ENGINEERING CAPSTONE PROJECT SPRING SEMESTER 2003

UTS

Faculty of Engineering

Emotion Detection of Speech in Telephony systems

by

Denis Bernard Ryan

Student Number: 97055507
Registration Number: S03 - 154
Major: Telecommunications Engineering

Supervisor: Anthony Kadi

This project is submitted as a mandatory requirement for the
Bachelor of Engineering
Telecommunications Major.

Credit Point Weighting: 9 points

CHAPTER 1 Project Description

1.1 Synopsis

This Project was undertaken as a mandatory requirement for the Bachelor of Engineering Telecommunications degree as awarded by the University of Technology Sydney. The requirement for the Emotion Recogniser project is linked to the UTS 'telecollab' project which centres on technologies that improve Human – Computer Interface which will provide faster, more affordable communication mediums and protocols.

My supervisor Anthony Kadi, proposed the Emotion recogniser project after I indicated I was willing to undertake a DSP assignment for my capstone project. The project was very suitable as my employment background largely comprised of voice and telephony solutions.

The project is being conducted alongside other capstone students who are developing visual or facial emotion recognition systems. The project is essentially research based with a demonstration element in the form of a simulation and later a telephony based implementation. The technology is in its infancy, however there is currently a great deal of research being undertaken in this area. The technology draws parallels with speech recognition and is up against the same development hurdles. The hurdles include: gender and cultural differences, noise and other interference and high processing requirements.

My goals are:

- To produce a well research report that is informative, accurate and useful as a tool for continued study in the technology.
- To run experiments using a local population to verify findings and observe performance and behaviour.
- Implement the system onto computer platform to help verify theoretical findings and prove that the emotion detector is a viable technological product.
- Interface the system to an IVR for demonstration purposes enabling easy phone access for interested parties.

1.2 Abstract

Human to machine interaction that is artificially intelligent, efficient and even empathetic is underway in its latest stage of evolution. Natural communication with machines will distance the past when people used toggle switches and punch cards to converse with their computers. This exciting technology will allow machines to analyse and determine the very trait that defines our humanity – our emotions.

This engineering thesis is based on research, experimentation and implementation of Emotion Detection in Speech technology. It is the result of 4 months research involving the gathering of publications and journals plus the undertaking of two voice capture experiments. There is also the design and implementation of a phone based prototype system that demonstrates the technology.

Although the technology is in its infancy, its applications in the new millennium are becoming very apparent and the push is to establish a reliable evolution of the technology. Experiments over a large cross section of people under a range of conditions help to highlight and define problems and obstacles, and pave the way to solving problems associated with this new form of human to computer interaction.

Emotion detection complements speech recognition and will prove to be a major benefit to business, society and to security services. Soon its development over the next five to ten years may let us experience machines being empathetic towards their human controllers.

1.3 Acknowledgments

This project would not have been possible if not for the support and technical assistance from several parties.

Telsis Pty Ltd

www.telsis.com

Telsis is a leading manufacturer and supplier of high quality telecommunications equipment. Their “fastIP” platform was used for the running of the experiments and demonstration services. Many thanks to the Sales Manager Kim Blacker, and technical support team, John Baily and John Szybowski for their undying technical input.

Dialect Solutions Pty Ltd

www.dialectsolutions.com

Australia's premier Web based Financial Transaction and Information Provision company. Dialect Solutions (formally News Connect) Supplies technology and services the clients globally. Many thanks to CEO Gareth Gumbly, technical manager Paul Rylands and operations manager Gabriel Kerninghan for the company's participation in the emotion capture experiments.

Table of Contents

Index of Tables

Table of Illustrations

CHAPTER 2. Introduction

2.1 Background

Electronic computing has been a part of industrial, technical and social culture for over 60 years. Computers assist the human species in solving problems, increasing productivity and help manage complexity.

Throughout the last 60 years of computing development, engineering energy has been spent in making computers more accessible to a broader spectrum of people. Human - machine interface is fundamentally difficult as the two entities are both complex and very different in design and nature.

Human beings are complex, emotive, biological organisms that process input via sensory processing regions in the temporal regions of the brain. Processing is based on interpretation of whats seen , heard, felt or smelt. Many of these inputs are coloured by associations and emotional states. On the other hand computers, while a complex entity, have rigid and well defined means of accepting and delivering information.

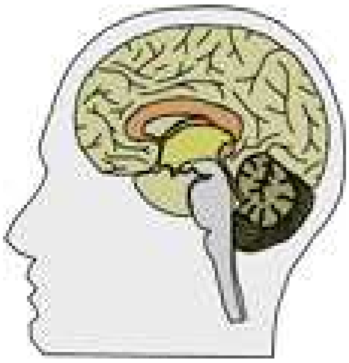


Illustration 1 The Human Brain. The ultimate supercomputer

Human – Machine interface has developed extensively, with information delivered originally by switches and load buttons then to punch cards, keyboards, pointing devices and more recently via speech recognition.

The latest technology of speech recognition provides a natural form of human machine interface. At last machines are able to interpret information contained in the words that are spoken to them, however they still cannot interpret other, possibly vital, information that is delivered with the spoken input. This extra information may be emotional stress depicting urgency, anger, confusion or elation. In many cases this input may be more important than the actual spoken words .

2.2 Project Outline

There is a number of ways we can detect a person's emotional state. Every day we, as human beings, assess the emotional state of others when we interact with them. At a conversational level whether its face to face or over the phone we can "listen" to emotion in their speech. How many times do we comment: - "You're amazingly chirpy today", "You sound stressed" or "Is something wrong – you sound so glum". We do not need to interact face to face with the person to analyse these states – over the telephone one can detect emotion simply by the level and intonation of the voice.

The chosen vocabulary also is a strong indicator of emotional alignment. An angry person may curse or use obscene language while an elated person may repeat words such as "great", "fabulous" or "wonderful". Everyday words we use can indicate many things depending on their context.

The project I have chosen to undertake for my engineering capstone is based on the development of a computer based system, designed to interpret human emotion via the spoken word. There has already been quite a lot of research in this area being conducted by individuals and specific organisations alike. The aim is to develop an economically viable system that can reliably interpret emotional state through voice, for the broadest spectrum of human users as possible.

The project will be conducted in two stages. First involves research and a simulation. The simulation will be able to process recorded .wav files and produce a reliable result based on the emotional quality of the speech recorded. Stage two involves the transfer of the system to a development platform and using an IVR telephony route as an input and produce a reliable emotional analysis of the calling party's speech.

2.3 Thesis Statement

Emotion Detection in speech is possible with modern computing systems and its application is pivotal in the evolution of human computer interface. Further to this, emotion detection in speech has many commercial applications especially in the arenas of telephony, security and entertainment.

2.4 Definition of terms

To get the best out of the technical explanation of the system a few definitions are required. Acoustics especially phonetics have many characteristics and definitions of these characteristics.. many of these definitions are misused or inaccurate. Below is a table of definitions pertinent to speech and speech acoustics that are used in this technical report.

Term	Description	Comment
IVR	Interactive Voice Response	Telephony equipment design to interact with Humans via the DTMF button on a telephone
SIR	Speaker Independent Recognition	Technology enabling the recognition of words from any speaker's spoken verse. Latest evolution known as Natural Language Recognition (NLR) which uses a catalog of phoneme sounds to construct words.
OR	Logical OR	A logic rule that delivers a true result when any one of two or more inputs is also true.
EDE	Emotion Detection Engine	The system which is comprised of the computer Hardware, operating System and Emotion Detection application as used in this research.
SR	Speech Recognition	The process where by a computer can translate words spoken by a human subject into unique symbols.
SRE	Speech Recognition Engine	A system which is comprised of the computer Hardware, operating System and Speech Recognition application.
DEL or PSTN	Direct Exchange Line or Public Switched telephony Network	The standard analogue telephone such as a home telephone.
E1 ISDN	Digital Telephony Interface over the ISDN Network	30 phone lines delivered via a 2 Mbit digital channel.
PRAAT	Speech Analysis Application	Down loadable free ware
Pitch	The Fundamental Frequency in relation to human voice studied in this research	

Term	***Description***	***Comment***
Fundamental Frequency	Generally acknowledged the same as Pitch. Also known as F0	
Jitter	Defined as the relative mean absolute 3rd order difference of the point process.	
Shimmer	Defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.	
Pitch Contour	The pitch contour describes the changing pitch level during the execution of speech.	
Rhythm	The patterned, recurring alternations of contrasting elements of sound or speech. [www.dictionary.com]	
Formant	Poles or points of resonance in FFT terms corresponding to resonant areas in the vocal tract. Contributes to the timbre or quality of the speech	In this research formants are the characteristics of each individual's throat, tongue, nasal passage and sinus cavities that contribute to the characteristic of speech.
Octave	A multiple of the Fundamental frequency or pitch	
loudness or Intensity	" intensive attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud" (The American National Standards Institute (1973))	May be the amplitude of the composite wave envelope or the energy associated with a particular frequency which the listener perceives as *loud*
AGC	Automatic Gain Control.	This mechanism is designed to compensate for low level speech by boosting gain and high volume through attenuation.

Term	***Description***	***Comment***
Glottal	"Of or pertaining to, or produced by, the glottis; glottic. Glottal catch, an effect produced upon the breath or voice by a sudden opening or closing of the glotts." [www.dictionary.com]	Major player in human speech and its manipulation contributes to the emotional message in speech.
Intonation	The pattern of pitch and intensity of the speech	
Prosody	The rhythm and intonation in respect to how an utterance is voiced.	

Table 1 Definition of Terms

2.5 Application of the technology

Initial reaction to research in the area of detection of emotion more often than not is: “that's nice but what would you use it for?”. The initial reaction to new and unexplored applications of technology undoubtedly receive such judgment because the need is not obvious. However one must understand the direction of human – machine communication which is heading down a natural communication path. The fundamental quality that is an underlying feature of our species is our demonstrable emotions which are a major player in our communication and our existence in general. It would therefore be unwise, even ignorant to ignore its significance in each and every facet of human life. The need always justifies the technology.

Entertainment has long been the prime driver for computer technology. Ever since the inception of the “adventure” game in the 1960's the computer has catered, and more recently been developed for entertainment applications such as 3D games. Sony's AIBO robot dog claims to be a substitute pet , replacing its biological counterpart by having its dog like qualities yet without the overhead of the care normal provided to a living creature.

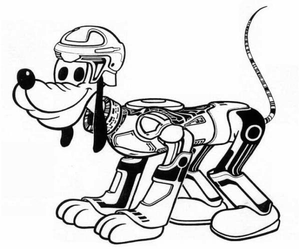


Illustration 2 Characterized Sony AIBO robot toy.

Courtesy Disney Productions

7 The latest incarnation of this product is presently been developed with emotion recognition technology at the forefront of its design. It is necessary that robotic pets can recognise emotions expressed by the humans who are interacting with them. (Oudeyer Pierre-Yves, 2002 [7]).

3 Other relevant studies have been done where real time recognizers , using neural networks have been developed for call centre applications (Lee, Narayanan, Pieraccini, 2002 [3]). In this type of application the requirement is for a categorization of negative and non negative emotions that may be acted upon automatically by the system. Angry or even suicidal callers call be prioritised and dealt with accordingly.

An *actor's training box* is one concept and a possible line of development after the research project is complete. The idea came from the running of experiments where trained actors were invited to participate. Not surprisingly trained actors tended to exhibit the best results, confirming that they are more capable of spontaneously generating emotion in speech than the average person. A firmware based processor with a small microphone and LCD display could be used to give instant feedback to a training actor. An icon such as a smiley or a grumpy would display illustrating the devices perception of the emotion in the speech.

Perhaps the technologies biggest application will eventually be the "naturalisation" of human to computer interaction. By exploiting the full message content of the spoken work interaction with machines will be faster, more reliable and possibly very enjoyable.

Illustration 3 Keyboard and pointing device of the future.

Courtesy: Telix Products inc.

CHAPTER 3. Literature Review

3.1 Outline

Research into speech based interaction with computers and human emotion recognition systems has been an ongoing study undertaken for many years. Research into human communication reveals a great deal about how and why we communicate. While the methods of communication vary, emotion is instrumental in tainting the meaning of prosodic verse. In reviewing a number of journals that are written about human speech analysis and emotion recognition, this paper summarises the key points under seven categories:-

1. Applications (To Emote or not to Emote)
2. Definition and qualification of emotions (Emotional Insight)
3. Technical Overview (Emotion Detection for Dummies)
4. Voice Features Extraction and Pattern recognition(Emotional Breakdown)
5. Algorithms for feature extraction used (Emotive Maths)
6. Mechanisms For Improving Reliability (Making it Work Better)
7. Results From Experiments conducted to date. (Emotions Quantified and Qualified).

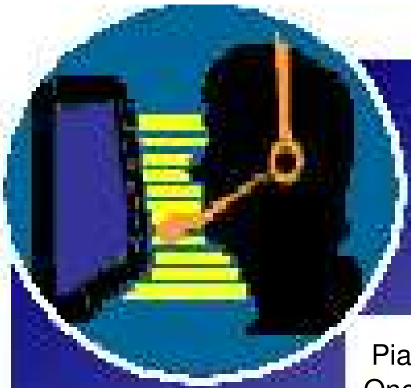
For each category I propose my own interpretation and add comments to the material described.

3.2 Review

3.2.1 To Emote or not to Emote

The concept that a computer being empathetic with a human subject may be difficult for many of us to imagine. However, the fact that a array of silicon chips and wires is not a self aware entity, does not undermine the possibility of it being capable of interacting with people in a natural and empathetic manner.

How can such systems it achieve this and what is the need for such an animal? What emotions do we need these devices to detect? Kollias and



Piat, [2] describe communications between humans as having two channels: One transmits explicit messages which may be about anything or nothing and the other transmits implicit messages about the speaker themselves.

This suggests that the emotional content of the message may contain information that would otherwise be lost if not detected. Polzin and Waibel put forward the argument that a speech recognition system is only partially effective if the system pays attention to *what's* being said but ignores *how* it was said([10] - pp1).

The goal of an emotion recognizer is to identify emotional states. One key commercial application maybe for call centres or “customer care” lines where irate or unhappy clients can be dealt with appropriately. ([3] – pp1)

Research done by Microsoft in China has revealed adequate reliability can be achieved with four emotions (happiness, sadness, neutrality and anger) with the technology being applicable to natural human-machine interface [4].

Also research into this technology has also been funded by DARPA and WRAIR (Walter Reed Army Institute of Research) concentrating on the detection of stress in speech. One possible application of the technology is the filtering of calls to detect terrorism or other critical situations [6].

The Sony corporation has also been investigating the application of emotion recognition for the novel use in its toy robot products. Pierre-Yves Oudeyer, an employee of the French subsidiary of the corporation, compiled a research paper on the technology and its possible applications. In particular the AIBO robotic dog product will pioneer the use of this technology to create a more realistic “pet” that responds accordingly to the owners commands and instructions.[8]

3.2.2 Emotional Insight

The characteristics of human emotion have been researched and analyzed for countless reasons. The reasons behind emotion are exceptionally important when dealing with systems that are required to interpret them. Many papers go into fair detail about human emotions, why they exist and what is the result of emotional feelings. This leads to the question which challenges the need for such technology, why is emotion detection important?

Polzin and Waibel describe speech as been coloured by prosodic as well as spectral information. Prosodic construction of the speech are features such as pitch (frequency) and intensity (loudness), while spectral information are particular acoustic qualities that make the voice sound “pleasant”. They report on an experiment conducted by P. Lieberman and S.B. Michaels in their journal published in 1962 [13] , that while humans receive 85% of information in normal communication, only 47% of information is received if spectral structure is striped out and only prosodic information is preserved.

The scope of human emotion is wide and varied and may be classified in several ways. A paper by Lee, Narayanan and Pieraccini simplifies emotions by placing them in one of two categories, focusing on *negative* and *non-negative* emotions ([3], pp1) and ([5], pp1). Possibly many commercial applications will require a high reliability detector for distressed or angry people. Distraught clients who are detected of being in an angry or frustrated state of mind may be isolated and dealt with accordingly.

The “activation – emotion” space is the entity that determines how a particular situation arouses the mind and the degree of influence causing the person to act in response. Two key themes in activation-emotion space are proposed by Kollias and Piat are valence and and activation level ([2], pp3). They describe valence as a central concern with the positive or negative *evaluations* of people and events with directly influence's a person's emotional state. Activation level involves a person's disposition to *act* in a certain way and the the strength of the disposition to take action rather than do nothing. The activation – evaluation space can be represented as a graph with the horizontal and vertical axis representing evaluation and activation respectively.

Investigations into the emotional state of the speaker has been widely studied in psychology and psycho-linguistics. Polzin and Waibel describe research conducted on how acoustic and prosodic features (speaking rate, intonation and intensity) can encode the emotional state of the speaker. They propose that when studying emotions one needs to make the following distinctions:-

- the emotional attitude of the speaker towards the hearer,
- the emotional attitude of the speaker towards the message, or
- the emotional state of the speaker.

The meaning of these observations has to be considered when contemplating that communication may possibly be a machine as the receptor rather than another human being.

3.2.3 Emotion Detection for Dummies

Many journals studied for this project propose methods and mechanisms for the implementation of a voice based emotion detection system. From a modular point of view many of these systems are uncomplicated, however the rather involved mathematical side is examined in Section 3.2.6 . The question is:- how does a machine know what emotional state a person is in?

The first step is to get speech into a form that we can study. This involves the breaking down into "features". Dellaert, Polzin and Waibel ([12] pp- 1) summarises voice features and the methods used to gain the useful information from them. This includes maximums and minimums as well as the distance between these extremes of features such as pitch and intensity. Information attained from signal slope and the speaking rate are also described as significant input information.

These features are compared to preset data points including contour curves and a "best fit" method is used to derive a result. There is also discussion regarding the search for better features to improve accuracy. Smoothing functions known as cubic splines offer enables the measurement of many new features based on the pitch, pitch derivatives and the behaviour of maxima plus minima over time. (Dellaert, Polzin and Waibel ([12] pp- 2).

Cubic splines are also explored by Nicholas van Rheede van Oudtshoorn ([11] pp 5 – 6) in his thesis study. He explores a process that used two different waveforms to compute features. The first is the standard wave representation of of the spoken utterance. Then to smooth out irregular features a second wave is created using cubic splines which enables the extraction of new features. The paper also discusses and tabulates 13 sub-features derived from pitch and amplitude and divides them into two separate groupings. One being features that are compared to their neutral counterparts and the other whose distance from similar features is calculated.

3.2.4 Emotional Breakdown

Human speech is as complicated as it is varied. Many acoustic aspects contribute to what we perceive as a person's voice. The human voice varies greatly from person to person and is further influenced by age, gender, culture and state of mind. This great diversity can make emotion detection challenging, however certain features appear common to particular emotional states. Many papers describe in detail the technical challenges posed to emotional detection and the voice features that may provide solutions.

Consensually pitch and intensity are looked upon as the two major voice features that provide the information for which emotions can be derived. Polzin and Waibel highlight these features in their paper sectioned "non-verbal information (Prosody)" ([10], pp2). Here the mean and variance of these features is offered as information related to emotion detection. Spectral information commonly known as "voice quality" also provides information on a person's mental state.

Although this is particularly difficult to map on a speaker independent basis due the extremely high diversity of voice qualities, a change in this state may provide valuable input for emotion recognition systems. Finally the journal described contextual or "verbal" information which is the meaning of words and how they may indicate the speaker's expressed emotion.

Nicholas van Rheed van Oudtshoorn proposes the use of extra features to improve certainty of detection. In his research thesis he has observed that many researchers have placed a high emphasis on pitch but less so on amplitude.([11] – pp1). The use of base amplitudes which is the intensity of voice when the speaker is emotionally neutral, may provide the solution to scaling the technology across a broad cross section of people. It may also be used as a predictive mechanism that determines a state change. Certain pitch and amplitude based features are considered for "distance" calculations

which essentially is a comparison to an emotional vector. The closer the utterances lines up with the predetermined vectorial template – the "closer" the speaker is to a particular emotion. Further to this Nicholas has investigated pitch contour and jitter as features that provide further differentiation and hence improving detection accuracy.

The goal of an automatic recogniser is to assign category labels that identify emotional states (Lee, Narayanan and Pieraccini [3] pp 1). Once specific voice features are extracted the recognition system has to reliably determine the emotional content.

Classification or pattern matching systems vary in their technique and performance. Lee, Narayanan and Pieraccini [3] compared classifier methods in performance based on error rates pertaining to key voice features. This comparison shows the strengths and weaknesses of different classifier technologies in relation to key voice features.

Results showed different performance of classification system in relation to male as opposed to female voices. Linear Discriminant Detection (LDC) generally surpassed k- NN (k – Nearest Neighbourhood) classifiers on most comparative tests. It was found female speakers encountered a lower error rate than their male counterparts, however for male speakers the base feature error rate was actually better with the k-NN classifier than the LDC classifier. ([3], pp 3).

3.2.5 Emotive Maths

Many complicated algorithms are used for feature extraction and pattern matching. Although this project used an utility with built in facilities to extract voice features, a background knowledge of how aspects of speech such as pitch are derived. Some publications provided algorithmic background describing extraction techniques.

A big problem with mathematics is generally that it is difficult to comprehend concepts if demonstrated in an obtuse or even arrogant manner. Holger Quast's journal [18] keeps to the facts and does an exceptionally good job of explaining feature derivation from the speech signal. Holger explains the make up of the speech signal s(t) in terms of the fundamental frequency p(t) (F0) , the impulse response of the vocal tract h(t) and the use of convolution in the time domain.

S(t) = p(t) convolve h(t) = integral (inf – 0) [p(t – t) * h(t)] dt

Feature extraction methodology is explained with discussion about “cepstrum” (an anagram of spectrum). Essentially cepstrum analyzes the signal in the frequency domain, where by it was windowed by 1 – cos2(x) before the FFT operation to remove strong spectral splatter that is associated with hash speech audio. The speech signal is the product of impulse response and excitation pulses. (excitation pulse originates as a short puff of air released through the glottis, the opening through the vocal cords)

Paul Boersma produced a paper in 1993 which gives detailed description on the methods used by the PRAAT speech analysis software to extract pitch from the voice signal. A four part summary that describes the 9 part algorithm implemented in the programme, is included as well as information to methods for autocorrelation and the determination of periodicity. Periodicity is the occurrence from local maxima and minima from which the pitch contour can be determined.

Another well explained paper is by Alain de Cheveigne [24] . He explores a mechanism to determine the fundamental frequency from one voice by searching th parameter space of two comb filters which yields an estimation of the component voices. The mechanism is computationally expensive but is believed to be highly reliable. Details of the process in obtaining F0 are well described in reasonably easy to understand technical format. Like the publication by Paul Boersma, this document would be exceptionally useful for the engineer who wants to write his or her extraction techniques or for those who just want some mathematical background on the subject.

