DEPARTMENT OF PHILOSOPHY,

LINGUISTICS AND THEORY OF SCIENCE

Prosody and emotion: Towards the development of an emotional agent

Emotional evaluation of news reports: production and perception experiments

Liina Tumma

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2018
Supervisor:	Dr. Charalambos Themistocleous
Examiner:	Dr. Staffan Larsson
Report number:	(number will be provided by the administrators)
Keywords:	emotion, speech, prosody,

Abstract

There is a recognised need for more research on the topic of emotion recognition from speech, and clear and defined methodology in this area is still lacking. Most studies in the field of emotional speech recognition and classification usually focus on acted speech as the data source; consequently, other methods that capture more natural speech are left aside. This study presents a novel perspective on corpus collection and emotion classification technique. The emotion from the perspective of an evaluation device is also highlighted. The aim of the study is to investigate the possibility to evoke happy, neutral and sad emotions from news reports, and to analyse the acoustic predictors that play a crucial role in the prediction of these emotions.

The thesis is based on three experiments: i. corpus collection by eliciting sad, happy and neutral emotional speech through news, and posterior statistical analysis of this data (Mixed-effect models); ii. automatic classification of these emotions by training Decision Tree (C5.0) classification models; iii. perception experiment to verify the findings from the previous experiments. Speech data obtained from 20 native speakers of Swedish is analysed. The participants were asked to summarize and give their personal opinion on 36 news reports about happy and sad events and read out loud 12 neutral Wikipedia short descriptions. To investigate emotion as an evaluation device, sad news reports are categorized following the Brandt Line division between Global North (developed countries) and Global South (developing countries).

Results indicate that news reports are suitable to be used as stimuli to evoke emotional responses of Swedish speakers. Decision Tree (DT) classifier reached an average accuracy of 70.88% (tested on validation data from 10-fold cross-validation). Final velocity, relative location of the *F0* peak, time of the *F0* peak and mean intensity are crucial attributes for the classifier. The perception experiment has also proved that Swedish speakers are capable of identifying and classifying these emotions, although machine learning outperforms the human evaluation. The findings do not show any clear difference between South and North news reports and therefore no evidence regarding emotion as an evaluation device in case of South and North news is found.

The findings can contribute to a better understanding of evaluation as a speech device and it also explores other possibilities regarding corpus collection and classification methods, such as using news reports as emotion stimuli and a Decision Tree algorithm for classification. The research results represent a further step towards developing an emotional agent.

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor Charalambos Themiscoleus, for his guidance, encouragement and help throughout the project. Furthermore, I am grateful to the Göteborg University staff for facilitate the booking of the rooms for the experiments and providing the access to Webropol Surveys. Special thanks also go to Peter Carlsson for being my experiment assistant and my personal Swedish advisor all along the project. I also wish to express my gratitude to my family and friends for giving me support and strength to complete the project, and specially to Tati for her patience and support.

1. Introduction **1**

2. Background **4**

2.1. News and emotions 4

2.2. Linguistic and emotional prosody 6

2.3. Emotions 7

2.4. Importance of emotional prosody 8

2.5. Corpora collection 9

2.6. Emotion detection and classification 9

2.7. Machine Learning Applications 12

3. Study 1: Production (priming experiment) **13**

3.1. Swedes, Swedish and Swedish dialects 13

3.2. Methodology 13

3.2.1. Speech material 13

3.2.2. Speakers 14

3.2.3. Procedure 15

3.2.4. Measurements 16

3.2.5. Statistical Analysis 17

3.3. Results 17

3.4. Discussion 24

4. Study 2: Classification **25**

4.1. Methodology 25

4.1.1. Decision Tree classification 26

4.2. Results 28

4.3. Discussion 28

5. Study 3: Perception experiment - emotion categorization **29**

5.1. Methodology 29

5.1.1. Speech material 30

5.1.2. Procedure 30

5.2. Results 31

5.3. Discussion 32

6. Final Discussion **33**

7. Limitations of the study **34**

8. Future work **34**

Bibliography **36**

Appendix A: Perception experiment - transcription of the used sentences **44**

Annex B: Perception experiment - demography survey **46**

B.1. Age 46

B.2. Sex 47

B.3. Education level 47

B.4. Residency of the participants 48

1. Introduction

Personal communication goes hand in hand with emotions. The role of emotions in speech interactions is crucial. However, for decades many studies did not give enough importance to this property (Davitz, 1964; Baber and Noyes, 1996; Scherer 2003) and research on this topic is still separated from the speech analysis literature (Murray and Arnott, 2008). There is clear evidence that people are capable of recognizing the emotional state of the speaker by hearing one short utterance (see Scherer, 2003). So for instance, the discourse marker "okay" might have different meanings depending on the emotional intonation used by the speaker, such as affirmation ("Let's go?"- "Okay"), delight ("and this is how you fix it" - "Okay!"), confidence ("Are you ready to play?" - "Okay") or confusion ("Now we need to add H_2SO_4" - "Okay…") (Louwerse and Mitchell, 2003). Therefore, in order to build an efficient and more natural human computer interaction machine, emotion cannot be overlooked.

Over the past few years, there have been several studies that explored emotional prosody from different perspectives (for more details see Scherer, 2003) and emotion classification of speech has an average reported accuracy of 70 to 80% (Neiberg et al., 2006; Yacoub et al., 2003). Curiously, the vast majority of published researches on emotional prosody use emotional samples produced by professional actors (acted speech) (Lee et al., 2006; Hoque et al., 2006; Seppänen et al., 2003; Yu et al., 2001; Mozziconacci, 1995; Banse and Scherer, 1996) and fewer researchers are focused on elicited or induced speech (McMahon et al., 2003; Batliner et al., 2004; Noroozi et al., 2017) and natural speech (Lee and Narayanan, 2005; Zhang, 2008). Koolaguda and Rao (2012) conclude that overall 60% of the collected databases are acted speech. Furthermore, acted speech has been criticized regarding the naturalness of emotion and some researchers consider that more relevant results are obtained when a database is consisted of as close to possible to spontaneous emotion speech samples (Tato et al., 2002; Stibbard, 2001). In this case, it is important to highlight the lack of more studies that focus on elicited and natural speech instead of acted speech.

Emotion can be evoked with help from many different stimuli such as music (Juslin and Laukka, 2004), utterances or text (Bao et al., 2011; Le Tallec, 2011; Strapparava and Mihalcea, 2007), pictures (Mikels et al., 2005) and even odor (Kadohisa, 2013). However, news reports could also be a potential data source for a stimulus to evoke emotion, since news is a media with expectations to evoke particular types of responses in readers. In fact, many researchers have their interest set on mass media psychology, focusing on the psychological process during media reception and the role of emotions (e.g., Schramm & Wirth, 2005; Zillmann, 1988; Nabi, 1999). Scherer (1998) concludes that although media-induced emotions are processed in the same way as naturally-evoked emotions, some changes might be noticed. The difference between the two scenarios is due to the less personal nature of the event in the case of mass media news, where more general experiences are treated. Moreover, it is shown that if the news reports present emotional pictures it evokes more feelings in the participants (e.g. Bucy, 2003; Wirth and Böcking, 2003).

From the perspective of sentiment analysis in text, several studies based on news corpora prove that mass media contains sources of emotion itself. For example, Bhowmick et al., (2009) and Strapparava and Mihalcea (2007) focus on the method to classify emotion in news headlines or sentences; and Mohammad and Turney (2010) focus on the creation of an emotion lexicon to

identify the emotional tone of larger units of text, such as newspaper headlines and blog posts. In the previously mentioned studies, annotators were used in order to label the emotion categories of the utterances or texts. Although mass media is proven to contain emotional information, as far as we recall, we are not aware of any study, which uses news articles as a stimuli to elicit emotions in order to investigate emotional prosody.

1.1 Aims and research questions

Based on the researches described above regarding emotions, corpus collection and the methodology for eliciting emotions, the following questions were asked:

1. **Will stimuli in the form of happy and sad news reports with an image affect the emotional state of Swedish speakers based on the speech?**
2. **Which prosodic factors convey emotional evaluation of news reports?**
3. **Can Swedish speakers perceive and classify emotions from the obtained corpora?**

To answer these research questions, three experiments were performed. First of all, data collection was conducted; with the aim to collect the emotional data using news reports evoking happy, neutral and sad emotions from Swedish participants. In total, 20 participants were asked to summarize and give their personal opinion on 36 reports and read out loud 12 neutral Wikipedia articles. Once the data was collected, the data was statistically analysed in order to understand the obtained data on a more detailed level. Duration, intensity, velocity and fundamental frequency related features were used, since they are the most important correlates of prosodic features (Murray and Arnott, 1993; Cowie and Cornelius, 2003; Banziger and Scherer, 2005). The second experiment focused on training a classification model algorithm to classify happy, neutral and sad emotions, and reveal the most useful factors to be used when training a classification model. Decision Tree (C5.0) was explored as a classification method, as it was very suitable for the type of data set, even though relatively little research has been carried out using the Decision Tree method to train an emotion classifier. The third study was a perception experiment in order to answer the third research question. We considered that in order to prove the existing difference between happy and sad emotions through news stimuli, the results should be also verified with the help of humans.

One of the most noticeable contributions of the thesis is the use of innovative experimental design, which is based on the data obtained from elicited speech, using news reports as stimuli. Elicited speech is collected in a controlled setting, instead of already collected and segmented corpora. Moreover, no assumptions or annotations were made of the emotions, as it has to be done when dealing with natural speech (e.g., Zhang, S., 2008). In other words, priming the desirable emotions in order to obtain an emotion database is an innovative approach and it allows us to obtain more natural speech compared to acted speech.

Emotion could also be seen as an evaluation device that is used by speakers to express their opinions about the given news report. As a result, when describing the news reports speaker's perception of social issues and cultural beliefs can also be detected (Cortazzi and Jin, 2000).To explore emotions as an evaluation device, a fine line between two types of sad news is drawn: sad reports that involve news about Global North (developed countries) and Global South (developing countries). This brings us to last research question:

4. **Can the Global North-South divide be distinguished when speakers summarize and evaluate sad news reports?**

This question is not purely related to language technology, although the intersection of humanities (and social sciences) and computing technologies is called Digital Humanities and Social Science (HSS) and the research in this new field is growing rapidly (Lazer et al., 2009; Berry, 2012), thus supporting collaborative and interdisciplinary research. However, as a new field, e-science for the HSS still requires more development and to overcome more obstacles in the future (Viklund and Borin, 2016).

So how does the division between Global North and Global South relate to this study based on emotion detection? We are expecting that participants can be more or less emotionally affected by news reports depending on how close geographically the event has happened, and therefore the emotional response of the participants might be different. It is clear that many factors are involved. First of all, it could be the geographical proximity to the home country of the participants, but it could be also related to the socio-cultural proximity, as people tend to care more about people from similar cultures. This social behaviour can be exemplified with the "missing white woman syndrome", as white women usually get higher coverage in the news compared to missing women of colour, and even less if they are from lower socioeconomic classes (Stillman, 2007; Liebler, 2010). People also tend to perceive atrocities that happen in "conflict" areas as an unavoidable part of life, whereas more peaceful and prosperous nations are seen as shocking news when a similar accident happens (Butler, 2009; Sontag, 2003).

The aim of the present work is, therefore, twofold. On the one hand, we are aiming to investigate whether news reports are suitable to be used as stimuli to evoke emotional responses, and also to detect the best attributes that predict sad and happy emotions. On the other hand, we focus on the strength of the emotions that are evoked by different types of news and how emotion is used as an evaluation device when dealing with news reports. We focus on the difference between Global South and Global North, and the effect that these two categories of news might have on the participants, and whether it is possible to distinguish between them. Therefore, this novel research on experimentally evoked emotions based on news stimuli can open a new range of research possibilities in multidisciplinary fields such as journalism, sociology, psychology and even marketing, to mention only a few. Through this study, we also aim to contribute to method development in e-science for the HSS, particularly in the area of media/global studies. Another novel perspective of this study is to support research for less explored languages such as Swedish in the field of emotional recognition and classification, since most of the research has been done investigating English speakers and listeners, with some research in German and Chinese (Ververidis and Kotropoulos 2006).

1.2 Structure of the thesis

The thesis is structured in the following way: In Section 1, the introduction along with the motivation and contribution of this thesis to the field in language technology was presented. In Section 2, the background on prosody and emotion is presented. Section 3 consists of the first study (priming experiment), followed by the discussion. Section 4 is aimed at describing and discussing the second study (classification) and Section 5 contains the third study (perception experiment). The last three sections focus on conclusions, limitations of the study and future work, respectively.

2. Background

This section delves into the concepts and the discussion of previous researches on prosody and emotions. Particular attention is paid to categories of emotion, corpus collection, and emotional prosody. On the other hand, the concept of Global South and North is presented in order to be able to understand the relevance of this matter to the current thesis.

2.1. News and emotions

How speakers feel and relate to particular news is very interesting from the point of view of evaluation in discourse. Evaluation is the expression of the speaker's stance or feeling about something the speaker is talking about (Thompson and Huston, 2000). Speakers do not use emotions for the sake of using emotions, but it is part of the discourse: speakers describe what they see, feel and think and it conveys emotions also. Evaluation is a major criterion of narrative; in fact, narrative without evaluation is simply a report or summary (Cortazzi and Jin, 2000). Different types of performance can be involved in the narrative such as gestures, laughter, stress, changes in the tone of voice, emotional variance, etc. Overall, narrative evaluation can use any linguistic, paralinguistic or non-verbal devices. The evaluation of narratives can also vary cross-culturally, since narrative itself reflects culture. Tannen (1980), for example, compared the stories of Greek and American women after they watched a film and the results showed that while the Greeks focused on personal involvement, the Americans focused on more objective details and context.

When analyzing evaluation in narrative (as written or spoken report of events), speaker's perceptions of the evaluation of social issues and cultural beliefs can also be detected (Cortazzi and Jin, 2000). Van Dijk (1987), for example, found in his study of group interviews about ethnic minorities in Holland that the evaluations of the speakers had negative or inferior connotations about these minorities. The author analysed how ethnic prejudice and racism is reproduced through everyday talk. This also implies that we also offer our evaluation of situations, events or objects to others.

The concept of a global North-South divide was first mentioned by Willy Brandt and extensively described in the Brandt reports (Brandt, 1980) when dealing with the distribution of wealth in the world. The Brandt Line (Figure 1) is a way of showing how the world is geographically divided into relatively richer (North) and poorer (South) nations.

Nowadays, this division is not as simplistic as it was in 1983, as many countries have experienced social and economical development. However, this division is still present from a more social and global view, and the concepts of "Global North" and "Global South" are very commonly found in social and global studies research, which usually focus on analysing the inequality between these two. This geographical north-south divide lives alongside the inequality and same type division that might occur within the country or city. However, we find useful to apply the simplified North-South division in this study for several reasons.

