.1). A CNN with one convolution layer and one fully connected layer is used to learn salient features from spectrograms of the speech signals. An SVM with a linear kernel is employed as an emotion classifier. The following results have been obtained for each of the databases: the classification accuracy on the SAVEE database is 73.6% with a standard deviation of 0.51, on EmoDB it is 85.2% with a standard deviation of 0.45, on DES it is 79.9% with a standard deviation of 0.53, on BHUDES it is 78.3% with a standard deviation of 0.61.

2.4 End-to end systems for speech emotion recognition

All the research works mentioned in previous sections describe the traditional approach to speech emotion recognition systems which include feature extraction from the speech signal. Recently, however, attempts have been made to build end-to-end speech emotion recognition models, utilising as little human a-priori knowledge as possible, i.e. deriving a representation of the input speech signal directly from raw, unprocessed data (Trigeorgis et al. 2016), (Tzirakis, Zhang, and Schuller 2018), (Zhao, Mao, and Chen 2018). The idea behind end-to-end systems is to learn feature extraction and do the classification in one jointly trained model for predicting the emotion.

In (Trigeorgis et al. 2016), the authors develop the following model. The raw waveform is segmented to 6 s long sequences, which are then fed into a two-layer CNN to extract fine-scale spectral information, followed by a two-layer LSTM for the classification of emotions. The model is trained and tested on RECOLA database (see Section 2.3.2.1), where emotions are represented in a 2D continuous emotion space with activation and valence dimensions. To evaluate the performance of the model concordance correlation coefficient (ρ_c) (L. I.-K. Lin 1989) is used. Concordance correlation coefficient measures the agreement between two variables, in this case, the agreement level between the predictions of the network and the gold-standard derived from the annotations. ρ_c ranges from *-1* to *1*, with perfect agreement at *1*. The authors report ρ_c equal to 0.686 for the activation dimension and 0.261 for the valence dimension.

In (Tzirakis, Zhang, and Schuller 2018), the model for speech emotion recognition is similar to the one described above, with the exception of using a deeper three-layer CNN and an input being segmented to 20 s long sequences to capture longer temporal dynamics. The model is also trained and tested on RECOLA database and concordance

correlation coefficient is utilized to measure its performance. For the activation dimension ρ_c obtained is 0.787, whereas for the valence dimension ρ_c is 0.440.

In (Zhao, Mao, and Chen 2018), the authors present a combined CNN-LSTM end-to-end speech emotion recognition system. In the model, there are four local feature learning blocks (LFLB), the core of which are a convolution layer and a pooling layer, which are followed by one LSTM and one fully-connected layers. This network is designed to learn features from raw audio clips. The solution was trained and tested on the Berlin speech database and the IEMOCAP database (see both in Section 2.3.2.1). Both speaker-dependent and speaker-independent experiments have been performed. In speaker-dependent experiments the average accuracies of 92.34% for the Berlin database and 67.92% for the IEMOCAP database have been obtained. In speaker-independent experiments the average accuracies have reached 86.73% for the Berlin database and 79.72% for the IEMOCAP database.

3 Methodology

This section describes the experiments performed to check the workability of some of the classifiers described in the Literature Review part (see Section 2), particularly, SVM, LSTM and DNN combined with Extreme learning machine (ELM) classifiers are discussed in this project. The section also includes the comparison of different feature sets as an input to the mentioned classifiers, as well as the description of the database used for the experiments and the environmental setup, where the experiments have been performed.

3.1 Speech emotion classifiers

3.1.1 SVM

3.1.1.1 General algorithm description

The experiment was performed on the solution from (Ruggieri n.d.). The following algorithm is applied: first, the short-term spectral and temporal features are extracted from the signal; the extracted features are then fed into a Support Vector Machine, which classifies the class of emotions. The scheme of the algorithm is presented in Figure 14.

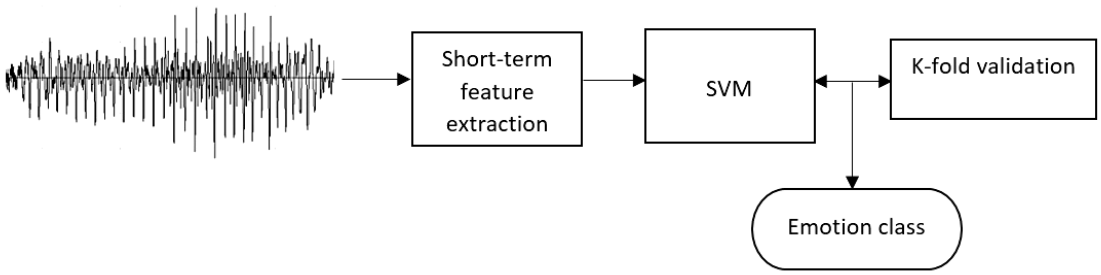


Figure 14. SVM algorithm overview

3.1.1.2 Feature extraction

The author uses pyAudioAnalysis (Giannakopoulos 2015) for extracting features – a python library for audio feature extraction, classification, segmentation and applications. In particular, short-term features are extracted from the signal, i.e. the input signal is split into frames (short-term windows) and a number of features are computed for each of them, so a sequence of short-term features is obtained for the whole signal.

Features from both time and frequency domain, as well as cepstral domain, are extracted; the total number of features extracted using pyAudioAnalysis is 34. The complete list of the features is shown in Table 3.

Index	Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9–21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22–33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Table 3. The list of features extracted with pyAudioAnalysis (taken from (Giannakopoulos 2015))

The time-domain features (features 1-3) are extracted from the raw signal, the frequency-domain features (features 4-8, 22-34) are obtained using the magnitude of the Discrete Fourier Transform and the cepstral-domain features, i.e. MFCCs (features 9-21) are computed after applying the Inverse Discrete Fourier Transform on the logarithmic spectrum (Giannakopoulos 2015).

3.1.1.3 Emotion classification

The author chooses Support Vector Machine as a classifier of emotional classes, which is implemented with scikit-learn (Pedregosa et al. 2012), python machine learning tools for data mining and data analysis.

C-Support Vector Classification (SVC) is used as a default classifier for classification, where soft margin C parameter and a number of kernel functions can be chosen. The author sets *C* parameter to 10 and uses radial basis function (RBF) kernel. scikit-learn tools also allow to employ Nu-Support Vector Classification (Nu-SVC), another

classification technique. Nu-SVC is similar to SVC with the exception of using a parameter *nu* to control the number of support vectors, which serves as an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors.

3.1.1.4 Original experiment

The author tests the solution on the Berlin Database of Emotional Speech (Burkhardt et al. 2005). The dataset contains 535 utterances pronounced by actors with different emotions: happy, angry, anxious, fearful, bored, disgusted and neutral. 80%-20% train-test data split is used in the experiment.

The input signal is sampled at 16,000 Hz, it is converted into frames of 40 ms with a frame period of 10 ms. Features are extracted for each frame and used in emotion classification, as described in Sections 3.1.1.2 and 3.1.1.3. The result of the test is not reported by the author.

3.1.2 Neural networks

3.1.2.1 Bidirectional LSTM

3.1.2.1.1 General algorithm description

The experiment is based on the solution from (R. Wang n.d.). The algorithm is built as follows. First, the short-term features are extracted from the input signal. Then, the extracted features are fed into a bidirectional LSTM (BLSTM) to classify the classes of emotions. K-fold cross validation is used to exclude the dependency on the data distribution. Figure 15 shows the scheme of the algorithm.