3.2.6 Making it Work Better

Of course its one thing to make things work but to make it work at a level that is regarded as acceptable is another matter. Many papers publish the results from experiments run on both computer and human emotion detectors with some surprising results in both cases. It seems, we humans are not at all perfect in this field. Unlike speech recognition systems which require an *exact* interpretation, emotion recognition systems will not be generally used to decipher verbatim the state of mind of its human subjects. Though this may be an application as the technology becomes fine tuned.

Major applications will probably involve the interception of any adverse , out of context or unexpected emotions for a given situation. This is not an exact science and is making the technology very difficult to qualify. However it begs the question:- How good does it have to be?

Dellaert, Polzin and Waibel discuss the use of pattern recognition techniques to maximise performance.. Three methods used in their research were Maximum Likelihood Bayes classifier, Kernel Regression (KR) and k-Nearest Neighbours (k-NN) with the k-NN returning the best results. They also attempt to improve classifier performance by using what is known as "Distance Metric Optimization".

The k-NN rule relies on a distance metric to perform its classification and it is expected that by changing this metric it will yield different and possibly better results. One needs to weigh each feature according to how well it correlates with the correct classification. A three dimension scaling map provides input to this technique.

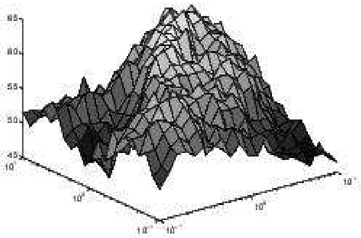


Illustration 4. K-NN weighting map

Oudeyer Pierre-Yves ([7] pp 26) included a mention of the PRAAT speech analysis software in relation to its measure of the pitch as being "known to be very accurate" and therefore ideal as the chief mechinism to an analysis engine.

3.2.7 Emotions Quantified and Qualified

Emotion recognition is regarded as leading edge technology which still is in research and experimental stages. Many of the papers review and tabulate results from test conducted on various emotion detection technologies.

In any experiment a reference or control is required to analyse results. Tests conducted using human subjects as well as computer based emotion detectors using speech combinations with and without spectral and prosodic information were studied by Polzin and Waibel. They concluded while accuracies based on automatic classification fell short of accuracies achieved by human subjects, they were significantly above chance level (33%)([10], pp 4).

In their research Polzin and Waibel [10 pp- 1] describe the mechanism they used in the automatic detection of expressed emotion. Prosodic features they studied include the mean and variance of the fundamental frequency with in utterance, which was normalised in respect to gender. Two features were used to approximate jitter which is a perturbation in the fundamental frequency (F0). Intensity was also factored in with the mean value and standard deviation being considered. Finally tremor which, like jitter, is a perturbation but of the intensity contour.

Prosodic features are multi functional - not only do they serve to express emotions but are found to serve a variety of other functions as well. The research found that the intensity and pitch contours varied considerably across speakers and proper normalisation was paramount.

The researchers used transcribed sections from English movies as data for the experiments([10] pp – 3). The data was re-synthesized with only pitch and intensity information from the original data, removing all other spectral information. The researchers found that the modified data returned accuracy of only 64% as compared to 85% they achieved with the full spectral quota.

More accuracy tests were conducted by Feng, Xu et al in the Microsoft Laboratory in China. They obtained results based on the accuracy of four separate emotions across a large group of people. Of these neutrality scored highly at 83.73%, next was anger at 77.16%, sadness ranked third at 70.59% with happiness trailing at 65.64%. They also compared three classifier mechanisms returning accuracy results between 27.28% for NN (Neural Networks) Happiness to 89.29% for k-NN for neutrality. Results from these tests agree with the results published by many other papers about this technology.([4], pp 3).

A comparison of the top 20 voice features in relation to the information gained from them was tabulated by Pierre-Yves Oudeyer of Sony CSL, Paris. The information gain was derived by a mean across 6 test speakers with the results depicting features such as median and *mean intensity low* as providing the greatest informational gains.([7], pp 33).

Probability vectors can be used to qualify decisions based on particular features. Nicholas van Rheed van Oudtshoorn portrays the implementation of a probability vector which is assigned to the return codes of a neutral pitch based analysis ([11], pp 3). The table is split into 3 possible result categories 0, 1 and -1 which indicate an equal, greater or lower result respectively. Against each feature the respective column contains the three possibilities of a hit against a return category for a neutral based pitch.

CHAPTER 4 Emotions and the Human Specimen

4.1 Definition of Emotions

Human emotion is an underlying mechanism in relation to the species behavior, social interaction and survival. Classic definitions observe that it is emotions that separate men from beast, however a more accurate, modern observation may describe Homo sapiens as the "most emotive" species.

1

The dictionary definition of emotion is given as "strong instinctive feeling" (Australian Oxford Dictionary). In essence this describes a trait of emotion but does not really define emotion itself. In fact scientists, philosophers and theologians throughout the years have attempted to describe and define emotions with little agreement. What emotion is and what embodies it in the makeup of man is still an issue of scientific pursuit and controversial debate.

4.1.1 Emotion in Speech

2

Constructing an automatic emotion recogniser depends in a sense on what emotion is. Trying to model an entity that is difficult to define provides many challenges in itself. Emotions, like any other physical phenomenon can be researched, monitored and explored with results from experiments being correlated into what may be described as "common behavior". However this too is controversial, as people from different cultures may react differently when experiencing the same emotions (Kollias, Piat, pp. 2, Section 2.1).

2

So where do we start? It is fair to assume that emotions are a rather complex and inconsistent human trait. However basic emotional traits may be catagorised into "fundamental categories" each fundamental emotion having its unique traits. For whatever reason an emotion is felt its one direct influence is "arousal"(Kollias, Piat, pp. 2, Section 2.1). This arousal shapes and conditions the persons psych altering or "colouring" her

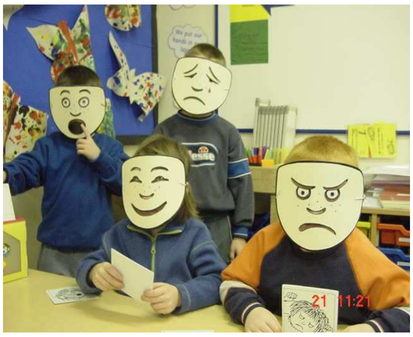


Illustration 5 We all learn to mask our emotions to help us succeed at life

or his actions. Still a point of controversy is what emotions are classified as fundamental and how many of them are there?

16 Visualisation of this can be achieved using *Descartes' Palette Theory of Emotion*, describing humans as having a number of basic emotions which, like the primary colours on an artist's palette, can be mixed into one of many possible emotional states (van Rheede van Oudtshoorn, 2003 [16]).

14 The first step in the study of expressive speech would be to determine what 'emotional speech' is? Intuitively emotional speech can be described as the speech coloured by the speaker's state of mind (Mozziconacci, S., 2003 [14]). Different prosodic cues, such as variations in pitch, intensity, speech rate, rhythm and voice quality are perceived by the listener as an emotional message , quite separate to, and in fact, possibly tainting the contextual message delivered in the spoken word.

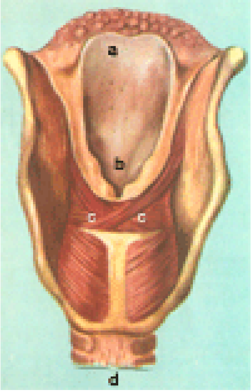
4.2 The Human Voice

4.2.1 Human speak

Human speech is one of evolution's marvels and quite possibly is the main reason why the species is dominant on Earth. Communication enables humans to work collaboratively by pooling resources and sharing loads. They can brainstorm, recall and explain historic events and send quite complex messages using relatively little energy.

4.2.2 How we speak

The opening of the larynx is called the glottis; when at rest, and with normal breathing, it is a triangular orifice through which the earth passes freely. When a sound is to be made the arytenoids muscles act so as to twist the arytenoids cartilages, and these stretch or slacken the vocal cords, which can assume about 170 different positions. At the same time the glottis itself is narrowed to a slit, the size of which regulates the rate at which air passes through the larynx from the lungs. This accurately controlled current of air makes the vocal cords vibrate and so sounds are produced. As produced by the vocal cords they would be too weak and faint to be heard, but the hollows of the respiratory system, in the trachea, the pharynx, the larynx itself, and the nose and mouth, act as resonators, which strengthen and modify the sound. The chest also acts as a resonator and amplifies the sound just as the body of a violin does.



a) epiglottis

b) aperture of the glottis

c) arytenoids muscles

d) trachea

Illustration 6. The human "voice box"

4.2.3 The voice

Voice is not the same as speech; many species have voices but only Human beings can speak. The sounds of the voice must be shaped to form words, which are made up of vowels and consonants. The shaping is done by the muscles of the mouth, palate, lips, and tongue. Vowels are accompanied by vibration of the larynx and the sound passes unobstructed through the mouth. Consonants are formed mainly by the alteration of the laryngeal sound by the tongue, teeth, lips and palate. A word cannot be made up of consonants alone, because most of these cannot be voiced unless a vowel precedes or follows them. Some consonants are called labials (Latin labia, lip) because they are formed by the lips; it is impossible to say b,p, f,m or v with your mouth open. Others (d,t,l,n,r,s, z,ch,j) are linguals requiring the use of the tongue (Latin lingua, tongue). G,q and k are gutturals, made with the back of the palate (Latin guttur, throat).

The pitch of the voice depends on the frequency of the vibrations of the vocal cords. If they are at normal tension the vibrations are about 80 per second (80 Hz); if the cords are tightly stretched they are more rapid, up to 1000 per second (1Khz).

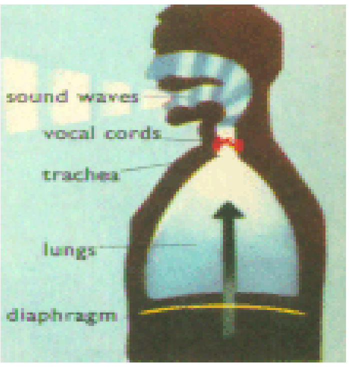


Illustration 7 Generation of sound through vocal tract

The diaphragm is pushed up against the lungs by the abdomen, like a piston. This forces the air in the lungs through the trachea, at the narrowest point of which lie the vocal cords. These break up the moving column of air by their vibrations, forming sound waves, which are transformed into normal speech by the hollows of the nose and mouth. (Extracted form “Know your body – Human Voice” [21]).

4.2.4 Psychology

The human psych is the underlying reason for emotion. Evolution has given us the ability to communicate our feelings both intentionally or sub consciously through our emotions. What we think is generally what we feel and the emotions we project are the window to these feelings and to our state of mind.

Human psychology is an exceptionally complex study beyond the scope of this journal, however it is important to consider psychological aspects and understand how they influence a person's speech

Speaking is a physiological process just like running, swimming and and even smiling. What makes speech so different is that it is the main tool in which we humans address and interface with the social mob. It is the basis for our survival and perhaps the reason our species dominates the Earth. A powerful but complex tool which requires thorough understanding to be able to completely analyse the total message content held within.

A major goal of this study is to derive a relation to the emotions contained in speech to physical aspects of the human voice. The better the understanding we have of these relationships, the more accurate and reliable recognition systems we can build.

4.2.5 Voice features

Voice features are the focal point of the experimental research outlined in this report. Essentially features are aspects of the voice that can be classified as independent entities that may or may not provide a consistent pattern during emotional speech. Some features provide more information that others in relation to extraction of emotions and it is these features that this report concentrates upon.

4.2.6 Variations

The are many variables to be considered when dealing with psychological elements. The sheer complexity of the human mind and the speech mechanism and the amazing diversity of humans and the places and cultures they come from further blur distinctions between what is or is not perceived as a particular emotion. One of these variations which is very close to home is gender. A person's sex dictates the pitch and versatility of their voice. Gender differences also plays and important part in the insertion of emotion into speech. Different cultures have slightly different ways of expressing emotion in speech. Another surprising find in this research is that different accents even differences between closely related accents, play havoc with an emotion matrix formulated on the one particular accent.

4.3 *Problems and Pitfalls*

4.3.1 Cultural differences

To most people it is obvious that people from different countries speak with different accents. The accent relates to the rhythm and emphasis and inflection of speech by the speaker. What maybe not so obvious is that emotion and its representation in speech audio is also heavily coloured by country of origin. In relation to the Shannon Weaver communication model which depicts communication as an exchange of messages between to entities – a source and a destination. Between these entities lie a transmitter and a receiver separated by a transmission channel.

In the case of human communication the source is the speaker and the destination is the listener. The transmitter is the speaker's voice and the receiver is the listener's ears. The medium my be the atmosphere for direct speech or possibly a telephone for electronically assisted communication. It is always the intention of the speaker to project the meaning of the conversation in a way that the listener interprets the meaning exactly the way he or she intended.

18 However the message at the sender end is not always interpreted as the same message by the listener.(Holger Quast, 2003[18]). For example a Swedish speaker that generates the impression of a happy extrovert person in a non-Scandinavian listener because of his or her fundamental frequency's strong modulation – which is a normal method of expressing language. A person laughing so hard that he breaks down into tears may be interpreted as unhappiness or grief. Even a human listener has to sometimes question the emotion they experience under these circumstances.

4.3.2 Noisy Public Houses

When conversing in a noisy environment such as an airport, public house or when working near machinery, the natural human reaction is to raise one's voice so the other party can hear us. This happens whether the conversation is face to face or takes place over the telephone. In exceptionally noisy environments we will raise our voice to the point of shouting. This may affect the pitch analysis and produce a false positive such as anger in place of neutrality. If intensity is used to analyse a phone conversation the loud volume generated by the speaker may also upset the applecart reducing the accuracy of the system.

Many Voice recording systems and IVRs perform a *Voice Measure* or *threshold check* when a call is first made to the system. This samples the ambient background noise to determine the threshold levels for automatic file recording and trimming. This parameter may be useful in adjusting (trimming) parameters in the emotion detection engine to compensate for the greater audio energy.

4.3.3 Automatic Gain Control

AGC is a feature found on many voice recording devices and conferencing systems. AGC simply adjusts the volume so soft speakers have their volume increased while loud speakers have their volume reduced. What eventuates is that speech intensity is maintained at near the one level. This causes a problem if intensity is used to help determine emotion. This problem is addressed in experiments 1 and 2 as the IVR uses AGC to regulate recording volume (See Section 7.6.1.4).

Another consideration is that people on speaker phones wander around rooms, and phone users sometimes move the microphone to and from their mouth whilst speaking. This produces the opposite effect to AGC where what should be a constant volume, actually varies dramatically. It is also very difficult to determine whether someone is speaking close to the microphone or if they are speaking form the other side of the room generating a similar problem.

It appears that using intensity to help analyse emotions in speech, especially in telephony systems, is quite dangerous, possibly producing misleading results. The research described in this report sidesteps intensity almost entirely, instead concentrating on 1st and 2nd order pitch features.

4.3.4 Foreign Subjects in Experiment One & Two

The emotion detection system performed poorly when dealing with subjects possessing strong accents such as Spanish or from the former Yugoslavia . This was more apparent when analysis with the simplistic 1st order analyser with much better results attained on the 2nd order analyser.

4.3.5 The SNATCH movie

During my investigations into this research I noted that many other researchers had resorted to obtaining experimental data from "English Movies." I still to this day am unsure whether this refers to the movies made in the English language or movies that were made in England. To play it safe I sampled script form an English movie made in England. The movie chosen was directed by Guy Ritchie and is titled "Snatch". The movie is based in Novae London and is blessed with highly charged emotional speech albeit in profound English accents.

An interesting observation was made when the speech data run through the emotion detection system. Verse that was neutral (the actor conversing in a calm , unexcited state) was interpreted by the analyser as being SAD! This would indicate that the primitive four featured analyser used for this exercise cannot deal with cultural variations. For such a lightweight application to work successfully it would need to be tailored to each region it was applied to. Research is required into a more robust mechanism, probably using more features and possibly pattern recognition which will be more forgiving across different accents.

4.4 Accuracy

How accurate is accurate? Before delivering a verdict to the is question we must take in account that the human listener is an imperfect emotion detector. The degree of accuracy varies from person to person. Some people are referred to as intuitive or have insight into people. This may be a quality they possess because they have superior mechanism to detect emotions projected by people. Other people project more emotion than others. Another old time saying is someone that can be read like a book. A person's intention may be entirely obvious because he or she either consciously or sub-consciously projects emotion.

This many variables will make it vary difficult to develop the 'perfect" emotion detection engine. What needs to be done is by using statistical information gathered from research build a model which deliverers the highest accuracy possible over the broadest range of people.

4.5 Summary

The human voice is a physiological phenomenon which is the net result of complex neurological and physical processes. It is a magnificent communication device, allowing the human species to communicate in a most efficient manner and its uniqueness helps define our individuality and persona.

Many features constitute a human voice, each possessing its own signature with can be identified with the emotional stresses projected in speech. These features are perceived by the human listener as an emotion and provides us with an idea about the state of mind of the speaker. These features can also be analysed via sophisticated audio analysis systems making it possible for computers to also interpret emotion encapsulated in speech.

A multitude of variations prevent this from being a straight forward process. Cultural and physical differences create a broad range of possible formulas that are required to reconstruct emotion. Different accents, cultural backgrounds and audio environments are create sizable hurdles which a high degree of sophistication is required to overcome.

The accuracy of a system that analyses human emotion is not easily quantifiable. Since the definition of emotion itself is rather blurred and the human receiver is generally far from perfect, a totally accurate and reliable emotion detection system may never by entirely achievable.

CHAPTER 5 Emotional Recognition Theory

5.1 Introduction

14

Communication amongst people is much more than an exchange of words. Linguistic, paralinguistic and non-verbal communication elements all comprise communication and they all convey meaning.

“Speech technology is a field in evolution” (14). With the development of our technical capabilities, science and engineering is capable of developing technologies that we previously regarded as unattainable even unimaginable.

5.2 Emotional Speech

14

Before we can begin the study of emotion in speech we have to first understand what emotion in speech is. Intuitively emotional speech may by defined as speech modulated by a speaker's state of mind.(Mozziconacci, S., [14]). However there is no commonly accepted definition and taxonomy of emotion. Contextually the term 'emotion' is the superclass of a large variety of expressions that a human being is capable of. The study in this report focuses on the prosody which constitutes the verbal expression in speech with labels such as joy, sadness, anger, surprise, disgust, frustration, interest, boredom and calm describing one or more possible emotional states.

5.3 Feature Extraction

5.3.1 Definition of feature

Human speech is a complicated mix of base tone harmonics, vocal cavity formants and glottal noises such as clicks and pops. Auditory elements combine to what we hear as a person's voice. The voice box or larynx is the underlying source of tone but is by no means the largest player in the creation of speech. Features are the individual voice components that can be extracted from speech and analysed on a stand-alone basis. That is components that have their own meaning independent from other speech features.

The speech is coloured by its underlying frequency, the change and standard deviation of the frequency, the position of the tongue, the contraction of neck muscles the vectoring of exhaled air through the nose and mouth as well as consonant sounds generated by tongue movement including rolling and clicking, perching or smacking of the lips.

The lungs as well as the size and shape of the cavities in the head provide resonance areas where one of many formants are created. The intensity of the sound which is a function of the air driven through the larynx and the standard deviation of the intensity are also relevant features.

Higher order features include the shape and pattern of frequency over time during the utterance of words. The shape of a speaker's pitch wave form known as the contour is a feature that is quintessential in adding emotional meaning to speech.

Other voice features including jitter and shimmer are other higher order features that relate directly to emotional stresses.

5.3.2 Methodology

Feature extraction from human speech is often a difficult process. Algorithms and techniques for extraction are complex and are contravened by added noise. Because these extraction techniques are so complex, computation is slow and real time analysis is therefore quite system intensive.

Pitch is a feature that is paramount to analysis of this nature. However pitch is more complex than one may imagine. It is not as simple as using a band pass filter or simple frequency domain analysis to attain the extract this feature successfully.

A number of methods may be used. Typically time domain methods are prone to windowing error causing problems in determining the height and position of the maximum This is important in determining the degree of periodicity (Harmonics to Noise Ratio) (Boersma, 1993). Usually frequency domain methods are used for extraction, however the “PRAAT” speech analysis software used in this research tackles the problem by utilising a robust and straightforward algorithm that works in the lag (autocorrelation) domain.

A summary of the complete 9 parameter algorithm as it is implemented into the PRAAT application is illustrated in the below extract (Boersma, Paul., IFA Proceedings 1993[15])

4 Algorithm

A summary of the complete 9-parameter algorithm, as it is implemented into the speech analysis and synthesis program *praat*, is given here:

Step 1. Preprocessing: to remove the sidelobe of the Fourier transform of the Hanning window for signal components near the Nyquist frequency, we perform a soft upsampling as follows: do an FFT on the whole signal; filter by multiplication in the frequency domain linearly to zero from 95% of the Nyquist frequency to 100% of the Nyquist frequency; do an inverse FFT of order one higher than the first FFT.

Step 2. Compute the global absolute peak value of the signal (see step 3.3).

Step 3. Because our method is a short-term analysis method, the analysis is performed for a number of small segments (*frames*) that are taken from the signal in steps given by the *TimeStep* parameter (default is 0.01 seconds). For every frame, we look for at most *MaximumNumberOfCandidatesPerFrame* (default is 4) lag-height pairs that are good candidates for the periodicity of this frame. This number includes the *unvoiced* candidate, which is always present. The following steps are taken for each frame:

Step 3.1. Take a segment from the signal. The length of this segment (the window length) is determined by the *MinimumPitch* parameter, which stands for the lowest fundamental frequency that you want to detect. The window should be just long enough to contain three periods (for pitch detection) or six periods (for HNR measurements) of *MinimumPitch*. E.g. if *MinimumPitch* is 75 Hz, the window length is 40 ms for pitch detection and 80 ms for HNR measurements.

Step 3.2. Subtract the local average.

Step 3.3. The first candidate is the unvoiced candidate, which is always present. The strength of this candidate is computed with two soft threshold parameters. E.g., if *VoicingThreshold* is 0.4 and *SilenceThreshold* is 0.05, this frame bears a good chance of being analyzed as voiceless (in step 4) if there are no autocorrelation peaks above approximately 0.4 or if the local absolute peak value is less than approximately 0.05 times the global absolute peak value, which was computed in step 2.

Step 3.4. Multiply by the window function (equation 5).

Step 3.5. Append half a window length of zeroes (because we need autocorrelation values up to half a window length for interpolation).