Sontag (2003) in "Regarding the Pain of Others" examines how the viewers and victims' positions can lead to particular privileges or preferences. In the book, she examines the existing inequality regarding the representation of death depending on country. She relates to the time when, especially in times of war, in North America and Europe it was forbidden to show the

faces of the deceased victims as a sign of dignity or "good taste" (2003:68). However, these rules were not applied for images that were arriving from postcolonial Africa. Sontag claims that the regularity and pervasiveness of suffering in death in postcolonial Africa that can be seen in North American and European news only leads to normalization of these events, and makes us interpret the events in the "benighted" or "poor" parts of the world as something inevitable (Sontag, 2003:71). Moreover, Butler (2009) argues that we see some people as less grievable than others and we dehumanize the others to live with the fact that we accept unacceptable suffering in their lives. The distinction between "us" and "them" (or North and South) is a political matter and also a matter of power and vulnerability.

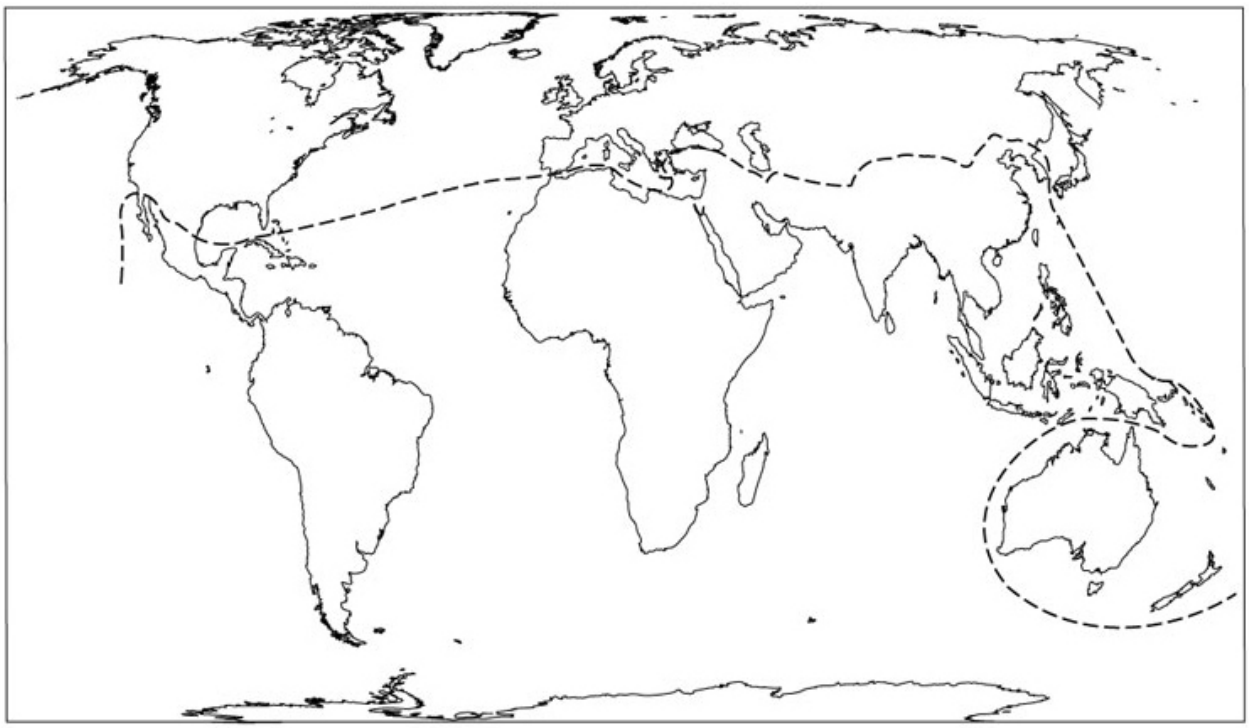


Figure 1: Map showing Brandt's North/South divide (from Smith and Warr, 1991: 245).

It is obvious that news has a clear emotional impact on readers or viewers and it has been widely investigated. This can be seen, for example, in the study of Unz, Schwab and Winterhoff-Spurk (2008) who examined the influence of violent television news based on viewers' facial expressions and emotional experiences, collected through viewers' ratings of subjective feelings. Their research proved that violence in TV news elicits negative emotions in viewers. The news also mirrors all the social prejudices and general behaviours regarding the South and North division. Adams (1986), for example, has discovered in his study about natural disasters and its coverage in North American news that two factors matter to predict the width of coverage of international disasters on U.S. news: how close geographically the event has happened and the number of U.S. tourists who visited the affected place. The social division between us and them, Global North and South, is detectable through news and therefore can have a clear impact on readers' emotions and the intensity of these emotions.

2.2. Linguistic and emotional prosody

Emotions are important for social interactions and for interpersonal relationships, and voice is one of the means to transport this information. Thus, the speech signal not only conveys linguistic information but also information about the speaker and his/her affective state (Laver and Trudgill, 1979). When talking, our emotional state is often clear based on our voice and face, and this information might be relevant to the meaning of the sentence. If somebody says "I have just received a message from my sister", the meaning and future interactions can vary depending on the emotion that was implied in this utterance. For instance, this person can experience fear, joy, anger, etc. Therefore, emotion expression can have a crucial impact on the listener and contribute to a successful communication.

Prosody refers to the melodic aspects of speech and conveys information that cannot be deduced from the lexical channel. In addition, prosody is employed to differentiate declarative sentences from questions, which are modified through pitch, but it also overlaps with emotion in speech. Some of the most important correlates of prosodic features, according to the literature, are pitch, energy, duration and their derivatives (Murray and Arnott, 1993; Cowie and Cornelius, 2003; Banziger and Scherer, 2005). Researchers have been studying for years what are the acoustic and prosodic features (e.g. speaking rate, intonation, intensity) that encode the emotional state of the speaker (Scherer et al., 1991). Prosody is still under discussion and many types of methodology and concepts can be found in different research. Particularly, some researchers find a division between linguistic and emotional prosody.

Intonation is one of the clearest features to detect emotion. Prosody is a domain of grammar; nevertheless, it interacts with other domains such as semantics and pragmatics. It can be divided into linguistic and emotional prosody. Linguistic prosody is used to express the intonation of the sentence, mark or disambiguate the internal organization of the sentence constituents (Ladd, 1996; Cutler and Clifton, 1999). Emotional prosody, on the other hand, provides listeners the needed information to understand speakers' emotions (Rigoulot and Pell, 2012). Both linguistic and emotional prosody employ the same prosodic features such as fundamental frequency, duration and intensity. Since prosodic features also vary with emotions, it is common that linguistic and emotional prosody overlap together (Mannell, 2007). Another significant difference between emotional and linguistic prosody is that linguistic prosody is language specific, whereas there is a prevalent assumption in the literature that emotions are not language specific.

The acoustic correlates of emotions that most of the researchers have focused on are fundamental frequency, duration, intensity and voice quality (Murray and Arnott, 1993). The authors present a short summary in the form of a table of different acoustic correlates for several emotions (Figure 2).

Even though much research has been done in the area of prosody and emotion, a predictive model of speech emotion (a model that can be used to accurately classify speech emotions) is still to be developed (Scherer, 2003; Chuenwattanapranithi, 2008). Apart from being a new multidisciplinary field involving psychology, sociology, cognition, linguistics, artificial intelligence and information technology, many researchers claim that the reason for this lack of development is the absence of a theoretical, unified background, and a clear and defined

methodology. More research supported by theory, models and methodology are needed to be able to improve existing models (Scherer, 2003; Xu, 2011; Chuenwattanapranithi et al., 2008).

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch average	Very much higher	Much higher	Slightly slower	Very much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice quality	Breathy, chest tone	Breathy, blaring	Resonant	Irregular voicing	Grumbled chest tone
Pitch changes	Abrupt, on stressed syllables	Smooth, upward inflections	Downward inflections	Normal	Wide downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

Figure 2. Summary of the most common correlates of emotions detected through speech (Murray and Arnott, 1993: 102)

2.3. Emotions

The lack of agreement on theoretical background in emotion detection and classification described in the previous chapter not only affects the definition of emotion itself but also the emotion labels which might vary widely from one study to another. As evidenced in different research, emotion might be understood as an attitude, mood or state of arousal. As Cowie and Cornelius (2003:6) claim, "In the meaning that people tend to feel is central, emotions are episodes that are relatively brief and highly distinctive. There is no generally agreed, compact term for these episodes." However, two main models should be taken into account. One theory understands emotions as categories as for example joy and sadness (Izard, 1977; Plutchik and Kellerman, 1980), with basic emotions and more complex combinations of the previous ones. Fear, anger, happiness, sadness, surprise, and disgust are 6 basic emotions that were proposed by Cornelius (1996), which he called the "Big Six".

Another approach is dividing emotions into dimensions. The most commonly used emotion dimension is activation (the degree of readiness to act), evaluation (positive and negative) and power (dominance and submission). Even though this simple description does not capture all the aspects of emotions, it presents a clear division between emotion categories, based on distance measurements and gradual representation of emotion (Schröder et al., 2001a). These features give a possibility to model weak emotions, which can improve the applications in speech synthesis that are mostly concentrated only on extreme emotion categories at the moment (Schröder et al., 2001a; Cowie, 2000). Emotional dimensions can be also combined to obtain more perspectives of particular emotions. As an example, Schröder et al., (2001b), instead of dealing with primary emotions (or emotional extremes) as most of the researchers, organizes emotions in a two-dimensional space as positive/negative (how positive or negative the emotion is) and active/passive (how much energy the person will experience during a precise emotion). Zei (2002), on the other side, proposes a three-dimensional space where emotion, cognition and

behaviour are merged, and this three-dimensional space contains valence, arousal and power for each dimension.

As we have seen so far, having different categories or dimensions for the emotions does not provide a clear consistency in the research field. Due to this difference, it is difficult to compile the data from different researchers (Douglas-Cowie et al., 2003). However, it also might be that by restricting and unifying the categories of emotion, some are lost and missed in the revision, and therefore this does not help to advance the field (Salmon, 2001). Lee and Narayanan (2005) also argues that even though the ability to recognize a large quantity of different emotions is attractive, it might not be practical or strictly necessary in the context of conversational interfaces, and in his study he analyses negative and non-negative emotions only in order to improve call centre applications.

2.4. Importance of emotional prosody

Minsky (2006) states that emotion is not that different from the process we interpret as "thinking" and it is necessary in artificial intelligence. Affective computing (Picard, 1995) is an interdisciplinary field within psychology, cognitive and computer sciences, and its goal is to " ... give computers the ability to recognize, express, and in some cases, 'have' emotions" (Picard, 1991:1). Computers that are aiming to interact naturally and intelligently with humans require the ability to recognize and express affect (Picard, 1995). Affective computing, therefore, is a new field to explore that can bring us to a more personal and user friendly computing for humans. Although affective computing has a lot of room for growth, it is starting by achieving speech emotion recognition as a first step towards the goal.

However, further research and data on emotional prosody can provide more information about the topic and therefore its advance (Rosis and Grasso, 2000). Apart from this, another interesting fact that should be taken into account can be found in Lee and Nass (2010), who suggest that people have the same social expectations on machines than humans. Moreover, Nass et al., (1994) claims that humans' interactions with machines are social and natural. In this case, machines should understand emotional responses in the same way they are understood by humans. Perhaps this might be one of the reasons that robots with emotions prove to be more believable (for more details see Bates, 1994).

Leaving behind the most obvious reason for emotion sensitive systems, which embrace naturalness, accuracy and effectiveness in speech human machine interaction (Schuller et al., 2004; Cowie et al., 2001), other possible useful applications of speech emotion recognition can be emphasized in the area of medicine, psychology, leisure, customer service (Herm et al., 2008; Lee and Narayanan, 2005), education (Brave et al., 2005; Yildirim et al., 2005) or accident preventing systems. In medicine and psychology, emotion detection can help to analyse the mental state of a patient and find any existing disorder and obtain an extra font of information regarding the case (France et al., 2000). Moreover, emotion recognition systems can be incorporated into on-board car driving systems in order to detect the emotional and mental state of a driver, and avoid the practice of the driver being agitated, stressed or emotionally unstable while driving (Schuller et al., 2004). Emotion recognition can help to improve the quality of the existing systems overall and make the experience more enjoyable and easy moving.

2.5. Corpus collection

Usually, one of the starting problems in the emotion prosody investigation is the data collection. It is hard to decide which is the best option since all three categories for data collection present their own pros and cons. The three corpora categories that are used to investigate emotional speech are:

- acted speech
- spontaneous speech
- elicited (or induced) speech

Acted speech is recorded with professional or not professional actors. Stibbard (2001) and Roberts (2011) question naturalness of the speech by claiming that the actors merely portray their stereotyped view of how emotions should be expressed. This fact might also mean that the actors exaggerate the emotion. Moreover, some authors such as Scherer et al, (2000) comment that this type of data, which has no dialog context, might not be suitable to generalize to human computer interaction scenarios. Other authors (Douglas-Cowie et al., 2007; Esposito and Esposito, 2012) also doubt about the authenticity of the emotions performed by actors. However, this type of corpora has very clear emotional categories, which are very useful to compare and analyse each emotion.

The spontaneous speech might be collected without speakers knowing they are recorded or through TV shows. This type of corpora is the most natural example of speech compared to acted or elicited speech, but its disadvantage is that the collection of the material might suppose ethical or copyright issues (Stibbard, 2001). The categories of emotion obtained from the spontaneous speech cannot be predefined either and posterior manual annotation of emotions is needed.

We can say that elicited speech is a middle way between acted and spontaneous speech. Elicited speech is based on inducing an emotion through particular stimuli, by using movies, pictures or even music. Comparing this category with acted speech, more natural speech is obtained, and only desired emotions are evoked. However, not everybody might react in the same way to the given stimulus and some emotions such as anger raise ethical issues. Some authors claim that induced speech is too mild. However, more complex emotions (found in real speech) can be elicited, which is an important pro comparing it to acted speech. There is also a risk in that speakers might feel inhibited by the recordings and not be expressive enough during the recordings. Studies which have used induced emotional speech are e.g. Iida et al., (1998), McMahon et al., (2003); Batliner et al., (2004); Noroozi et al., (2017).

In general, there is an enormous diversification among the databases from the perspective of languages, emotions, methods of database collection, type and number of speakers, etc. (Koolagudi and Rao, 2012; Scherer, 2003). In particular, most experiments that focus on emotional speech were produced by actors and actresses (for overview see Scherer, 2003).

2.6. Emotion detection and classification

In recent years, new techniques and methods are used to detect emotion through speech, however researchers are still dealing with disagreements regarding corpus collection, methodology and

categories of emotion. Researchers proposed several classification algorithms for the classification of emotions in speech.

Nowadays, different methods and classifiers are used to perform emotion recognition, such as Hidden Markov Model (HMM), Neural Networks (NN), Gauss Mixture Model (GMM), Support Vector Machines (SVM), Bayesian Networks (BNs) or Decision Trees (DT). Combining the classifiers to achieve better results and also comparing classifiers is a common practice. Even when the corpora and its characteristics vary widely in each study, the results do not always differ significantly. This can be illustrated briefly by studies such as Noroozi et al., (2017) who classified 6 emotions with an average recognition rate of 66.28% with DT and Random Forest (RF), and Kwon et al., (2003) reached 70.1% for 4-class speaking style classification using Gaussian SVM, comparing it with other classifiers such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and HMM. The size of the corpora might also also vary considerably in each study. Fifteen utterances were selected for positive and negative emotions using acted corpora, to obtain an average of 83.33% accuracy using the combination of Principal Component Analysis (PCA) and LDA (Hoque et al., 2006). On the other hand, Kao and Lee (2006) achieved 90% accuracy with SVM model to detect 5 emotions (neutral, anger, happiness, sadness and fear), although only 2 sentences were used for each emotion.