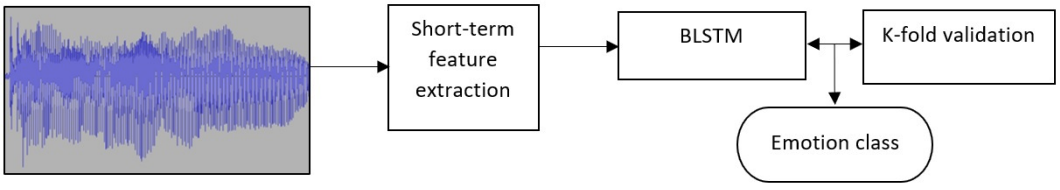


Figure 15. BLSTM algorithm overview

3.1.2.1.2 Feature extraction

The authors use pyAudioAnalysis (Giannakopoulos 2015) for extracting features (described in detail in Section 3.1.1.2). Apart from the features extracted with pyAudioAnalysis, the authors also compute harmonic ratio and pitch, which are then

added to the already extracted features. So, for each frame a vector of 36 features is obtained.

3.1.2.1.3 Emotion classification

The classification of emotions is performed by BLSTM combined with a weighted-pooling strategy as in the experiment carried out in (Mirsamadi, Barsoum, and Zhang 2017). The architecture of the network is presented in Figure 16.

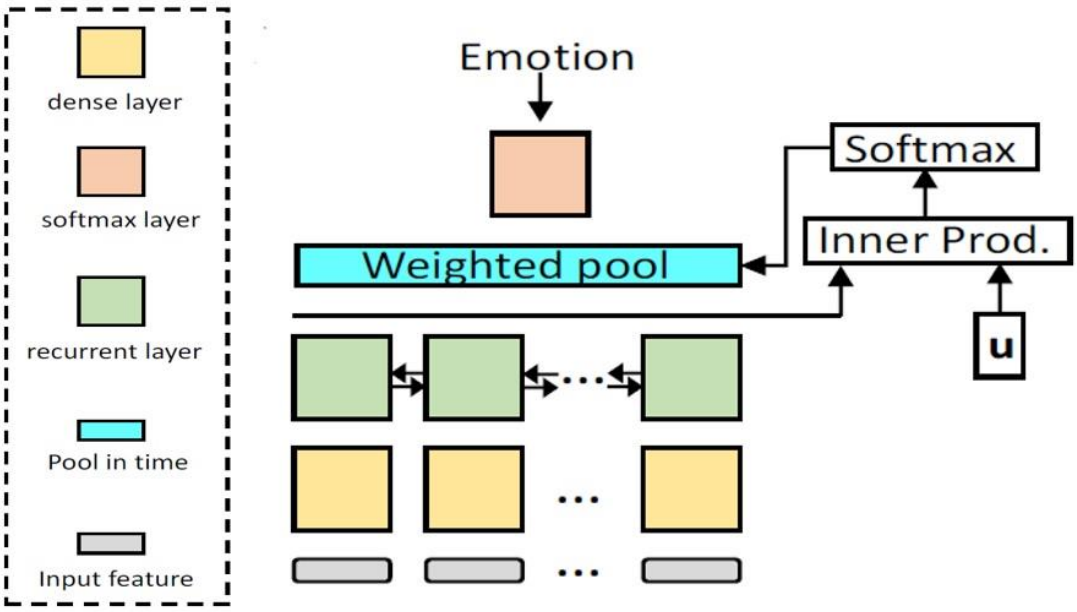


Figure 16. BLSTM architecture (taken from (Mirsamadi, Barsoum, and Zhang 2017))

The authors describe the solution as follows. The inner product between the attention parameter vector and the LSTM output is computed at each time frame, which is set as a score for the contribution of that frame to the final utterance-level representation of the emotion. A softmax function is then applied to the results to obtain a set of final weights for the frames. The obtained weights are used in a weighted average in time to get the utterance-level representation. The result is finally passed to the output softmax layer to get the probabilities for each class of emotion.

3.1.2.1.4 Original experiment

The authors test their solution on the Berlin Database of Emotional Speech (described in Section 2.3.2.1). They use 80% of the dataset for training the network and the remaining 20% for testing.

The input signal is sampled at 16000 Hz, it is converted into frames of 25 ms with a frame period of 10 ms. Features are extracted as described in Section 3.1.2.1.2 and then fed into the neural network for training and classification, as described in Section 3.1.2.1.3.

The authors report an average accuracy of 68.6% with a standard deviation of 1.88 across 10 folds.

3.1.2.2 DNN combined with ELM

3.1.2.2.1 General algorithm description

The solution from (Han, Yu, and Tashev 2014) was used for the experiment. The algorithm the authors use is as follows. First, the signal is divided into frames and segment-level spectral and temporal features are extracted. The extracted features are then fed into a DNN, which computes the emotion state distribution for each segment. Next the utterance-level features are formed from segment-level features and transferred to train an ELM, serving as an emotion classifier for the whole utterance. The algorithm is shown in Figure 17.

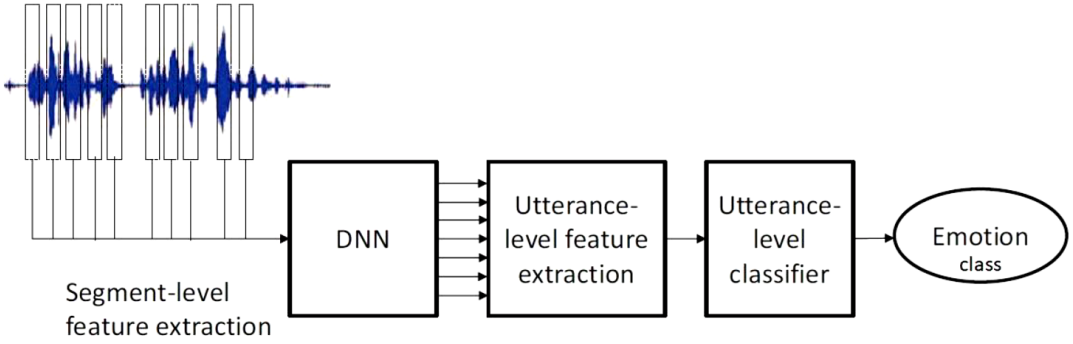


Figure 17. DNN+ELM algorithm overview (taken from (Han, Yu, and Tashev 2014))

3.1.2.2.2 Segment-level feature extraction

The input signal is sampled at 16000 Hz and converted into frames of 25 ms length with overlapping windows of 10 ms. The feature vector extracted for each frame consists of MFCC features, pitch-based features, and their delta feature across time frames. The pitch-based features include pitch period and the harmonics-to-noise ratio (HNR). HNR is computed as

$$HNR\ (m) = \ 10\log\frac{ACF(\tau_0(m))}{ACF(0)-ACF(\tau_0(m))}, \qquad (8)$$

where m is a frame, $\tau_0(m)$ is a pitch period and $ACF(\tau)$ denotes the autocorrelation function at time τ.

3.1.2.2.3 Training of a DNN

A DNN is trained with segment-level features to predict the probabilities of each emotion state. A DNN can be treated as a segment-level emotion classifier. The number of input units of the DNN is equivalent to the number of segment-level features. It uses a softmax output layer, whose size is set to the number of possible emotions and there are three hidden layers in the DNN. The trained DNN produces a probability distribution over all the emotion states for each segment.

3.1.2.2.4 Utterance-level features

The features in the utterance-level classification are computed from statistics of the segment-level probabilities. Four utterance-level features are computed: minimal, maximal and mean of segment-level probability of each of the emotion class over the utterance, and the percentage of segments which have high probability of each emotion class.

