Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Spoken Affect Classification: Algorithms and Experimental Implementation

A thesis presented in partial
fulfilment of the requirements
for the degree of
Master of Science
in Computer Science

at Massey University,
Palmerston North, New Zealand.

Massey University

Donn Alexander Morrison
2005

Para Consuela, mi fiel furgoneta

Abstract

Machine-based emotional intelligence is a requirement for natural interaction between humans and computer interfaces and a basic level of accurate emotion perception is needed for computer systems to respond adequately to human emotion. Humans convey emotional information both intentionally and unintentionally via speech patterns. These vocal patterns are perceived and understood by listeners during conversation. This research aims to improve the automatic perception of vocal emotion in two ways. First, we compare two emotional speech data sources: natural, spontaneous emotional speech and acted or portrayed emotional speech. This comparison demonstrates the advantages and disadvantages of both acquisition methods and how these methods affect the end application of vocal emotion recognition. Second, we look at two classification methods which have gone unexplored in this field: stacked generalisation and unweighted vote. We show how these techniques can yield an improvement over traditional classification methods.

Acknowledgements

I would like to thank my supervisor, Dr. Ruili Wang for putting faith in me and allowing me to pursue this degree under scholarship. Without this financial help, it would have been unfeasible. His tireless direction and advanced motivational techniques also helped keep my focus.

My co-supervisors, Dr. Liyanage C. De Silva and Dr. Peter Xu, also lended support throughout the research. Their extensive experience was indispensable at times when I needed support.

I also owe gratitude to Pete Morrison of Mabix International for his unique insight into the research. He provided the data used in this research, and without it, it could not be possible.

The second speech database was provided by Tin Lay Nwe of the National University of Singapore. This database was collected and compiled by her and was graciously provided to aid in this research.

And to my partner Jen, who endured countless late nights, filled me with confidence when I lacked it, fed me when I didn't have time to feed myself, bathed me when...well, you get the picture.

Of course, I thank my family, Bill, Debb, Bugs, Michael, Todd, and Jodybird, for their constant support and love, all the way across the Pacific Ocean.

To my postgraduate friends at Massey University, who helped create a fun and relaxed working environment: Cath, Frank, Matthew, Michael, Stefan, and Yiming, among others. And Francis, for always seeming to be in Australia when I needed it the most.

Last, I would like to thank the developers of the many free and open-source software tools I used during the research. Packages such as LaTeX, Gnuplot, OpenOffice, Graphviz, Vim, Octave, donnrisk, Praat, The Speech Filing System, WEKA, The GIMP, Mozilla Firefox, and most of all Debian GNU/Linux were where I spent most of my time this past year and were instrumental in the development and completion of this work.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

With the ever-increasing importance and reliance on computers in our society comes the unnatural burden of interacting with those systems. This increase in human-computer interaction has, in turn, led to a marked increase in research on modelling such systems against human behaviour in an effort to enable more natural interaction. For this to succeed, these systems must have at least a basic level of *emotional intelligence*.

Emotional intelligence is defined by Salovey *et al.* (2004) as having four branches: the perception of emotion, emotions facilitating thought, understanding emotions, and managing emotions. These will be discussed below, with the exception of emotions facilitating thought, as this assumes the ability to think independently, which current computer systems cannot.

The *perception* of emotion is the ability to recognise emotion in oneself and others. These perceptions generally come from three channels: sight, sound, and language or contextual information present in text or prose. For example, a person may recognise that his or her friend feels distraught by the expression in the face or the tone of the voice. The perception of emotion also covers the recognition of emotion in oneself. An emotionally intelligent being is aware of the emotions expressed in itself at any time.

Following perception, an emotionally intelligent being must be able to *understand* emotions and emotional characteristics in order to correctly process and respond to emotional information. This consists of the knowledge of how emotions relate to one another, what causes them, what follows them, etc. Take, for example, a person who becomes angry at him or herself by missing the bus to work before an important meeting. The ability to determine the causes of this anger (e.g., the bus that is missed) is a critical part of emotional intelligence. An emotionally intelligent being will be aware of emotional changes and their nature.

Emotional understanding is a prerequisite for *managing* emotions. An emotionally intelligent being is one that can be open to all types of emotion, reflect on them, manage them in oneself, and engage, prolong, or detach from an emotional state in oneself or others (Oatley, 2004). A hypothetical situation may involve a doctor tending to a critically injured relative. The doctor must manage his or her emotions in order to operate in an effective manner.

Humans feel most natural communicating with other humans because the extra information conveyed in their emotional expressions can be recognised, processed, and reflected. This information is conveyed through several modes: facial expressions, vocal properties, bodily gestures, and behaviour. This added information helps people understand each other and interact more naturally and efficiently.