According to many studies (Picard, 1995; Murray and Arnott, 1993; Schröder et al., 2001b), pitch and intensity are relevant acoustic parameters for emotion detection, as well as velocity and duration. Many studies have focused on these features among many others. In "Prosody-based classification of emotions in spoken Finnish", Seppännen et al., (2003) employed more than 40 prosodic features with K-Nearest Neighbors algorithm (KNN) classifier to reach between 60% and 80% of average accuracy rate. Through prosodic features, Yu et al., (2001) achieved an accuracy of 73% on four distinct emotions such as anger, happy, sadness, and neutral.

When using a natural speech database, Zhang (2008) reached the accuracy of 76% classifying 4 emotions. An elicited database is used in Batliner et al., (2004), although the elicited data had to be annotated before performing the analysis of 4 emotions (anger, boredom, joy and surprise) and the overall recognition rate reached 69.8%. Several researchers reported that the task of emotion annotation is very complex (Strapparava and Mihalcea, 2007) since the perception of the emotions is variable (Le Tallec, 2010). Annotation can also limit the categories; it is subjective and the categorization is always open to discussion.

Table 1 collects a review of studies mentioned above in order to demonstrate the wide variety of different techniques and methods used to detect and classify emotions. Although this table is by no means an exhaustive review of all the previous studies regarding emotion detection and classification, it still justifies the use of acoustic features to detect emotion through speech. As can be observed in Table 1, the accuracy varies and it depends on various factors such as number and categories of emotion, features, type of data and classifiers. It also highlights the wide use of acted speech corpora. After the review of recent literature on emotion recognition speech, Koolagudi and Rao (2012) concluded that overall, 60% of the collected databases are acted speech. Moreover, some researchers (Tato et al., 2002) consider that more relevant results are obtained when the database is consisted of as close to possible to spontaneous emotion speech samples.

Table 1. An overview of the studies in emotion detection through speech with focus on acoustic features

Authors and year of publication	**Title**	**Data collection method**	**Language**	**Model**	**Emotions**	**Features considered**	**Accuracy rate**
Kao and Lee (2006)	Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language	Acted	Mandarin	SVM	Neutral, anger, happiness, sadness, and fear	Pitch and power based features are extracted from frame, syllable, and word levels	90.0%
Hoque, Yeasin, and Louwerse (2006)	Robust recognition of emotion from speech	Acted (from movies)	English	21 classifiers selected from WEKA toolbox, divided into five stand-alone models (with rules, trees, meta, function and bayes classifiers) and data projection techniques are used such PCA and LDA	Positive and negative	Spectral and prosodic features (features related to fundamental frequency (F0), energy, rhythm, pause and duration)	83.33 %
Kwon, Chan, Hao, Lee (2003)	Emotion recognition by speech signals	Simulated and actual stress (Text-independent SUSAS database) Non-actor speakers, with context action before the emotion (speaker-independent	English	SVM, LDA, QDA and HMM	Angry, bored, happy, neutral, sad	Pitch, log energy, formant, mel-band energies, and mel frequency cepstral co-efficients (MFCCs) as the base features, and added velocity/acceleration of pitch and MFCCs to form feature streams	96.3% for stressed/neutral style classification and 70.1% for 4-class speaking style classification using Gaussian SVM

		AIBO database)					
Batliner et al. (2004)	You stupid tin box children interacting with the Aibo robot: a cross-linguistic emotional speech corpus	Elicited: robot is used to elicit emotions. Emotion annotation	German and English	NN	Anger, boredom, joy, and surprise	Prosodic and 30 part-of-speech features	69.8%
Seppänen et al. (2003)	Prosody-based classification of emotions in spoken Finnish	Acted (professional actors)	Finnish	kNN	Neutral, sad, angry,happy	More than 40 prosodic features	60-80%
Noroozi et al. (2017)	Vocal-based emotion recognition using random forests and decision tree	Elicited (actors)	English	RF and DT	Happiness, fear, sadness, neutral, surprise, and disgust	Pitch, intensity, the first four formants, the first four formants bandwidths, mean autocorrelation, mean noise-to-harmonics ratio and standard deviation	66.28%
Zhang, S. (2008)	Adaptive and optimal classification of speech emotion recognition	Natural speech	Mandarin	SVM	Anger, joy, sadness and neutral	Prosody and voice quality based features	76%
Yu et. al. (2001)	Emotion Detection From Speech To Enrich Multimedia Content	Acted speech: data captured from movies teleplays	English	SVMs, binary classifiers	Anger, happy, sadness, and neutral	Prosodic features	73%

3. Study 1: Production (priming experiment)

To elicit emotion, we invited 20 participants to summarize and give their personal opinion on 36 news reports. Their voice was recorded during the experiment. Each experiment lasted for about half an hour.

3.2. Methodology

The basic design for the first experiment was to obtain emotional data using news reports to evoke happy, neutral and sad emotions from Swedish participants. As was mentioned in the previous chapter, three main different methods exist to obtain vocal expression samples: natural speech, induced speech and acted speech (Scherer, 2003). Taking in mind the advantages and disadvantages of each method, induced speech is used as a method for the first experiment. This method allows us to obtain less affected data; in other words, more natural speech compared to simulated emotions by actors can be obtained.

3.2.1. Speech material

The material consisted of 36 news reports with pictures and 12 definitions or short descriptions from Wikipedia. Whereas news reports were expected to trigger some particular emotion, the definitions were added as a neutral input. Moreover, 3 distractors were distributed between the articles and one extra news article was added at the beginning of the experiment followed by the first distractor. The distractors were added in order to provide a distraction and break the monotony of the experiment, to prevent the speakers from getting too tired and, consequently, less engaged with the news.

Reading Wikipedia definitions or short descriptions reading style as neutral stimuli was chosen for several reasons. As a neutral stimulus, reading a definition presents several advantages over summarization and picture description as in the rest of the categories:

- the same picture can evoke a different reaction in each participant, so choosing a picture as neutral is subjective
- reading style can be a fixed way to force all the participants to have a specific pace and style
- read out Wikipedia articles have very specific length for all the participant
- read out Wikipedia articles will also distract participants from the experiment; participants will not be emotionally involved in describing a particular event and picture.
- reading style is easy to replicate

Based on these factors, read out Wikipedia articles is the most appropriate as neutral stimuli. The emotion can be controlled better and the reading style can be used as a baseline for the experiment.

Different types of news were selected for this experiment. Common Swedish newspapers were used (Dagens Nyheter, Aftonbladet, Svenska Dagbladet, Expressen, Metro, GP), in order to preserve the particular journalistic style and vocabulary. A wide scanning of articles that implied sad or happy news was performed and best matches were selected. Different newspapers were used for the selection in order to preserve a hegemonic style and type of the news reports.

Different newspapers will also allow us to cover more topics and countries, in order to avoid the repetition of the news from the same country. The image of each news event is also included, since the news reports that present emotional pictures evoke more feelings in the participants (e.g. Bucy, 2003; Wirth and Böcking, 2003).

In general, the reports chosen for the stimuli were selected according to the following three criteria: (1) The reports had to be "typical" for daily Swedish newspapers (regarding the presentation and language); (2) Only 3-4 sentences had to be used with the key information, usually the lead paragraph was selected; (3) The affected country could appear only once for each emotional category. Table 2 represents the type of news reports that were selected. A few news reports that affect animals were also included, as Unz, Schwab and Winterhoff-Spurk (2008) found that news about animals lead to more intense feelings, comparing to news that affect humans.

Two pilot experiments were conducted before in order to measure the suitable length for the experiment and to verify if all the articles were easy to understand. Some of the articles were replaced based on the output of the speakers and on the posterior analysis of the data. The number of articles was limited to 12 news reports for each emotion in order to have an experiment with a length of 30 minutes only. This limit has been imposed to prevent the speakers from getting too tired and, consequently, less engaged with the news.

Table 2. Speech stimulus.

Quantity	Type of news	Details
12	Neutral	Read out Wikipedia articles about water, density, intonation, graphite, free radical, etc.
12	Happy	Free summarisation and discussion of events such as volunteering, saving somebody's life, successful occurrences all over the world (Denmark, Sweden, Thailand, etc).
12	Sad (North)	Free summarisation and discussion of events such as natural disasters, accidents, terrorist attacks in Sweden, England, USA, Greece, Spain, France, Norway, Italy, Russia, Finland, Belgium and Australia.
12	Sad (South)	Free summarisation and discussion of events such as natural disasters, accidents, terrorist attacks in Somalia, Syria, Guatemala, Argentina, India, China, Mexico, Egypt, Haiti, Afghanistan, Sudan and Venezuela.

The articles were randomized for each participant to ensure that no pattern existed. Pseudo-randomization was used in this experiment, since it was important that the same categories of stimuli were not next to each other and interspersed with distractors and neutral stimuli. This pseudo-randomization creates an illusion of more heterogeneous stimuli and avoids participants getting too tired of the same category of stimuli that might appear again and again otherwise.

3.2.2. Speakers

Twenty typical native Swedish speakers, 10 female and 10 male, participated in the production experiment. They were 22-34 years old (M=27.6) and regular readers of print or digital media.

The speakers reported no hearing or articulatory problems. Although it was not an initial requirement for the participants, it turned out that almost all the participants had an university education and all of them were studying and/or working in Göteborg at the time of the recording. Most speakers spoke Standard Swedish (rikssvenska). More details regarding the speakers' dialect can be found in Table 3.

Table 3. Participants and their dialects.

Number of participants	**Dialect**
15	Neutral (*rikssvenska*)
2	Göteborg dialect
2	Southern dialect
1	Småländska dialect

3.2.3. Procedure

The experiment took place in a quiet room at the University of Göteborg, where the recordings were least affected by the background noise. While conducting the experiment, particular criteria were followed such as the degree of participation of the researchers, the position and location of the recording system, participant and the researchers. The procedure was performed in a very similar way to not interfere in the data collection. Zoom H2n was used as a recorder during the experiments. All the speakers were unaware of the purpose of the experiment. Prior consent was given by the participants to be recorded and they had been informed that we were interested in studying different linguistic phenomena related to vowels and consonants. Once the experiment was finished, the real aim of the experiment was revealed.

A Swedish assistant gave the instructions for the experiment in Swedish to the subjects, in order to preserve the most natural environment for the participants. The participants were asked to briefly summarize what the article was about to the assistant and explain to him how they felt about the particular news article. In the case of Wikipedia definitions, their task was only to read the text aloud. Each stimulus was presented in one slide in Microsoft PowerPoint 2010. The speakers manually changed the slide after each recording. Apart from the assistant, whose main task was to listen and assist the participants during the experiment, the main investigator was present during the recording phase too, and the tasks were primarily auxiliary: to start the microphone, to correct speaker's distance from the microphone, etc. Figure 5 represents a typical experimental setting during the recordings.

The open source software Praat 6.0.14 was used for the acoustic analysis (Boersma and Weenink, 2016). The first stimulus was considered preparatory and it was removed from the data. The distractors were removed from the data too. Using a segmentation script (Lennes, 2011), each file of the participant was divided to obtain one separate file for each news article, in order to facilitate the posterior analysis. This segmentation script used pauses to divide these files and the length of the pauses was adjusted manually, since each speaker had different speeds and the pauses between each news report varied. Since emotion and more specifically pitch accent has a global effect on the utterance, there is no need to split the utterances, but in our case we find it useful to have shorter utterances to facilitate the analysis and also to obtain more data samples.

Figure 5: A typical experimental setting during the onset of the experiment. A laptop computer is used to show news reports to the speakers. The laptop is located a little bit to the right of the participant, in order to facilitate the eye contact and to have more natural arrangement with the assistant, who is sitting in front but a little bit to the right. The participant will be more emotionally engaged with a news report if he/she is talking about it to someone.

For this reason, each of the obtained files was also divided into three parts. All the parts were not totally equal in duration since the goal was to have complete words without splitting them in two, and therefore once Praat TextGrid boundaries were automatically applied, manual adjustment was needed to review the splitting. For the first experiment, 2880 utterances (20 speakers, 48 main utterances for each news/article, 3 parts) were obtained.

3.2.4. Measurements

All the recordings were analysed using Praat software with a Praat script called ProsodyPro (Xu, 2005-2011), a script for large-scale systematic analysis of continuous prosodic events. The script was run on the recordings. In total, 11 prosodic features were extracted, listed in Table 4. Most of the extracted features are focused on fundamental frequency, which is the lowest frequency of a periodic waveform that provides the rhythmic and tonal properties of the speech. It is commonly accepted that the fundamental frequency (*F0*) contour is a key feature to detect and differentiate emotional information (Murray and Arnott, 1993; Hoque et al., 2006; Tao and Kang, 2005; Cowie and Cornelius, 2003; Banziger and Scherer, 2005). Other features such as duration of the utterance, intensity and velocity are also included.

Table 4. List of prosodic features.

Number	Feature name	Explanation
1	Max *F0*	Maximum fundamental frequency
2	Min *F0*	Minimum fundamental frequency
3	Mean *F0*	Mean fundamental frequency
4	Excursion size	Excursion size is a property of fundamental frequency (*F0*) variation
5	Final *F0*	Final fundamental frequency
6	Mean intensity	Average number of loudness of the intervention
7	Duration	Duration of the utterance
8	Max velocity	The maximum rate of velocity of the intervention
9	Final velocity	Final velocity of the intervention that is detected
10	Max *F0* loc ms	Time of the *F0* peak relative to the onset of an interval in milliseconds
11	Max *F0* loc ratio	Relative location of the *F0* peak as a proportion to the duration of the interval

3.2.5. Statistical Analysis

Linear Mixed-Effects Model is a type of regression model that takes into consideration variation that is not generalizable to the independent variables. This variation might include variation across different speakers or utterances (Baayen et al., 2008).

Linear mixed effect model helps to solve several disadvantages that traditional approaches to random effects modelling present. Methods for estimating LMMs have addressed a solution for several existing problems (Baayen et al., 2008):

- deficiencies in statistical power when dealing with repeated observations
- the lack of a flexible method of dealing with missing data
- disparate methods for treating continuous and categorical responses

Unlike the univariate ANOVA or ordinary least squares regression, it can account for multiple random effects at a time, and is not just limited to a random effect of subject.

A mixed-effects model consists of two parts, fixed effects and random effects (Baayen, 2008):

- Fixed effects: factors with repeatable levels that we expect will have an effect on the dependent variable. In our case, we are interested in making conclusions about how emotion impacts the prosodic events. So emotion is a fixed effect and prosodic events are the dependent variables.
- Random effects: factors with levels randomly sampled from a much larger population. A lot of the time we are not specifically interested in their impact on the response variable. Additionally, the data for our random effect is just a sample of all the possibilities. We have a dependent variable, the prosodic events and we are attempting to explain part of the variation in prosodic events through fitting emotions as a fixed effect. But the dependent variable has some residual variation (i.e. unexplained variation) associated

with speaker, utterance and also the position (as a reminder: the utterance was divided into 3 parts). By using random effects, we are modelling that unexplained variation through variance.