3.1.2.2.5 Training of an ELM

The utterance-level statistical features are fed into a classifier for emotion recognition of the utterance, in this particular experiment the authors use ELM (Ding et al. 2015), (Yu and Deng 2012) for this purpose, which was claimed to achieve promising results when the training set is small. ELM is a single-hidden-layer neural network. The utterance-level features serve as an input to an ELM and the output of the ELM for each utterance is a vector corresponding to the scores of each emotion state. The emotion class with the highest score from the ELM is chosen as the recognition result for the utterance.

3.1.2.2.6 Original experiment

The authors use IEMOCAP database (detailed in Section 2.3.2.1) for their experiment. The database contains audio-visual data from 10 actors, but only the audio track is employed. Recordings from 8 speakers are used as training data, while the recordings from the 2 remaining speakers are left for testing the solution, thus 80%-20% train-test

split is used in the experiment. The authors report 54.3% weighted accuracy and 48.2% unweighted accuracy for the classification of seven emotions.

3.2 Database description

The experiments on the solutions described above were carried out on the data from Neosound Intelligence, a Netherlands-based company working on speech and emotion recognition.

The dataset comprised of 17 telephone calls made by call-centre customers with various purposes: some were inquiring about some services, others were making complaints or reporting issues they were experiencing. Each call was made by a different customer and answered by a different call-centre agent, making it 34 speakers.

All in all, the length of the dataset is 84 minutes 50 seconds. The calls were made in the Latvian language.

The dataset was annotated by three people, possessing some knowledge of Latvian. They were asked to label parts of the recordings with angry or neutral emotion. Only the recordings where all the three annotators agreed on the label were taken for the experiments, so as to avoid having invalid data. This process left 1,986 seconds of speech, which is about 33 minutes, as the final dataset for training and testing.

The recordings were then cut into separate files to have a single emotion per file. 563 files were obtained in total: 285 files with ‘neutral’ label and 278 files with ‘angry’ label, making it respectively 51% and 49% of each class. Out of 278 files marked as having an ‘angry’ label, however, the annotators identified 113 as ‘somewhat angry’, which may be viewed as a separate class of emotion. To check if such a separation is valid, three types of dataset split is made in the experiments: 2 classes with all ‘angry’ labelled files in one class, 2 classes with only strictly ‘angry’ labelled files, ‘somewhat angry’ labelled files were left out of the training and testing, and 3 classes, where ‘angry’ and ‘somewhat angry’ labelled files are considered different emotions. A full description of the dataset splits is shown in Table 4.

Dataset split	Number of classes	Emotion class	Number of files		
			absolute	relative, %	total
Dataset 1	2	Neutral	285	51	563
		Angry (‘angry’ + ‘somewhat angry’)	278	49	
Dataset 2	2	Neutral	285	63	450

		angry (pure ‘angry’)	165	37	
Dataset 3	3	Neutral	285	51	563
		angry	165	29	
		somewhat angry	113	20	

Table 4. Description of the dataset splits used in the experiments

The experiments were run with 75%-train and 25%-test split. 10 K-fold cross-validation was used on all the experiments, i.e. the experiments on all the solutions were repeated 10 times with different train-test distributions, to ensure that the values of accuracy are robust, and do not depend on the data distribution.

3.3 Feature sets

The features extracted to represent a speech signal have been described in detail in the sections above. For convenience and future reference, they are grouped in sets, see Table *5*.

Feature set	Number of features	Reference in text
Feature set 1	34	Section 3.1.1.2
Feature set 2	36	Section 3.1.2.1.2
Feature set 3	30	Section 3.1.2.2.2

Table 5. Features sets used in the experiments

3.4 Environmental setup

The experiments were performed on a machine with Windows OS. As there are known problems of compatibility of Windows and tensorflow, Anaconda environment was additionally installed to facilitate the process of installation of necessary dependencies as well as running and training neural networks.

4 Findings

This section presents the results from the experiments performed on the solutions described in the previous chapter, as well as the database analysis, on which the experiments have been carried out.

4.1 Dataset analysis

4.1.1 Duration analysis

The analysis of the duration of the three emotion classes in the database (see Figure 18, Figure 19 and Figure 20) shows that the length of the segments in all the three classes is similar: most of the recordings are less than 6.5 seconds long, moreover, more than 67% of samples in each class do not exceed 4.5 seconds, i.e. the duration of the expressed emotion is quite fast. In each class there are also few short segments (less than 0.5 seconds): 2% – in 'neutral' and 'angry' and less than 1% – in 'somewhat angry', as well as long segments (longer than 10 seconds): 2% – in 'neutral' and 4% – in 'angry' and 'somewhat angry' classes. Therefore, in this particular database the emotion classes may be difficult to differentiate based on the duration of the segments.