Step 3.6. Append zeroes until the number of samples is a power of two.

Step 3.7. Perform a Fast Fourier Transform (discrete version of equation 15), e.g., with the algorithm `realft` from Press et al. (1989).

Step 3.8. Square the samples in the frequency domain.

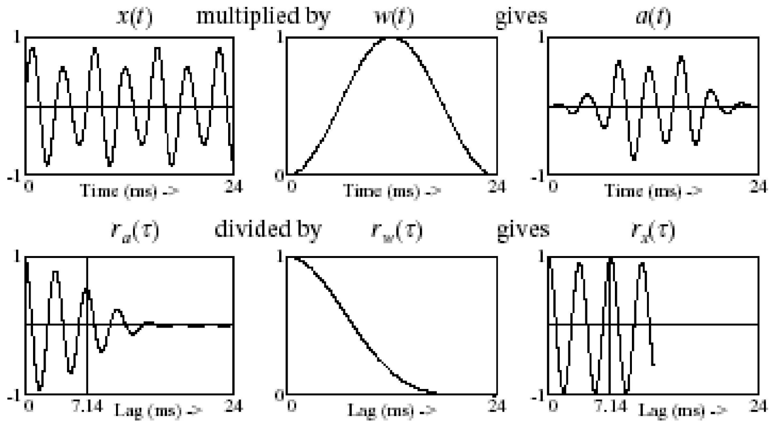


Fig. 1. How to window a sound segment, and how to estimate the autocorrelation of a sound segment from the autocorrelation of its windowed version. The estimated autocorrelation $r_x(\tau)$ is not shown for lags longer than half the window length, because it becomes less reliable there for signals with few periods per window.

Step 3.9. Perform a Fast Fourier Transform (discrete version of equation 16). This gives a sampled version of $r_a(\tau)$.

Step 3.10. Divide by the autocorrelation of the window, which was computed once with steps 3.5 through 3.9 (equation 9). This gives a sampled version of $r_x(\tau)$.

Step 3.11. Find the places and heights of the maxima of the continuous version of $r_x(\tau)$, which is given by equation 22, e.g., with the algorithm `brent` from Press et al. (1989). The only places considered for the maxima are those that yield a pitch between *MinimumPitch* and *MaximumPitch*. The *MaximumPitch* parameter should be between *MinimumPitch* and the Nyquist frequency. The only candidates that are remembered, are the unvoiced candidate, which has a *local strength* equal to

$$R = VoicingThreshold + \max\left(0, 2 - \frac{(local\ absolute\ peak)/(global\ absolute\ peak)}{SilenceThreshold/(1 + VoicingThreshold)}\right) \quad (23)$$

and the voiced candidates with the highest (*MaximumNumberOfCandidatesPerFrame* minus 1) values of the local strength

$$R = r(\tau_{max}) - OctaveCost \cdot {}^{2}\log(MinimumPitch \cdot \tau_{max}) \quad (24)$$

The *OctaveCost* parameter favours higher fundamental frequencies. One of the reasons for the existence of this parameter is that for a perfectly periodic signal all the peaks are equally high and we should choose the one with the lowest lag. Other reasons for this parameter are unwanted local downward octave jumps caused by additive noise (section 6). Finally, an important use of this parameter lies in the difference between the acoustic fundamental frequency and the perceived pitch. For instance, the harmonically amplitude-modulated signal with modulation depth d_{mod}

$$x(t) = (1 + d_{mod} \sin 2\pi Ft) \sin 4\pi Ft \quad (25)$$

has an acoustic fundamental frequency of F, whereas its perceived pitch is $2F$ for modulation depths smaller than 20 or 30 percent. Figure 1 shows such a signal, with a modulation depth of 30%. If we want the algorithm's criterion to be at 20% (in order to fit pitch perception), we should set the *OctaveCost* parameter to $(0.2)^2 = 0.04$; if we want it to be low (in order to detect vocal-fold periodicity), say 5%, we should set it to $(0.05)^2 = 0.0025$. The default value is 0.01, corresponding to a criterion of 10%.

After performing step 2 for every frame, we are left with a number of frequency-strength pairs (F_{ni}, R_{ni}), where the index n runs from 1 to the number of frames, and i is between 1 and the number of candidates in each frame. The *locally* best candidate in each frame is the one with the highest R. But as we can have several approximately equally strong candidates in any frame, we can launch on these pairs the *global path finder*, the aim of which is to minimize the number of incidental voiced-unvoiced decisions and large frequency jumps:

Step 4. For every frame n, p_n is a number between 1 and the number of candidates for that frame. The values $\{p_n \mid 1 \le n \le \text{number of frames}\}$ define a *path* through the candidates: $\{(F_{np_n}, R_{np_n}) \mid 1 \le n \le \text{number of frames}\}$. With every possible path we associate a *cost*

$$cost(\{p_n\}) = \sum_{n=2}^{numberOfFrames} transitionCost(F_{n-1,p_{n-1}}, F_{np_n}) - \sum_{n=1}^{numberOfFrames} R_{np_n} \quad (26)$$

where the *transitionCost* function is defined by ($F = 0$ means unvoiced)

$$transitionCost(F_1, F_2) = \begin{cases} 0 & \text{if } F_1 = 0 \text{ and } F_2 = 0 \\ VoicedUnvoicedCost & \text{if } F_1 = 0 \text{ xor } F_2 = 0 \\ OctaveJumpCost \cdot \left| {}^2\log \frac{F_1}{F_2} \right| & \text{if } F_1 \neq 0 \text{ and } F_2 \neq 0 \end{cases} \qquad (27)$$

where the *VoicedUnvoicedCost* and *OctaveJumpCost* parameters could both be 0.2. The globally best path is the path with the lowest cost. This path might contain some candidates that are locally second-choice. We can find the cheapest path with the aid of dynamic programming, e.g., using the Viterbi algorithm described for Hidden Markov Models by Van Alphen & Van Bergem (1989).

For stationary signals, the global path finder can easily remove all local octave errors, even if they comprise as many as 40% of all the locally best candidates (section 6 presents an example). This is because the correct candidates will be almost as strong as the incorrectly chosen candidates. For most dynamically changing signals, the global path finder can still cope easily with 10% local octave errors.

For many measurements in this article, we turn the path finder off by setting the *VoicedUnvoicedCost* and *OctaveJumpCost* parameters to zero; in this way, the algorithm selects the locally best candidate for each frame.

For HNR measurements, the path finder is turned off, and the *OctaveCost* and *VoicingThreshold* parameters are zero, too; *MaximumPitch* equals the Nyquist frequency; only the *TimeStep*, *MinimumPitch*, and *SilenceThreshold* parameters are relevant for HNR measurements.

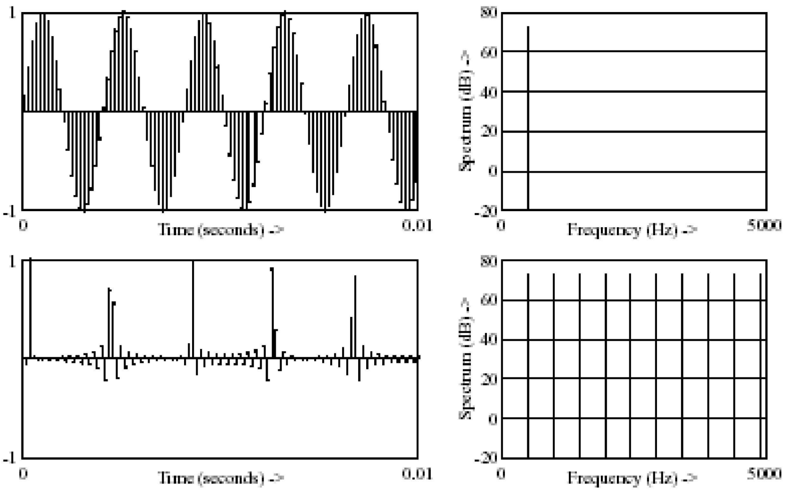


Fig. 4. At the left: two periodic signals, sampled at 10 kHz: a sine wave and a pulse train, which was squarely low-pass filtered at 5000 Hz (acausal, phase-preserving filter). Both have a fundamental frequency of 490 Hz. At the right: their spectra.

5.4

5.4.1 Patterns

Pattern recognition techniques may be employed to help analysis speech and improve reliability of a 1st order analyser. Techniques vary in their complexity and function.

5.4.2 Methodology

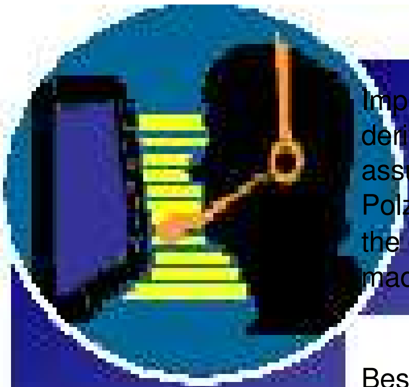
18 One solution is a dynamic approach which analyzes the F0 content as a function of time compares the pitch contour to a pre-defined template. (Holgar Quast[18]). This contour of the utterance is represented by a “best fit” sinusoidal. This system does not achieve the desired accuracy unless used with visual aided detection.

Training a network based on a large sample of non verbal vocal content to ensure non vocal cues are left out , makes reducing the cues to F0 and loudness values computationally impractical. However the framework can be generalised to recognise sentences of different structures is possible. When an input is reduced to a number of syllables (voiced intensity peaks) in a given utterance, their positions, the fundamental frequencies at these instances, a loudness value and a spectral histogram, the process becomes viable with high specification computers.

18 The pitch contour holds part of the information necessary to classify the recording influencing the receiver's *impression* in the determination of expressive emotions such as angriness or happiness. By using the pitch F0 contour a lot of information can be preserved without having to store a lot of data (Holger Quast[18]).

Historically pattern recognition techniques have been used by researchers such as the Bayes classifier (MLB), Kernel Regression (KR) and K-Nearest Neighbour) k-NN.

12 The MLB classifier is a parametric method where its assumed the probability density function P(x|w) of each class can be sufficiently described by a multivariate Gaussian centred around a prototype variable. The maximum likelihood estimation of the Gaussians is calculated from the training data.. The class chosen is the one with the maximum posterior probability P(w|x) which can be calculated from P(x|w) using Bayes theorem. The MLB classifier returned the poorest results when compared to other classifiers (Dellaert, Polzin and Waibel,[12]).



Improved performance was archived using Kernel Regression which is derived from non-parametric method which does not make strong assumptions about class PDFs (Probability Distribution Functions). (Dellaert, Polzin and Waibel,[12]). Essentially KR places a Gaussian Kernel at each of the data points to get an estimate for P(x|w), and the classification again, is made by using Bayes rule.

12 Best results were achieved using a k-NN classifier. This method approximates the local posterior probability P(w|x) of each class by the weighted average of class membership over the K nearest neighbours. Choosing a class with the highest estimated posterior probability is equivalent to taking a majority vote over these neighbours. Cross validation is used to select an appropriate K. (Dellaert, Polzin and Waibel,[12])

3 Other research featured Linear Discriminant Classifier (LDC) which assumes each class has a Gaussian distribution which further improved the results attained by k-NN by between 5% - 8%. (Lee, Narayanan, Pieraccini [3]). Error rates achieved for such systems were as low as 22.5% for female speakers and reduced to 21.25% if a feature reduction technique (Principle Component Analysis) was implemented.

5.4.2.1 **Bayes Rule*

Mathematically, Bayes' rule states

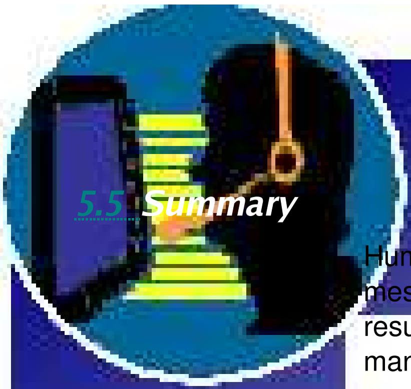
```
             conditional likelihood * prior
posterior = ------------------------------
                   likelihood
```

or, in symbols,

```
             P(e | R=r) P(R=r)
P(R=r | e) = -----------------
                   P(e)
```

where P(R=r|e) denotes the probability that random variable R has value r given evidence e. The denominator is just a normalizing constant that ensures the posterior adds up to 1; it can be computed by summing up the numerator over all possible values of R, i.e.,

```
P(e) = P(R=0, e) + P(R=1, e) + ... = sum_r P(e | R=r) P(R=r)
```



5.5 Summary

Human speech is an exceptionally complex mechanism for conveying messages. The audio signal produced consists of several features that are a result of the design of the individual speaker's vocal tract, psychological mannerisms and tainted by his or her state of mind.

Extracting these features for analysis is algorithmically complex task, especially when reliability and accuracy are paramount. Many researchers have developed very sophisticated analysis and synthesis systems that deal specifically with speech and is used by speech pathologists and therapists.

PRAAT is such a system that is widely regarded as the most sophisticated as well as the most intuitive. Using its script language, near real time analysis can be performed on speech to extract features. Features are then assessed to determine the embedded emotion.

To further reliability and accuracy pattern recognition systems have also been employed to complement speech feature selection. Results from such composite systems are very good especially for female subjects.

CHAPTER 6 Using Voice Features to Detect Emotions

6.1 Introduction

One of the major tasks facing automatic emotion detection is automatic recovery of relevant voice features. (Kollias, Piat , 2003 [2]). This can be quite involved, with extraction requiring sophisticated techniques containing algorithms that tend to be computationally intensive. It is also important to identify which features provide the best results under a variety of conditions. Conditions such as relatively high background noise provide significant hurdles in the clear extraction of these features.

6.2 Relevant Feature Types

6.2.1 Pitch

Definition:- The property of sound that varies with variation in the frequency of vibration. [www.dict.die.net]

2 Voice pitch is certainly a key parameter in the detection of emotion (Kollis, Piet , 2003 [2]). The extraction of pitch from voice is a complicated task involving the harmonic structure, compensating for short term instabilities and fitting continuous pitch contours to instantaneous data points. Pitch is generally recognised as the fundamental frequency (F0) except in some pathological conditions.

6.2.2 Intensity

Definition - Intensity is the energy or power of the speech. It is what is perceived as 'loudness'.

2 Voice level is one of the most intuitive indicators of emotion (Kollias, Piat , 2003 [2]). This is generally a direct function of microphone voltage output. A problem exists that it is difficult to determine whether a person is speaking softly close tot he microphone or speaking loudly at a distance. In either case the loudness perceived by the system will be the same. The same problem will exist in a telephony channel where an AGC will adjust audio levels to a set level. It is envisaged Neural networks may be adapted to distinguish between these conditions.

6.2.3 Voice Quality

2 A wide range of phonetic variables contribute as what is subjectively perceived as 'voice quality' (Kollias, Piat , 2003 [2]). Spectral properties can be used to characterise quality as well as inverse filtering that can recover the glottal waveform. Jitter and shimmer is a voice quality that is examined in this research.

Jitter: - Defined as the relative mean absolute third order difference of the point process.

Shimmer: - Defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

6.2.4 Temporal Structure(Pitch Contour)

2 The aspect of Temporal Structure studied through this research is the Pitch Contour. The pitch contour describes the changing pitch level during the execution of speech. This contour is divided into simple movements – rises, falls and level movements (Kollias, Piat , 2003 [2]). The pitch contour is used in conjunction with first order features such as pitch and intensity to help discriminate between emotions in an utterance.

6.2.5 Rhythm

Definition:- The patterned, recurring alternations of contrasting elements of sound or speech. [www.dictionary.com]

Rhythm is known to be an important aspect of speech (Kollias, Piat , 2003 [2]). Generally rhythm is a difficult feature to measure with the alternation between speech and silence being a reasonable indicator. Research has shown stressed subjects show shortened switching pauses (pausing before taking their turn to speak) and lengthened internal pauses (pauses during the course of their speech during their turn.

6.2.6 Formant

Definition:- Poles or points of resonance in FFT terms corresponding to resonant areas in the vocal tract. Contributes to the 'timbre' or quality of the speech.

Speech Formants are the physical phenomenon that gives each of us our individual voice. Formants are created in the vocal tract by the cavities and shapes in the lungs, mouth ,nose and other cranial orifices. The material and texture of the vocal tract also play an important part in the characteristic of formants. Formants can be easily recognised as “poles” in the frequency domain indicating the frequency of resonance of each formant.

6.2.7 Octaves

22 Definition:- An octave is the relationship between two pitches whose frequencies are related 2:1. Physically it is the interval between a fundamental and its first harmonic. (Pitkow, Xaq., 2000[22]).

Vocally men and women speak one octave apart. In music terms ,based on the Western system, it is an eighth relationship in an increasing series of intervals (A,B,C,D,E,F,G). When men and women sing together they sing one octave apart yet at the same note creating harmony. The definition of an octave can vary from discipline to discipline.

6.3 Methodology

6.3.1 Praat Speech Analyser

The PRAAT speech analysis is an exceptionally capable and easy to use pre-packaged speech analyser. It uses a sophisticated windowing mechanism to extract the voice pitch contour and other features such as FFT formants, intensity, plus second and third order functions such as jitter and shimmer.

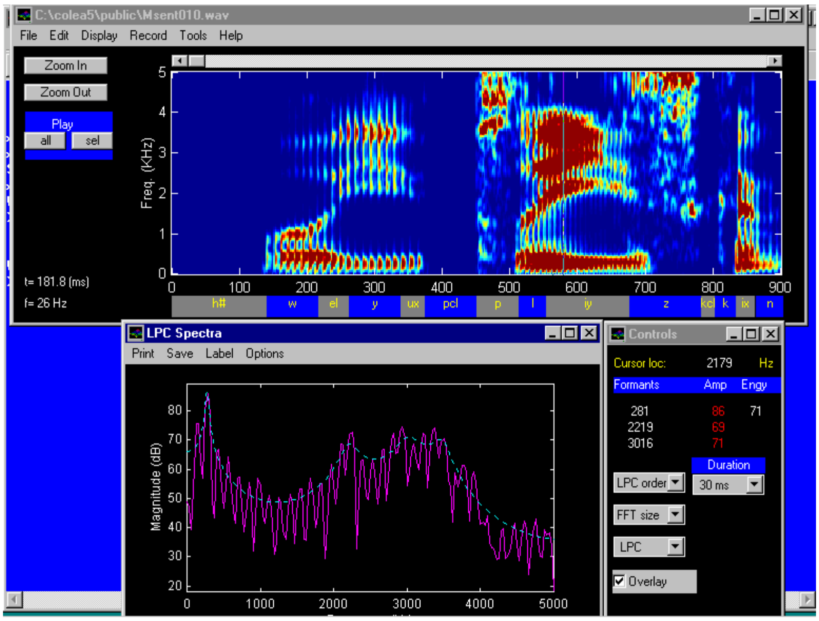
The advantage of using such a system as the base for an emotion detection system is the rapid development time. The engineer does not need to write his or her own algorithmic extraction techniques and has the added bonus of a powerful scripting language that provides an interface between the PRAAT analysis and other processes that are utilised in the analysis.

In this research the PRAAT system was elected because of this reason. The script language is flexible and has to ability to communicate via TCP/IP enabling calls to other processes acting as an interface between the emotion detection adjunct processor and the Demonstration service running on the IVR.

6.3.2 Matlab COLEA toolbox

www.utdallas.edu/~loizou/cimplants

“Colea” is a Matlab utilities tool designed specifically for speech analysis. It was originally written for the Matlab 4.x package as part of the COchLEA implants toolbox.



t= 181.8 (ms)
f= 26 Hz

Cursor loc: 2179 Hz

Formants	Amp	Engy
281	86	71
2219	69	
3016	71	

Duration: 30 ms

☑ Overlay

Illustration 8. COchLEA Analysis Package

The package includes full pitch and formant analysis plus wave editing tools. It was not elected as a tool for this report because the PRAAT utility already has a highly regarded pitch analysis tool built in. The Matlab package does have the advantage of low level filter programming which is compatible with many DSP environments. Development using Matlab would provide an easier porting solution to an DSP platform based application.

6.4 Four Feature Detection

Initial Investigations into the detection of emotion revealed that there were 4 voice features that varied significantly from emotion to emotion. These four features varied consistently with the 4 emotions initial tested while the other derived features, although, exercising some degree of consistency in variation, could not be used to reliably distinguish emotions.

A summary of these four features :-

Feature	***Description***	***Comments***
Mean Intensity	The average intensity of speech generally varies for different emotions.	Speakers tend to be louder when angry and softer with in a sad or depressed state.
Intensity Standard Deviation	The deviation of speech intensity is the amount in which the speaker's volume changes over time	The level change may help differentiate sadness from neutrality
Mean Pitch	This is the fundamental voice frequency which is controlled by the tightening of the larynx.	This tends to be higher when the speaker is happy
Pitch Standard Deviation	This is the variation in pitch during speaking.	Speakers subconsciously “dress” their speech to embody speech with emotional messages.

Table 2. Four voice features used for detection

6.4.1 Pitch Contour

A fifth feature that contributed highly to the differentiation of emotions was pitch contour. This research has revealed that the importance of pitch contour to be second only to pitch standard deviation in producing consistent results. Like pitch standard deviation, pitch contour is a subconscious dressing of speech used to convey emotional meaning.

Table 2 demonstrates the relationship of contours and how their shape relates to emotions embodied in speech. While speech contour is valuable input into emotion detection, its analysis is somewhat difficult and hence initial demonstrations of the technology during this research did not take advantage of it.

6.5 Analysis of pitch contour.

To achieve a universally reliable system the analysis of pitch contour has become necessary. While the four featured system explored in this research delivered a reasonably reliable result for some speakers, there is still a drawback that some emotions share characteristics with others reducing resolution. The test system has a tendency to muddle happiness with anger and sadness with neutrality. Pitch contours would prove to be the deciding factor in many cases. Happiness and anger have reversed pitch contour slopes, sadness has downward sloping curves while neutrality is generally flat.

In an effort to improve the performance of the emotion detection system, an analysis programme was written which processes pitch data generated by the PRAAT analyser and determines the general direction of the pitch contour as well as producing an indication of the amount of unvoiced time during the utterance. Non spoken regions or “quite times” are another indicator of a speaker's emotional state. (See section 6.5).

6.5.1 Zero voiced regions or “Quiet Time”

The degree a to which a person varies the silence or gaps between spoken words can be related to their emotional state. The rate or rhythm of speech is heavily influenced by anxiety , anger or elation. Sadness and confusion also contribute to variance in these quiet times. A simple analysis routine was written in Perl for the second generation analyser to count the number of samples that contained quite times as compared to the total sample count (See Section 6.5.1.1) . Although the accuracy was limited because non voiced regions occurred due to the pitch going outside the nominal 75 – 600 Hz voice range *as well* as due to silence, the sample count did provide information that was particularly helpful and determining the anger emotion.