While we still want to know whether there is an association between emotion and prosodic events, we want to know if that association exists after controlling for the variation in speakers, position and utterance. We will fit the random effect using (1|variableName), where "1" means the random intercept and in our case it will vary among variables such as speakers, position and utterance:

Prosodic events ~ emotion + (1|Utterance) + (1|Position) + (1|Speaker)

R (version 3.4.1) was used for the statistical analysis (R Core Team, 2012) and lme4 (Bates et al., 2012) to perform a linear mixed effects analysis for each one of the response variables of prosodic events obtained from collected corpus. Nonlinear transformation of the response (log transformation) on max F0, final F0 and duration was performed for normality.

3.3. Results

The linear mixed effects models for prosodic events are shown in Table 5. There are several details in this tables that should be clarified first.

As it can be observed the happy emotion in Table 5 for Mean *F0* is not shown in the list; this is called the reference level or intercept, which means that all the other are referenced back to it. In other words, the value of the intercept is the happy emotion. The intercept is the predicted value of the dependent variable (prosodic events) when all the independent variables (emotions) are 0. To calculate the means for the other groups we need to sum the value of the **estimate** reference level with the slopes. For example neutral is 167.39 + 4.29 = 171.68. To sum up, the estimate for "(Intercept)" is the estimate for the "happy" category, and the estimate for "neutral" is the estimate for the difference between happy and neutral categories. In other words: the mean fundamental frequency is higher in neutral emotion than in happy emotion, by about 4.29 Hz.

Then, the **standard error** (SE) associated with the slope is given, followed the **number of degree of freedom** (df), which represents the number of values in the final calculation of a statistic that are free to vary. **T-value** is simply the estimate divided by the standard error and it is also used to compute p-values. The last column of the tables represents the **p-value (Pr(>|t)).** The p-values correspond to tests if each estimate coefficient is non-zero (Winter, 2013). Since the estimate coefficient is a pairwise comparison of each of the categories with the estimate of the intercept, the p-value shows the significance of the results in relation to the reference level. Meaning for example that neutral is significantly different from happy (reference level or intercept) and the p-value is 0.03. In general, the difference is significant when p-value is below 0.05. In general, the difference is significant when p-value is below 0.05. Note that three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis, which allows us to conclude that there is a relation between emotion and particular prosodic event. Clearly, since the p-value of intercept is compared to the 0, this value is not very interesting. The important values are those of the other categories with respect to the intercept.

Table 5. Results of the Linear Mixed Models with happy as intercept.

		Estimate	SE	df	t value	Pr(>|t|)
Mean *F0*	**Intercept**	**167.35**	**8.87**	**20.36**	**18.88**	**2.26e-14 *****
	Neutral	4.29	1.92	44.00	2.23	0.03 *
	Sad-north	-4.20	1.92	44.00	-2.20	0.03 *
	Sad-south	-3.43	1.92	44.00	-1.79	0.08 .
Min *F0*	**Intercept**	**88.82**	**2.53**	**24.16**	**35.11**	**<2e-16 *****
	Neutral	-1.18	1.11	43.99	-1.06	0.30
	Sad-north	0.46	1.11	43.99	0.41	0.68
	Sad-south	-0.79	1.11	43.99	-0.71	0.48
Max *F0*	**Intercept**	**5.73**	**0.06**	**18.36**	**102.02**	**<2e-16 *****
	Neutral	0.08	0.03	43.98	2.46	0.02 *
	Sad-north	-0.03	0.03	43.98	-0.77	0.45
	Sad-south	0.02	0.03	43.98	0.54	0.60
Final *F0*	**Intercept**	**5.08**	**0.06**	**21.30**	**75.90**	**<2e-16 *****
	Neutral	0.01	0.02	2855.00	0.61	0.54
	Sad-north	-3.53e-02	0.02	2855.00	-2.12	0.03 *
	Sad-south	-3.58e-02	0.02	2855.00	-2.15	0.03 *
Mean intensity	**Intercept**	**45.60**	**1.23**	**9.36**	**37.01**	**1.83e-11 *****
	Neutral	3.23	0.22	43.96	14.73	< 2e-16 ***
	Sad-north	-0.70	0.22	43.96	-3.17	2.74e-03 **
	Sad-south	-0.89	0.22	43.96	-4.06	1.94e-04 ***
Duration	**Intercept**	**8.78**	**0.07**	**10.93**	**132.73**	**< 2e-16 *****
	Neutral	-0.27	0.04	44.00	-6.55	5.22e-08 ***
	Sad-north	0.10	0.04	44.00	2.34	0.03 *
	Sad-south	0.13	0.04	43.97	3.28	2.2e-03 **
Time of the *F0* peak	**Intercept**	**3359.70**	**223.07**	**40.66**	**15.06**	**< 2e-16 *****
	Neutral	-858.08	204.04	43.99	-4.21	0.0001 ***
	Sad-north	489.51	204.04	43.99	2.40	0.02 *
	Sad-south	476.20	204.04	43.99	2.33	0.03 *
Location of the *F0* peak	**Intercept**	**0.49**	**0.02**	**4.64**	**22.20**	**6.81e-06 *****
	Neutral	-0.01	0.02	44.06	-0.48	0.64
	Sad-north	0.01	0.02	44.05	0.40	0.70
	Sad-south	-0.01	0.02	43.06	-0.37	0.71
Max velocity	**Intercept**	**-1.84**	**9.73**	**252.20**	**-0.19**	**0.85**
	Neutral	12.27	13.59	2858.40	0.90	0.37
	Sad-north	2.00	13.59	2858.40	0.15	0.88
	Sad-south	-6.53	13.59	2858.40	-0.48	0.63
Final velocity	**Intercept**	**0.96**	**0.89**	**11.79**	**1.07**	**0.31**
	Neutral	-0.55	0.95	44.02	-0.58	0.56
	Sad-north	-0.57	0.95	44.01	-0.60	0.55
	Sad-south	0.20	0.95	44.02	0.25	0.83

"***" p < 0.001, "**" p <0.01, "*" p <0.05, "." p < 0.08

Fundamental frequency (mean F0, min F0, max F0, final F0). In case of mean fundamental frequency that we can observe in Table 5, the p-values of the neutral and sad-north show the significance of the results in relation to the happy category. The estimate coefficient of the neutral category is 4.29Hz higher with respect to the happy category, and sad-north category is 4.20 Hz lower with respect to the happy category. Sad-south category is not a significant predictor of mean *F0*. For max *F0* the neutral is significantly different from happy and for final *F0* the significance only between happy and sad-north, and happy and sad-south can be detected. Overall, we can see that neutral can be differentiated from happy category through mean *F0* and max *F0*. In case of sad-north, the significance of the results in relation to the reference level can be seen for mean *F0* and final *F0*. Regarding sad-south, the significant p-value can be only observed for final *F0*. However, it is important to highlight that the estimate coefficient of sad-south category is close to sad-north category, with the estimate coefficient of 5.0442Hz for sad-south and 5.0447Hz for sad-north with respect to happy.

Mean intensity. The mean intensity appears to be a very representative feature since the p-values of neutral, sad-north and sad-south categories show the significance of the results in relation to the happy category. The coefficient estimate for happy category is 45.60dB, higher for neutral category with 48.83dB and lower with respect to the happy category for sad-north (44.9dB) and sad-south (44.71dB) categories. It is important to notice that although sad-north and sad-south present significant values with respect to happy category, the estimate coefficient of these two categories is very similar. More information and discussion about this case will be provided based on Table 6.

Duration. For duration, the findings show that neutral, sad-north and sad-south are significantly different from happy. Neutral presents the shortest estimate of duration with respect to the happy category, with 8.51ms, which can be explained by users reading the neutral stimulus. The mean estimate of duration for sad-north and sad-south is higher in comparison with happy.

Time and location of the F0 peak. The time of the *F0* peak presents significant results for neutral, sad-north and sad-south categories with respect to happy. No relationship between emotion and the location of the *F0* peak was found.

Velocity. Emotion does not present any significant effect on the velocity of the speech. The findings did not show significant results for maximum and final velocity, since the p-value is too high to conclude that there is any relation between the given emotions.

In the previous results (Table 5), with happy as the reference level (intercept) we could observe that the coefficient estimates for sad-south and sad-north were very similar. Therefore another linear mixed effects model analysis was performed to detect any existing difference between these two categories, with sad-north category as the reference level. This analysis can be observed in Table 6. Since the relation between happy and sad-north categories has been discussed in Table 5, our main focus in this case is the pairwise comparison between sad-north and sad-south, and sad-north and neutral.

Closer inspections of Table 6 for each prosodic event show that sad-south category does not present any significance of the results in relation to the sad-north category. This means that sad-south is not significantly different from sad-north. This explains the similar estimate values for

these categories in Table 5 for prosodic events that contained a significant value of sad-south and sad-north respect to happy, such as final *F0*, mean intensity, duration, time of the *F0* peak. For all these prosodic events, sad-south category presented a significant p-value with respect to happy emotion only because sad-south is very similar to sad-north category.

Regarding the pairwise comparison between sad-north and neutral, the following prosodic events show a significant difference between sad-north and neutral: mean *F0*, max *F0*, final *F0*, mean intensity, duration and time of the *F0* peak.

Table 7 presents the results of the linear mixed effect model analysis, with sad-south category as the reference level. Although, it was proved in the previous tables that sad-north and sad-south present very similar values, the comparison between sad-south and neutral needed to be included, in order to complete the pairwise comparison for all the categories. The following prosodic events show a significant difference between sad-south and neutral: max *F0*, final *F0*, mean intensity, duration and time of the *F0* peak.

Table 6. Results of the Linear Mixed Models with sad-north as intercept.

		Estimate	SE	df	t value	Pr(>|t|)
Mean *F0*	**Intercept**	**163.15**	**8.87**	**20.36**	**18.40**	**3.71e-14 *** **
	Happy	4.19	1.92	44.00	2.19	0.03 *
	Neutral	8.48	1.92	44.00	4.43	6.17e-05 ***
	Sad-south	0.77	1.92	44.00	0.40	0.69
Min *F0*	**Intercept**	**89.28**	**2.53**	**24.16**	**35.29**	**<2e-16 *** **
	Happy	-0.46	1.11	43.99	-0.42	0.68
	Neutral	-1.64	1.11	43.99	-1.48	0.15
	Sad-south	-1.26	1.11	43.99	-1.13	0.26
Max *F0*	**Intercept**	**5.70**	**0.06**	**18.36**	**101.57**	**<2e-16 *** **
	Happy	0.03	0.03	43.98	0.77	0.45
	Neutral	0.11	0.03	43.98	3.22	2.39e-03 **
	Sad-south	0.04	0.03	43.98	1.30	0.20
Final *F0*	**Intercept**	**5.05**	**0.06**	**21.30**	**75.37**	**<2e-16 *** **
	Happy	0.04	0.02	2855.00	2.12	0.03 *
	Neutral	0.05	0.02	2855.00	2.73	6.30e-03 **
	Sad-south	-4.628e-04	0.02	2855.00	-0.02	0.98
Mean intensity	**Intercept**	**44.89**	**1.23**	**9.36**	**36.45**	**2.09e-11 *** **
	Happy	0.70	0.22	43.96	3.17	2.74e-03 **
	Neutral	3.93	0.22	43.96	17.91	< 2e-16 ***
	Sad-south	-0.20	0.22	43.96	-0.89	0.38
Duration	**Intercept**	**8.88**	**0.07**	**10.97**	**134.21**	**< 2e-16 *** **
	Happy	-0.10	0.04	43.98	-2.32	0.02 *
	Neutral	-0.37	0.04	43.98	-8.88	2.29e-11 ***
	Sad-south	0.04	0.04	43.98	0.93	0.36
Time of the *F0* peak	**Intercept**	**3849.21**	**223.07**	**40.66**	**17.26**	**< 2e-16 *** **
	Happy	-489.51	204.04	43.99	-2.40	0.02 *
	Neutral	-1347.59	204.04	43.99	-6.61	4.36e-08 ***
	Sad-south	-13.32	204.04	43.99	-0.07	0.95
Location of the *F0* peak	**Intercept**	**0.49**	**0.02**	**4.64**	**22.53**	**6.38e-06 *** **
	Happy	-7.42e-03	0.02	44.05	-0.39	0.70
	Neutral	-1.65e-02	0.02	44.05	-0.87	0.39
	Sad-south	-1.45e-02	0.02	43.05	-0.76	0.45
Max velocity	**Intercept**	**0.17**	**9.73**	**252.20**	**0.17**	**0.99**
	Happy	-2.00	13.59	2858.40	-0.15	0.88
	Neutral	10.27	13.59	2858.40	0.76	0.45
	Sad-south	-8.53	13.59	2858.40	-0.63	0.53
Final velocity	**Intercept**	**0.38**	**0.89**	**11.79**	**0.43**	**0.68**
	Happy	0.57	0.95	44.01	0.60	0.55
	Neutral	0.02	0.95	44.01	0.03	0.98
	Sad-south	0.78	0.95	44.01	0.82	0.42

"***" p < 0.001, "**" p <0.01, "*" p <0.05, "." p < 0.08

Table 7. Results of the Linear Mixed Models with sad-south as intercept.