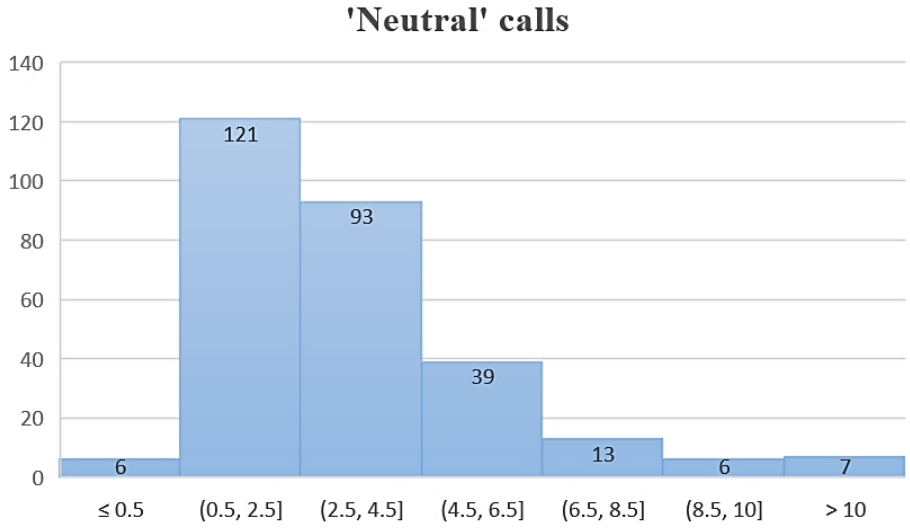


Figure 18. Duration histogram of the recordings with 'neutral' label

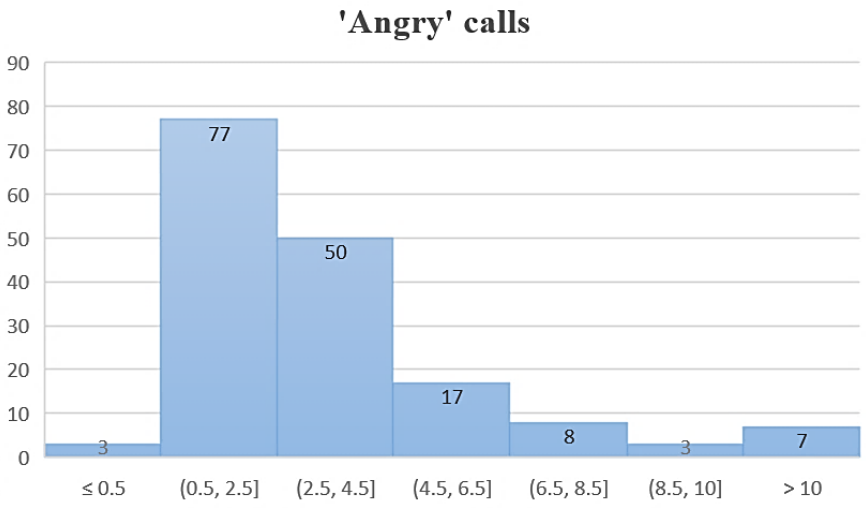


Figure 19. Duration histogram of the recordings with ‘angry’ label

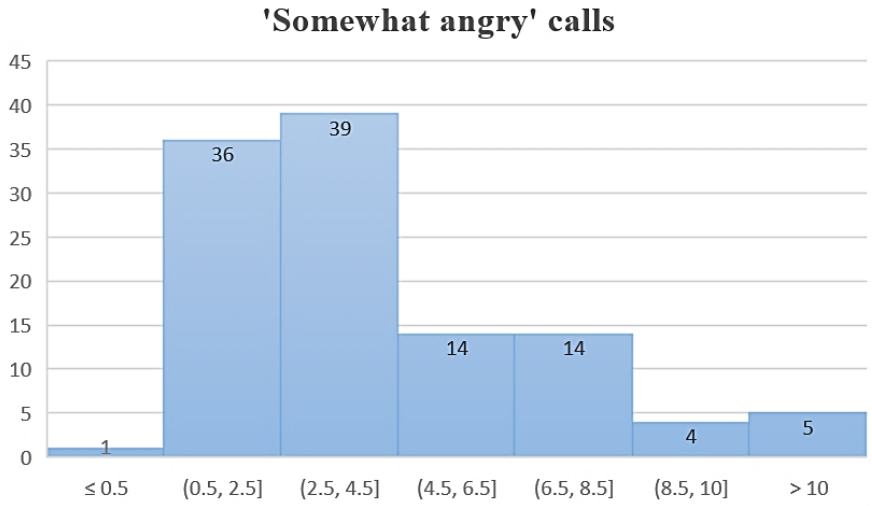


Figure 20. Duration histogram of the recordings with ‘somewhat angry’ label

4.1.2 Principal component analysis

The analysis of the segment duration in the dataset has not shown any particularities, therefore, it has been decided to perform principal component analysis, which is a technique used to highlight variation and emphasize strong patterns in a dataset. PCA has been performed on the three splits of the dataset and is shown in Figure 21, Figure 22 and Figure 23. In Dataset 1, where the ‘angry’ class includes segments labelled both as ‘angry’

and ‘somewhat angry’, there is quite a big overlap between the classes, therefore, in this case 2 component PCA is not representative. In Dataset 2, where ‘somewhat angry’ segments are left out, the separation between the classes is visible enough in spite of some overlap. In Dataset 3, as in the case with Datasets 1, 2 component PCA is not representative, the borders of the classes are not detectable. From this we can conclude that the third class, labelled as ‘somewhat angry’, may be considered as a separate class, not as part of the ‘angry’ class, however, it is hard to identify it as well.

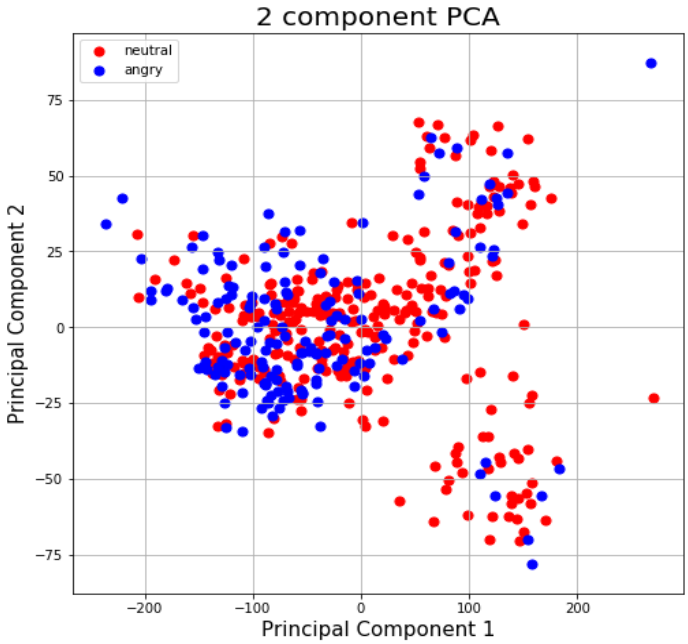


Figure 21. PCA for Dataset 1

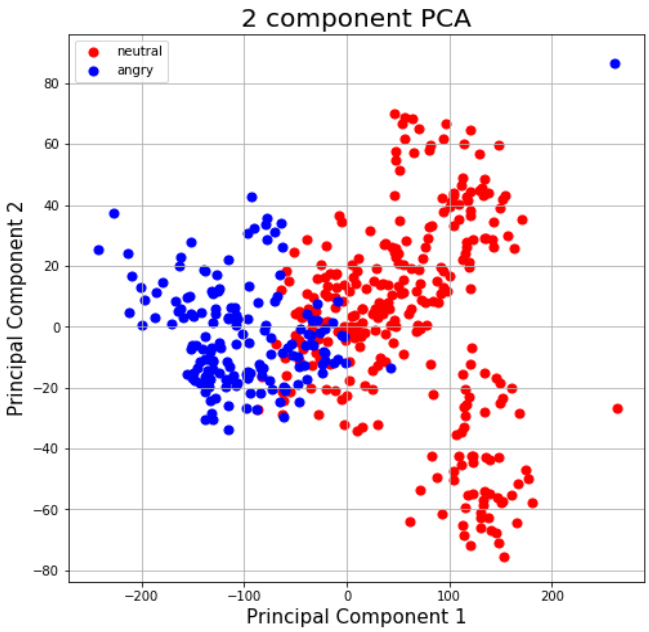


Figure 22. PCA for Dataset 2

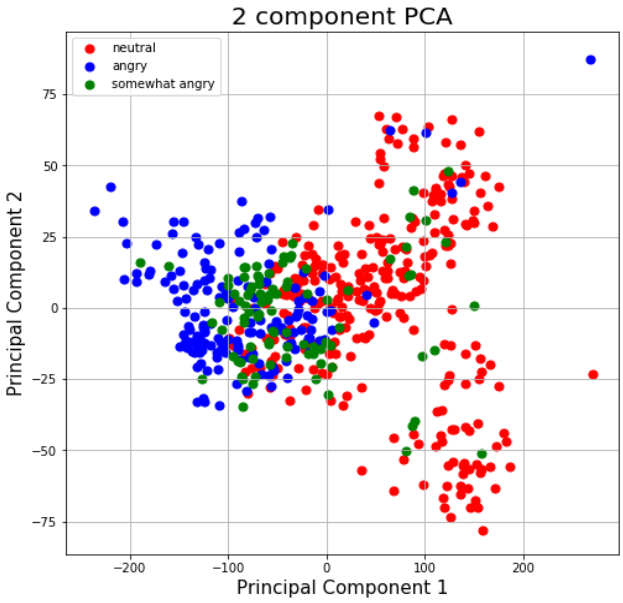


Figure 23. PCA for Dataset 3

4.2 Experiments

4.2.1 Support Vector Machine

The system architecture of the SVM used to perform the experiments is described in detail in Section 3.1.1.

The first experiment was the replication of the original experiment on the Berlin database. The mean accuracy across 10 folds is 71.12% with a standard deviation of 2.99. As the authors do not report the results they obtained in their experiment, we cannot compare them. The second experiment using the solution was performed on the Berlin database with Feature set 3. The mean accuracy obtained across 10 folds is 78.56% with a standard deviation of 3.57. Both results can be considered quite successful, as there are seven emotion classes for classification.