PITCH CONTOUR

Illustration of pitch contours

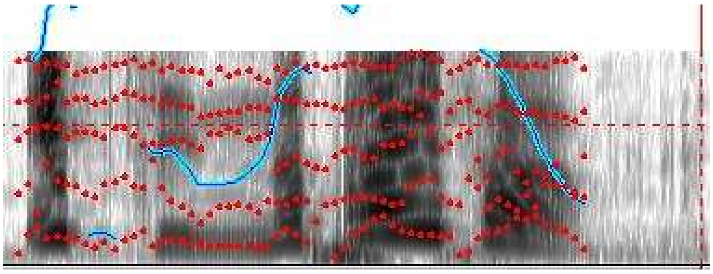
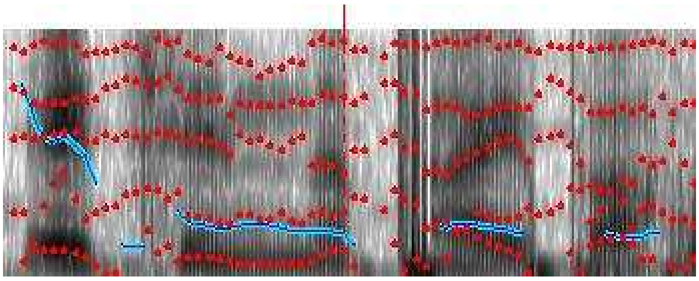
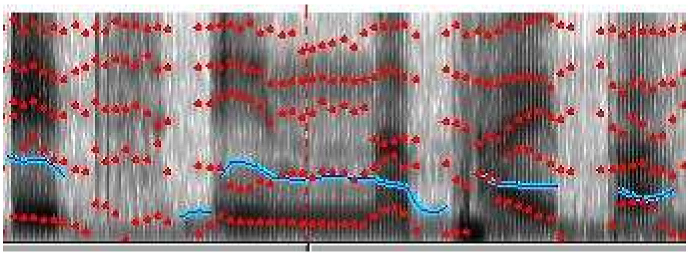
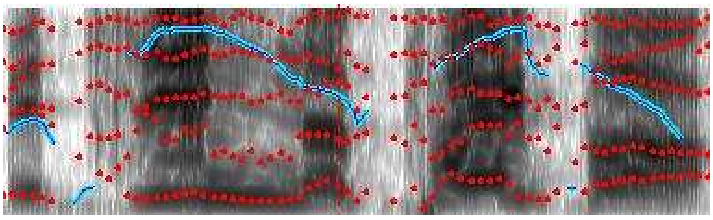
Happiness Pitch starts off low and slopes upwards. High Pitch STD DEV	
Sadness Pitch starts off medium an slopes down. Medium Pitch STD DEV	
Neutrality Pitch stays generally flat low Pitch STD DEV	
Anger Pitch starts off higher an slopes down. Medium Pitch STD DEV	

Table 3. Pitch contours for different emotions

6.5.2 Four Emotions

For this project the research and implementation of the emotion detection system is actually based on 3 primary emotions with the 4th state, 'calmness' or 'neutrality' defining a state with no emotion or a state lacking arousal.

I chose the three emotions that are specifically documented in the experiments and the report because they are easily definable and broadly experienced. They are also sufficiently different in terms of definition to qualify separate explanation and experimentation.

Together, calm, anger, happiness and sadness form the basis for this documents discussion. Other emotions such as interest, confusion and frustration may be of interest for ongoing research.

6.5.3 Limitations

With any system that professes to be a natural interface, limitations exist because of the pure complexity of human communication. Humans are incredibly perceptive receivers being able to "read lips", "read between the lines" and somehow able to communicate perfectly in a deafeningly loud night club. To develop a level of perception and the learning capability to even mildly compare to a human subject is a big call for any computing system. Many researchers are working on Artificial Intelligence to help speed up the development and evolution of the perception process.

Also people use a daunting number of labels to describe emotions (Kollias, Piat, [2]). Emotions are experienced by different people for different reasons. The physiological mechanism that translates into a manifest of speech is also broadly varied. We often say that we need to know someone before we can tell how they feel. In essence, a system, whether biological or electronic needs to be trained to be proficient at reading emotions Neural networks that are self learning or can learn in an unsupervised manner may be suited in addressing this problem.

6.5.4 New emotions that can be differentiated

23

As previously mentioned human emotion proves to be exceptionally difficult to define and be categorised. Many researchers studying emotion categories emotions into a main set of "base emotions" and define a subset of emotions that are any combinations of the base emotions. (Cowie, Roddy., 2000[23]) Four example if four emotions are used as the primary or base emotions, such as happiness , sadness, anger and jealousy then a person's temperament may be described using all or non of these emotions. For example someone who is confused may exhibit the primary emotions of sadness with a touch of anger.

These descriptions of emotion vary greatly from author to author and the subject has been heavily debated since the dawn of philosophy. The emotion detection system featured heavily in this report is designed to analyse four states of mind:- Happiness, Sadness, Anger and Neutrality. It is arguable if neutrality is an emotion, perhaps it is correct to say it is the *lack of any emotion* as perceived by the system.

6.6 Refining the system

6.6.1 Jitter and Pitch Contour - 2nd Generation Analyser

Telephony applications using this technology provide a major hurdle in the form of information loss due to restricted bandwidth and mechanical noise. The problem is further exacerbated by AGC mechanisms on telephony applications such as multi-party conferencing which nullifies any intensity based feature detection.

As discussed in section 6.5 the pitch contour is a second order feature which provides useful information pertaining to emotional content. The second order analyser used a Perl based application to determine upward, downward movements of pitch as well as zero voice regions as described in section 6.5.1. The net upward movements were compared to downward and zero voice regions in experiment two to determine patterns fundamental to each emotion. Jitter which is described as a third order function is analysed and reported by the PRAAT application and proved to be very useful in the isolation of sadness from other emotions.

6.6.1.1 *Perl program written to extract pitch contour from speech data*

A simple Perl script was written in an effort to extract the pitch contour and determine the relative number of unvoiced regions. The script sampled speech data delivered by the PRAAT analysis package. The routine simply logs the number of samples that are moving downwards, upwards and the regions where no determination could be made. This non-occurrence is either a result of no voice activity or the system unable to attain a fundamental between 75 and 600 Hertz. In this case the audio is unlikely to be regular speech anyhow. See Section 22.5 in **Appendix D** for Source code of routine.

6.6.2 The importance of Amplitude or Intensity

Voice intonation features (pitch and intensity) provide the most information relating to emotion extraction from speech. Pitch is the prime candidate , but intensity and especially intensity standard deviation is useful in differentiating expressive emotion such as happiness and anger from neutral speech. Happiness and anger were found to have a high intensity standard deviation in initial investigation (See table 4, Section 7.3). Since this project concentrates on telephony applications of emotion detection, intensity is not used in the second generation analyser. However it can be enabled by the setting of a flag in the code header. **See Appendix E** – Emotion Detection Engine Source Code.

6.7 *Summary*

Voice features that comprise the human voice are quintessential elements in the detection and analysis of emotion in speech. The methods for extracting key features such as pitch and intensity tend to be complex and computationally expensive. Four emotions: happiness, sadness, anger and neutrality are studied in this report but many more can be detected by correlating feature patterns with particular emotional responses. Pitch contour is a second order operation that is paramount in providing decisional information. Intensity and its standard deviation also provides useful information but is not available for use in some telephony applications.

CHAPTER 7 Experimental Research

7.1 Introduction

Although the measure of emotion through computing devices is a relatively new field – the study of human emotion and why we humans perceive them as such has been conducted since the origins of philosophy. There is countless publications and research journals based on emotions, the associated characteristics and how they influence speech. Actors and entertainers study emotions and the physiological mechanisms that deploy them. From these fields of study researchers have a wealth of information and data to aid their study.

Illustration 9 Aristotle

"The things we have to learn before we can do them, we learn by doing them."

7.2 History

7.2.1 Work done so far

Emotion recognition by computer has recently been the subject of many research projects and thesis. The research is generally limited to research students or smaller research teams, but major inroads have been made to transforming the idea into reality.

Experiments based on emotional speech and voice features have revealed consistent results pertaining to the technology. Research centres around the actual speech and the analysis of voice features, but some research also has been conducted taking into account human psychology.

10 Several experiments have displayed that humans extrapolate their interpersonal interaction patterns onto their computers. Humans are polite to their computers and are flattered by them. They have emotional experiences interacting with them. (Polzin, Waibel, 2003[10]). This may also explain the latest phenomenon known as *computer rage*. Computers become friends or pets, we trust them and they are often our lifeline. When they fail they betray us – and the users may act unpredictably.

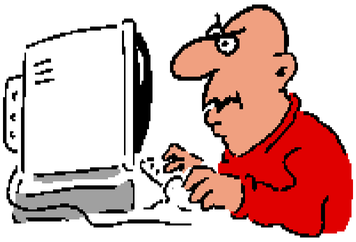


Illustration 10 The bond between man and machine is not always a polite one

10 Many experiments conducted deal with the analysis of emotion itself and portray statistical evidence of the performance of such analysis. An experiment conducted by Polzin and Waibel concentrated on the performance of human subjects as emotion detectors rather than on the technology itself. It seems we humans are not perfect emotion detectors with scores ranging between 60% and 80% depending on the emotion.

17 Results took a turn for the worse when prosodic and/ or spectral information where removed from the speech(Polzin, Waibel, 2003[10]). Studies also revealed that this accuracy further diminishes across cultures. Louise Bosch performed a study involving Japanese and American subjects which showed people are better able to determine emotion in speech if the speaker is from there own culture (Louis Bosch, 2000[17]).

3 Many research projects relay statistical analysis on the accuracy of emotion detection systems. These systems are generally quite sophisticated using several multi-order featured systems and pattern recognition techniques. Lee Narayanan and Pieraccini used a ten feature analyser with Linear Discrimination (LDC) and alternatively k-Nearest Neighborhood (k-NN) classifiers to detect emotion from speech recorded by actors. (Narayanan and Pieraccini, 2003[3]). They achieved results of 70% – 77.5% for female speakers and 56.65% - 74.19% for male speakers. Results from the experiments in this study also revealed considerably greater accuracy with female as compared to male speakers.

7.2.2 Actor's Training

To date many studies have been made involving prosodic and vocal aspects of actor's voices. Understanding about the voice and how it projects emotion is very important to the thespian who wishes not only to get his or her lines correct but sound convincing. Unless actual stimulants are present the physiological mechanism of conditioning speech with emotion is exceptionally difficult. One must understand the individual features that consolidate a voice and how they may be manipulated.

As a result there is a plethora of books and journals that investigates the human voice and subsequently how it can be trained to create the impression of real emotion without the emotional stimulus. This knowledge is also directly applicable to the engineer who develops emotion recognition systems. Knowing what happens and how it modifies voice features is an important stepping stone in the design and creation of such systems.

To the Reader.
This Figure, that thou here seest put,
It was for gentle Shakespeare cut,
Wherein the Graver had a strife
with Nature, to out-doo the life :
O, could he but have drawne his wit
As well in brasse, as he hath hit
His face ; the Print would then surpasse
All, that was ever writ in brasse.
But, since he cannot, Reader, looke
Not on his Picture, but his Booke.

Ben Jonson's
Commendation of the
Droeshout engraving
First published 1623.

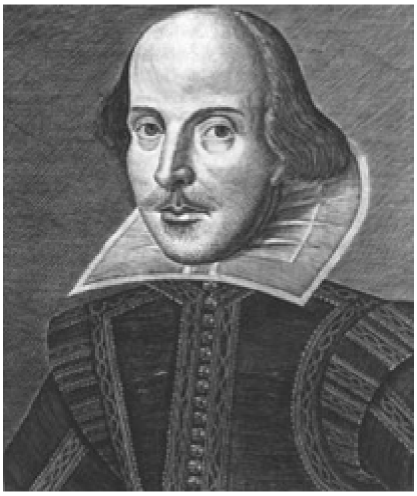


Illustration 11. William Shakespeare 1564 - 1616

7.3 Preliminary Investigation

.www.fon.hum.uva.nl/praat/

After attaining an acceptable Software package that would break speech done into a number of key features, the task was then to apply this software in investigating which speech features are conducive to differentiating emotions in speech. The PRAAT Speech analysis Software package was elected to do the job as it is free to download, is extremely intuitive and by all accounts is the most accurate system for reliably extracting pitch from speech. Written by two Dutch Speech Specialists: Paul Boersma and David Weenink, the PRAAT system includes a powerful scripting language, making it ideal for real time applications when run on a suitable processor.

12

5

9

Initial investigations involved the use of my wife and my own voice to model emotional characteristics. A small script was devised and simple phrases where recorded and analyzed with the PRAAT software. See table 4 on page 58. Research conducted by Dellaert [12], Lee & Narayanan[5] and Polzin & Waibel[9] show results depicting that intonation (pitch and intensity) is the base feature through which analysis can be conducted. Standard Deviation, mean, range , maximum and minimum values of pitch and intensity play a key role in differentiation of key emotions.

Based on these results I extracted these plus several more features to study the relative change in regard to different emotional stresses. I made up two short phrases for my wife and I to say in four different tones. I chose to analyse on Happiness, Sadness and anger as I believe they will have the biggest commercial appeal. Calmness or neutrality may be defined as *no emotion* and was also included in the analysis for reference purposes.

Two other emotions considered for their commercial appeal were fear and confusion. Both are distinctive enough to be technically viable, however they may only cloud results at these initial stages.

Table of results – Initial Investigation

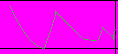
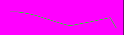
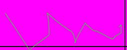
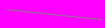
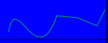
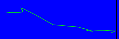
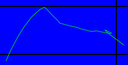
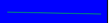
	EMINA – "The material is ripped"				DENIS -"No way Macca"			
	Happy	Sad	Angry	Neutral	Happy	Sad	Angry	Neutral
Mean intensity	70.34 db	76.22	87.77	75.48	84.7	80.3	85.4	77.66
Minimum Pitch	119 Hz	158.3	157.54	119	121.6	110.3	1169.9	118.92
Maximum Pitch	495Hz	495.07	441.58	165.66	367.5	248.5	2916.6	145.38
Minimum Intensity	65.04 dB	62.92	59.99	66.09	89.11	65.12	62.09	62.89
Maximum Intensity	91.95 dB	87.65	92.41	81.04	60.45	78.16	92.74	85.44
Rhythm	1.145 s	1.43	0.861	0.978	1.017	1.35	0.87	0.846
Inter-phoneme Space	Med-Lrg	Large	Small	Medium	Med Lrg	Large	Small	Med-Sml
Inter-phoneme Space Deviation	Med-Lrg	Large	Med - Sml	Medium	Large	Med-Lrg	Small	Small
F1variance	142.2 Hz	17.73	459.2	446	80.8	199.2	85	95
F2 variance	195.56Hz	138	667.57	443	1135.9	473.24	363.3	1387.7
Jitter	0.061	0.0268	0.0438	0.025	0.0615	0.0128	0.0165	0.01206
Shimmer	0.0998	0.0947	0.0796	0.100	0.0675	0.0594	0.0516	0.105
Pitch Contour								
Mean Pitch	252.9 Hz	215.74	281.54	128.81	252.46	144.4	230.9	137.32
Intensity Std deviation	9.138 dB	7.9	12.228	7.7	11.09	8.27	10.95	7.12
Pitch Std Deviation	114.85 Hz	39.1	75.26	248.82	77.2	33.9	53.37	12.66
Pitch 2nd Quantile (median)	196.68 Hz	218.7	248.82	129.86	269.6	127.07	210.75	134.65

Table 4 Results from initial analysis

From these initial investigations it was apparent that four features in particular contributed the most toward differentiating between emotions. From this a small demonstrator application was written based on an investigation undertaken by Oudeyer Pierre-Yves from Sony in Paris [7] using a simple decisional logic (decision matrix). The supplied .wav file is analysed through this simple matrix which delivers a verdict based on the voice features.

```
If 'pitchSD' < 30
            emotion = Neutral
elsif 'pitchSD' > 75
            emotion = Happiness
elsif 'intensitySD' <9
            emotion = Sadness
else
            emotion = Anger
endif
```

Simple decisional logic to determine emotions

7.4 The Pitch Contour

The way in which we project our voice and change the pitch in words spoken directly relates to the emotional message we suggest on our speech. The pitch contour, although difficult to analyse, plays an important part in emotion detection. The table below illustrates the relationship between the four emotions and the recorded pitch contour in the experiment. Refer to the "Analysis of the Pitch Contour " (Section 6.5).

	Mean Slope	***Slope Direction***
Happiness	High	Upwards
Sadness	Low	Downwards
Anger	High	Downwards
Neutrality	Low	Flat

Table 5 Pitch contour of different emotions

7.5 Conclusion

From this preliminary Investigation it is evident that there are certain physical aspects of a person's speech that indicate the emotional state of the speaker. The four features in particular derived form the experiment with two subjects are: 1. Mean Intensity, 2. Intensity Standard Deviation, 3. Mean Pitch and 4. Pitch Standard Deviation.

Other features that may contribute to accurate differentiation are Jitter Shimmer and Pitch Contour. It is evident from the the pitch contour that emotions follow a predictable shape defining the mean slope and upward or downward sloping characteristics of the contour line.

As discovered with ongoing research the four features do not completely guarantee a high reliability detection of emotion. However they do provide a very good indicator and provide a basis for more comprehensive analysis.

7.6 Experiments conducted

7.6.1 Experiment 1 - Single Word Utterance Experiment

7.6.1.1 Outline

A major consideration in developing a technology that interacts with human beings in the fact that we, as a people, vary considerably due to our genetic makeup, or culture and our gender. This variability provides many obstacles when analysing human emotion in speech.

One method of dealing with such an issue is to analyse speech from a random selection of people who might represent a reasonable cross-section of English Speakers living in Sydney Australia.

The first experiment conducted in this report was conducted in the offices of Dialect Solutions Pty Ltd, in Sydney. The staff are composed of technical, sales, management and call centre people, so of who have a theatrical bent. In all 12 staff members contributed to the experiment – six men and six women.

7.6.1.2 The Experiment.

The 1st research experiment was written on a Telsis Hi-Call IVR located in the offices of Telsis Pty Ltd North Sydney. The experiment is accessible via telephone and records the participant's responses to certain prompts. The experiment takes about seven minutes to conduction and uses scenarios to generate an emotive response from the caller. Below is a table of the required responses from each participant.

Experiment 1 Recordings	
No	Neutral
No	Happy
No	Frustrated
No	Angry
Yes	Neutral
Yes	Happy
Yes	Frustrated
Yes	Angry
Manager	Neutral
Manager	Happy
Manager	Frustrated
Manager	Angry

Table 6. Experiment 1 recordings

Recorded audio from this experiment was converted into .wav files and individually titled according to the actual utterance, person speaking and associated emotion. The .wav files where then put through a PRAAT speech analyser and broken down into individual voice features. Since there where 144 separate files a Unix script was written to auto mate this task and generate a tab delimited data file. Next the datafile was loaded into excel and the various features for individual speakers plus their emotions where charted and manually compared.

7.6.1.3 Objective

Derived results are compared to see what voice features are unique to a particular emotion. The sample of participants would hopefully provide a wide spectrum of results. This would provide enough information to create a high level emotion matrix which can be used in the Emotion Detection Engine.

7.6.1.4 Problems Encountered

The majority of the problems encountered were related to the set -up of the IVR service itself. The script required a fair bit of thought needing to be both instructive and entertaining. Encouraging people to use the service was not as difficult as I first thought. The response was encouraging.

Aside from this one other problem is that the telephony system and the IVR itself use an AGC (Automatic Gain Control) to pull up low audio and attenuate the loud. This had a tendency of delivery the same audio energy despite the emotion. The system had to be designed to maximise pitch variation. Unfortunately this is difficult to work around in any most communication situations as most telephony channels use AGC to optimize conversation levels.

7.6.1.5 Results

The results from the first run of this experiment provided valuable information about people's emotional speech characteristics. It must be noted that the experiment was conducted in a working office environment which inhibited the participants from really characterising the required emotion. This provided results with narrow margins, but despite this definite characteristics were able to be derived.

Overall results where calculated by manually deciding on what characteristics contributed to the differentiation of the four emotions of Happiness, Sadness, Anger and Neutrality (It may argued Neutrality is not an emotion but a lack of emotion). If the data population was large enough a ROC (Receiver Operator Characteristic) analysis may be performed on the data to try and confirm "true positives" to ensure accuracy and reliability of the system. However this operation was not a success for the data I had retrieved due to a relatively low population and a high number of wayward outliers.

7.6.1.6 Analysis by Four Feature analyser

Results from the experiment where put through the four feature analyser. Results varied from perfect scores to 15% accuracy. The test audience varied in age gender and theatrical training. Those who scored poorly were young male , theatrically untrained participants. The best score was the my own voice on which the four featured analyser was originally devised from. Other top scoring subjects where female and theatrically trained. Participants with foreign or non-Australian accents also performed poorly. See table of results posted in **Appendix C**

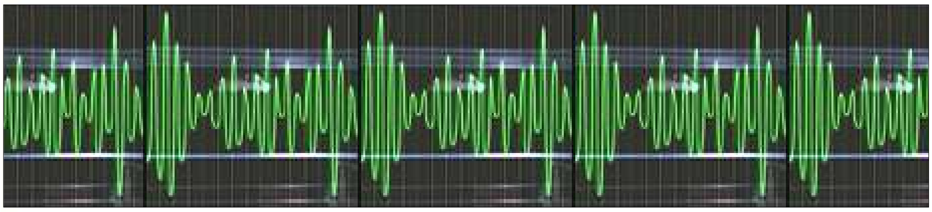
7.6.2 Conclusion for Experiment 1

The results attained in Experiment 1 show that the basic four feature analyser can make a fair analysis of some of the subject speaker's emotion in speech. The positive results were attained despite the experiment being conducted on subjects who are not theatrically trained and in an environment not particularly conducive for recording emotional verse.

The results for subjects who were trained actors are good , with the analyser achieving between 66% and 100% accuracy. The analyser performed poorly on subjects with non –Australian accents which indicates that voice features that are bound to emotion in speech vary from culture to culture. Even the closely related English accent derived erroneous results on the system.

However for the system to be commercially applicable the reliability over a broad client base must be researched and improved especially when used in multi-cultural environments such as Sydney. More voice features including pitch contour analysis may provide improved recognition and accuracy.