		Estimate	SE	df	t value	Pr(>|t|)
Mean *F0*	**Intercept**	**163.92**	**8.87**	**20.36**	**18.48**	**3.38e-14 *** **
	Happy	3.43	1.92	44.00	1.79	0.08 .
	Neutral	7.72	1.92	44.00	4.03	2.18e-04 ***
	Sad-north	-0.77	1.92	44.00	-0.40	0.69
Min *F0*	**Intercept**	**88.03**	**2.53**	**24.16**	**35.79**	**<2e-16 *** **
	Happy	0.79	1.11	43.99	0.72	0.48
	Neutral	-0.38	1.11	43.99	-0.34	0.73
	Sad-north	1.26	1.11	43.99	1.13	0.26
Max *F0*	**Intercept**	**5.75**	**0.06**	**18.36**	**102.34**	**<2e-16 *** **
	Happy	-0.02	0.03	43.98	-0.54	0.59
	Neutral	0.06	0.03	43.98	1.92	0.06 .
	Sad-north	-0.04	0.03	43.98	-1.30	0.20
Final *F0*	**Intercept**	**5.05**	**0.07**	**21.30**	**75.37**	**<2e-16 *** **
	Happy	3.58e-02	0.02	2855.00	2.15	0.03 *
	Neutral	4.60e-02	0.02	2855.00	2.76	5.79e-03 **
	Sad-north	4.63e-04	0.02	2855.00	0.03	0.98
Mean intensity	**Intercept**	**44.70**	**1.23**	**9.36**	**36.29**	**2.17e-11 *** **
	Happy	0.89	0.22	43.96	4.07	1.94e-04 ***
	Neutral	4.912	0.22	43.96	18.80	< 2e-16 ***
	Sad-north	0.20	0.22	43.96	0.89	0.38
Duration	**Intercept**	**8.92**	**0.07**	**10.97**	**134.80**	**< 2e-16 *** **
	Happy	-0.13	0.04	43.98	-3.26	2.20e-03 **
	Neutral	-0.41	0.04	43.98	-9.81	1.21e-12 ***
	Sad-north	-0.04	0.04	43.98	-0.93	0.36
Time of the *F0* peak	**Intercept**	**3835.89**	**223.07**	**40.66**	**17.20**	**< 2e-16 *** **
	Happy	-476.20	204.04	43.99	-2.33	0.02 *
	Neutral	-1334.27	204.04	43.99	-6.54	5.44e-08 ***
	Sad-north	13.32	204.04	43.99	0.07	0.95
Location of the *F0* peak	**Intercept**	**0.48**	**0.02**	**4.64**	**21.87**	**7.33e-06 *** **
	Happy	7.08e-03	0.02	44.06	0.37	0.71
	Neutral	-1.98e-03	0.02	44.06	-0.10	0.92
	Sad-north	1.45e-02	0.02	43.05	0.76	0.45
Max velocity	**Intercept**	**-8.36**	**9.73**	**252.10**	**-0.86**	**0.40**
	Happy	6.53	13.59	2834.30	0.48	0.63
	Neutral	18.80	13.59	2834.30	1.38	0.17
	Sad-north	8.53	13.59	2834.30	0.63	0.53
Final velocity	**Intercept**	**1.16**	**0.89**	**11.79**	**1.30**	**0.22**
	Happy	-0.21	0.95	44.02	-0.22	0.83
	Neutral	-0.76	0.95	44.02	-0.80	0.43
	Sad-north	-0.78	0.95	44.01	-0.82	0.42

"***" $p < 0.001$, "**" $p < 0.01$, "*" $p < 0.05$, "." $p < 0.08$

3.4. Discussion

We hypothesized that news stimuli will be suitable for corpus collection to produce emotional responses from speakers and it will be possible to identify neutral, happy and sad emotions. From the linear mixed effect models, the results suggest that various analysed features are clearly useful to detect happy, neutral and sad emotions. Table 8 summarizes the pairwise comparison between each category to display which prosodic features present significant results for these pairs.

Table 8. Significant results (based on the p-value) of pairwise comparison for each prosodic event, extracted from Tables 5, 6 and 7. Categories: happy (H), neutral (N), sad-north(SN) and sad-south (SS)

Prosodic event	**H-N**	**H-SN**	**H-SS**	**SN-N**	**SS-N**	**SN-SS**
Mean F0	✓	✓		✓	✓	
Min F0						
Max F0	✓			✓		
Final F0		✓	✓	✓	✓	
Mean intensity	✓	✓	✓	✓	✓	
Duration	✓	✓	✓	✓	✓	
Time of the *F0* peak	✓	✓	✓	✓	✓	
Location of the *F0* peak						
Max velocity						
Final velocity						

The results demonstrated that mean intensity and duration were the clearest features to distinguish happy, neutral and sad-north categories. These features are proved to be useful for emotional speech classification according to other researchers (Tao and Kang, 2005). Time of the *F0* peak is also a significant predictor to distinguish between happy, neutral and sad-north categories.

The features related to fundamental frequency (maximum, minimum, final, mean) are also important to demarcate and distinguish the emotions. In case of the final F0, significant values were found between happy and sad-north, happy and sad-south, sad-north and neutral, and sad-south and neutral. Max F0 presents significant difference between happy and neutral, and sad-north and neutral; and mean F0, between happy and neutral, happy and sad-north, sad-north and neutral, and sad-south and neutral. Similar to these results, Tao and Kang (2005) report that the most important features for emotion classification are *F0* features (mean, maximum, range) in their study where 5 emotions were analysed (neutral, fear, sad, angry, happy) on acted speech data. However, in this study, min F0 does not present any significant difference between categories.

Also, regarding the Global North and South news, the results show that sad-south category does not present any significance of the results in relation to the sad-north category. The results prove that sad-north and sad-south categories have very similar values and this fact explains why sad-south category presented significant p-values with respect to happy and neutral categories when sad-north category also had significant values.

It could be debated that all differences related to neutral category might be caused by the reading style versus summarization and personal opinion. However, from the statistical analysis results

we can observe that neutral category follows a very similar pattern compared to the rest of the categories, since the most representative prosodic events that can be used to distinguish between happy and sad are the same for neutral category. Mean intensity, duration and time of the *F0* peak were the clearest features to distinguish all the categories: happy, neutral and sad. Nonetheless, we do not the reject the idea that the difference between neutral and the rest of the categories might be also (partly or not) to the reading style of neutral emotion.

Velocities, both maximum and final, and the location of the *F0* peak do not show any clear effect between the emotions. These findings were very surprising since several researchers have shown that different emotional dispositions of a person are clearly expressed in his/her speaking rate (Philippou-Hübner et al., 2012). Koolagudi and Krothappalli, (2011) use speech rate as the first stage of classification of emotion, by categorizing the emotions into three main groups: fast, normal and slow. On the other hand, the perception experiments such as Breitenstein et al., (2001) proved that German and American listeners associated slow rate alternated stimuli with sad vocal emotion and fast rate was classified most usually as angry, frightened or neutral.

With the help of statistical analysis we could see patterns between neutral, sad and happy emotions exist. This analysis presents a general overview of the data and the next step is to train a machine algorithm in order to see if these emotions can be also classified based on the given data.

4. Study 2: Classification

In this section, we describe the procedure and results of training a classification algorithm on the obtained data. We first describe the method in detail and then present the classification results.

4.1. Methodology

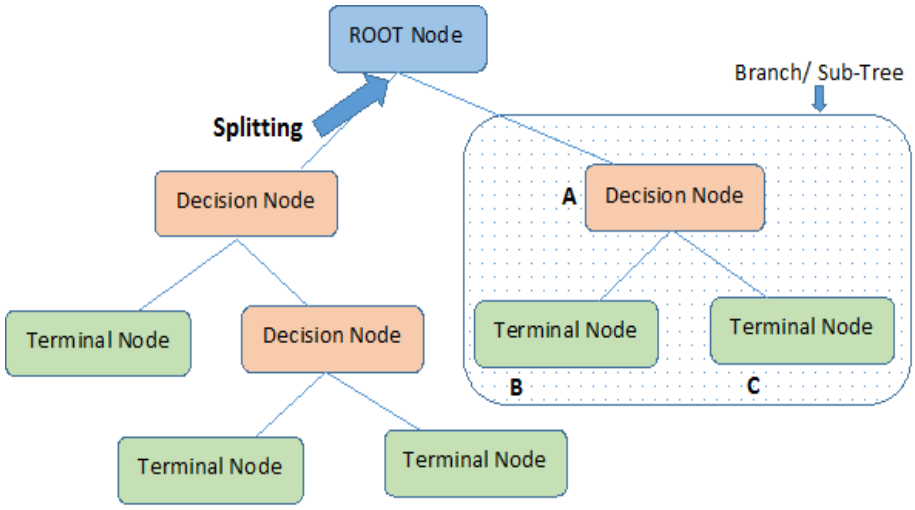
The same features were employed from the data set collected in Experiment 1. Decision Tree (DT) is the classification model algorithm used to classify happy, neutral and sad (South and North) emotions.

4.1.1. Machine Learning Applications

In our research, the main aim was to discover if the methodology of the corpus collection through news reports is suitable for emotion detection and recognition, and in order to understand its effectiveness, a classification method is needed. One of the reasons that the DT classification is chosen among others is because this classification method is robust and easily applicable for categorical values: values that have categories such as female/male, sad/happy/neutral, etc. The classification tree model is also easily interpretable and is capable of managing missing data and many predictors (Kuhn and Johnson, 2013). When DT is used as an acoustic model, it offers additional advantages opposed to generative models, such as Naive Bayes, HMM or GMM because it makes no assumptions about the distribution of underlying data (Akamine and Ajmera, 2012). Akamine and Ajmera (2012), for instance, explored these advantages through experiments on gender and context-dependent acoustic models and found that DT-based models are more compact and effective to GMM-based acoustic models.

This study employs C5.0 Decision Tree model, developed by Ross Quinlan (1993). Several studies have showed that the C5.0 Decision tree model outperforms the others when working with acoustical data. Themistocleous (2017) provided a classification model of two modern Greek dialects (Athenian Greek and Cypriot Greek) using vowel acoustic parameters. In his study, the comparison between LDA, FDA, and C5.0 was investigated, with the latter providing the highest classification accuracy. Vieru et al., (2011) discovered the C4.5 algorithm (an earlier iteration of C5.0) to be the most effective to characterize and identify foreign accents in French, employing acoustic parameters such as duration and voicing for consonants, the first two formant values for vowels, word-final schwa-related prosodic features, etc. In the study, overall 20 classification algorithms were considered (through the Weka data mining software (Witten and Frank, 2005)), such as BNs, Logistic Regression Models (LR), Multilayer Perceptrons (MLP), SVM and RF. To the best of our knowledge, this is the first study that employs C5.0 only for emotion classification.

DT is a supervised machine learning algorithm and unlike linear models, DT can deal with non-linear relationships too. Usually, DT are mostly used in classification problems. Figure 3 represents the functioning of the learning algorithm, which classifies the data, splitting it into branches for different values of predictors, until the terminal node or the final decision is obtained.



Note:- A is parent node of B and C.

Figure 3: representation of how DT performs the classification (from Analytics Vidhya Content Team, 2016)

Following this architecture, the DT will take the given features as input and predict events based on these features, forming the structure of a tree. Using C5.0, the output obtained in the form of the tree is helpful to understand the acoustic predictors and their interactions in speech.

4.1.2. Decision Tree classification

To be able to understand how the data is classified using DT, the Figure 6 shows one small branch or subtree where the features are split and the percentage of one particular emotion is given in the terminal node. In case of this image, the amount of data was reduced to 20 samples. More data creates a more compact and clearer DT.

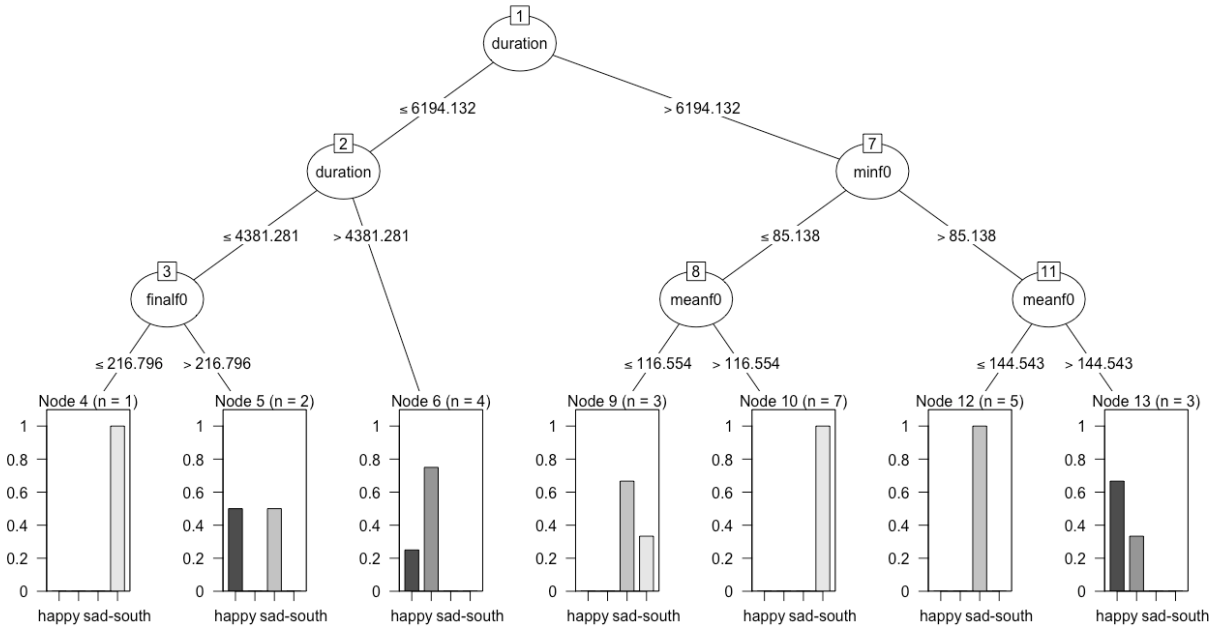


Figure 6. One branch of the DT.

The C5.0 DT algorithm (Kuhn et al., 2015) and caret package (Kuhn, 2016) were used to train the classifier with 2880 samples with 10 predictors and 4 classes ("happy", "neutral", "sad-north", "sad-south"). Although, during the training of the model, it was decided to merge sad-north and sad-south classes at the end, due to the findings in the first experiment, where no significance was found between these two categories.

In order to improve the performance of the model, we attempted to adjust the model options, which is also called parameter tuning. Scaling data makes a significant difference in the performance of the classifier, which helps to increase the accuracy. Adaptive boosting is an incorporated feature of C5.0. The role of boosting is to generate several classifiers (in the form of Decision Trees in this case) instead of just one, and when a sequential classifier is obtained, more focus is given to the problematic cases to get them right and so on. When a new case is to be classified, each classifier gives a vote for its predicted class and the votes are counted to determine the final class (Freund and Schapire, 1999). Overall, boosting improved the performance of the classifier. The predetermined number of iterations or trials was set up to 10 trials. Each line represented one trial until the last (10th) classifier and its error percentage rate is obtained. Some of the classifiers have high error rate. Eventually, when the trees are combined by voting, the final predictions have a lower error rate.

Ideally when training the classifier, the data should be split into training and testing samples. However, if the number of samples is not large, a test set should be avoided, since having all the samples is desirable to train the model. Plus, in the case of small data, the testing set might be too small to obtain reasonable judgements (Kuhn, 2016). In the book “Applied Predictive Modeling” (Kuhn and Johnson, 2013), the authors proposed to use resampling techniques such as cross validation to overcome the problem of small data. Since resampling techniques evaluate many alternative versions of the data, the performance estimates are better than with a single test set. The collected data in this research was not big enough to be divided into training and testing sets,

and therefore a resampling technique was used to estimate the performance of the model. In k-fold cross-validation, the data is randomly divided into k equal size subsamples. One of these subsamples is kept as the validation data to test the model. The rest of the subsamples are used as training data. The cross-validation process is repeated k times, where each of the subsamples is used as the validation data. Finally, the final estimation is given combining all the results obtained from testing the model. In this study, 10-fold cross-validation is used.

4.2. Results

Table 9 shows the confusion matrix of the results obtained from C5.0 DT.

Table 9. Confusion matrix of the classifier converted into percentages.

Class	Happy	Neutral	Sad
Happy	**63.43**	32.92	3.65
Neutral	34.16	**60.63**	5.21
Sad	1.04	2.81	**96.15**

The model has a classification accuracy of 70.88% as an average accuracy tested on validation data from 10-fold cross-validation (mean SD=2.48%, mean kappa 0.56%, mean kappa SD=0.04%). In Table 10, we can notice that attributes such as final velocity, relative location of the *F0* peak, time of the *F0* peak and mean intensity have a high percentage and therefore present a crucial juncture for the classifier. Final velocity and the relative location of the *F0* peak are used on full (100%) capacity. The attributes such as time of *F0* peak and mean intensity present important usage features for the classifiers, with 90% and 83.09% of attribute usage respectively. Final *F0* is used only 63% and the rest of the attributes present low attribute usage below 37%.