A number of experiments were performed on the Neosound data, on the three dataset splits (see Section 3.2 for details) with various parameters: different *C* parameter values as well as different kernels and different feature sets (detailed in Section 3.3). Another classification technique, Nu-SVC with different kernels, was also employed to see if it could improve the classification results. The mean accuracies from all the experiments with SVM are presented in Table 6.

In Dataset 1 the highest mean accuracy of 88.67% with a standard deviation of 2.61 was achieved with a linear-kernel SVM with the *C* parameter set to 1, when Feature set 3 was used a speech signal representation. It should be mentioned, however, that the results of the tests performed with other kernel functions on the same feature set do not lag far behind: the mean accuracy of 87.73% (+/-3.36) and 87.59 (+/-3.26) with an RBF and sigmoid kernels, respectively. The polynomial kernel gave the worst results for all features sets and parameters. The classification performed with Nu-SVC technique also shows similar results on Feature set 3: 88.16% (+/-3.52) – with a linear kernel, 87.73% (+/-2.95) – with an RBF kernel and 87.52% (+/-3.31) – with a sigmoid kernel.

The highest mean accuracy of 93.63% with a standard deviation of 1.99 was obtained in Dataset 2 with an RBF-kernel SVM using Nu-SVC classification technique on Feature set 3. There was not much variance in the results performed with other parameters and feature sets, however, for example, an RBF-kernel SVM achieved an accuracy of 92.83% (+/-1.35) using SVC classification technique. Setting C parameter to different values did not show any improvement in SVC classifications. As in the case with Dataset 1, polynomial kernel performed the worst on all the feature sets.

In Dataset 3 the highest mean accuracy of 84.26% with a standard deviation of 1.39 was achieved with an RBF-kernel SVM using Nu-SVC classification technique on Feature

set 3. As in the case with Dataset 1 and Dataset 2, the tests performed with other classification parameters, except for a polynomial kernel, showed similar results.

In general, the tests performed on Feature set 3 showed, on average, a 3% increase in performance for all the datasets. A better performance of the SVM on Dataset 2 than on Dataset 1 may also indicate that 'somewhat angry' segments do not belong to the 'angry' class indeed.

Parameters		Accuracy, %	Parameters		Accuracy, %
Dataset 1, Feature set 1					
kernel	RBF	82.41 (+/-2.85)	kernel	RBF	84.89 (+/-3.17)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	83.4 (+/-3.62)	kernel	sigmoid	83.4 (+/-3.16)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	78.56 (+/-2.03)	kernel	polynomial	77.24 (+/-3.32)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	82.48 (+/-3.34)	kernel	linear	83.05 (+/-2.92)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	rbf	84.61 (+/-2.74)	kernel	polynomial	73.37 (+/-6.26)
classification method	SVC		classification method	SVC	
C	1		C	1	
kernel	sigmoid	83.26 (+/-3.21)	kernel	linear	81.7 (+/-3.57)
classification method	SVC		classification method	SVC	
C	1		C	1	
Dataset 1, Feature set 2					
kernel	RBF	81.63 (+/-2.67)	kernel	RBF	82.41 (+/-3.06)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	82.62 (+/-3.23)	kernel	sigmoid	81.91 (+/-2.44)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	75.96 (+/-4.33)	kernel	polynomial	76.7 (+/-3.38)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	

Parameters		Accuracy, %	Parameters		Accuracy, %
kernel	linear	81.06 (+/-4.03)	kernel	linear	82.06 (+/-2.41)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
Dataset 1, Feature set 3					
kernel	RBF	84.82 (+/-3.16)	kernel	RBF	87.73 (+/-2.95)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	85.75 (+/-3.37)	kernel	sigmoid	87.52 (+/-3.31)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	83.26 (+/-4.65)	kernel	polynomial	83.12 (+/-3.68)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	85.68 (+/-1.94)	kernel	linear	88.16 (+/-3.52)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	rbf	87.73 (+/-3.36)	kernel	polynomial	76.74 (+/-4.23)
classification method	SVC		classification method	SVC	
C	1		C	1	
kernel	sigmoid	87.59 (+/-3.26)	kernel	linear	**88.67 (+/-2.61)**
classification method	SVC		classification method	SVC	
C	1		C	1	
Dataset 2, Feature set 1					
kernel	RBF	91.95 (+/-3.02)	kernel	RBF	91.78 (+/-1.91)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	91.95 (+/-2.26)	kernel	sigmoid	92.04 (+/-2.7)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	81.78 (+/-3.36)	kernel	polynomial	89.47 (+/-3.05)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	91.42 (+/-2.36)	kernel	linear	92.04 (+/-2.67)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	

Parameters		Accuracy, %	Parameters		Accuracy, %
Dataset 2, Feature set 2					
kernel	RBF	92.39 (+/-2.67)	kernel	RBF	92.39 (+/-2.18)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	92.03 (+/-2.13)	kernel	sigmoid	92.75 (+/-3.09)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	83.1 (+/-3.62)	kernel	polynomial	89.21 (+/-3.36)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	88.76 (3.34)	kernel	linear	92.83 (+/-2.84)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
Dataset 2, Feature set 3					
kernel	RBF	92.83 (+/-1.35)	kernel	RBF	**93.63 (+/-1.99)**
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	92.3 (+/-1.56)	kernel	sigmoid	92.66 (+/-2.57)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	89.83 (+/-2.61)	kernel	polynomial	87.44 (+/-2.94)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	91.6 (+/-1.92)	kernel	linear	92.66 (+/-2.13)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
Dataset 3, Feature set 1					
kernel	RBF	80.39 (+/-2.6)	kernel	RBF	80.66 (+/-1.63)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	78.75 (+/-1.45)	kernel	sigmoid	76.91 (+/-2.03)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	72.13 (+/-2.6)	kernel	polynomial	75.81 (+/-1.91)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	

Parameters		Accuracy, %	Parameters		Accuracy, %
kernel	linear	79.56 (+/-2.63)	kernel	linear	80.15 (+/-2.52)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
Dataset 3, Feature set 2					
kernel	RBF	80.89 (+/-2.38)	kernel	RBF	81.47 (+/-2.1)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	79.35 (+/-1.75)	kernel	sigmoid	77.88 (+/-1.95)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	73.24 (+/-2.43)	kernel	polynomial	75.74 (+/-2.23)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	76.95 (+/-3.09)	kernel	linear	80.07 (+/-2.0)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
Dataset 3, Feature set 3					
kernel	RBF	84.12 (+/-2.32)	kernel	RBF	**84.26 (+/-1.39)**
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	sigmoid	80.81 (+/-3.03)	kernel	sigmoid	80.59 (+/-2.33)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	polynomial	81.32 (+/-2.33)	kernel	polynomial	81.52 (+/-2.17)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	
kernel	linear	83.53 (+/-2.2)	kernel	linear	82.8 (+/-2.03)
classification method	SVC		classification method	Nu-SVC	
C	10		nu	0.4	

Table 6. SVM experiments results

4.2.2 Neural Networks

4.2.2.1 Bidirectional Long Short-Term Memory

The architecture of the model used to perform the following tests is described in detail in Section 3.1.2.1.

The first experiment on the solution was the replication of the original experiment carried out by the authors (detailed in Section 3.1.2.1.4) to check if their results are close to the ones we obtained. The mean accuracy achieved across 10 folds is 65.89% with a standard deviation of 2.89, which is close to the result obtained in the original experiment.