The work in this thesis is dedicated to the *perception* of human emotion from the prosodic properties of speech. In other words, this thesis aims to build a system that can capture and interpret the vocal expression of emotion in humans. More specifically, we seek to improve on traditional emotional speech classification methods using ensemble or multi-classifier system (MCS) approaches. We also aim to examine the differences in perceiving emotion in human speech that is derived from different methods of acquisition. For example, how is the perception of acted emotion different from that of spontaneous or naturally occurring emotion?

1.2 Research motivations and applications

There are wide-ranging applications for emotionally intelligent systems in real-world situations. Taking advantage of the emotional information in speech allows more effective processing of the contextual (language) information and a much more natural interaction between humans and machines. The following are some examples of how emotion recognition can yield improvement in the field of human-computer interaction. Figure 1.1 shows the relationships between vocal emotion recognition and potential application areas.

1.2.1 Health and public safety

Situations in which public safety is a major issue would greatly benefit from real-time automatic affect recognition. For example, such a system could be placed in the cockpits of airliners, oceanliners, and buses, where one or two principal operators control the fate of the vessel. These systems would be used to detect pilot boredom, inattention, or fatigue (Pantic and Rothkrantz, 2003). In private vehicles, detection of anger could reduce incidents of road rage by alerting the driver and trying to make them aware of the situation (Fragopanagos and Taylor, 2005).

Affect recognition could avoid concerns of having observers constantly monitoring or recording in situations where security or safety is of concern. For example, in hospitals, closed-circuit security systems, prisons, etc. (Pantic and Rothkrantz, 2003). These systems could alert personnel to certain situations such as disputes, accidents, riots or fighting.

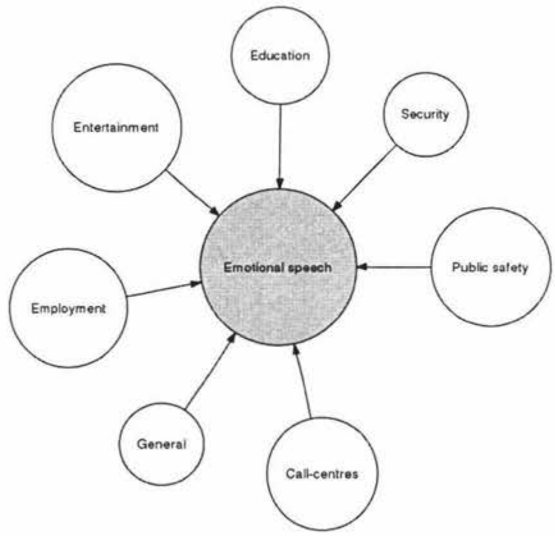

Figure 1.1: Applications of vocal emotion recognition

1.2.2 Education

Perception of human affect is important in areas where subjects are being taught or instructed. Human teachers can recognise student boredom, fatigue, and confusion and are then able to take steps to revive attention levels, or perhaps terminate the instruction if too many students are unable to process effectively.

Emotion and affect recognition from speech would be beneficial in an automated tutoring environment. The system could determine the affective states of the students and depending on how well they appear to be learning, or based on feedback (levels of frustration, confusion, boredom, fatigue, etc.), adjust the rate at which the information is presented to make the learning as efficient as possible (Picard, 1997).

1.2.3 Fraud and crime prevention

Voice profiling is directly related to vocal affect recognition. Voice profiling aims to classify speech samples according to predefined psychological profiles. These profiles can be generated or trained on pathological examples.

The use of voice profiling for fraud detection can be a useful measure to reduce the number of fraudulent insurance claims for insurance companies. The time needed to process claims can be reduced if claims that are potentially fraudulent are eliminated early on in the process. A system could be easily developed that allows claimants to provide information about their claim over the telephone with a disclaimer stating that their voice profile will be analysed for signs of fraud. If the analysis comes back positive for possible fraud, the customer can be notified of the result and offered an opportunity to retract their claim without penalty. Such a system does have obvious drawbacks, for example people may be discouraged from submitting a valid claim over fears of a false-positive from the voice profile analysis.

Another practical use of voice profiling would be for police and security in interviewing suspects for criminal cases. Suspects could be interviewed and their speech analysed by profiling software that could detect pathological patterns correlating to lying or nervousness. As with the above scenario, however, there are many ethical issues relating to this application and its output would have to be used only as one of many sources of information during interrogation.