Table 10. Attribute usage.

Usage(%)	**Feature**
100.00%	Final velocity
100.00%	Relative location of the *F0* peak
90.00%	Time of the *F0* peak
83.09%	Mean intensity
66.01%	Final *F0*
36.35%	Duration
34.93%	Max *F0*
31.94%	Max velocity
30.28%	Excursion size
9.72%	Mean *F0*

4.3. Discussion

With a main accuracy of 70.88%, tested on 10-fold cross-validation for the performance of the classification model, we can conclude that sad emotion is the most clearly detected with 96.15%, followed by happy with 63.43% and neutral with 60.63 % being the least accurate classified emotion. An interesting observation is that the neutral and happy emotions are misclassified more regularly between each other. This misclassification might be due to the closer position of neutral to happy emotion than sad.

All the acoustic variables that were used for the classification contribute to the overall classification of the emotions. Specifically, the machine learning algorithm C5.0 employed in this study demonstrated that final velocity, relative location and time of the *F0* peak, mean intensity and final *F0* play a significant role in emotion classification. The attribute usage of features such as final velocity, relative location and time of the *F0* peak or mean intensity is more than 83% in the classifier. Based on previous studies, it was expected that *F0* related features and intensity would play a crucial role to detect happy, sad and neutral emotions. As outlined by Nwe et al., (2003) and Murray and Arnott (1993), angry and happy emotions have both ascending pitch contours, increased average pitch, raised intensity, increased speech rate and higher intensity compared to neutral. Sad emotions, on the other hand, present low pitch contours, slightly slower average pitch, lower intensity, slightly slower pitch rate and lower intensity. Therefore these changes are perceptible to the system very notoriously.

The classification of positive and negative emotions was also investigated by Hoque et al., (2006) based on spectral and prosodic features (features related to fundamental frequency (*F0*), energy, rhythm, pause and duration), and comparable results were achieved: 83.33% of average accuracy. In their study, PCA and LDA were applied. Noroozi et al., (2017) have made use of Random Forest and Decision Tree methods based on pitch and intensity (among others), and the average recognition rate was 66.28% to classify 6 emotions.

5. Study 3: Perception experiment - emotion categorization

Perception helps us to select the relevant information among all other features that are not as relevant for the purpose of communication, and not all the perceivable elements of speech melody have a communicative purpose either. It is important to highlight that a perception approach is subjective per se. Therefore, reliable experimental techniques, which can guarantee the replicability, are necessary (‘t Hart et al., 1990).

In order to confirm that humans can also identify the produced emotions in the first experiment, a listening experiment was designed and conducted. The main motivation to perform this study was to investigate the human perception of positive and negative emotions through voice, and verify the findings of statistical analysis and the machine-learning model from a perceptual point of view. Xu (2011), in his review of speech prosody methodology, points to the disconnection between production and perception oriented strategies, and suggests more linking and integration between these two as more beneficial for the interpretation of the data. Various researchers consider it necessary to use human evaluation as a benchmark for the machine results (Vieru et al., 2011; Janssen et al., 2013, Schuller et al., 2003).

The emotions were represented by 24 utterances in total, selected from Experiment 1. Eight of the utterances were expressing a happy emotion, eight a sad emotion and eight utterances were neutral.

5.1. Methodology

Identification of emotions was used as an experimental design for this study. The main research issue in this study was: how accurate can a set of listeners correctly identify happy and sad emotions from a speech database without other forms of context (such as linguistic, visual, etc.). This experiment sought to test whether participants could discriminate three distinct emotions on

the stimuli: happy, sad and neutral. Previous research has shown that audio-visual perception had better results when categorizing emotions, but the overall results in correctly-identified emotion through only hearing a voice are expected to be around 84.2% (de Oca, 2009). It is important to mention that in this study only prosodic phrases whose meanings were not emotionally-laden were extracted from the utterances. It was done in order to avoid words that might give any clue about the emotional context of the speaker.

The study of Silva et al., (2016) about cross-cultural and cross-linguistic perception of authentic emotions through speech with Brazilian and Swedish listeners was used as a reference guide to design the perceptual experiment through an online survey. As a method of measurement, participants judged the emotion of the Swedish speakers from the recordings on 5-point Likert scales. Listeners rated one emotional dimension of the utterances on a scale from "very negative" to "very positive", passing through “neutral” emotion. A score of 1-2 indicated a negative mood, 3 represented a neutral mood, and 4-5 represented a positive mood.

Some studies suggest that antonyms behave as two distinct affects and therefore do not represent opposite values of a scale for the listeners (Schimmack, 2001). In the case of this study, however, it is important to highlight that “positive” and “negative” itself represent the evaluation dimension of the perceived emotion only; words such as “sad” and “happy” were not used to avoid ambiguities.

5.1.1. Speech material

The stimuli were obtained from the utterances recorded during the production experiment (Experiment 1). All the recordings were reviewed and random stimuli were selected for each category, having in mind the length and semantic neutrality of each. From the psycholinguistic perspective, we should take in mind that the stimuli should have fixed parameters such as length. The length of the selected stimulus was approximately the same: it diverged slightly between 2.5-3.5 seconds. In case of the utterances themselves, it is important to mention that it was impossible to use the same syntactic sentences, however it was crucial to keep the emotion-neutral semantic content of the utterances. Therefore, 2 evaluators gave a second opinion about the neutrality of the meaning for each stimulus before the final selection of the stimulus for the experiment. The evaluators were asked to read the sentence (no audio sample was provided, but only the transcribed sentence) and give it one positive and one negative possible context. For those utterances that the evaluators could not give both contexts, the stimuli were deleted.

In total, 24 recordings were used as stimuli for the experiment (the transcribed recordings and their translation can be found in Annex A). The same number of male and female speakers was presented for each emotion (happy, sad and neutral) with 8 different recordings per each emotion. Overall, the stimuli consisted of recordings from 10 different speakers, who appear in a maximum of 2 utterances. The speaker was not repeated more than once within each emotion.

5.1.2. Procedure

An online questionnaire was prepared using Webropol Surveys. The test consisted of 2 parts: the first was a social geographical survey and the second part was the experiment itself, including the recordings and the emotional questionnaire. The utterances selected from the first experiments were extracted and encoded as 128kb/s MP3 at 44.1 KhZ, 16 bit.

The test was Internet-based and consequently the participants were able to participate wherever and whenever they wanted. Participants were approached via social media and e-mail. First, after opening the link in their usual browser, a small introduction was given. Instruction, questions and all the text was presented in Swedish, except for predefined survey buttons. The goal of the experiment was given in the introduction, which was to investigate the emotion identification in utterances. After the description, participants were asked their sex, age, education, place of birth and place of residency. Following this, the first recording was used as a sound check, in order to adjust the volume of their equipment. Each sample was located on a separate page and the user could start playing the recording when ready. After selecting the desirable value on the scale, the user had an option to go to the next page. On the bottom of the page, the user could see the progress bar of the survey. The survey was also cross-platform which allowed the listeners to visualize the survey properly from computer, phone or tablet. Figure 7 is an example of the survey seen on participant's the phone.

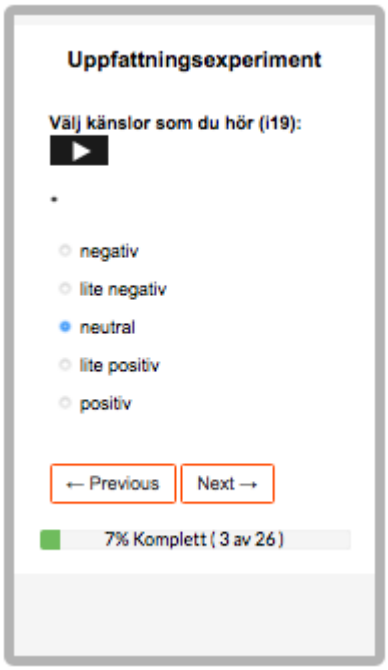


Figure 7: Screenshot of the screen form for stimuli evaluation seen on a phone.

Only native Swedish speakers were asked to participate, having as a reference Abelin and Allwood (2000) who suggested that listeners can recognize better the emotions expressed by speakers of the same language. If desired, the participants had an option to repeat the utterance as many times as they needed. Overall, completion of the experiment took approximately 5-6 minutes.

5.2. Results

120 participants took part in the experiment. Around 80% of the participants were between 18 and 34 years old. Almost an equal number of females and males participated: 52% female and 47% male. Regarding the residency of the participants: more than 60% lived in Göteborg. The participants came from all over the country. In total, 60 different cities were registered, with 24% of participants from Göteborg and 9% from Stockholm. Educational levels varied considerably,

between grundskola (primary school):1%, gymnasium (secondary school): 24%, högskola (high school): 16%, university: 57% and PHD: 2 (for more detailed information see Annex B).

Table 11 present the participants' classification of neutral, happy and sad stimuli. The total rate of correct answers is therefore 58.8%, where sad stimuli is better classified comparing to the other emotions, with 63.65% of accuracy. It is interesting to notice that compared to the results of the classifier (Table 9), the most accurately classified emotion, both by machine and humans, is sad, followed by happy and then, neutral.

Table 11. Correctly classified responses (%) for happy, neutral and sad stimuli.

	Happy	Neutral	Sad
Happy	**57.3%**	24.9%	17.8%
Neutral	19.6%	**55.6%**	24.8%
Sad	5.10%	31.25%	**63.65%**

5.3. Discussion

This experiment was set up to test the third hypothesis, stating our expectation that Swedish participants will be able to perceive and identify happy, sad and neutral emotions through the sequences of short recordings extracted from previously collected corpora. Based on the results, we have found evidence supporting this hypothesis. As mentioned earlier in this chapter, the higher accuracy was not expected due to several reasons. First of all, the stimuli were extracted from a larger production and secondly, stimuli without emotional-laden meaning were selected. Moreover, since the investigation is not based on acted but on elicited speech, each stimulus diverged from one another. This factor discarded the possibility to hear the same sentence in a happy and sad mood, as it is common with acted speech corpus. Furthermore, the human participants did not have the same quantity of data as the machine learning model.

Similar results were obtained in the research by Tóth et al., (2008), where the human emotion perception reached the accuracy of 60%. In their research, the authors were working with 8 emotions (happy, sad, angry, surprised, disgusted, afraid, nervous and neutral), which were produced by actors. The same accuracy was achieved in Vieru et al., (2011) study, where six foreign accents were investigated from perception, among others, perspectives. Although, higher results were reported recently in another perception study by Kim and Sumner (2017), where the non-emotional word "pineapple" was uttered with emotional prosody by non-professional actors. However, this study cannot be fully compared to the experiment performed in this thesis, since various crucial factors diverged. First of all, the corpus in this study is based on elicited speech and not acted speech. This characteristic leads us to the second difference: the stimuli used in this study were different in each case, which complicated the perception task, but at the same time the experiment presented a more natural condition for emotion perception. Finally, compared to Kim and Sumner's (2017) study, which excluded participants based on poor performance, this study included all the participants for the reason to obtain more representative data with all possible perception cases.

It is worth noting that similar results were obtained in Vieru et al., (2011). In their study, perception, production and automatic processing experiments were designed in order to

investigate six foreign accents. Perceptual tests on read and spontaneous speech achieved 60%, while automatic processing techniques reached 74%, outperforming therefore human perception. Janssen et al., (2013) in their research compared machine with human emotion recognition through facial expression, speech and physiological signals in order to establish a benchmark in the performances of machine emotion recognition systems. In the study, they conclude that the machine outperformed humans, comparing SVM (with 82% of recognition accuracy using physiological and facial features only) against 75 humans (with 62.9% accuracy). His study is one of the most closely related studies that can be compared to this current study since the data consisted of elicited speech of self-reports of experienced emotion. Although, humans annotated the emotions of each performance and more emotions were analysed (happy, sad, angry, relaxed and neutral). Another example where machines outperform humans is in the contribution of Schuller et al., (2003), where the recognition of seven emotions with Hidden Markov Models exceeded 86% recognition rate, while the human classification of the same corpus was at 79.8%. High accuracy was achieved by Seppänen et al., (2003) with 76.9% classification accuracy, although it is worth mentioning that the used data included 56 professionally acted monologues, each about one minute in length.

6. Final Discussion

The present study was designed to investigate whether news reports are suitable to be used as stimuli to evoke emotional responses of Swedish speakers. Unlike earlier studies that generally use actor-based corpora, this study focused on elicited speech and the importance of more natural corpora, not only for the classification method but also for perception experimentation. Through the classification and linear mixed effect models, we demonstrated that the collected data presented clear patterns of happy, sad and neutral emotions and therefore news stimuli can be used as a data generation method to detect emotions through speech.

From the linear mixed effects models, the results showed that mean intensity, duration and the time of the *F0* peak were the clearest features to distinguish happy, sad and neutral emotions. Emotions had also some effect on prosodic features such as mean *F0*, Max *F0* and Final *F0*. Decision Tree (DT) classifier reached an average accuracy of 70.88% (tested on validation data from 10-fold cross-validation). Final velocity, relative location of the *F0* peak, time of the *F0* peak and mean intensity present a crucial juncture for the classifier.

Regarding the Global North and South news, the results from statistical analysis show that sad-south news reports do not present any clear difference in relation to sad-north news reports. It can be observed from the results that sad-north and sad-south categories present very similar values and therefore no significant difference exists. This leads us to conclude that no evidence regarding emotion as an evaluation device in case of South and North news is found in this study. However, based on the limits of the experiments, it is questionable if the results can be generalized. More research to confirm this finding needs to be performed in the future.

In this research work, we did not only focus on machine detection but human emotion perception was also examined, to verify and contrast the previous findings. The perception experiment demonstrated that the evaluation dimension (positive and negative) of the emotion in speech could be identified by human participants, although the human participants do not reach the

machine accuracy in emotion discrimination. As previously mentioned, machine is more accurate than humans evaluators in similar studies. Different factors that affect human performance should be taken in mind; as for example the quantity of the given samples, since human participants did not analyse all the data, but only a small portion of it.

Overall, we believe that the research on experimentally evoked emotions based on news stimuli can open a new range of research possibilities in different correlated fields and this novel approach to elicit emotion brought a fresh perspective on this multidisciplinary topic. More research needs to be done to delve into emotion detection through speech using news report-based elicited corpora.

7. Limitations of the study

Although it is impossible to compare recognition accuracies obtained in this study with other studies due to the general variety of methodologies in other studies (regarding corpora, features, classifiers, emotions, etc.), the methods implemented in this study are clearly promising. The recognition accuracies obtained using DT reach our expectations and the classification between happy, sad and neutral was performed to a good level.

However, the main limitation of the study is related to the size and nature of the corpora: more data would provide even more accurate results, as the size of the database used for speech emotion recognition plays a crucial role in deciding the properties such as generalizability, reliability and scalability (Douglas-Cowie et al., 2003). Due to the size of the database, the testing had to be performed using cross-validation techniques, since it is the best approach when dealing with small data, but given a bigger corpus, the evaluation of the classifiers could be performed on the test data set.