A few experiments were performed with the solution on the Neosound data: the experiments were performed on three dataset splits (Dataset 1, Dataset 2 and Dataset 3) with Feature set 2 as a signal representation. Then, an additional layer was added and the solution was trained and tested on Dataset 1. However, as the highest mean accuracy reached only 79.21%, moreover, the training of the solution takes at least five days, it has been decided not to perform further tests. The results of the experiments are presented in Table 7.

Parameters	Accuracy
Dataset 1, Feature set 2	
Default	70.71 (+/-4.1)
default + 1 fully connected layer	64.04 (+/-9.58)
Dataset 2, Feature set 2	
Default	**79.21 (+/-8.76)**
Dataset 3, Feature set 2	
Default	65.44 (+/-5.44)

Table 7. BLSTM experiments results

4.2.2.2 Deep Neural Network + Extreme Learning Machine

The architecture of the solution used to perform the experiments is detailed in Section 3.1.2.2.

The original experiment was not replicated in this project due to technical problems. The IEMOCAP database requires at least 20 Gb space on the disk and the training of the model takes up equivalent amount of space. Our server cannot deal with such amount of data. However, the solution has been trained and tested on the Berlin database with Feature set 1 and Feature set 3 as signal representations. The mean accuracy across 10 folds for Feature set 1 is 28.54% with a standard deviation of 3.72. The mean accuracy across 10

folds for Feature set 3 is 50.32% with a standard deviation of 3.63. Considering the number of emotion classes (seven, in this case), the result for Feature set 3, is satisfactory, however, much better results have been reported for the classification of the emotions on the Berlin database, described in the Literature Review part (see Section 2).

A number of experiments were performed on the Neosound data with the DNN+ELM model. The results of the experiments are presented in Table 8.

In Dataset 1, the highest mean accuracy of 82.38% with a standard deviation of 2.88 was achieved when an additional layer was added to the DNN with Feature set 3 as its input. However, the DNN with default 3 layers and additional 2 layers gave close results of 81.57% (+/-3.38) and 80.95% (+/-3.27), respectively. Feature set 1 as an input gave worse results, the same goes for the experiments with SVM, therefore, it has been decided to run other dataset splits only with Feature set 3. The mean accuracy of 87.98% with a standard deviation of 4.17 was obtained for Dataset 2, which proves the hypothesis that 'somewhat angry' segments may belong to a separate class, rather than being a part of the 'angry' class. The mean accuracy in Dataset 3 was 65.45% with a standard deviation of 5.75.

Parameters	Accuracy
Dataset 1, Feature set 1	
Default	74.56 (+/-7.66)
Dataset 1, Feature set 3	
Default	81.57 (+/-3.38)
default + 1 layer	**82.38 (+/-2.88)**
default + 2 layers	80.95 (+/-3.27)
Dataset 2, Feature set 3	
Default	87.98 (+/-4.17)
Dataset 3, Feature set 3	
Default	65.45 (+/-5.75)

Table 8. DNN+ELM experiments results

5 Conclusions and future works

In this document, we presented the results of the work obtained for this master's thesis. This section will revise the completed tasks and give an outline of the work that can be done in the future.

5.1 Conclusions

This master's thesis has described several speech emotion classification models, which are currently used in speech emotion recognition field, and presented the results of the experiments testing some of the available ones.

The theoretical review has shown that the most successful speech emotion recognition systems include SVM, neural networks, RNN and CNN, in particular, and end-to-end systems. Different features representing a signal is another concern when building a speech emotion recognition system, and according to numerous researchers, the ideal set of features has not been found yet. However, end-to-end systems may release the researchers of this task.

In this work, SVM, recurrent LSTM and DNN combined with ELM, as well as three sets of features have been tested on the issue of speech emotion recognition. The systems have been trained and tested on the dataset containing spontaneous speech. The task was to classify two, or three emotion classes, depending on the dataset split (see Section 3.2 for more details). The fact that the experiments were performed on spontaneous speech brings a special value of this project, since the acquisition of such data is challenging, as well as classification of emotions from spontaneous speech is more difficult than from acted speech (Vogt and André 2005). The highest accuracy was achieved with an SVM classifier both for the classification of two and three emotion classes: 93.63% and 84.26%, respectively. The closest results were obtained with DNN+ELM classifiers: 82.38% accuracy for the classification of two emotion classes and 65.45% accuracy for the classification of three emotion classes. The worst results have been shown by the LSTM classifier, which showed 79.29% accuracy for the classification of two and 65.44% for the classification of three emotion classes, however, these results are not that worse than the results from DNN+ELM classifiers. The best results on SVM and DNN+ELM classifiers were obtained with the feature set including MFCC, pitch, harmonics-to-noise ratio and their delta across time frames.

All the classifiers have also been trained and tested on the Berlin speech emotional database to check if the tendency is the same with the classifiers when tested on non-spontaneous speech. For two of the solutions, SVM and LSTM, this has been a replication of the original experiment, as the authors have also tested their systems on this database, i.e. Feature set 1 (detailed in Section 3.3) has been used as a speech signal representation. The classification accuracy for an SVM classifier is 71.12% and for an LSTM classifier is 65.89%. DNN+ELM classifiers have been trained and tested on the database with both Feature set 1 and Feature set 3. For Feature set 1, the classification accuracy is 28.54%, for Feature set 3 – it is 50.32%. As Feature set 3 proved to be more successful for the Neosound data, SVM classifier, being also the most accurate one, has also been trained and tested on the Berlin database with Feature set 3, as a speech signal representation. In this case, SVM classification has shown increase to 78.56% accuracy. This proves that the SVM classifier combined with Feature set 3 is the best of the tested solutions.

5.2 Future work

As has been mentioned above, the accuracy obtained with an SVM classifier for the classification of two emotions in spontaneous speech is rather high: 93.63%. Some of the classifiers described in the Literature review section have shown better results, some – worse. Although the results described in the Literature review and the results obtained in this project are incomparable, since the experiments were performed on different databases and different features, this reference is made only to know that the results of speech emotion recognition can be quite successful.

The dataset, on which the experiments were performed, has a certain particularity: the annotators labelled some of the segments as ‘somewhat angry’ instead of simply ‘angry’, therefore, it was decided to check how well the classifiers can identify these classes. When ‘somewhat angry’ segments are part of the ‘angry’ class, the classification accuracy is 88.67%, when ‘somewhat angry’ segments are left out of the training and testing dataset, the classification accuracy raises to 93.63%, when ‘somewhat angry’ segments are added to the training and testing dataset as a separate class, the classification accuracy is 84.26% (the mentioned classification accuracies describe the results of the tests performed with an SVM, being the best classifier from the tested ones). As has been mentioned in the above sections, a better performance on the two classes, when ‘somewhat angry’ segments has been left out, may indicate that these segments indeed belong to a separate class, nonetheless, the classification accuracy of the two classes,

when these segments are part of the ‘angry’ class, is higher than the accuracy of the three classes classification, which may indicate the opposite conclusion. Therefore, it may be possible to introduce a term ‘degree of emotion’, i.e. instead of simply telling which emotion it is, the classifier also adds information of how strong the emotion is. This may be one of the directions of the future work for the project. It should be mentioned, however, that the lower accuracy of the classification of three classes is normal than that of the two, since the task is more difficult.

Another direction is to test other speech emotion classification systems, which have been described in the Literature review section, mainly CNN and end-to-end systems, and compare the results with the ones obtained in the project.

In (Niu et al. 2017), the authors present how DAARIP feature processing algorithm, borrowed from image processing, can help increase the accuracy in speech emotion classification. The authors tested their solution on nonspontaneous speech, therefore, it wold be useful to test their approach on spontaneous speech, which may be one of the other future works for the project.