7.6.3 Experiment 2 – Short Phrase Utterance



7.6.3.1 Outline

The second experiment in this research deals directly with a wide sample population across a single utterance rather than a selection words. Experiment 1 helped define the effect of pronunciation of three different words. The second experiment concentrates on a single repeated phrase.

Once again the experiment was conducted in the offices of Dialect solutions in Sydney but was also made available for after hours calls for friends and relatives of Dialect staff. Some members of staff are closely tied with theatrical organizations and a sample of such a population should deliver rewarding results.

7.6.3.2 The Experiment

Similar to experiment 1 the second experiment is conducted on an IVR device permitting access to anyone with a access to a telephone to participate. The service uses voice prompts to guide the participant through its course, this time expressing a single phrase “Not now- I'm in the garden” with four different emotions. (See table 5). Utterances are captured for analysis via the IVR's voice recording mechanism. Experiment 2 also introduces the more practical emotion of “sadness” which reflectively is more practical than the “frustrated “ emotion which it replaces in Experiment 2.

Experiment 2 Recordings	
Not now I'm in the garden	Happy
Not now I'm in the garden	Sad
Not now I'm in the garden	Neutral
Not now I'm in the garden	Angry

Table 7. Test utterance and the 4 emotions

7.6.3.3 Objective.

The key objective of experiment 2 is to get as large a sample population as possible that has definite emotional stress to help aid analysis. Experiment 1 was hindered by a low sample population, restrained participants and variations induced by the utterances of different vocables. The concentration on the same speech pattern plus a larger population is intended to produce an adequate quantity of results so ROC analysis can be performed on the data. It is also the intention to use as many thespian subjects as possible to create a data pool of 'true' emotional stresses.

7.6.3.4 Problems Encountered

This experiment was very similar to experiment one and used the same IVR code. Scripting was also very similar so setup time was very quick. Once again Participants were a little slow to start but volunteered graciously. Many trained voices also responded.

Once again AGC restricted intensity variation so pitch is to be the main consideration in this analysis. The results correlation was complicated by the adjustment of the original Perl and Unix scripts that take the wave files and create a .CSV file for spreadsheet style processing. The new scripts had to run an adjunct process (also written in Perl) that determines the pitch contour and number of non-voiced samples. This ended up being a major operation for modest programming skills but eventually a result was attained.

The correlation of data and the search for patterns through trial and error as well as through deductive investigation proved to be extensive and mentally tiring.

7.6.3.5 Results

With initial investigation a definite correlation was found between Pitch Standard Deviation and emotion. Other observations noted was that the ratio of downward pitch contour samples to non voiced samples was generally above 4:1 for angry emotion while between 3:1 and 4:1 for happiness. Neutrality was singled out by a very low Pitch SD below 25 Hz. Sadness was more difficult to differentiate, however a definite correlation between a medium Pitch SD and a low Average Pitch can be used to give fairly reliable results here. Jitter also proved invaluable in isolating sadness. The Emotion Detection engine must determine whether the speaker is male or female by comparing average pitch to a set value as the two genders are about an octave apart.

The captured sound files were used as input to two different emotion analysers. The 1st analyser was the original four featured analyser as used in the initial investigation. Due to the AGC inherent in the recording IVR intensity Standard Deviation and mean were not useful in determining results. The 1st order analyser returned a reasonable result of 24/72 or 33% accuracy.

Better performance was achieved from the 2nd analyser which incorporated pitch contour and jitter in its analysis. Jitter was especially useful in determining sadness which was the most difficult emotion to detect with the 1st order analyser. Markedly improved results were obtained with this system scoring 29/72 or 40% accuracy. The anaylser works on a more complex "points" system where the program awards points to each emotion depending on the information contained in each feature. The emotion granted the most points is the determined result. Unfortunately some emotions tie equal first especially Happiness with Anger. Happiness proved to be the most difficult emotion to reliably distinguish. See **Appendix D** for Experiment 2 results.

7.6.4 Conclusion for experiment 2

To vastly improve the performance of the system, a feature has to be found that defines happiness as apart from the other 3 emotions. If we can reliably isolate happiness to the degree of the other emotions the system may achieve a reliability of about 66% based on the current results.

7.7 Experiment 3

7.7.1 Correctness assessment experiment (Part of DEMO 1)

This experiment was to be conducted on an IVR in conjunction with the “How are you” demonstration (Demo 1). This experiment records the accuracy of the Emotion Detection Engine when used real time in a possible real world application.

The table below logs the results from the first demonstration. The caller after greeted is prompted to respond in an emotion of either Happiness, sadness, anger or calm. The demonstration service then replies with a response applicable to the emotion it has perceived the caller is expressing. At this point the services asks if it got the correct emotion. If incorrect not asks the caller to enter in the emotion they did intend to portray. The results are logged in a table with not only logs the number of successful hits but also cross references the emotion that the system mistook for the actual emotion portrayed . This gives the researcher a matrix of what emotion was mistaken for another as well as a ratio of successful entries.

	Happiness	*Sadness*	*Anger*	*Calm*
Hit				
Mistaken for Happiness				
Mistaken for Sadness				
Mistaken for Anger				
Mistaken for Calm (neutrality)				

RESULTS DUE DECEMBER 2003

Table 8. Table results for experiment 3

7.7.2 Conclusion for Experiment 2

Unfortunately due to technical issues the demonstration service was not available before the publishing of this report so the last experiment of this research will be published separately in an addendum to this report. Expected delivery of results mid December 2003.

CHAPTER 8 Improving the Product

9.1 *Improving Reliability*

By itself the four feature analyser performs fair to moderately well depending on the speaker and on the environmental conditions. The four first order features do a sterling job of differentiating basic emotions with a considerable degree of accuracy and reliability. However human speech is a rather complex and the speaker's manipulation of it when producing emotional verse exceeds the capability of first order measurements. As described in section 6, the pitch contour or the pattern of pitch change during speech has been demonstrated by this research as being a exceptionally important in the detection and confirmation of certain emotional stresses. In addition to the existing analysis techniques a supplementary routine that analysis pitch contour is required to improve the system's reliability.

Initially a good think about the problem at hand was required. The research to date has revealed two things. One was that emotion detection was possible if but not overly accurate on a basic first order analyser. Secondly the requirement for second order analysis was required to work with the first order results to improve performance. It was decided that a independent pitch contour analyser could be drafted in Perl scripted using pitch data produced by the PRAAT speech analyser system. Perl was the language of choice due to its programming simplicity, compatibility across Linux and Microsoft platforms and simple Socket programming directives.

Accuracy constitutes a major part of reliability for this type of application. The more accurate the detection to more reliable on the whole, the system is perceived to be.

9.1.1 Extra Features

The old saying goes “the more the merrier”. This is applicable in the case of resolving emotions from speech audio. By extracting and analysing as many well defined and independent features as possible the accuracy of the system should increase. Further to this, pattern matching and feature removal techniques such as those discussed in section 5.4.2, may also have an impact on the overall reliability of the system. The major disadvantage of using a large number of features for detection is the computational power required. This would be especially apparent in real time applications of the technology.

9.2 Configuration

9.2.1 Configuration Application

It is important to be able to access and improve the engine of the emotion detection system, taking into account the vast number of variables that affect the system's operation. An independent application that interacts to the main system process will enable engineers and operators to attain feedback and consequently tune the system to suit local requirements.

9.2.2 Technical Outline

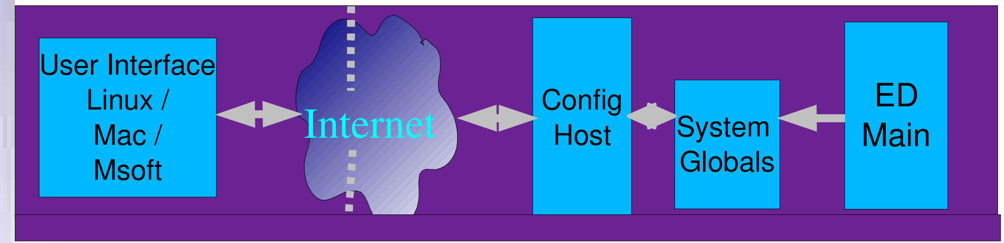
The requirement is for a GUI based application designed to initially configure the emotion detector and provide a maintenance interface that allows the operator to fine tune the system based on user feedback. The programming language of choice is Sun Micro-system's Java as it is perfect for both browser based and terminal based applications as well as being platform independent.

The configuration application would need to work directly on global variables used by the emotion application permitting instant changes to be made so real time diagnosis of performance and fine tuning can be performed.

The key configurable agents are variables containing vectors to trim the points awards generated by the second generation emotion detection engine. (See Section 5.6.3.5). Also the interface has to permit the operator to select one or more of any of the emotions featured by the system. For example the operator may only want the system to detect anger.

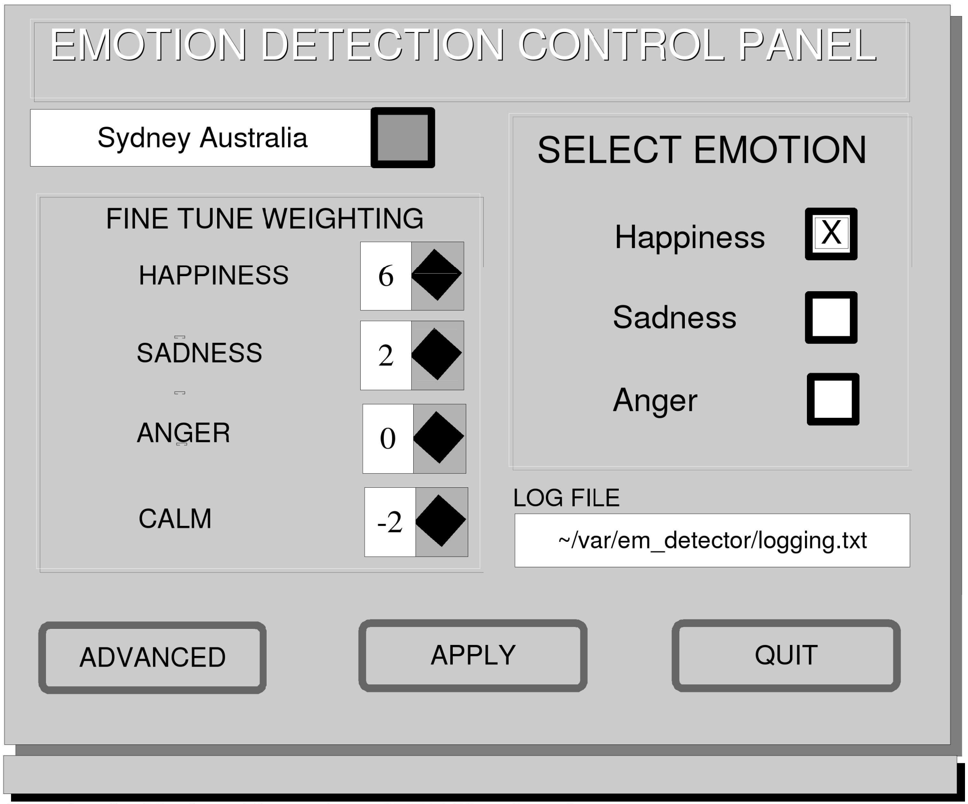
The interface should also be able to deliver a log of the emotions interpreted with the confidence level in percent. From this the operator can analyse which emotion/s the system is having problems in deciphering.

The Java application would most likely communicate to a configuration host process via TCP/IP which allows the remote setup and maintenance via an intranet or the Internet. A configuration host will interact with the emotion detector either via modification of system globals or by interprocess messaging.



Drawing 1. ER diagram for EDE User Interface

9.2.3 Example Interface



EMOTION DETECTION CONTROL PANEL

Sydney Australia

FINE TUNE WEIGHTING

HAPPINESS 6

SADNESS 2

ANGER 0

CALM -2

SELECT EMOTION

Happiness ☒

Sadness □

Anger □

LOG FILE

~/var/em_detector/logging.txt

Drawing 2. Mock up illustration of EDE User Interface

The illustration above depicts a possible user interface that performs four functions. First it allows the operator to select a region so the system would apply a pre-determined template to optimize perform for uses of the selected locality. This is similar to timezone selection when installing an operating system providing the operator with a selection of all localities catered for by the application.

The interface will also provide a check box selection giving the operator the ability to scan for one or many of the possible emotions offered by the system. A trim or fine tuning mechanism is also provided to allow the weighting of each emotion based on user feedback. For example if the system continuously returns a false positive of happy when the users are talking in a neutral state, then the happiness trim-wheel can be adjusted to reduce the false triggering on the happiness emotion. Finally the directory and filename for an appending log file can be nominated so the operator can analyse the system's behaviour.

9.3 Cultural Differences

A problem for the emotion detector is that it is difficult to achieve compatibility across different accents as described in section 4.3.1. Essentially a broad based emotion detector, unless using exceptionally complex algorithmic processing needs to be configured or trimmed to work optimally with a specific culture or accent. This same situation exists for present technology Natural Language Recognition (NLR) systems.

A configurable application interface such as the one described in section 8.2.3 may include a pull down menu that offers a choice of regions to help the technology zero in on a more accurate result. Like speech recognition, the engineers responsible for the emotion detection system in the respective localities need to research and configure the mechanism appropriately to suit local requirements.

9.4 Artificial Intelligence

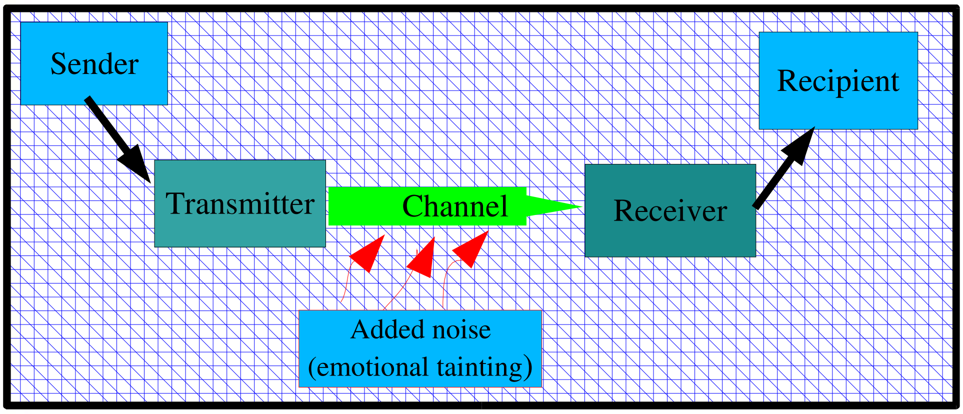
Learning algorithms have been a hot topic for developers over the last 20 years. Learning schemes range from simple decision trees to pattern recognition systems as discussed in Chapter 6 right through to very complex neural networks. Meta-learning schemes like AdaBoostM1 have been employed by some researchers (Oudeyer, Pierre-Yves., 2003[7]).

Name	***Description***
1-NN	1 nearest neighbour
10 – NN	Voted 10 nearest neighbours
Decision Tree/C4.5	C4.5 decision tree
Linear regression	Classification via linear regression
LWR	Classification via locally weighted regression
Voted perceptions	Committee of perceptions
SVM 1	Polynomial (deg 1) Support Vector machine
VF1	Voted features interval
M5Prime	Classification by M5Prime regression method
Naïve Bayes	Naïve Bayes Classification algorithm
AdaBoostM1/C4.5	Ada Booted version of C4.5

Table 9. List of 10 common learning schemes

9.5 The icing on the cake – speech recognition as a context analyser

The Shannon – Weaver model of communication describes a message delivered from the sender via a transmitter through a channel to a receiver then to the recipient



The emotion embedded in speech colours the spoken word and alters the meaning of what is said. Emotion can be paralleled to the noise interference in Shannon Weaver's transmission model above. In communication terms it may be described as semantic noise. It distorts the message possibly giving it meaning beyond the informational capacity of the spoken words themselves. The research described in this document deals with this *noise* by analysing the emotional content rather than the message itself.

The context or wording of speech can reveal much about a speaker's intention or state of mind. For instance a bank robber could not rob a bank unless he or she actually utters the words “this is a stick up”. In fact no amount of emotion expression will convince the bank teller otherwise. This contextual principle can be applied by the implementation of a speaker independent voice recognition (SIR) system. This technology has been in development for the last 20 years and is now proficient enough to be used in commercial applications such as phone banking, results services and airline reservation systems.

Used in conjunction with the emotion detection engine a SIR system can be programmed to intercept words used out of context in any 'expected' situation. The biggest target would be obscene or swear words which are generally indicate emotional negativity especially in a commercial customer facing situation. The fact is that most people are polite on the phone when they are calm. If the person is any way upset or angry politeness normally is put to one side and obscene language is often directed at the recipient. A word recognition system will be able to flag such word usage and act as a auxiliary monitor to the emotional detector.

There are two main types of voice recognition systems. Speaker Dependent Recognition (SDR) require training so it can train of the speaker's voice. These systems are affordable and are available through the open source community. They cannot, however, deal with more than one speaker without re-training. The Speaker Independent Recognition system has a predefined template which does not require training and works across a range of speakers. It can ofter work across different accents however performance suffers where there is a noticeable accent deviation. SIR systems are suitable for this application but tend to be very expensive to buy, implement and maintain.

Originally this research was to formulate an implementation of SIR in parallel with the EDE (see design specification in **Appendix B**). The system is easily implemented using one of the many SIR and SDR systems available. The modern SIR systems such as those provided by Philips, Speechworks and Nuance deliver excellent reliability over a broad range of speakers.

9.6 Summary

The voice features that comprise a person's voice contain certain features that both give character to our voice and reveals the state of a person's mind. Each feature has special qualities that computer technology can capitalize on and use to make analysis. Several mechanisms are available to the engineer to devise and develop such analytical devices. By using more features and more processing power the accuracy of these systems may be further improved. Artificial intelligence is the ideal mechanism to help evolve the the technology based on certain learning criteria. To further improve reliability an independent adjunct voice recognition system may be employed to scan for words that are obscene or out of context.

CHAPTER 9 Implementation of the Emotion Detection System

10.1 Outline

I considered the implementation of the technology as extremely important part of conducting research into emotion detection technology. Without prototyping, interested parties are unable to fully appreciate its operation and its applications in the real world. A mind loaded only with the theory of its operation will probably be unable to visualise how to apply such technology to the every day facets of life. Two phone based demonstrations and an off line laptop based application have been created to demonstrate the technology in what are, perhaps, some of its future applications.

10.2 Technical Overview

The Demonstration system is designed to illustrate how the technology works in a telephony environment. The system is comprised of :-

- A telephony grade IVR and switch connected directly to the PSTN network via ETSI ISDN.
- An Intel based PC 'adjunct' running Version 9.1 Mandrake Linux OS.
- PC applications written in Perl and Unix Script
- PRAAT Speech analysis software and script language.
- TCP/IP over Ethernet connection between the IVR and adjunct PC.

The demonstrations appear to run real time however the IVR actually gates the speaker's audio to the adjunct PC which records a snippet as a 2 second (.wav) file in standard telephony (PCM) format. The file is read into the analyser and then removed ready for the next sample. Total processing time for each utterance is about 3 seconds. I envisage this as being an acceptable lag for most real time voice applications. For a full explanation refer to the Design Specification in **Appendix B**

10.3 Emotion detection Using Simulation systems

A simple PC based application was created that prompts the user to record a two second utterance, analyzes it then reports to the screen the interpreted emotion. This portable application uses the same emotion detection engines as the IVR application. It is also a modular design , permitting the easy upgrade or renewal of the emotion detection engine. The stand alone application will also run on an open systems platform such as Linux or Microsoft's propriety Operating Systems, Windows 9X, XP and 2000.

10.4 Telephony Accessible Demonstration lines

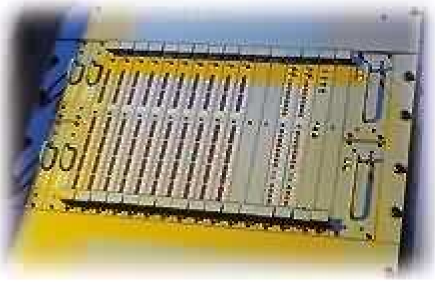


Illustration 12. Telsis FastIP IVR

Demonstration services where written on commercial quality Telsis FastIP Interactive Voice Response units (IVR). The Telsis equipment provides a high reliability platform with multiple open voice, telephony and data interfaces that are conducive to the design and implementation of such prototypes.

10.4.1 The "How are You?" Demonstration

The first demonstration is an IVR based service which greets the caller and politely asks them "How are you feeling today". The system waits in silence for four seconds for the caller to reposed. If it does not detect anything the system will state that it cannot hear and will ask the question again. This will continue until the caller either hangs up or utters a response. On uttering the response the system then analyzes the tonal aspects of the utterance to try and determine the emotional stress placed on it by the caller.

The service responds with an appropriate remark after which the caller is asked whether the service correctly interpreted the intended emotion OK. The call enters a response by either selecting '5' for OK or '1-4' representing the emotion that they intended to project if the service made an incorrect analysis. This information collecting part of the service constitutes the third experiment which provides a scale of reliability when the technology is used in a real world application. Please refer to page 94 in **Appendix B** for design specifications of the service.

10.4.2 The "Intercept" Demonstration

Multi-party conferences are common place in business and entertainment circles. Many of these are moderated by a human operator who has the power to listen in and intercept any party perceived as been rude, abusive or causing a general disturbance. However many conferences such as chat lines do not constantly have a mediator present. In some countries these conferences may continue totally unsupervised.

An application that would position a monitoring resource on each voice channel entering the conference would react if an individual became hostile, overly depressed or just frustrated. This functionality would certainly give more credence to the term “auto attendant”.

The demonstration service places two callers into a conference call. The maximum number in the conference is two for the sake of the demonstration. The emotion detection “listens” in on the conversation on a continuous basis. If anger is detected for greater than 3 second phrase segments the system breaks the conference and intervenes announcing to both parties that the conference has been terminated due to mis-conduct. Both parties are cleared down. Please refer to page 95 in **Appendix B** for design specifications of the service.

10.5 Summary

Most technologies require a tangible prototype to demonstrate its capabilities and its “touch and feel”. It is hard to appreciate the operation of an entity such as an emotion detector unless people are encouraged to use the system in one or more possible (commercially viable) applications. Without need there is no motivation. The demonstration applications created for this exercise are designed to explore the use of the technology in both the realm of entertainment (“how are you” demonstration) and the the commercial world (telephony intercept demonstration).

CHAPTER 10 Conclusion

The detection of emotions in speech is a technology in its infancy. Recently there has been a fair amount of research directed at this technology driven by commercial requirements and the need to further evolve human – machine interaction. Much of the research conducted to date has revealed that systems can perform with reasonably high accuracy and possibly as well as some of its human counterparts.