8. Future work

The results of the study open paths to several new directions. Investigating the efficiency of other features can be interesting to see if the results improve. More features include: acoustic pause parameters and their position (Bhargava and Polzehl, 2009); a combination of acoustic, lexical, and discourse features (Lee and Narayanan, 2005); or Bio-informational Dimensions (BID) which include body-size projection, dynamicity, audibility and association (Xu et al., 2013). These can be analysed as possible options to enlarge the feature list, since the results in the aforementioned studies are promising and may help to improve the overall accuracy.

It seems straightforward to extend the variety of classification techniques to investigate the obtained results. Even though the DT classifier provides satisfying results, we do not discard the combination of two classifiers in order to improve the system, such as SVM and DT which perform in a very positive way according to the researchers (Nguyen and Bass, 2005) or NN-C5.0 classification method (Javidi and Roshan, 2013). Apart from these suggestions, other supervised learning algorithms can be tested and compared in future work.

It is not possible to conclude from our data that the pattern of emotion expression regarding South and North is universal. A more detailed study needs to be designed in order to provide more specific information regarding this topic, such as dividing the articles by location and proximity to see if any particular pattern affecting emotion is evidenced or comparing the findings to other nationalities, for example. Moreover, in the future work, it might be interesting to examine this method with other languages to see if there are any particular characteristics that differentiate from Swedish and the findings that are reported in this study.

Bibliography

Abelin, Å., and Allwood, J. (2000). Cross Linguistic Interpretation of Emotional Prosody. *Proc. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 110-113. Newcastle.

Adams, W. C. (1986). Whose Lives Count?: TV Coverage of Natural Disasters. *Journal of Communication*, vol 36, Issue 2, 113–122.

Akamine, M., and Ajmera, J. (2012). Decision tree-based acoustic models for speech recognition. *EURASIP J Audio Speech Music Process* 2012(1):10.

Analytics Vidhya Content Team. (2016). [Untitled illustration of a decision tree classification dog]. Retrieved December 15, 2017 from https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R.H., Davidson, D.J., Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390- 412.

Baayen, R. H. (2012). Mixed-effects models. In Cohn, A. C., Fougeron, C., and Huffman, M.K. (Eds.), *Handbook of Laboratory Phonology*, 668–677. Oxford: Oxford University Press.

Baber, C., and Noyes, J. (1996). Automatic speech recognition in adverse environments. *Hum. Fact.* 38 (1), 142–155.

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol,* 70 (3), 614-636.

Batliner, A., Hacker, C., Steidl, S., Noth, E., Archy, D. S., Russell, M., and Wong, M. (2004). You stupid tin box children interacting with the Aibo robot: a cross-linguistic emotional speech corpus. *Proc. language resources and evaluation (LREC 04)*, Lisbon.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. (2011). Mining Social Emotions from Affective Text. *IEEE Transactions on Knowledge & Data Engineering*, vol. 24, no., 1658-1670.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM* 37(7), 122–125.

Bates, D.M., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes.

Bhargava, M., and Polzehl, T. (2009). Improving automatic emotion recognition from speech using rhythm and temporal feature. *ICECIT*, 2229–3116.

Bhowmick, P., Basu, A., Mitra, P., and Prasad, A. (2009). Multi-label Text Classification Approach for Sentence Level News Emotion Analysis. *Pattern Recognition and Machine Intelligence. Lecture Notes in Computer Science*, vol. 5909, 261- 266.

Berry, D. M. (2012). The Computational Turn: Thinking About the Digital Humanities. *CultureMachine12*. Available at: http://www.culturemachine.net/index.php/cm/ article/view/440/470.

Boersma, P., and Weenink, D. (2016). Praat: Doing Phonetics by Computer. Version 6.0.14. Available at: http://www.praat.org/.

Bozkurt, E., and Erzin, E. (2009). Improving automatic emotion recognition from speech signals. In 10th annual conference of the *Interspeech*, Brighton, UK, 324–327.

Brandt, W. (1980). North-South: a programme for survival: report of the Independent Commission on International Development Issues. Cambridge: MA

Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Internat. J. Human–Comput. Stud,* 62 (2), 161–178.

Breitenstein, C., Van Lancker, D., and Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, 15 (1): 57–79.

Bucy, E. P. (2003). Emotion, presidential communication, and traumatic news. *Harvard International Journal of Press/Politics, 8*(4), 76-96.

Butler, J. (2009). *Frames of War: When is Life Grievable?* Brooklyn, NY: Verso

Bänziger, T., and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46/3-4, 252–267.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code. A perceptual investigation. *Phonetica,* 65, 210–230.

Cornelius, R. R. (1996). The science of emotion. Research and tradition in the psychology of emotion. Upper Saddle River (NJ): Prentice-Hall.

Cortazzi, M. and Jin, L. (2000). Evaluating Evaluation in Narrative. In Thompson, G. and Huston, S. (Ed.), *Evaluation in text. Authorial Stance and the Construction of Discourse*, 1-27. Oxford: Oxford University Press.

Cowie, R., and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40/1-2, 5–32.

Cowie, R. (2000). Describing the Emotional States Expressed in Speech, I*SCA Workshop on Speech & Emotion*, 11-18. Northern Ireland.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Process*, vol. 18(1), 32–80.

Cutler, A., and Clifton, C. (1999). Comprehending Spoken Language: A Blueprint of the Listener. In Brown, C.M. and Hagoort, P. (Eds.), *The Neurocognition of Language*, 123–166. Oxford: Oxford University Press.

Daun, Å. (1996). *Swedish Mentality.* University Park: Pennsylvania State University Press.

Davitz, J. R. (1964). Auditory correlates of vocal expression of emotional feeling. In Davitz, J.R. (Ed.), *The communication of emotional meaning*, 101–112. New York: McGraw-Hill.

De Oca, A.M., Cook, T., Dias, J., and Rosenblum, L. (2009). Speech Perception and Emotion: Linguistic vs. Emotional Content. UCR: University of California.

Douglas, B., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1),1–48.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40, 33–60.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., and Mcrorie M. (2007). The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In Paiva A. R., Prada R., Picard R. (Eds.), *Affective Computing and Intelligent Interaction*, 488–500. Heidelberg: Springer Berlin.

Esposito, A., Esposito, A. M. (2012). On the recognition of emotional vocal expressions: motivations for a holistic approach. *Cogn. Process.* 13(Suppl. 2), 541–550.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., and Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7), 829–837.

Freund, Y., and Schapire, R. E. (1999). A Brief Introduction to Boosting. *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99)*, vol. 2, 1401–1406. Stockholm, Sweden.

Frijda, N.H. (2007). *The laws of emotion.* Mahwah, NJ: Erlbaum.

Herm, O., Schmitt, A., and Liscombe, J., (2008). When calls go wrong: How to detect problematic calls based on log-files and emotions. *Proc. Interspeech*, 463–466.

Hoque, M. E., Yeasin, M., and Louwerse, M. M. (2006). Robust recognition of emotion from speech. In *Intelligent virtual agents*. *Lecture notes in computer science,* 42–53. Berlin: Springer.

Iida, A., Campbell, N., and Yasamura, M. (1998). Design and Evaluation of Synthesised Speech with Emotion. *Journal of Information Processing Society of Japan*, 40 (2).

Izard, C. E. (1977). *Human emotions.* New York: Plenum Press.

Janssen, J.H., Tacken, P., Vries, J.J.G. de, Broek, E.L. van den, Westerink, J.H.D.M., Haselage, P., IJsselsteijn, W.A. (2013). Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. *Human-Computer Interactions,* 28, 479-517.

Javidi, M. M., and Ebrahim F. R. (2013). Speech emotion recognition by using combinations of C5. 0, Neural Network (NN), and Support Vector Machines (SVM) classification methods. *J. Math. Comput. Sci*, 6, 191-200.

Juslin, P. N., and Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3), 217-238.

Kadohisa, M. (2013). Effects of odor on emotion, with implications. *Frontiers in System Neuroscience*, 7(66),66.

Kao, Y. H., and Lee, L. S. (2006). Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. *INTERSPEECH - ICSLP*, 1814–1817.

Kim, S. K., and Sumner, M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *The Journal of the Acoustical Society of America*, 142(1), 49-55.

Koolagudi, S. G. and Krothapalli, R.S. (2011). Two stage emotion recognition based on speaking rate. *Int J Speech Technol*, 14, 35-38.

Koolagudi, S. G., and Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15(2), 99-117.

Kuhn, M. (2016). *Caret: Classification and Regression Training*. R package version 6.0-68.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling.* New York: Springer.

Kuhn, M., Weston, S., Coulter, N., and code for C5.0 by R. Quinlan, R. (2015). *C50: C5.0 Decision Trees and Rule-Based Models.* R package version 0.1.0-24.

Kwon, O., Chan, K., Hao, J., Lee, T. (2003). Emotion Recognition by Speech Signals. *Eurospeech*. Geneva.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Laver, J. and Trudgill, P. (1979). Phonetic and linguistic markers in speech. In. Scherer, K.R. and Giles, H. (Eds.), *Social markers in speech*, 1-32. Cambridge: Cambridge University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science*, vol. 323, issue 5915 (6 February), 721-723.

Le Tallec, M. (2011). EmoLogus: Presentation of a linguistically based model for emotion detection and adaption to another context. *IEEE International Conference on Mechatronics and Automation (ICMA),* 1303 -1308.

Lee, B., Kao, E., and Soo, V. (2006). Feeling ambivalent: A model of mixed emotions for virtual agents. *6th International Conference on Intelligent Virtual Agents (IVA'06)*, 329–342. Marina del Rey, CA. Springer.

Lee, C.M., and Narayanan, S.S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process,* 13 (2), 293–303.

Lee, J.-E. R., and Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In Latusek, D. (Ed.), *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*. IGI Global.

Liebler, C. M. (2010). Me(di)a Culpa?: The 'Missing White Woman Syndrome' and Media Self-Critique. *Communication, Culture & Critique*, 3, 549-565.

Louwerse, M., and Mitchell, H.H. (2003). Towards a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes*, 199-239.

Mannell, R. (2007). Introduction to Prosody: Theories and Models. Retrieved October 5, 2017, from http://clas.mq.edu.au/speech/phonetics/phonology/intonation/prosody.html.

McMahon, E., Cowie, R., Kasderidis, S., Taylor, J., and Kollias, S. (2003). What chance that a DC could recognize hazardous mental states from sensor inputs? *Tales of the disappearing computer*. Santorini, Greece.

Meyer, E. (2015). Getting to Si, Ja, Oui, Hai, and Da. *Harvard business review,* 93(12), 74-80.

Minsky, M. (2006). The Emotion Machine. Simon & Schuster.

Lennes, M. (2011). SpeCT - The Speech Corpus Toolkit for Praat". Available at: https://lennes.github.io/spect/.

Mikels, J., Dredrickson, B., Larkin, G.R., Lindberg, C., Maglio, S., and Reuter-Lorenz, P,: (2005). Emotional category data on images from the international affective picture system. *Behav. Res. Methods* 26(4), 526-630.

Mohammad, S. M., and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. *Proceedings of the NAACLHLT 2010 workshop on computational approaches to analysis and generation of emotion in tex*t. Los Angeles, CA.

Mozziconacci, S.J.L. (1995). Pitch variations and emotions in speech. *Proceedings of the 13th International Congress of Phonetic Sciences, ICPhS-95,* August 13-19, 1, 178-181. Stockholm, Sweden.

Murray, I.R., and Arnott, J.L., (1993). Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93, 1097–1108.

Murray, I. R., and Arnott, J. L. (2008). Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer Speech and Language*, 22(2), 107–129.

Nabi, R. L. (1999). A cognitive-functional model for the effects of discrete negative emotions on information processing, attitude change, and recall. *Communication Theory,* 9(3), 292-320.

Nass, C., Steuer, J. S., Henriksen, L., and Dryer, D. C. (1994). Machines and social attributions: Performance assessments of computers subsequent to "self-" or "other-" evaluations. *International Journal of Human-Computer Studies*, 40, 543-559.

Neiberg, D., Elenius, K., Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. *Proceedings of Interspeech*.

Nguyen, T., & Bass, I. (2005). Investigation of combining SVM and decision tree for emotion classification. *Proc. 7th IEEE international symp. on multimedia*, 540-544.

Noroozi,F., Sapinski, T., Kaminska, D., Anbarjafari, G. (2017). Vocal-based emotion recognition using random forests and decision tree. *I. J. Speech Technology*, 20(2), 239-246.

Nwe, T. L., Foo, S. W., and Silva, L. C. D. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*, 603-623.

Philippou-Hübner, D., Vlasenko, B., Bock, R., and Wendemuth, A. (2012). The performance of the speaking rate parameter in emotion recognition from speech. *Proceedings of IEEE ICME*, 248–253.

Picard, R. W. (1995). Affective Computing. *MIT Media Laboratory Perceptual Computing* Section Technical Report No. 321. Media Lab. Massachusetts Institute of Technology, Cambridge Univ.

Plutchik, R. y Kellerman, H. (1983). *Emotion: Theory, Research and Experience*. Academic. New York.

Quinlan, R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, http: //www.rulequest.com/see5-unix.html

Rigoulot, S., and Pell, M. D. (2012). Seeing emotion with your ears: emotional prosody implicitly guides visual attention to faces. *PLoS ONE* 7:e30740.

Roberts, L. (2011) Acoustic effects of authentic and acted distress on fundamental frequency and vowel quality. *Proceedings of the 17th International Congress of Phonetic Sciences*, 1694–1697. Hong Kong.

Rosis, F. de., Grasso, F. (2000). Affective natural language generation. In Paiva, A. (Ed.), *Affective interactions*, Lecture notes in Artificial Intelligence, vol. 1814, Springer-Verslag.

Salmon, P. (2001). Effects of Physical Exercise on Anxiety, Depression and Sensitivity to Stress - A Unifying Theory. *Clinical Psychology Review*, vol. 21(1), 33-61.

Stillman, S. (2007) 'The missing white girl syndrome': disappeared women and media activism. *Gender & Development*, 15(3), 491-50.

Scherer, K. R. (1998). Emotionsprozesse im Medienkontext: Forschungsillustrationen und Zukunftsperspektiven [Emotion processes in the context of the media: Illustrative research and perspectives for the future]. *Medienpsychologie,* 10, 276-93.

Scherer, K. R. (2000). A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. *Proc. ICSLP*, 379–382. Beijing, China.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, 227-256.

Scherer, K. R., Banse, R., Wallbott, H. G. and Goldbeck. T. (1991). Vocal cues in emotion encoding and decoding. *Motivation & Emotion*, 2(15), 123-148.

Schimmack, U. (2001). Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive? *Cogn. Emot,* 15, 81-97.

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden markov model-based speech emotion recognition. *Proceedings of IEEE ICASSP*, 2, II–11.

Schuller, B., Rigoll, G., and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *Proc. IEEE int. conf. acoust., speech, signal processing*, 577– 580. New York: IEEE Press.

Schröder, M., Cowie R., and Cowie, E. (2001a). Emotional Speech Synthesis: A Review. *Eurospeech*.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001b). Acoustic correlates of emotion dimensions in view of speech synthesis. *Proceedings of Eurospeech*, 87–90. Aalborg, Denmark.