6 Bibliography

Atal, Bishnu. 1974. “Efiectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification.” *The Journal of the Acoustical Society of America* 55: 1304–22. https://doi.org/10.1121/1.1914702.

Aubrey, Andrew J, David Marshall, Paul L Rosin, Jason Vandeventer, Douglas W Cunningham, and Christian Wallraven. 2013. “Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations.” In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/CVPRW.2013.48.

Averill, James R. 1980. “A Constructivist View of Emotion.” In *Theories of Emotion*, 305–39. Academic Press. https://doi.org/10.1016/B978-0-12-558701-3.50018-1.

Awad, Mariette, and Rahul Khanna. 2015. “Hidden Markov Model.” In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, 81–104. Berkley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_5.

Ayadi, Moataz El, Mohamed S. Kamel, and Fakhri Karray. 2011. “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases.” *Pattern Recognition* 44 (3): 572–87. https://doi.org/10.1016/j.patcog.2010.09.020.

Badshah, Abdul Malik, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. 2017. “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network.” In *International Conference on Platform Technology and Service (PlatCon)*. https://doi.org/10.1109/PlatCon.2017.7883728.

Banse, Rainer, and Klaus R. Scherer. 1996. “Acoustic Profiles in Vocal Emotion Expression.” *Journal of Personality and Social Psychology* 70 (3): 614–36. https://doi.org/10.1037/0022-3514.70.3.614.

Bou-Ghazale, S E, and J H L Hansen. 2000. “A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress.” *IEEE Transactions on Speech and Audio Processing* 8 (4): 429–42. https://doi.org/10.1109/89.848224.

Breazeal, Cynthia, and Lijin Aryananda. 2002. “Recognition of Affective Communicative Intent in Robot-Directed Speech.” *Autonomous Robots 2*.

Brownlee, Jason. n.d. “A Gentle Introduction to Convolutional Layers for Deep Learning Neural Networks.” Accessed August 23, 2019. https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/.

Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. “A Database of German Emotional Speech.” In *Proceedings of Interspeech, Lissabon*, 3–6.

Busso, C, S Lee, and S Narayanan. 2009. “Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection.” *IEEE Transactions on Audio, Speech, and Language Processing* 17 (4): 582–96. https://doi.org/10.1109/TASL.2008.2009578.

Busso, Carlos, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database.” *Language Resources and Evaluation* 42 (4): 335–59. https://doi.org/10.1007/s10579-008-9076-6.

Calvo, Rafael A., and Sidney D’Mello. 2010. “Affect Detection : An Interdisciplinary Review of Models , Methods , and Their Applications Affect Detection : An Interdisciplinary Review of Models , Methods , and Their Applications.” *Affective Computing, IEEE Transactions on Affective Computing* 1 (1): 18–37. https://doi.org/10.1109/T-AFFC.2010.1.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Supprot-Vector Networks.” *Machine Learning* 297 (20): 273–97. https://doi.org/10.1111/j.1747-0285.2009.00840.x.

Cowie, Roddy, Ellen Douglas-Cowie, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor. 2001. "Emotion Recognition - IEEE Signal Processing Magazine." *IEEE Signal Processing Magazine*, no. January.

Darwin, Charles. 1872. *Expression of the Emotions in Man and Animals*. London: John Murray.

Ding, Shifei, Han Zhao, Yanan Zhang, Xinzheng Xu, and Ru Nie. 2015. "Extreme Learning Machine: Algorithm, Theory and Applications." *Artificial Intelligence Review* 44 (1): 103–15. https://doi.org/10.1007/s10462-013-9405-z.

Ekman, Paul. 1971. "Universals and Cultural Differences in Facial Expressions of Emotion." In *Proc. Nebraska Symp. Motivation*, 207–83.

Engberg, Inger Samso, and Anya Varnich Hansen. 1996. "Documentation of the Danish Emotional Speech Database Des." *Internal AAU Report, Center for Person Kommunikation, Denmark*, 22.

France, D J, R G Shiavi, S Silverman, M Silverman, and M Wilkes. 2000. "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk." *IEEE Transactions on Biomedical Engineering* 47 (7): 829–37. https://doi.org/10.1109/10.846676.

Fu, Liqin, Xia Mao, and Lijiang Chen. 2008. "Speaker Independent Emotion Recognition Based on SVM / HMMs Fusion System." In *International Conference on Audio, Language and Image Processing*, 61–65.

Giannakopoulos, Theodoros. 2015. "PyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis." *PLoS ONE* 10 (12): 1–17. https://doi.org/10.1371/journal.pone.0144610.

Gunes, Hatice, and Björn Schuller. 2013. "Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions." *Image and Vision Computing* 31 (2): 120–36. https://doi.org/10.1016/J.IMAVIS.2012.06.016.

Han, Kun, Dong Yu, and Ivan Tashev. 2014. "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine." In *Interspeech 2014*, 223–27. https://doi.org/10.1109/ICASSP.2013.6639346.

Haq, S, and P J B Jackson. 2009. "Speaker-Dependent Audio-Visual Emotion Recognition." In *Proc. Int'l Conf. on Auditory-Visual Speech Processing (AVSP'08), Norwich, UK*.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1–32.

Huang, Zhengwei, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. "Speech Emotion Recognition Using CNN Categories and Subject Descriptors." In *ACM International Conference*, 801–4.

James, William. 1884. "What Is an Emotion?" In *Mind*, os-IX:188–205. https://doi.org/10.1093/mind/os-IX.34.188.

Johnstone, Tom, and Klaus R. Scherer. 2000. "Vocal Communication of Emotion." In *Encyclopedia of Personality and Individual Differences*, edited by M. Lewis and J. Haviland, 226–35. New York, NY: Guilford. https://doi.org/10.1007/978-3-319-28099-8_562-1.

Kecman, Vojislav. 2014. *Support Vector Machines – An Introduction Support Vector Machines – An Introduction*. https://doi.org/10.1007/10984697.

Kemper, Theodore D. 1991. "Predicting Emotions from Social Relations." *Social Psychology Quarterly* 54 (4): 330–42. http://www.jstor.org/stable/2786845.

Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. 2019. "Automatic Speech Emotion Recognition Using Machine Learning." In *Social Media and Machine Learning*. IntechOpen.

https://doi.org/10.5772/intechopen.84856.

Knill, Oliver. 2009. *Probability Theory and Stochastic Processes with Applications*. *Probability Theory and Stochastic Processes with Applications*. Overseas Press. https://doi.org/10.1142/10029.

Kossaifi, Jean, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Kam Star, Elnar Hajiyev, and Maja Pantic. 2019. "SEWA DB : A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild."

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, 1097–1105.

Lalitha, S, Shikha Tripathi, and Deepa Gupta. 2018. "Enhanced Speech Emotion Detection Using Deep Neural Networks." *International Journal of Speech Technology* 0 (0): 0. https://doi.org/10.1007/s10772-018-09572-8.

Lang, Peter J. 1994. "The Varieties of Emotional Experience: A Meditation on James-Lange Theory." *Psychological Review* 101 (2): 211–21.

Lee, Chul Min, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. 2004. "Emotion Recognition Based on Phoneme Classes." *8th International Conference on Spoken Language Processing, ICSLP 2004*, no. 1: 889–92.

Li, Ya Feng, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. 2017. "CHEAVD: A Chinese Natural Emotional Audio–Visual Database." *Journal of Ambient Intelligence and Humanized Computing* 8: 913–24.

Lim, Wootaek, Daeyoung Jang, and Taejin Lee. 2016. "Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks." In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*.