The technology may be paralleled to speech recognition twenty years ago, where initially the technology was labeled as too complex, computationally too expensive and not fitting any real commercial requirements. Like speech recognition, emotion detection has become a requirement in the entertainment and telecommunications markets. The research performed by large companies such as Sony and Microsoft are testament to this.

To engineer a computer based system that is capable of detecting human emotions in speech, a sound understanding of physiological processes that stimulate emotions and the effect they have on speech need to be researched and qualified. The study of emotion itself has been a grey area from the dawn of civilisation to the present time. Building a good recognition system first requires a thorough understanding of what is being recognised. Emotion recognition is a mosaic of computer science, speech pathology and human psychology.

Emotions and their influence on speech vary from individual to individual and from culture to culture. Interaction with human subjects presents hurdles of sizable magnitude, with the technology needing to be sophisticated and intelligent enough to “tune in” to a large cross section of users. Research into artificial intelligence and pattern recognition will pave the way to acceptable performance in much the same way as it has done for speech recognition systems.

Results attained from experiments conducted during this research confirm what other researchers have determined. Pitch and its second and third order derivatives are the most valuable features. Intensity too is important but can be a dangerous measure in light of variabilities out of the control of the system in an end to end scenario. Regardless of this the prototype analyser created as a part of this research, demonstrates a reasonable degree of accuracy and has potential to be improved considerably by involvement of extra voice features and possibly the inclusion of an user interface allowing the system to be tuned to local conditions. It is reasonable to assume that the technology is not only feasible as demonstrated by this research but practical in the sense that it has tremendous commercial potential.

It is also fair to say that human – machine interaction using speech recognition will be vastly improved by making the receptacle device empathic to the speaker. Predictably, in the near future, the will be a massive push to research and develop machines that behave in a human like fashion, changing the way we interact with computers – good material for an active imagination!

Emotion detection technology is a reality and has real commercial and social applications. It is difficult to imagine human – machine interaction evolving with out this technology being at the forefront of design.

APPENDIX A - BIBLIOGRAPHY

1. The Australian OXFORD Dictionary – Second Edition

2. Kollias, G., Piat, F.,"Future prospects for neural networks.", *NEuroNet Journal 2003*, Retrieved 16th August 2003.

3. Lee, C. M., Narayanan, S., Pieraccini, R., "Recognition of negative emotions from the speech signal", Paper - *Dept. of Electrical Engineering and IMSC, Uni of Southern California 2002*, Retrieved 16th August 2003.

4. Yu, T., Chang, E., Xu, Y.Q., Shum, H.Y., "Emotion Detection from speech to enrich multimedia content", *Microsoft Research Journal, China,* Date Unknown, Retrieved: 30th August 2003.

5. Lee, C. M., Narayanan, S., "Emotion recognition using a data-driven fuzzy inference system", *Extract from Eurospeech 2003, Geneva,* Retrieved: 26th August 2003.

6. Mandar, A., Rahurkar, J., Hanson, H. L., "Frequency band analysis for stress detection using a Teager energy operator based feature.", *Paper from Centre for Spoken Language Research, University of Colorado*. Date unknown, Retrieved:- 5th September 2003.

7. Oudeyer, Pierre-Yves., "The production and recognition of emotions in speech: Features and algorithms.", Paper*: Sony CSL Paris*, 25 November 2002, Retrieved: 10th September 2003.

8. Oudeyer, Pierre-Yves., "Novel useful features and algorithms for the recognition of emotions in human speech.", Paper*: Sony CSL Paris*, Date unknown, Retrieved 10th September 2003.

9. Polzin, Thomas S., Waibel, Alexander., "Detecting emotion in speech", Paper*: Carnegie Mellon University,* Date unknown, Retrieved 5th September 2003.

10. Polzin, Thomas S., Waibel, Alexander., "Emotion sensitive human-computer interfaces", Paper: *Carnegie Mellon University*, Date unknown, Retrieved 5th September 2003.

11. Van Rheed van Oudtshoorn, Nicholas., "Using pitch and amplitude to increase emotional certainty.", *extract from thesis conducted and the University of Western Australia,* 5th June 2003, Retrieved: 29th August 2003.

12. Dellaert F., Polzin, Thomas S., Waibel, Alexander., "Recognizing emotion in speech.", Paper: *Carnegie Mellon University*, Date unknown, Retrieved 16th September 2003.

13. Lieberman, P., Michaels, S. B., ""Some aspects of fundamental frequency and envelope amplitude as related the the emotional content of speech", *Journal of the acoustic society of America*. 34:922-927, 1962

14. Mozziconacci, Sylvie., "Emotion and attitude conveyed in speech by means of prosody", Paper*, Leiden University, the Netherlands,* Date unknown.

15. Boersma, Paul., "Accurate short term analysis of the fundamental frequency and the harmonics to noise ratio of a sampled sound", Paper– *Institute of Phonetic sciences, University of Amsterdam,* 1993

16. Van Rheede Van Oudtshoorn, Nicholas., "An emotional literature review", Literature Review – *Thesis University of WA, June 2003.* Extracted September 2003.

17. Louis ten Bosch., "Emotions: What is possible in ASR framework.". *ISCA Workshop on Speech and Emotion, Northern Ireland*, Newcastle 2000.

18. Holger Quast - "Robust Machine Perception of Nonverbal Speech",*Paper– Speech Research*, 2003.

19. Johnstone, Tom., Banse, Rainer., and Scherer, Klaus., "*Acoustic profiles in prototypical vocal expressions of emotion",* Dept. of psychology, University of Geneva, Date Unknown.

20. Ang, Jeremy., Dhillon, Rajdip,. Krupski, Ashley., Shriberg, Elizabeth., Stolcke, Andreas., *"Prosody- based automatic detection of annoyance and frustration in human – computer dialog"*, International Computer Science Institute, Berkly CA, Extracted October 2003.

21. "Know your body – Human Voice". *Web page provided by www.Fundooz.com, 2000.* Extracted November 2003.

22. Pitkow, Xaq., "Why do octaves sound the same?", *Preliminary Qualifying exam for Harvard Biophysics*, Spring 2000.

23. Cowie, Roddy., "Describing emotional states expressed in speech", *ISCA workshop on Speech and Emotion, Northern Ireland, 2000.* Extract October 2003 [www.qub.ac.uk/en/isca/proceedings/pdfs/cowie.pdf.

24. de Cheveigne, Alain., "A mixed speech F0 estimation algorithm", *Paper, Laboratoire de Linguistique Formelle, CNRS-University Paris 7, Paris, France. ,* Extracted November 2003.

APPENDIX B Design Specification

Capstone Project

Engineering Design Document

(Addendum to Engineering Report)

SPEECH BASED EMOTION DETECTOR

Denis Ryan

V0.04

14.1 Document Control

Document Version	Date and Author	Comment
Version 0.00	3-8-2003 Denis Ryan	Draft Release
Version 1.00	6-8-2003 Denis Ryan	Final Release
Version 1.10	10-10-2003 Denis Ryan	Updated to suit changed requirements. PRAAT S/W Optional Speech Rec.
Version 1.11	12-11-2003 Denis Ryan	Format revised – integrated into main report

Table 10. Design Specification Revisions

Introduction

15.1 Background

Electronic computing has been a part of industrial, technical and social culture for over 60 years.   Computers assist the human species in solving problems, increasing productivity and help manage complexity.

Throughout the last 60 years of computing development, engineering energy has been spent in making computers more accessible to a broader spectrum of people. Human - machine interface is fundamentally difficult as the two entities are both complex and very different in design and nature.

Humans are complex emotive biological organisms that process input via sensory processing regions in the temporal regions of the brain. Processing is based on interpretation of whats seen , heard, felt or smelt. Many of these inputs are coloured by associations and emotional states. On the other hand computers, while a complex entity, have rigid and well defined means of accepting and delivering information.

Human – Machine interface has developed extensively, with information delivered originally by switches and load buttons then punch cards, keyboards, pointing devices and more recently via speech recognition.

The latest technology of speech recognition provides a natural form of human machine interface. At last machines are able to interpret information contained in the words that are spoken to them, however they still cannot interpret other , possibly vital, information that is delivered with the spoken input. This other information may be emotional data representing, urgency, anger, confusion or elation. In many cases this input may be more important than the actual spoken words .

Project Outline

16.1 Project Description

There is a number of ways we can detect a person's emotional state. Every day we as human beings assess the emotional state of other humans when we interact with them. At a conversational level whether its face to face or over the phone we can "listen" to emotion in there speech.. how many times do we comment: - "You're amazingly chirpy today", "you sound stressed" or "Is something wrong – you sound so glum". We do not need to interact face to face with the person to analyse these states – over the telephone one can detect emotion simply by the level and intonation of the voice.

The chosen vocabulary also is a strong indicator of emotional alignment. An angry person may curse or use obscene language while an elated person may repeat words such as "great", "fabulous" or "wonderful". Everyday words we use can indicate many things depending on their context.

The project I have chosen to undertake for my engineering capstone is based on the development of a computer based system, designed to interpret human emotion via the spoken word. There has already been quite a lot of research in this area being conducted by individuals and specific organizations alike. The aim is to develop an economically viable system that can reliably interpret emotional state through voice, for the broadest spectrum of human users as possible.

The project will be conducted in two stages. First involves research and a Matlab simulation. The Matlab simulation will be able to process recorded . wav files and produce a reliable result based on the emotional quality of the speech recorded. Stage two involves the transfer of the system to a TI DSP development platform and using an IVR telephony route as an input, produce a reliable emotional analysis of the calling party's speech.

16.2 Document Description

16.2.1 Audience

This document is public engineering specification for the information of the client, (UTS project sponsor), other interested parties and a reference for the designing engineer.

16.3 Scope

This specification covers client requirements, data, functional, and behavioral aspects of the project as well as containing a full regression test plan, verifying the operation of the system.

16.4 Acronyms

Acronym	Description
EDE	Emotion Detection Engine.
IVR	Interactive Voice Response
SIR	Speech Independent Recognition
OR	Logical Function where a system has two binary inputs and one binary outputs, where a one on either or both input causes an output of one. Else the output is zero.
SR	Speech Recognition – Act of recognising speech
SRE	Speech Recognition Engine
DEL or PSTN	Direct Exchange Line – standard analogue pair , same as your home phone line.
E1 ISDN	2 Meg Telephony interface with ISDN overlay
ED System	The composite product which is a system comprising of the SRE and the EME to perform reliable emotion detection.
PRAAT	Open Source Voice Analysis Software. Used to derive speech characteristics, including formant, pitch, intensity etc.

Table 11. Common Acronyms

Software Requirement Specification

17.1 Narrative Reflecting Product Requirement

A technology needs to be developed as a part of UTS collaborative technologies project that is capable of discerning human emotion through speech. The technology needs to work across a wide range of speakers with a reasonable diversity in cultural backgrounds. The technology must demonstrate reasonable reliability and consistency and a prototype system must be created to demonstrate the technology.

17.2 Non-Functional Project requirements

The project need to be demonstrated in seminar before academics and leading members of the technology industry. The demonstration will need to be telephony based which gives access to the general public as well as demonstrating the system in a real world application. It is envisaged that telephony based industries. will be a major consumer of the technology

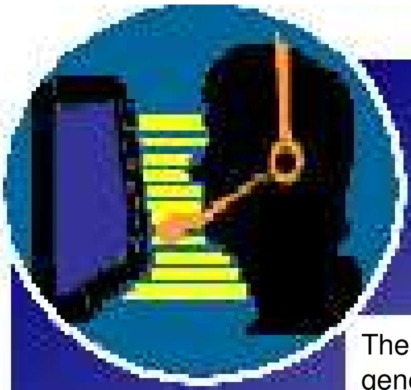
17.3 Executive Technical Summary

Emotional content of speech can be analyzed on two levels. Low level analysis is accomplished by filtering speech into constituent components and applying logic to decide on emotional state. A higher level analysis can be achieve using a speech recognizer to filter out particular words such as swear words or words used out of context.

The project will use a two stage detection system will will consist of a speech recognition based filter, referred to as "speech recognizer" form this point on, for the 1st stage and an DSP based filter ("emotion recognizer") as the final stage. What the speech recognizer does not pick up, the emotion detector may.

The construction of the filter involves simultaneously presenting the utterance to both the speech and emotion recognizers. Each recognizer analysis the input and produces either a true or false output or a fractional value between 0 and 1 known as a confidence level. Ideally the binary output from each recognizer are tied in a logical OR so if a true result is produced by either or both recognizers, a true result would be produced.

In its most fundamental functionality the system will be able to detect negative emotions, namely anger, within speech. In the case of this project the recognizer returns a "true" when a negative emotion is detected. The system could also be enhanced to detect a number of emotions such as happiness or sadness, however this project will concentrate solely on anger detection.



The speech recognizer will be a word based detector rather than the newer generation phoneme based detector. The reason is that the detection is only for a small subset of words (for example swearwords). Under these conditions word based recognizers work exceptionally well , are easier to set up and are less complex. A template of chosen words will be set for the recognizer with a cross section of accents representing the cultural distribution of Sydney Australia.

In order to achieve reliable emotion detection through speech DSP based detectors must be designed to analyze fundamental speech patterns. Th composite of these speech characteristics make it possible for a computing system to detect emotion. Four specific speech qualities can be analyzed to achieve this.

1 Pitch
2 Energy
3 Duration
4 Formant – Any of several Frequency regions of intensity in the sound spectrum which determines the characteristic qualities of sound.

This project will concentrate mainly on energy levels, pitch and duration to decipher emotional state. DSP Digital Filtering will be used to break down utterances into these key characteristics which are compared to stored templates. Results from this pattern matching operation are analyzed and inserted into an “Emotion Matrix”, resulting in the derived emotion associated with the utterance. The Emotion Matrix is the key mechanism in this project and will lend itself to different levels of complexity varying from simple preset vectors to Artificially Intelligent self learning systems. For this project this engine will remain fundamental so most energy can be expended in the reliability and accuracy of the system.

In essence, an affective computer would utilize the “Emotion Matrix” onto which the emotions are plotted against all the speech characteristics. There are a number of ways of constructing such a model.

The energy recognizer function will initially be simulated with Matlab functions which will process supplied audio files, analyze characteristics through a suitable digital filter simulation and generate an result that corresponds to the emotion detected in the audio.

At an abstract level emotion in speech can be categorized into two levels, negative and non negative emotions. Negative emotions include anger, frustration and despair while non-negative emotion includes, happiness, neutrality and delight. The separation of speech into these two emotional categories, while lacking the fidelity of full emotion analysis , will provide import information to any system processing spoken input. For example non-negative speakers may be well left alone while negative speakers may qualify for external assistance or support.

17.4 Requirements Validation

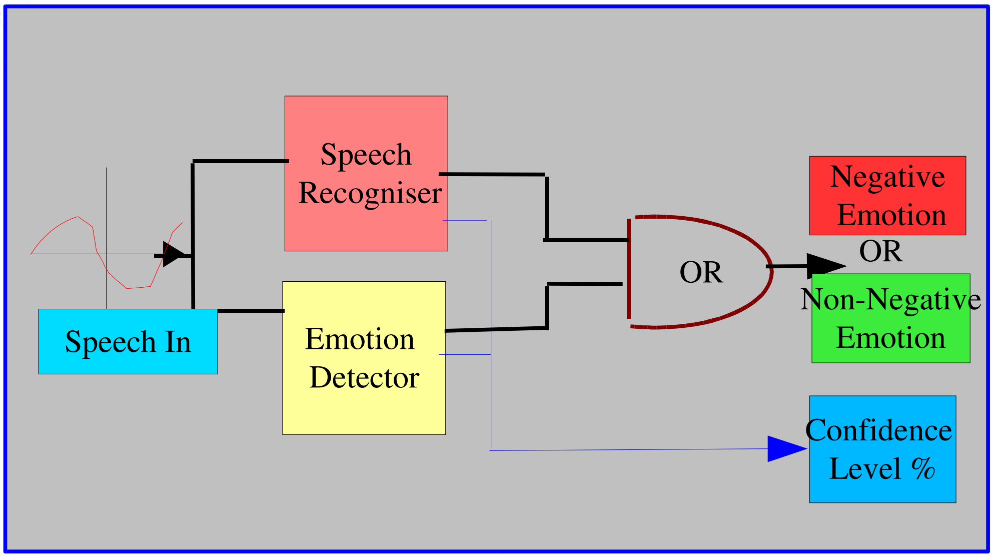
Below is a table of requirements that have been requested by the client.

Requirement Number	***Requirement***	***Requested By***
A1	Prime Objective: Matlab simulation of a DSP Filter system that can differentiate between negative and non-negative emotions in human speech.	A Kadi (Client)
A2	Inputs must be .wav files sampled at 8KB/s with 16 bit quantisation resolution.	D Ryan (Engineer)
A3	Output must be binary TRUE or FALSE where TRUE represents negative emotion. There must also be a confidence level returned where a certainty level between 0 and 1, where 1 represents 100% certainty of the derived result.	To Be confirmed by Client
A4	Reliability must exceed 75% for speech derived from English speaking candidates who have resided in Australia for more than 10 years. A successful result is one where the correct result is obtained with a certainty greater than 75%.	
A5	Submission of Report that adheres to the requirement as stated in the Spring 2003 Student Guide to Capstone Projects. Submission is required before 12th November 2003.	A Kadi

Requirement Number	***Requirement***	***Requested By***
A6	Bonus Objective Speech Recognition pre-filter to perform word scan to determine emotional state via context. Speech Detection scans for: Swearing Obscene words Loud Banging Huffing / Heavy Breathing. Other out of context words. Accuracy must be in excess of 90% which is that of many of the open sourced systems presently available.	A Kadi
B7	Bonus Objective: Implement Matlab simulation onto TI 6000 Development platform and perform in-line Emotion Detection on real time speech input.. Output must be binary TRUE or FALSE where TRUE represents negative emotion. There must also be a confidence level returned where a certainty level between 0 and 1, where 1 represents 100% certainty of the derived result. Overall performance should be within 90% of the simulated system.	A Kadi
C8	Bonus Objective. Implement on telephony system such as an IVR to permit easy demonstration. IVR system will also be capable of switching audio prompts that relay the perceived emotional state of any speech input.	D Ryan. To be confirmed by client

Table 12. Requirements

17.5 Context Diagram



Drawing 3. Context Diagram

17.5.1 Main Components

Component	***Description***
Analogue Input (For Requirement B6)	Analogue to Digital Sampling system. Speech pre-filtered between 300 and 3400 Hz and sampled at 8Khz @ 16bits.
Speech Recogniser	Open Source Base Speaker Independent Recogniser (SIR). Used to scan recognisable words or sounds that may indicate emotional state.
Emotion Detector.	DSP based filter that monitors speech energy and pitch to determine emotional alignment.
System Output	Log file and/ or messaging system containing information on emotional alignment during real time operation.
IVR	Interactive voice response that couples Emotion Detection Engine to telephony system. Utterances from caller phone directed to input of EME and messages returned from EME processed by IVR script which switches appropriate audio.

Table 13. List of main components

17.6 Description of main Components

17.6.1 Analogue Input

For the secondary objective of the project it is required that real time speech is processed. Microphone and amplification equipment produce an analogue input that is either 600 ohm balanced or 10 K ohm unbalanced. I will elected to use the more versatile 10K interface for this project. The Ti 6000 Development platform is also equipped with a 10K interface. This interface is not applicable for the Matlab simulation which will process pre-recorded .wav files.

17.6.2 Speech Recogniser

A HTK Speech Recogniser will be employed for this project. The HTK SR is an Open Source system originally developed by Cambridge university , but recently has fallen into the hands of Microsoft Corporation. The product has managed to remain open source for the time being and is largely used by developers.

The speech recogniser will operate in a somewhat special mode for this project. The SR will support a very small template of words and noises that are associated will agitated or angry client. The recogniser will serve to detect emotion by recognising such words in a person's speech. These words include:-

- Swear words
- Obscene Words
- Huffing / heavy breathing,
- Banging noises.
- Other words out of context.

17.6.2.1 *Emotion Detector*

The emotion detector will use digital filter technology to produce metrics of speech energy, tempo and pitch. The results of these measurement s will be fed into a tunable matrix which is instrumental in deciding on the emotional alignment of the speaker. Analysis will be performed on 5 – 10 second chucks of audio at a time. For the simulation a Praat speech analyser will be employed to extract particular features from supplied utterances. These features include:-

1. Vocal Formants F0 – F4
2. Intensity
3. Duration – pause length

17.6.2.2 *System Output*

Ideally the EDE must permit its analysis to be made available to adjunct equipment or processes that can perform action based on data received. The EDE will support output to an appending log file as well as a propriety messaging system. See section 17.8 for more detail.

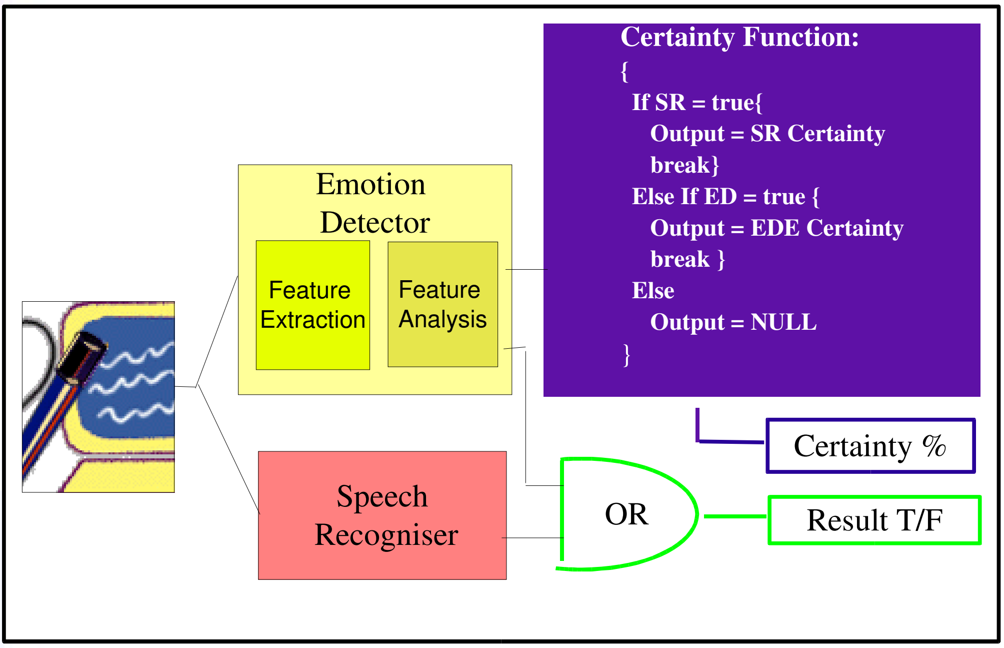
17.6.2.3 *IVR*

The IVR is an audiotex device that has a Direct Exchange Line (DEL) or a Digital telephony interface that permits routing to the EM test system via the PSTN. As well as this the IVR has the capacity to switch audio prompts back to the calling party based on messages received from the adjunct ED system. This permits the system to manifest human like responses such as “don't get angry please”, or “please calm down” when negative emotions are detected. This will serve to package the technology and provide a realistic environment for its application.