Seppänen, T., Väyrynen, E., and Toivanen, J. (2003). Prosody-based classification of emotions in spoken finnish. *Proc. 8th Eur. Conf. on Speech Communication and Technology*, 717-720. Geneva, Switzerland.

Silva, W., Barbosa, P. A., and Abelin, Å. (2016). Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic-phonetic study with Brazilian and Swedish listeners. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 32(2), 449-480.

Smith, P. M. and Warr, K. (1991). *Global Environmental Issues*. London, Hodder & Stoughton.

Sontag, S. (2003). *Regarding the Pain of Others.* New York: Farrar, Straus & Giroux.

Stibbard, R. M. (2001). *Vocal Expression of Emotions in Non-laboratory Speech: An Investigation of the Reading/Leeds Emotion in Speech Project Annotation Data.* PhD thesis. University of Reading, UK.

Strapparava, C., and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*. Prague, Czech Republic.

't Hart, J., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody.* Cambridge: Cambridge University Press.

Tannen, D. (1980). A comparative analysis of oral narrative strategies: Athenian Greek and American English. In Chafe, W. L. (Ed), *The pear Stories: Cognitive, Cultural and Linguistic Aspects of narrative production*,51-87. Norwood, NJ: Ablex.

Tao, J.H., Kang, Y.G. (2005). Features importance analysis for emotional speech classification. *Proceedings of lecture notes in computer science*, 449-457. Springer.

Tato, R.S, Kompe, R., Pardo, J.M. (2002). Emotional Space Improves Emotion Recognition. *ICSLP*, 2029-2032.

Themistocleous, C. (2017). Classifying linguistic and dialectal information from vowel acoustic parameters. *Speech Communication*, 94, 13–22.

Thompson, G. and Huston, S. (2000). Evaluation: An Introduction. In Thompson, G. and Huston, S. (Ed.), *Evaluation in text. Authorial Stance and the Construction of Discourse*, 102-120. Oxford: Oxford University Press.

Tóth S.L., Sztahó D., and Vicsi K. (2008). Speech Emotion Perception by Human and Machine. In Esposito A., Bourbakis N.G., Avouris N., and Hatzilygeroudis I. (Eds.), *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Lecture Notes in Computer Science, vol 5042. Springer, Berlin, Heidelberg.

Unz, D. C., Schwab, F., and Winterhoff-Spurk, P. (2008). TV News – The Daily Horror?: Emotional Effects of Violent Television News. Journal of Media Psychology Theories Methods and Applications 20(4), 141-155.

Van Dijk, T. A. (1987). Communicating Racism. Ethnic Prejudice in Thought and Talk. Newbury Park, CA: Sage.

Ververidis, D., and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48, 1162–1181.

Vieru,B., de Mareüil, P. B., Martine, A. D. (2011). Characterisation and identification of non-native French accents. *Speech Commun*., 53, 292-310.

Viklund, J. & Borin, L. (2016). How Can Big Data Help Us Study Rhetorical History? *Selected Papers from the CLARIN Annual Conference 2015*, 123, 79–93. Linköping University Electronic Press.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf]

Witten, I. H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. (2nd edition). San Francisco: Morgan Kaufmann Publishers.

Wirth, W., and Böcking, S. (2003). Emotionalized News Stories and Attribution of Relevance. *ICA Conference.* San Diego.

Wirth, W., and Schramm, H. (2005). Media and emotions. *Communication Research Trends*, 24 (3), 3–39.

Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences,* 1, 85-115.

Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7-10. Aix-en-Provence, France.

Xu, Y., Kelly, A., and Smillie, C. (2013), Emotional expressions as communicative signals. In Hancil, S., and Hirst, D. (Eds.), Prosody and Iconicity, 33–60. John Benjamins Publishing Company.

Yacoub, S., Simske, S., Lin, X., and Burns, J. (2003). Recognition of Emotions in Interactive Voice Response Systems. *Proceedings of Eurospeech*, 1–4.

Yildirim, S., Lee, C.M., Lee, S., Potamianos, A., and Narayanan, S., (2005). Detecting politeness and frustration state of a child in a detecting politeness and frustration state of a child in a conversational computer game. *Proc. Eurospeech*. Lisbon, Portugal.

Yu,F., Chang, E., Xu, Y. Q., and Shum, H. Y. (2001). Emotion Detection From Speech To Enrich Multimedia Content. *Second IEEE Pacific-Rim Conference on Multimedia*, 24-26. Beijing, China.

Zei, B. (2002). A place for affective prosody in a unified model of cognition and emotion. In Bel, B., Marlien, I. (Eds.) *Proceedings of 1st International Conference on Speech Prosody*, 17-22.

Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In Sun, et al. (Eds.), *Advances in neural networks*. Lecture notes in computer science, 457–464. Berlin: Springer.

Zillmann, D. (1988). Mood management through communication choices. *American Behavioral Scientist, 31*, 327-340.

Appendix A: Perception experiment - transcription of the used sentences

Table 15. Transcription of the sentences used sentences.

	Original	**Translation**
1	De kände inte varandra	They did not know each other
2	Och se en bild på mannen och hans kats	And see a picture of the man and his cat
3	Bli större och större varje år	become bigger and bigger every year
4	Det handlar om en karneval i Hammarkullen	It is about carnival in Hammarkullen
5	Cch ska flygas tillbaka till afrika	And will fly back to Africa
6	Det bli en gigantisk nyhet	To become gigantic news
7	Och på bilden ser man hästen	And on the photo you can see the horse
8	Och säljer dessa för 1.4000 miljoner	And sell these for 1.4 million.
9	Bild på en person som bar en annan	Photo of a person, carrying another [person]
10	I bakgrunden ser man människor som stod	There were people standing in the background
11	På en gata i oslo	On a street in Oslo
12	Man vill förstå hur hon känner	You want to understand what she feels
13	[eeeh] var en dam på bilden	[eeeeh] was a lady on the picture
14	Man tänker på familjen också	You also think about the family
15	På pendlare och studenter och turister	of commuters, students and tourists
16	Olika människor som betalar	different people that pay
17	Vanligen lever i vatten	Normally lives in water
18	Med kristaller av hexagon struktur	With crystals of hexagon structure
19	Ironi och sarkasm	Irony and sarcasm
20	Som anvands for beräkningar	Which are used for calculations

21	Flera olika namn	Many different names
22	Implantat medel i människokroppen	Implants in the human body
23	Fast form och i gasform	In the form of solid or gas
24	Som är relativ stabila	Which are relatively stable

Annex B: Perception experiment - demography survey

B.1. Age

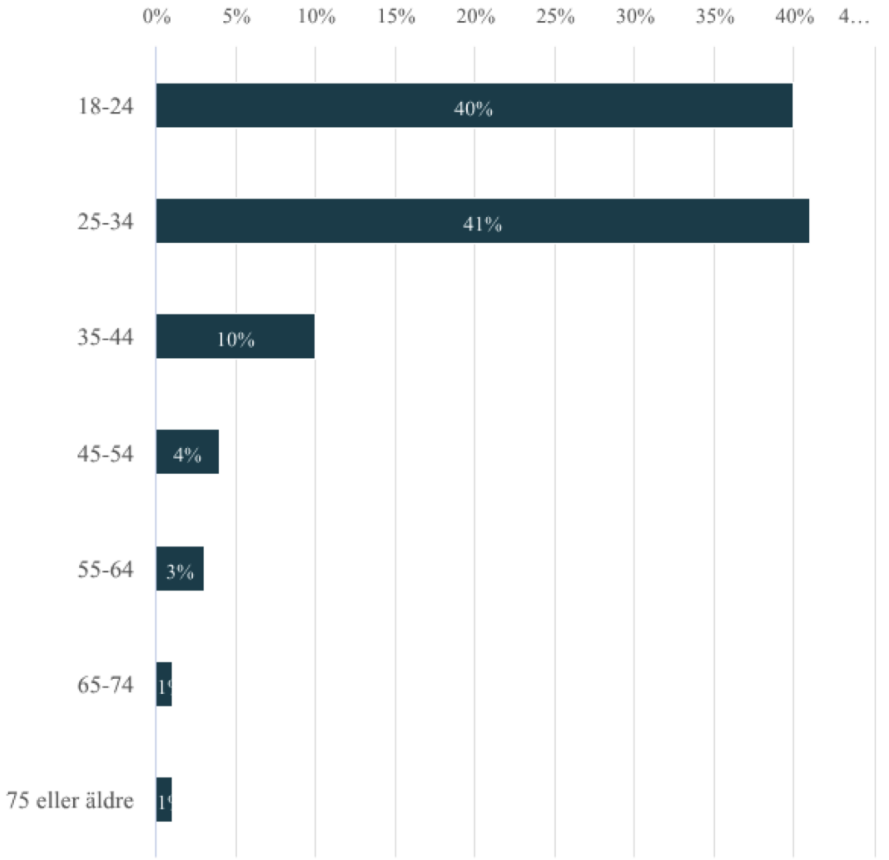


Figure 8. Age of the participants.

B.2. Sex

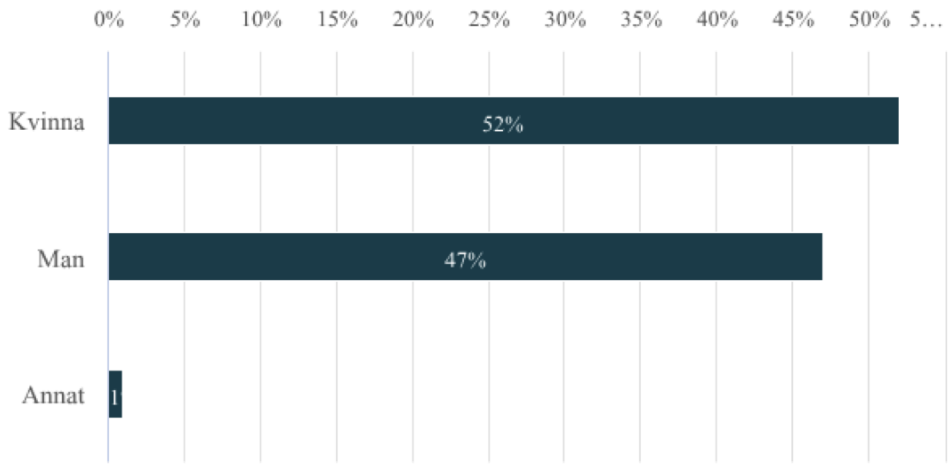


Figure 9. Sex of the participants.*
*translations: kvinna(woman); annat(other)

B.3. Education level

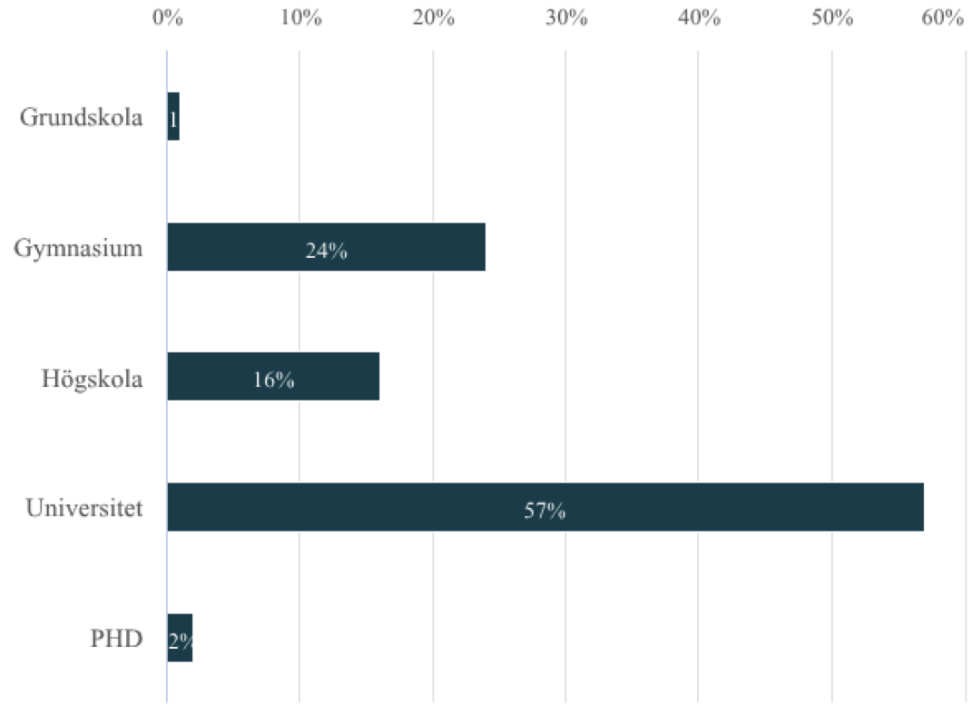


Figure 10. Education level of the participants.*
*translations: grundskola (primary school); gymnasium (secondary school); högskola (high school)

B.4. Residency of the participants

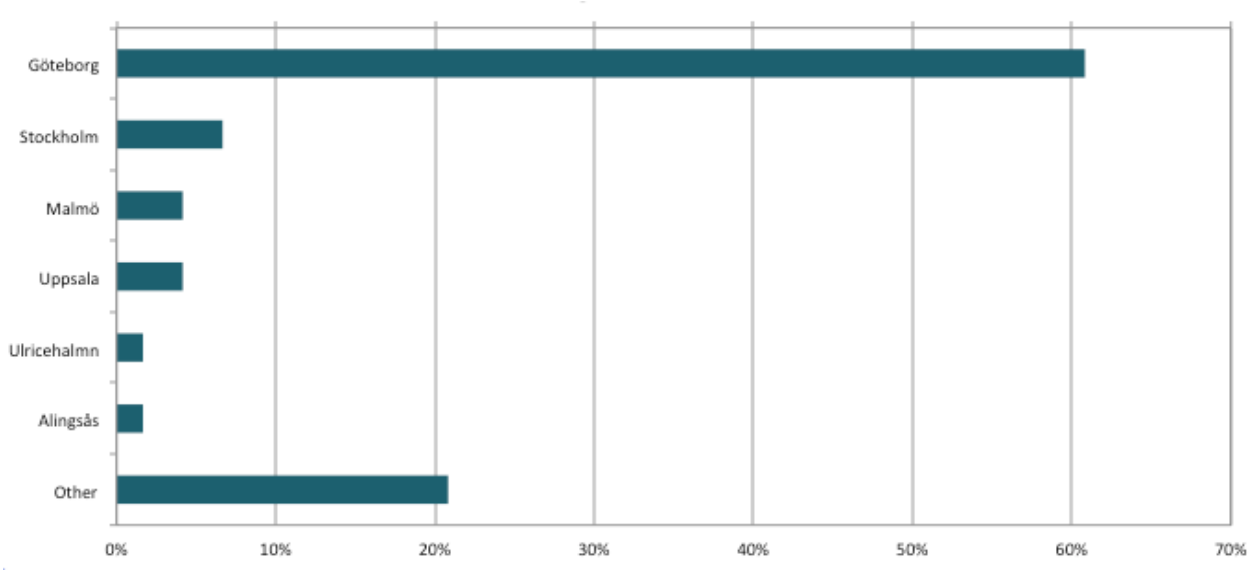


Figure 11. Residency of the participants.