Lin, Lawrence I-Kuei. 1989. "A Concordance Correlation Coefficient to Evaluate Reproducibility." *Biometrics* 45 (1): 255–68. http://www.jstor.org/stable/2532051.

Lin, Yi Lin, and Gang Wei. 2005. "Speech Emotion Recognition Based on HMM and SVM." *2005 International Conference on Machine Learning and Cybernetics, ICMLC 2005*, no. August: 4898–4901.

Mao, Xia, Lijiang Chen, and Bing Zhang. 2007. "Mandarin Speech Emotion Recognition Based on a Hybrid of HMM / ANN." *International Journal of Computers* 1 (4): 1–4.

McCulloch, Warren, and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–33.

Mckeown, Gary, Michel Valstar, and Roddy Cowie. 2007. "The SEMAINE Database : Annotated Multimodal Records of Emotionally Coloured Conversations between a Person and a Limited Agent." *IEEE Transactions on Affective Computing* 3 (1): 1–14.

Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. 2017. "Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention Center for Robust Speech Systems , The University of Texas at Dallas , Richardson , TX 75080 , USA Microsoft Research , One Microsoft Way , Redmond , WA 98052 , USA." *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2017*, 2227–31. https://doi.org/10.1109/ICASSP.2017.7952552.

Mori, Shinya, Moriyama Tsuyoshi, and Shinji Ozawa. 2006. "Emotional Speech Synthesis Using Subpace Sonstraints in Prosody." In *IEEE International Conference on Multimedia and Expo*, 1093–96.

Murray, Iain R, and John L Arnott. 1993. "Toward the Simulation of Emotion in Synthetic

Speech: A Review of the Literature on Human Vocal Emotion." *Journal of the Acoustical Society of America*. US: Acoustical Society of American. https://doi.org/10.1121/1.405558.

Nielsen, Michael A. 2018. "Neural Networks and Deep Learning." Determination Press. http://neuralnetworksanddeeplearning.com/.

Niu, Yafeng, Dongsheng Zou, Yadong Niu, Zhongshi He, and Hua Tan. 2017. "A Breakthrough in Speech Emotion Recognition Using Deep Retinal Convolution Neural Networks," 1–7.

Nwe, Tin Lay, Say Wei Foo, and Liyanage C De Silva. 2003. "Speech Emotion Recognition Using Hidden Markov Models." *Speech Communication* 41 (4): 603–23. https://doi.org/10.1016/S0167-6393(03)00099-2.

Öster, Anne-Marie, and Arne Risberg. 1986. "The Identification of the Mood of a Speaker by Hearing Impaired Listeners." *Speech Transmission Lab. Quarterly Progress Status Report 4* 7: 79–90.

Parthasarathy, Srinivas, and Ivan Tashev. 2018. "Convolutional Neural Network Techniques for Speech Emotion Recognition." In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 1–5.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2012. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*, no. January. http://arxiv.org/abs/1201.0490.

Rabiner, Lawrence, and Biing-Hwang Juang. 1986. "An Introduction to Hidden Markov Models." *IEEE ASSP Magazine* 3 (January).

Ringeval, F, A Sonderegger, J Sauer, and D Lalanne. 2013. "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions." In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–8. https://doi.org/10.1109/FG.2013.6553805.

Roseman, Ira, Martin S. Spindel, and Paul Jose. 1990. "Appraisals of Emotion-Eliciting Events: Testing a Theory of Discrete Emotions." *Journal of Personality and Social Psychology* 59: 899–915. https://doi.org/10.1037/0022-3514.59.5.899.

Rosenblatt, Frank. 1962. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

Ruggieri, Mario. n.d. "Emotion-Recognition-from-Speech: A Machine Learning Application for Emotion Recognition from Speech." GitHub. Accessed July 27, 2019. https://github.com/MarioRuggieri/Emotion-Recognition-from-Speech.

Schuller, Björn. 2018. "Speech Emotion Recognition. Two Decades in a Nutshell, Benchmarks, and Ongoing Trends." *Communications of the Acm* 61 (5). https://doi.org/10.1145/3129340.

Schuller, Björn, Stephan Reiter, and Gerhard Rigoll. 2006. "Evolutionary Feature Generation in Speech Emotion Recognition." In *IEEE International Conference*, 5–8. https://doi.org/10.1109/ICME.2006.262500.

Seehapoch, Thapanee, and Sartra Wongthanavasu. 2013. "Speech Emotion Recognition Using Support Vector Machines." In *5th International Conference on Knowledge and Smart Technology*, 86–91.

Slaney, Malcolm, and Gerald McRoberts. 2003. "BabyEars: A Recognition System for Affective Vocalizations." *Speech Communication* 39 (3–4): 367–84. https://doi.org/10.1016/S0167-6393(02)00049-3.

Stankovic, Igor, Montri Karnjanadecha, and Vlado Delic. 2011. "Improvement of Thai Speech Emotion Recognition By Using Face Feature Analysis." In *International Symposium on Intelligent Signal Processing and Communication Systems (LSPACS)*.

Steidl, Stefan. 2009. "Automatic Classification of Emotion-Related User States in Spontaneous

Children's Speech." Friedrich-Alexander-Universität Erlangen-Nürnberg.

Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. "Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network." *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2016-May (October 2017): 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472669.

Tzirakis, Panagiotis, Jiehao Zhang, and Bjorn W. Schuller. 2018. "End-to-End Speech Emotion Recognition Using Deep Neural Networks." *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2018-April: 5089–93. https://doi.org/10.1109/ICASSP.2018.8462677.

Vignolo, Leandro D., S.R. Mahadeva Prasanna, Samarendra Dandapat, H. Leonardo Rufiner, and Diego H. Milone. 2016. "Feature Optimisation for Stress Recognition in Speech." *Pattern Recognition Letters* 84 (i): 1–7. https://doi.org/10.1016/j.patrec.2016.07.017.

Vogt, Thurid, and Elisabeth André. 2005. "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition." *IEEE International Conference on Multimedia and Expo, ICME 2005*. https://doi.org/10.1109/ICME.2005.1521463.

Wang, Rayan. n.d. "Speech_emotion_recognition_BLSTM: Bidirectional LSTM Network for Speech Emotion Recognition." GitHub. Accessed July 27, 2019. https://github.com/RayanWang/Speech_emotion_recognition_BLSTM.

Wang, Z, and I Tashev. 2017. "Learning Utterance-Level Representations for Speech Emotion and Age/Gender Recognition Using Deep Neural Networks." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5150–54. https://doi.org/10.1109/ICASSP.2017.7953138.

Williams, Carl E, and Kenneth N Stevens. 1981. "Vocal Correlates of Emotional States." *Speech Evaluation in Psychiatry*, 221–40.

Xu, Min, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. 2004. "HMM-Based Audio Keyword Generation," 566–74. https://doi.org/10.1007/978-3-540-30543-9_71.

Yang, B, and M Lugger. 2010. "Emotion Recognition from Speech Signals Using New Harmony Features." *Signal Processing* 90 (5): 1415–23. https://doi.org/10.1016/j.sigpro.2009.09.009.

Yu, Dong, and Li Deng. 2012. "Efficient and Effective Algorithms for Training Single-Hidden-Layer Neural Networks." *Pattern Recognition Letters* 33 (5): 554–58. https://doi.org/10.1016/j.patrec.2011.12.002.

Zhao, Jianfeng, Xia Mao, and Lijiang Chen. 2018. "Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks." *Biomedical Signal Processing and Control* 47: 312–23. https://doi.org/10.1016/j.bspc.2018.08.035.

Zheng, W Q, J S Yu, and Y X Zou. 2015. "An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks." In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 827–31.