17.6.3 Speech Recogniser

Not implemented in first design phase

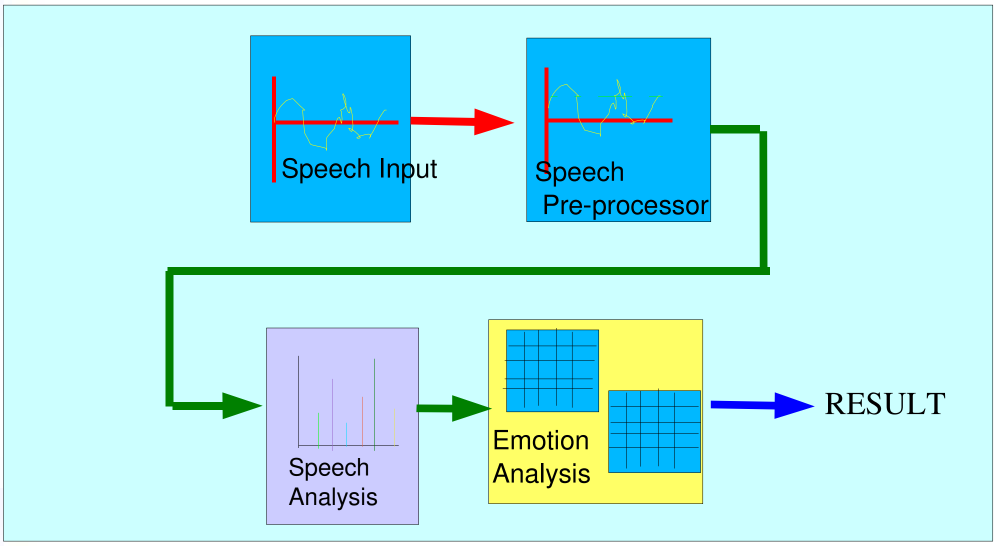
17.7 ER Diagram of Major Components



```
Certainty Function:
{
 If SR = true{
   Output = SR Certainty
   break}
 Else If ED = true {
   Output = EDE Certainty
   break }
 Else
   Output = NULL
}
```

Drawing 4. Entity Relationship Diagram of main components

Schematic Diagram of Subsystems



Drawing 5. Schematic of Sub Systems

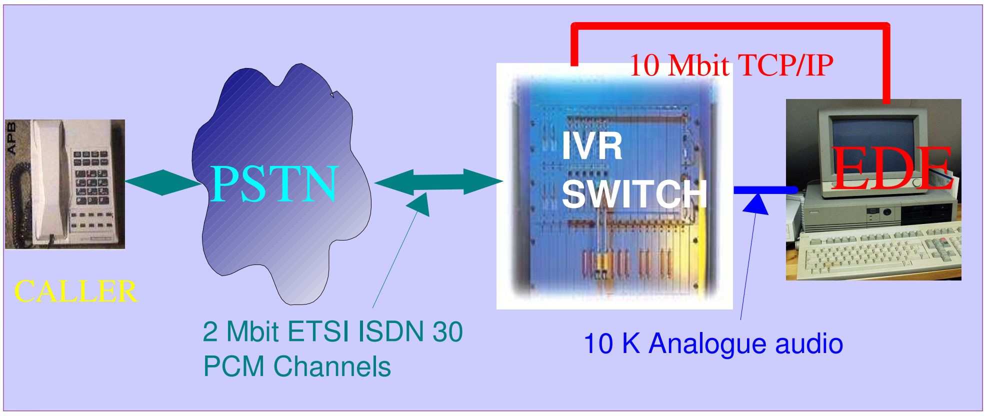
17.8 Demonstration Services

Demonstration services have been provided to demonstrate the viability and performance of the technology. There will be two services that are IVR based.

17.8.1 Demonstration 1 - "How are You" Service

The first Demonstration service will be known as the "How are you service", which ,as the name suggests, asks the caller about there state of mind. The caller's response to that question is processed by the emotion Detector which is interfaced adjunct to the IVR system via a TCP/IP connection. The EDE supplies the IVR with its interpretation by which to IVR plays out to the caller an appropriate comment. The demonstration also has a statistics gathering facility where by the caller who has had their emotions analysed is asked whether or not the system successfully assessed their state of mind.

17.8.2 Schematic



Drawing 6. Schematic of Telephony and Data processors in Exp. 1.

17.8.3 Technical Outline

Calls to the system are made via the PSTN which connects to the Telsis FastIP switch via an ETSI ISDN 2 MBit Interface. Calls are handled by the IVR switch which has the ability to switch pre-recorded audio to the caller and process DTMF tones as input to menu selection. The FastIP IVR interfaces to an adjunct Intel PC running a LINUX OS via TCP/IP over Ethernet. There is also an analogue audio feed via a 10K interface from the IVR to a Sound card fitted to the PC.

Calls made into the IVR are greeted with pre-recoded audio. The audio generated by the IVR and spoken by the caller are gated through the audio feed to the PC in analogue form via the 10K interface. A UDP message is sent from the IVR to the adjunct PC to start recording the utterance which the caller is prompted to say. Unix SOX application is used to record and store the audio in 8000 bit sample rate 16 bit resolution mono as a windows PCM format (.wav).

After recording is complete (2 seconds adjustable record time) a Perl script is initiated to pass the recorded file to the PRAAT analyser which subsequently generates a result communicated back to the IVR via a UDP message. The IVR then responds to the caller by playing out the appropriate audio indicating the analysis of the emotion by the system. The caller is then invited to answer a menu driven survey that collects information pertaining to the accuracy of the system.

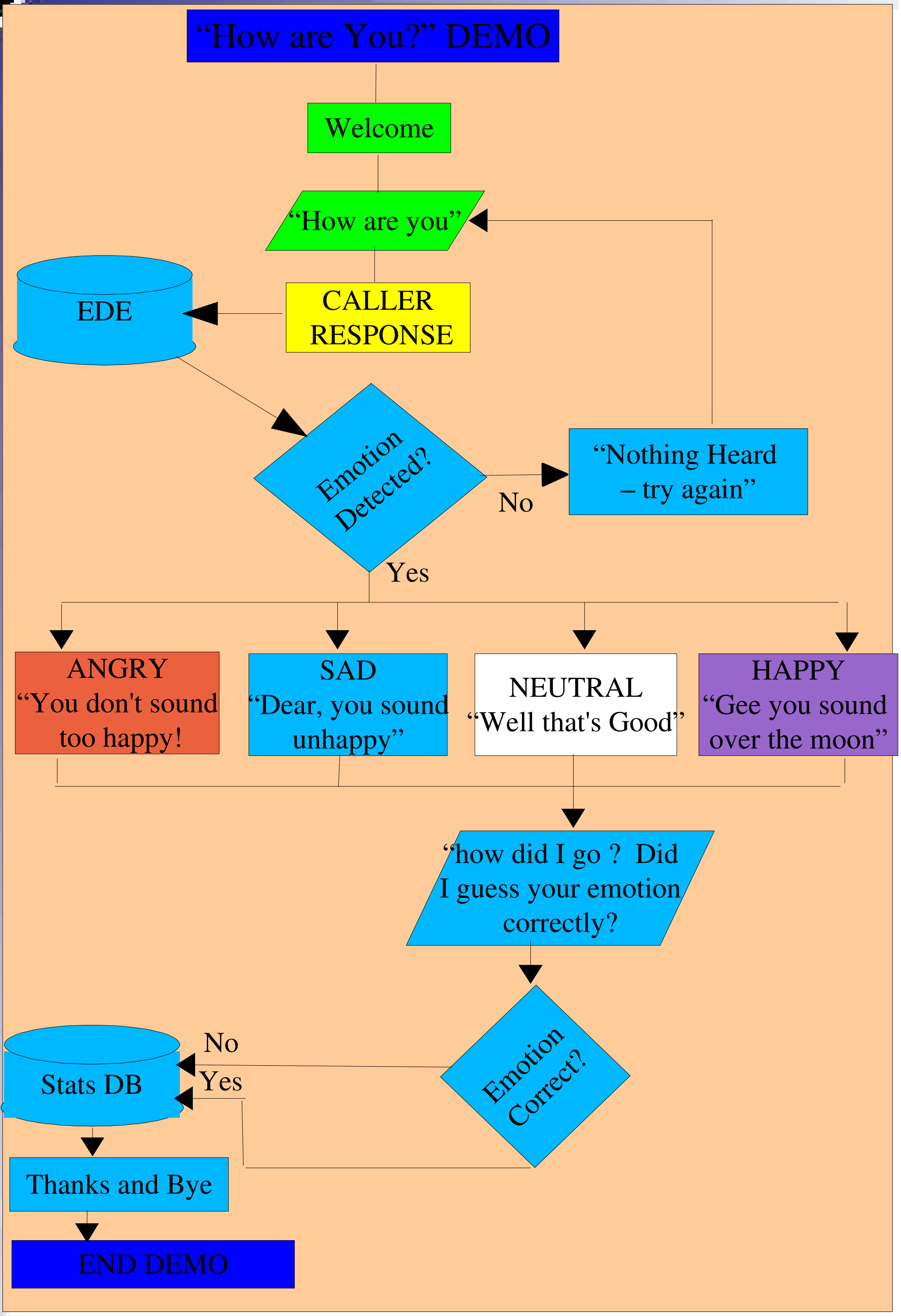
17.8.4 Equipment Description

IVR – The FastIP is a 120 port IVR Connected to a 480 port Intelligent switch. The system can handle 120 calls simultaneously and is completely non-blocking by design. It has ESTI E1 Telephony interface a 10MbIt Ethernet LAN running TCP/IP and 12 analogue 10K audio ports accessible by all voice channels.

ADJUCT PC (EDE) – Intel based “PC” computer running 2.2 GHz Celeron processor. Equipped with 100 MBit Ethernet Interface and Sound card with 10K line port.

17.8.5 Flow Chart of “How are You?” DEMO Service

DEMO 1 FLOWCHART


“How are You?” DEMO
Welcome
“How are you”
EDE
CALLER RESPONSE
Emotion Detected?
No
“Nothing Heard – try again”
Yes
ANGRY “You don't sound too happy!
SAD “Dear, you sound unhappy”
NEUTRAL “Well that's Good”
HAPPY “Gee you sound over the moon”
“how did I go ? Did I guess your emotion correctly?
Emotion Correct?
No
Yes
Stats DB
Thanks and Bye
END DEMO

Drawing 7. Flowchart for Demonstration 1.

17.8.5.1 Script for “How are you?” Service

The table below is the script of the recorded audio that is required to be preloaded onto the IVR system for the demonstration. Files are recorded in Windows PCM format and are converted to the Telsis proprietary .AXR format. The files are stored on the IVR as 32Kbit ADPCM Fixed point Alaw files.

File Offset	***Prompt***
1	Good morning,
2	Good Afternoon
3	Good Evening
4	Welcome to the emotion detection system. How are you feeling today?
5	Sorry I didn't hear that -. So, how do you feel today ?
6	Well you certainly sound on top of the world – what happened – win the lottery?
7	Oh dear, it really sounds like we have our tail between our legs. Cheer up – it bound to get better.
8	That's great. Good to hear your as calm and relaxed as always
9	Wow! Get out of the wrong side of the bed did we??? Promise not to bite my head off if I speak again?
10	Well how did I do? Did I detect your emotional state of mind correctly? If I did please dial 5 or if I got it wrong dial the appropriate key reflecting the emotion you thought you were projecting If you were supposed to be happy dial 1 If you were sad, dial 2 If you were just neutral dial 3 Or if your were supposed to be angry dial 4.
11	Sorry that's an invalid option, please try again
12	Thanks a million for trying this demonstration. If you have any comments you would like to leave regarding the demo or the technology in general please log onto www.denisryan.com click on the Leyland P76 picture and post your comment in the Emotion detection forum. Good Bye.
13	Please wait while I decide on your mental state..
14	Golly I cannot hear a thing – maybe its me – you have to try me later – maybe from a different telephone.
15	Sorry there has been a technical error, please contact www.denisryan.com and post a email message with the webmaster to alert me of the problem, thank you'll

Table 14. Demonstration 1 Script

17.8.6 Program Code.

17.8.6.1 *Outline*

No less than four different script languages where employed for the creation of this service. The scripts run on both the IVR and the adjunct PC with IP communications and interprocess signalling managing the co-ordination of the system.

17.8.6.2 *PDL – IVR Main Script*

The FastIP is programmed in a script language known as Programme Definition Language (PDL) which is a high level language purpose designed for IVR Voice applications. The IVR processes telephony calls switch s and records audio and manages IP communications via PDL commands. A PDL process runs for every port on the machine independently providing a robust non blocking environment for multi caller services.

```
\**********************************************

\ "How are you" Emotion Detection DEMO

\

\ Denis Ryan

\

\ 8-10-2003

\

\ Version 0.05

\***********************************************

Main:

  G 1

  USER 8

  @100 = 30    \TCP Timeout (seconds)

  @101 = X    \Feed port number

  @105 = 0  \Local IP Address

  @106 = 0

  @107 = 0

  @108 = 0

  @109 = 2000

  @110 = 192  \Target IP Address for Adjunct

  @111 = 168
```

```
@112 = 0
  @113 = 70
  @114 = 2000 \Port # for adjunct application
 \
 \  @115 = 192  \Listening IP Address
 \  @116 = 168
 \  @117 = 0
 \  @118 = 70

 \  @119 = 2002 \Listening port for response
\
\  @200 Emotion detected
\          0 = No Detection
\          1 = Happy
\          2 = Sad
\          3 = Neutral
\          4 = Angry
\
  @333 = 12337 \Check emotion command
  @334 = 12338
  @889 = 1    \fool system into tone train
  @943 =943    \menu sub
\
\  @400 - @403 \Source IP address from client
\
  ZCLR, CloseSocket
   ABASE = 2999
  @891 = 2999       \File base

   USER 8
   JSR ConnectA                              \Start with TCP
   Q %H > 11, NotMorning
   P.ABASE 1     \'Good Morning
   J Begin
```

```
\
NotMorning:
    Q %H > 17,NotArvo
    P.ABASE 2      \,Good arvo
    J Begin
\
NotArvo:
    P.ABASE 3   \'Good evening
\
Begin:
    \ROUTE UNIT.Y > UNIT.Y

    \G 1         \Turn gate 1 on
    P.ABASE 4      \'How are you today?
    B 4
\
    \R = 4
   \ ROUTE UNIT.Y > TA.0
    H 4
    WAIT
    JSR StartRec \ready recogniser
    T = P + 4    \Set Time
HERE: ZDUR, FInRec
    \G 1
    \H 32        \ Silence for response
    J Here
FinRec:
    P.ABASE 14  \'please wait
    JSR GetEmotion
\
\
    BLOCK
    #Z = #Z + 1 \inc counter
    C = #Z
    !C = @200   \Load system's guess
\
    Q@200 = 999,TechError
    G = @201-12330  \get offset
    P.ABASE G   \Play response
\
    @890=11        \'How did I do??
    @801 = 0
    @802 = 5 \highest jmenu option
    @803 = 1 \lowest menu option
    @804 = 3 \3 secs timeout
    @805 = 3  \MAX attempts
    JSS @943    \Menu SUB
    BLOCK
    C = #Z      \reload in case of corruption
    !C = @1     \load choice into U8
    A#,CloseandReturn
    P.ABASE 13 \'Thanks and bye
                  D
TechError:
    P.ABASE 15 \'tech error email denis
\ ********************************************
\  SUBS
\
\Conect to Adjunct
ConnectA:
\
\   #G=0
    DO IP.SOCKET A, #S, 4   \Create Socket      **XMIT **
WaitSock:Q A< 0,WaitSock
   Q A <> 1,Beep
  \ DO IP.BIND A, #S, @105  \Bind to local address
\WaitBind:   Q A<0 ,WaitBind
  \ Q A <> 1,Beep2
\
    B=10
 DO IP.RECVFROM @99, #S, @200, B,@110 \Got Connect now wait

\con: DO IP.CONNECT A, #S, @110 \Connect to Adjunct
\  SKIP A < 0
   JRET
 \ J con
Beep:  B32
   WAIT
   D
Beep2:
   B2
   H 2
   B2
   D
SendErr:
   #T = 666
   #S = A
   B 2
   H 2
   B 2
\ Send Command
\
StartRec:\D=0
   WAIT
StartD:

    DO IP.SENDTO A, #S, @333, 2,@110 \Send command 1
WaitSend:Q A< 0,WaitSend
   Q A <> 1 , SendErr

EndD:
 \DO IP.CLOSE A, #S
\WaitClose:   Q A < 0,  WaitClose
\
JRET
\----------------------------------------------
\ Then Await response from sub Process
\
GetEmotion:
 \  DO IP.SOCKET A, #T, 4   \Create Socket
 \  DO IP.BIND A, #T, @105  \Bind to local address
 \  DO IP.LISTEN A, #T      \Wait for Response
   T = P + @100            \Set TCP timeout
   ZDUR, TCPRcvError
\
\AC: DO IP.ACCEPT A, #S,@400,#T
 \  SKIP A <> 1
  \  J GotSomething
   \ J AC
 \
GotSomething:b1
 \
h 2
    Q @99 < 0,GotSomething
    Q @99 <> 1,TCPRcvError
    ZDUR, OFF
\
   \DO IP.CLOSE A, #S
   \DO IP.CLOSE A, #T
JRET
\
TCPRcvError:ZDUR,OFF
  \  @200 = 999            \Comms Error
    #T = 667
    #S = @99
    JRET
CloseSocket: WAIT
   ZOFF
```

```
DO IP.CLOSE A,#S
  #S = 0
  D

CloseAndReturn:
  DO IP.CLOSE A,#S
  #S = 0
   J Main

\=============================================
```

17.8.6.3 IVR – IP FEED Service – Maintenance task.

The IP Background Process Creates and Binds an ID to a TCP socket . The process monitors errors generated by the Main caller tasks and resets the socket if necessary.

```
\  UDP Socket Feed Routine
\  Version 1.00
\****************************************
\
USER 8
\
  @105 = 0  \Local IP Address
  @106 = 0
  @107 = 0
  @108 = 0
  @109 = 2000
  @110 = 192  \Target IP Address for Adjunct
  @111 = 168
  @112 = 0
  @113 = 70
  @114 = 2000 \Port # for adjunct application
 \
#T = 0  \Reset to start
#S = 0
```

```
I = 1
Do IP.CLOSE A, I
I = I + 1
Q I > 99,Main
\
Main:
    DO IP.SOCKET A, #S, 4   \Create Socket       **XMIT **
WaitSock:Q A< 0,WaitSock
    Q A <> 1,ErrSocket
    DO IP.BIND A, #S, @105  \Bind to local address
WaitBind:   Q A<0 ,WaitBind
    Q A <> 1,ErrBind
    H 24
    WAIT

MonLoop:
    Q #T < 2,MonLoop
    Q #T = 666,ErrSend
    Q #T = 667,ErrRcv
Reset:                          \Kill sockets and restart
    DO IP.CLOSE A, #S
    #S = 32000
    #T = 32000
    H 16
    WAIT

    #S = 0
    #T = 0
    J Main
\
ErrSocket: #S = 997   \Flag Socket Error
    H 16
     WAIT
     #S = 99
```

```
H 16
    WAIT
   J ErrSocket
\
ErrBind: #S = 998     \Flag Bind Error
   H 16
    WAIT
    #S = 99
    H 16
    WAIT
   J ErrBind
ErrSend:D = 5      \Flag Send Error
ES: #S = 996
   H 16
    WAIT
    #S = 99
    H 16
    WAIT
    D = D - 1
    SKIP D < 1
   J ES
   J Reset
\
ErrRcv:    \Flag receive Error
   D = 5
ER: #S = 995
   H 16
    WAIT
    #S = 99
    H 16
    WAIT
    D = D - 1
    SKIP D < 1
   J ER
   J Reset
```

17.8.6.4 Perl Script

Perl was chosen as the major scripting language in this project because of its simplicity and its powerful yet easy to implement TCP sockets. The Script Awaits for a UDP datagram from the IVR and after checking the command goes to execute then original Unix script that performs the record and analysis of the utterance provided by the IVR.

```
#!/usr/bin/perl
#$counter = 0;
while (1 == 1)
 {
#   $counter = $counter + 1;
   $port = 2000;
   use IO::Socket;
#Set socket and listen for client connection from Hi-Call
   $server = IO::Socket::INET->new
   (
                        LocalPort=> $port,
                        Type     => SOCK_DGRAM,
                         Domain   => PF_INET,
                         Proto    => 'udp',
                        #Listen  => 1,
                        #Reuse   => 1,
   ) || die "Bind failed\n";
#Accept connection - client is the descriptor for the client
#machine that initiated the request
  # if ( $client = $server->accept()){
#                       print "\n Accepted Connection from Hi-Call";
# Wait for Connection
 while ( $client = $server -> recv($line,128,0))
   {
     if ($line eq "01")
     {
 #Start getSound script
         print "\nGot Hi-Call Command to GO with command $line\n";
```

```
        system("./getsound1.sh");
    }
    else{
         print "\nGot a command but it was $line\n ";
     }
  }
close ($server);
}
```

17.8.6.5 Unix Script.

The Unix script was originally used to run inbuilt Unix applications such as SOX record and play. The is script is initiated by the Perl script that listens to the command from the IVR. The script starts the REC process as a background task then ,sends a CNT Break to the Process ID to end recording and kill the task. The PRAAT analysis routine is then called to analyse the recorded .wav file.

```
#!/bin/bash
#getsound1.sh
cd /home/denis/Documents/UNI/Thesis/PRAAT
    socketID = $1  #load socket info for Comms


                            #echo "Start Recording!"
                            #/usr/bin/rec -r 8000 /home/denis/Documents/UNI/Thesis/PRAAT/test_sound.wav&
                            echo $! " = Rec task number"
                            echo
                            /bin/sleep 1
                            #echo "Recording Complete"
                            #/bin/kill -TERM $! # sends SIGTERM (CRTL-C) to last background job

                            pid=$!
                            ((pid = $pid + 4))
                            #/bin/kill -INT $pid # sends SIGTERM (CRTL-C) to last background job +4
                            echo $pid " = killed task"
                            #/bin/kill -9 $pid # sends SIGTERM (CRTL-C) to eradicate pid
                            /bin/sleep 2
```

```
                    ./praat check_emotion     #PRAAT Script


                    echo
                    echo "PRAAT task number = "$!
                    echo
                    #/bin/kill -9 $! # sends SIGTERM (CRTL-C) to last background job
                    #/bin/kill -TERM $! # sends  (CRTL-Z) to last background job
                    #rm /home/denis/Documents/UNI/Thesis/PRAAT/test_sound.wav
                    #echo "removed /home/denis/Documents/UNI/Thesis/PRAAT/test_sound.wav"


#INT = CNTL C
#TERM = CNTL Z
```

17.8.6.6 PRAAT Script

Perhaps the most crucial script here is the PRAAT script which directs the interrogation of the recorded utterance and calls one of five possible Perl scripts that returns the result in the form of a UDP datagram directly back to the IVR. The IVR then plays the appropriate response.

```
form Enter Filename excluding the .wav extension
   text  Filename "test_sound"
endform
ip_type$ = "UDP"  #
# UDP or TCP


pitchSD = 999
Read from file... /home/denis/Documents/UNI/Thesis/PRAAT/'filename$'.wav
 select Sound 'filename$'


# select LongSound test_sound


 To Intensity... 100 0
 intensitySD = Get standard deviation... 0 0
 select Sound 'filename$'
 To Pitch... 0 75 600
```

```
pitchSD = Get standard deviation... 0 0 Hertz
 pitchQuant2nd = Get quantile... 0 0 0.5 Hertz
echo filename = 'filename$'
echo pitchSD = 'pitchSD'

if pitchSD = 999
   emotion = 9
else
 if pitchSD < 30
                            emotion = 1
 elsif pitchSD > 75
                            emotion = 2
 elsif intensitySD <9
                            emotion = 3
 else
                            emotion = 4
 endif
endif

 if emotion = 1
                            echo "Neutral"

 elsif emotion = 2
                            echo "Happy"

 elsif emotion = 3
                            echo "Sad"

 elsif emotion = 4
                            echo "Angry"


 else
        echo "No detection"
 endif
 if emotion = 1
                    echo "Neutral"
                    system cd /home/denis/Documents/uni/thesis/PRAAT; perl TCP_neutral.pl

 elsif emotion = 2
                    echo "Happy"
                    system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl TCP_happy.pl

 elsif emotion = 3
                    echo "Sad"
                    system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl 'ip_type$'_sad.pl

 elsif emotion = 4
                    echo "Angry"
                    system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl TCP_angry.pl
 else
      system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl TCP_no_activity.pl
                    echo "No activity"
 endif

#system rm 'filename$'.wav
exit
```

17.8.7 "In-line Monitoring" service

The second demonstration service is the "in-line monitoring" service which demonstrates a possible commercial application of the technology. The in-line monitor listens to a conversation on an one of a number of voice channels and may be programmed to interrupt the conversation if it detects a specific emotion. A Java based client will provide a GUI User Interface allowing the system to be configured to a number of personal preferences.