1.2.4 Leisure and entertainment

An area ripe for new applications of emotion perception is that of leisure and entertainment. Here, the technology is applied in anecdotal ways. An example is the Sony ERS-7 Aibo Entertainment Robot. This robotic pet dog learns from interaction with its "owner" and can express different emotional states.

Computer video games are the result of billions of dollars of research and development investment aimed at making the player feel like he or she is experiencing reality. Emotion detection and synthesis in these games could greatly improve the gaming experience. Online games such as Everquest where human players interact with other human and computer players can benefit from both emotion recognition and synthesis to enhance the experience. Interaction with computer characters is often unnatural due to the lack of emotional understanding on the part of the computer character. Adding an affective element to these characters would introduce an entire new level to the gaming experience, providing a much more natural environment that would more closely model reality. This can be accomplished by integrating speech and facial expression recognition using cameras and microphones to measure the human player's affect. This affect can then be transmitted to other human or computer players in the game (Nakatsu *et al.*, 1999).

The research of Breazeal and Aryananda (2002) has primarily focused on the integration of a multi-modal emotion classification system in a robot. This robot, named Kismet, responds to caretakers by way of sight and sound. An integrated affective intent classification system allows the basic recognition and modelling of primary emotions. The robot approximately models an infant that responds to affirmation, prohibition, attention and soothing. After more research, this could be extended to a more full set of emotions or affective states allowing the robot to interact naturally with human operators.

1.2.5 Employment

Voice profiling can help streamline the processing of job applicant interviews. By interviewing applicants through an automated telephone system, the responses can be analysed for specific qualities which can be mapped to different positions within the company. For example, if a company is screening applicants for job openings in multiple departments, e.g., sales or customer support, the applicants can be automatically sorted into groups based on how their voice profile fits the target profile for each category. Job positions where an employee is constantly interacting with customers may require specific voice qualities. An applicant with a monotone pitch contour can be screened out automatically, and an applicant with a melodic pitch contour can be placed in a sales category for further inspection. Such a system would not be designed to completely take the place of human interviewers, but can greatly reduce the time requirements for selecting candidates.

1.2.6 Call-centres

Last, we look at applications of emotionally intelligent systems in call-centres. This is the primary focus of the end result of this research. Call-centres often have a difficult task of managing customer disputes. Ineffective resolution of these disputes can often lead to customer discontent, loss of business and in extreme cases, general customer unrest where a large amount of customers move to a competitor. It is therefore important for call-centres to take note of isolated disputes and effectively train service representatives to handle disputes in a way that keeps the customer satisfied (Petrushin, 2000).

Additionally, a team lead or manager may want to inquire on the status of any currently active calls in order to help coach new or inexperienced CSRs. Additionally, a manager can use the information provided by a spoken affect recognition system in several other ways. First, if such a system is deployed with each CSR, then a manager or senior member of staff can preview the emotional states of every caller at once, having an "overview" snapshot in real time. Other uses include the generation of statistics on the number of angry or upset callers each CSR has or whether any CSRs are being angry at the customers. This can lead to action to correct this behaviour or find things that a CSR can improve on and in turn help the call-centre more effectively manage the customer base.

Automated telephone systems are another potential application area that humans find themselves interacting with more and more. These systems have speech recognition units that process user requests through spoken language. A spoken affect recognition system can help process callers according to perceived urgency. If a caller is detected as being angry or confused in the automated system, their call can be switched over to a human operator for assistance. This could be particularly useful for the elderly who can often be disoriented when interacting with automated telephone systems. Petrushin (2000) built a system to monitor voice-mail messages in a call-centre and prioritise them with respect to emotional content. Such systems can make interaction with automated call-centres more efficient and less daunting.

1.3 Methodology

In this section we present the methodology followed during the development of this thesis. Figure 1.2 shows a flow diagram describing the methodology. Because the research focus is primarily a classification problem, that being the classification of different emotions, the methodology followed is much like any other classification problem. The first step is a review of the literature relevant to the field. Previous research on automatic emotion recognition was surveyed to build a knowledge of the state of the art.

Once a general knowledge of the state of the art was achieved, data had to be collected. Fortunately, a natural speech database was provided through the partner company for this project. A second speech database was collected from a previous study on emotion research (Nwe, 2003). Unlike the natural set, this database used actors and actresses. This provided a way to compare the classification methods on different types of data as well as investigate inherent differences between the two datasets. To gain a ground truth on the natural database, a system was developed to allow human listeners to judge the emotions present in the database.

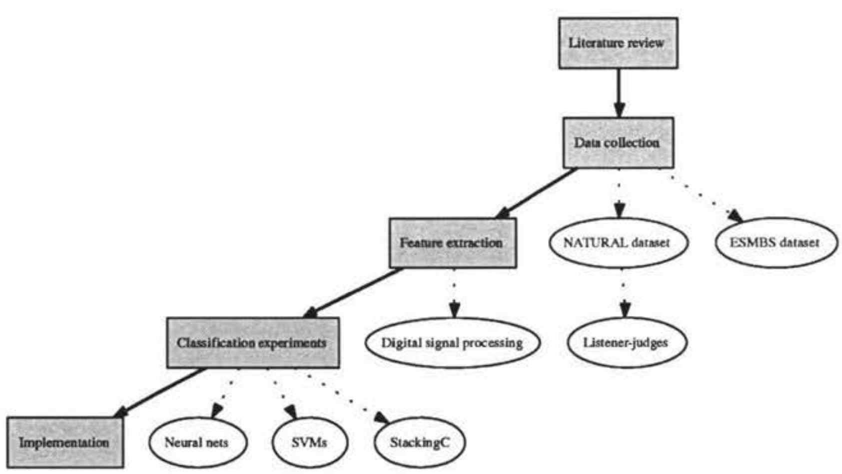

Figure 1.2: A flow diagram of the methodology followed for this thesis.

Next, characteristics of emotional speech from the existing literature were reviewed. Prominent psychologists such as Klaus Scherer who have explored emotion research for many years provide a strong basis for this area. These characteristics were extracted and compiled into feature vectors. These feature vectors describe the most relevant characteristics of emotional speech. Briefly, these include the fundamental frequency, energy, and formant frequency contours as well as features relating to rhythm such as the rate of speech.

Classification algorithms were then reviewed. As a starting point, artificial neural networks were experimented with, as they have proven quite useful in previous studies. These are subsequently improved upon using support vector machines. Feature selection techniques such as forward selection, genetic search, and principal component analysis were compared to reduce dimensionality in the feature space.

We then tested novel ensemble classification approaches in this field of using stacked generalisation and a simple voting scheme. Stacked generalisation takes as input base-classifier predictions and target classes and attempts to predict when the base-classifiers are incorrect. The voting scheme takes the predicted classes from each base-level classifier and determines the class with the greatest popularity.

The last step was to build an implementation of the theoretical system. This took all previous steps, the algorithms for endpoint detection, feature extraction, the use of the feature selected sets, and classification and brought them together into a single, modular system. This application reads input from a microphone or WAVE file and outputs a prediction based on the recorded speech sample. A modular artificial neural network functions as a plug-in to facilitate efficient replacement. Figure 1.3 shows the data flow for the emotion recognition system.

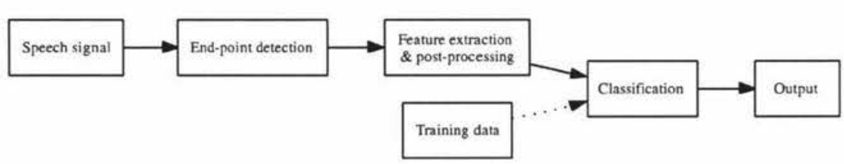

Figure 1.3: A data flow diagram of the real-time emotion recognition system.

1.4 Structure of the thesis

This thesis is organised as follows. In Chapter 2, a brief history of emotion research and theoretical representations of emotion are presented. This chapter also introduces the expression of emotion in humans and lists previous work in automatic spoken emotion recognition. Some areas which require additional attention are defined.

Chapter 3 presents the three data acquisition methods that are applied to vocal emotion research. Next, the two emotional speech datasets used in this research are introduced. The first database is collected from a call-centre and consists of natural interactions between humans. The second database is collected from non-professional actors and actresses. The advantages and disadvantages of each collection method and how it affects the research are discussed in detail.

Different emotions induce different physiological changes in the body, which in turn directly affect prosodic patterns in speech. Chapter 4 formalises and reviews correlations and characteristics of emotional speech.

Building on Chapter 4, Chapter 5 explores features chosen to describe emotional content contained in speech. These features are taken from previous research and experimental features based on the formant frequencies are investigated.

Chapter 6 introduces several classification algorithms used in this research. These algorithms are compared against each other in an attempt to reveal the most efficient and suitable candidate for use in the system. Feature selection methods are also compared. Next, we introduce two ensemble techniques: stacked generalisation and unweighted vote.

Chapter 7 presents the experimental results based on the classification and feature selection algorithms described in Chapter 6. This chapter also offers an in-depth look at the building of a prototype emotion classification system. The system is developed using existing algorithms and is brought together using C and C++. It functions in real-time and performs automatic classification via a modular artificial neural network.

Finally, Chapter 8 presents a conclusion and directions for future work.