Another two demonstrations will follow. One where the caller dials in an option (1 or 2) if the caller has guess her or his emotion correctly, and the other which will interrupt the speech channel and interject if the system detects you are angry. I my even make this a full telephony route so the system will work in a live phone call!

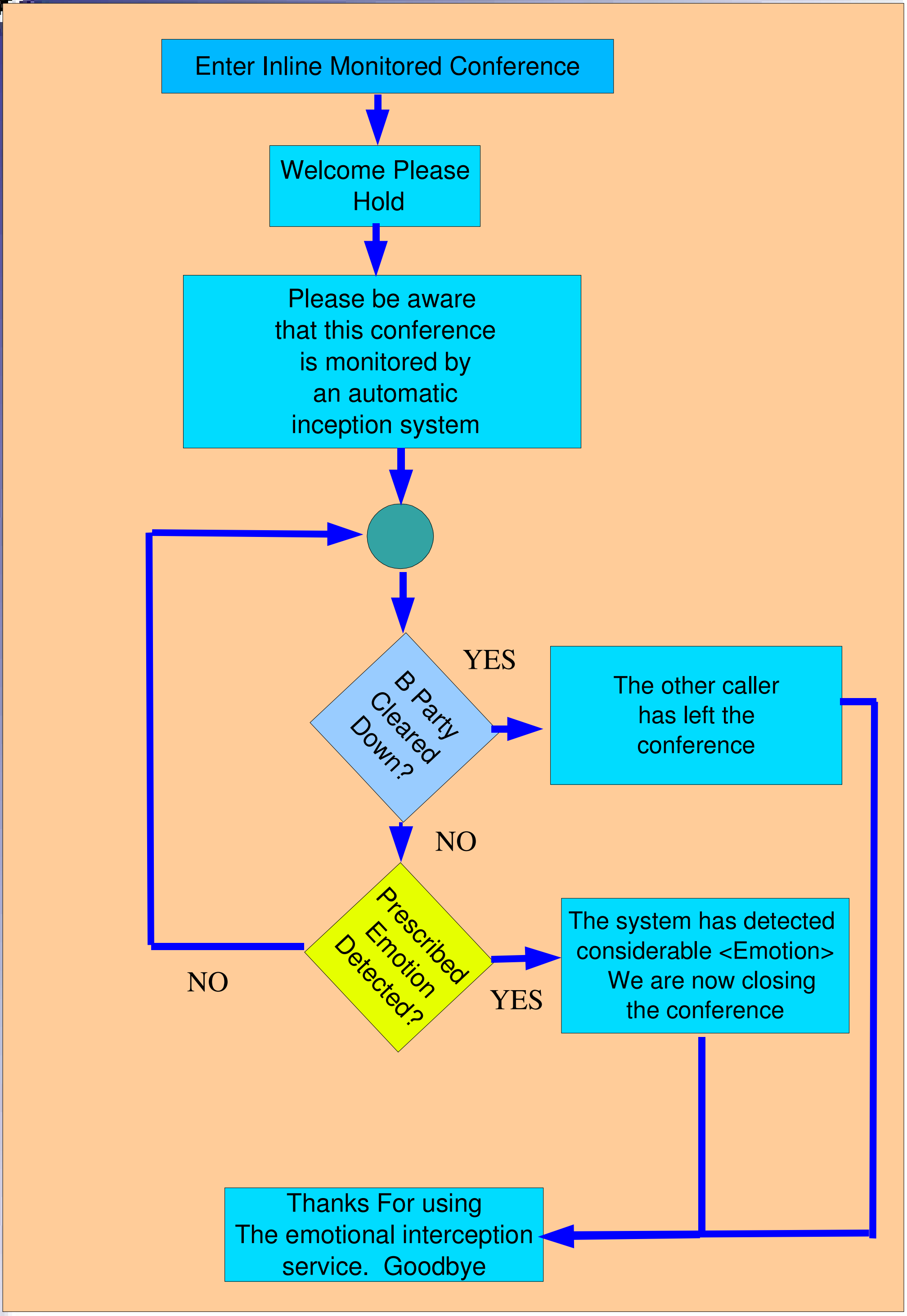
17.8.8 Technical Outline

The interception demonstration is based on the same hardware as the first demonstration. It is front-ended by an IVR connected directly to the PSTN and connects to an adjunct PC via an analogue audio feed and TCP/IP interface. Calls are made into the IVR which greets the callers individually then proceeds to route the two callers together.

In actuality the demonstration should work like a calling card system where a caller dials in, enters a destination phone number them the IVR dials and and connects through. However this was not possible as the host company would have to pay out dial costs.

Once connected the two callers are free to converse to each other in a standard telephony fashion apart from the fact that the emotion detection system is monitoring their conversation. On detecting three consecutive utterances that contain the selected emotion, the IVR is commanded by the adjunct to intervene. The intervention is executed by breaking the conversational route and the IVR announcing to both parties that the conversation has been ended due to detection of a particular emotion. Both parties are then cleared down.

DEMO 2 FLOWCHART


Enter Inline Monitored Conference
Welcome Please Hold
Please be aware that this conference is monitored by an automatic inception system
B Party Cleared Down?
YES
The other caller has left the conference
NO
Prescribed Emotion Detected?
NO
YES
The system has detected considerable <Emotion> We are now closing the conference
Thanks For using The emotional interception service. Goodbye

Drawing 8. Flowchart for Demonstration 2.

Functional Modelling

18.1 Technical Explanation

This section explains how the EDE works from start to finish. A general overview that is a short explanation of how the system works leading to a breakdown of system facets , each with its own technical explanation. The overview is a generalised summary which does not distinguish between the simulated and the DSP based product.

Technical Overview

The Emotion Detection system is composed of two main analysis engines:- 1. A Speech Recognition Engine and 2. a Emotion Detection Engine. Both of these systems are utilised in parallel to provide adequate detection properties. The SRE is the system's first line of defense, designed to intercept obscene, unsuitable or out of context words that enter the voice channel. The SRE used standard speech recognition technology that is easily available.

Words that are judged suitable by the SRE a then interrogated by the EDE. The EDE bases its decision on emotional alignment by analysing certain aspects of speech. Next speech is delivered to pre-processor routine determines background noise and sets a threshold level used which is required for analysis. Aspects such as intensity, pitch, formant, silence duration are extracted using a speech analysis routine. The data derived is then delivered to a decision making routine that compares derived vectors to predefined “emotion matrix”. The degree of fit to this matrix determines the emotion and certainty of the emotion detected.

Speech Recogniser

Audio Preprocessor

The audio Preprocessor has two main functions. The first is to adjust the amplitude threshold to account for background noise. Background noise must be assessed and used to gage the amplitude of the input speech. For example a speaker may have high volume, not because of any emotional state but because they are speaking within a loud environment. A high amplitude may otherwise give us false information deducing the caller is angry rather than trying to speak above noise.

The next function of the preprocessor is to normalize gender. This is because a female voice is about one octave higher than their male counterparts.

18.1.1 Speech Quality Matrix

Based on derived speech characteristics the table below depicts the speech qualities associated with different emotions. Emotions investigated are:

1. Neutral
2. Elation
3. Sadness
4. Frustration
5. Anger / Rage

These emotions are further split into two groups:-

1. Negative
2. Non-negative.

NB: Although sadness is generally regarded as a negative emotion, in this case it will remain unclassified due to its difficulty in differentiating from neutrality and because general commercial requirements disregard sadness ass a negative emotion.

Table Of Emotions and associated Voice characteristics

EMOTIONAL MATRIX ***Properties***	***Positive (non-negative) Emotions***		***Non Classified***	***Negative Emotions***	
	Elation	**Neutrality**	**Sadness**	**Frustration**	**Anger**
Mean F0 (pitch)	High	Med – high	Low		High
F0 Floor					
F0 Range	high		low		high
Intensity (Volume) Max	medium	Low-medium	low	Medium - high	high
Intensity Range	high	medium	low	medium	low
Pitch Contour	rising	rising	falling	falling	falling
Syllable accenting	few	few			many
Last syllable	accented	accented			Not accented
Av. Phoneme Duration	medium	medium	Long - medium	long	short
Phoneme Duration Range	low	low	high	medium	low
Rhythm (speed)	Medium - fast	medium	slow	slow	fast
Average Smoothed Pitch					
Smooth pitch Maximum					
Smoothed Pitch Minimum					
Smoothed pitch range					

Table 15. Emotion matrix - properties

18.2 Operation via Simulation Applications

18.2.1 Input File specifications

Files will be have 8 bit sampling at 8000 Hz to be compatible with telephony 64000 KBit Channels. Simulation will load .WAV files recorded at these specifications.

18.2.2 Praat Feature Derivation

The Praat acoustic analyser is a key tool in the analysis of speech. It is capable of isolating the facets of speech that we may view as unique pertaining to emotional analysis. The following parameters will be examined and used as an input to a Matlab based decision making matrix.

18.2.2.1 *Speech Amplitude.*

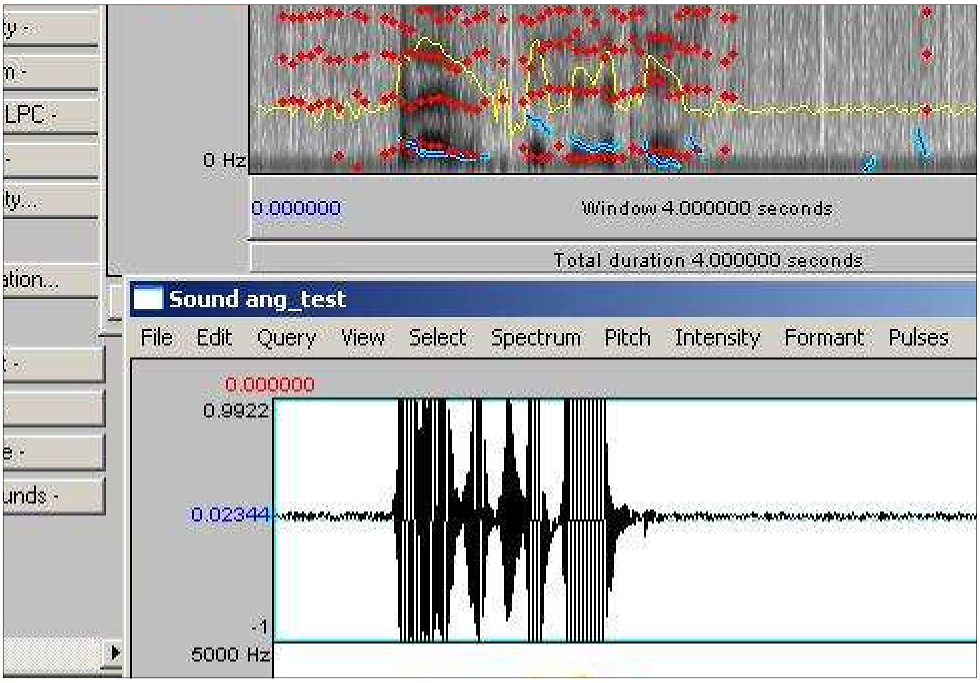
Although this will vary from person to person and be affected by line attenuation, this is still a key indicator of emotional state. A huge swing from low to high amplitude may indicate anger or rage. A low amplitude may indicate sadness or even confusion.

18.2.2.2 *Formant.*

Formant represents the the base frequency and harmonics associated with the generation of sound. Musical instruments have particular qualities and irregularities that create formant tones over and above there base resonant frequencies. The same applies to human speech. The position of the tongue , direction of breathing an shape of the glottis regions contribute to additive formant tonnes that create the total voice quality of human speech. By studying these formant tones and analysing there prevalence during different emotional states will help to provide data in deciding on emotional state.

18.2.2.3 *Pitch.*

This has many definitions in acoustics but when referring to speech it is the composite harmonic that makes the sound of the human voice. The pitch varies as we talk due to variations in the fundamental oscillation in the larynx as well as formant variation as discussed above.


0 Hz
0.000000
Window 4.000000 seconds
Total duration 4.000000 seconds
Sound ang_test
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
0.000000
0.9922
0.02344
-1
5000 Hz
LPC -

Drawing 9. PRAAT waveform display and editor

18.2.3 PRAAT Feature Analysis

Voice feature vectors received from the Praat analyser are compared with reference templates to determine emotional condition. The closeness of the match to any one template represents the confidence returned by the Matlab program.

18.2.3.1 Outputs

One possible variant of the product the emotional detection will deliver one of two possible results. Negative emotions include anger, rage and frustration and will return a TRUE result. Emotions of neutrality, happiness and others such as boredom and sadness will return a FALSE result.

The multi-emotion detection will produce an output that indicate which one of the multiple possible emotions has been detected. Accompanying this result can be a confidence level where the system delivers a certainty of the result produced based on the fit of the speech sample to the pre-defined emotion matrix.

18.3 Operation via DSP Development board:

A DSP or embedded system implementation is the final step in the development of the emotion analyser. This will provide a robust system that can be run efficiently on purpose build wave processor architecture. This will enable the technology to be integrated with current voiced based technologies such as SIR and IVR which are all based on DSP architecture.

18.3.1 Functional Description of input and output interfaces

18.3.1.1 *Software based application.*

Two possible scenarios exits for the detection of emotion in speech. A store an forward system may be employed for the off-line or batch processing of recorded speech. This may be used in analysis off line for the statistical processing of a large number of recordings or for speech therapy. The other application is real time processing or instantaneous analysis which processes and delivers a response instantly for immediate processing. The technology in this form is ideal for Call Centre and robotic applications.

18.3.1.2 *Input / output for batch style application.*

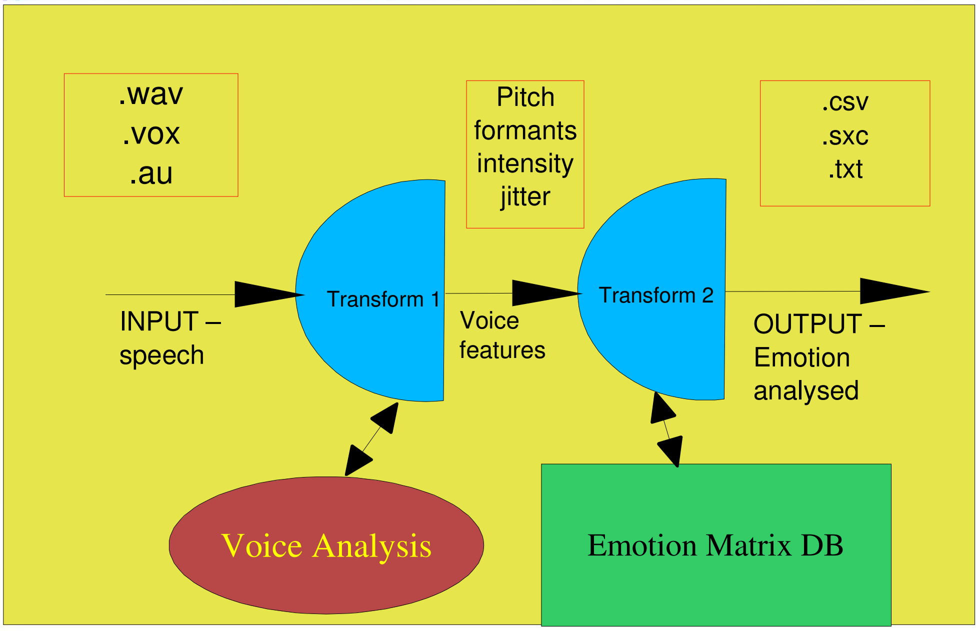
Ideally the batch processing system should be able to process industry standard .wav and .vox files and produce a generic output file (eg .csv) that can be transferred or uploaded into another application for further processing. The files may be loaded in a specified directory before the command is entered or the event “clicked “ on a GUI (Graphical User Interface).

18.3.1.3 *Input / output for real time application.*

For an instantaneous application of the emotion detector input to the system generally has to be directly from the point of source – namely the human month. The system therefore requires a microphone or telephone interface enabling it to “listen” directly to the spoken word. The output from the system has to instantly available and reliable for many real time applications. A communications protocol such as TCP/IP is recommended because it is fast, reliable and industry standard.

18.4 Functional Summery of EDE

18.4.1 Information Flow Diagrams



Drawing 10. Information Flow for Emotion Detection System

18.4.2 EDE Development Project Standalone

Using the PRAAT speech analysis software and simple Unix script is used to analyse a supplied .wav file and determine the emotion. The emotion determined is printed to a window on the standard Output device. This standalone application is written for both Unix and Microsoft environments and can be demonstrated with a portable device such as a laptop or a Pocket PC.

The application is designed to demonstrate the technology using the same emotion detection engine as the real time demonstration. The portable computer must be at least a Pentium processor with 16 Mb RAM running either Linux, Solaris for Intel, Win 9X, Win 2000 or XP. The application software and code occupy about 5 megabytes of disk space.

Improved performance can be achieved by upgrading processor speed. For example a Pentium 100 laptop with 16MB of MM takes approximately 4 seconds to return a result on the first generation analyser. Performing the same exercise on an Athelon 1.8 GHz with 512 MB of main memory takes less then one second. The analysis software appears to be very processor intensive.

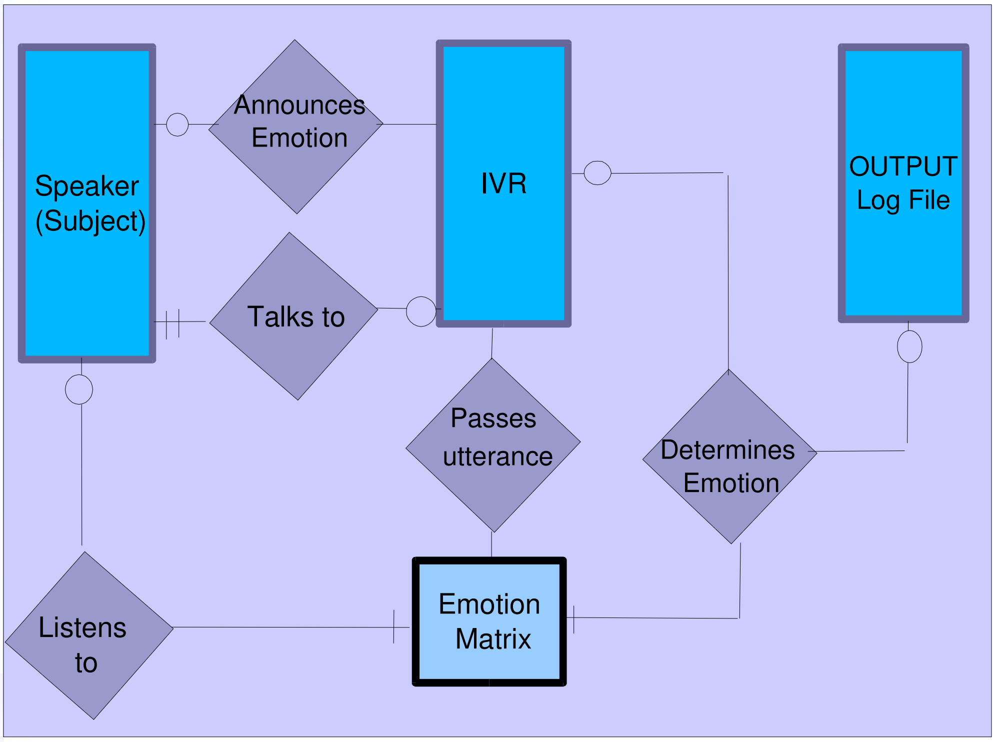
18.4.3 EDE coupled to IVR

18.4.3.1 Outline

The Emotion detection engine can be demonstrated with real time demonstration if connected to a system which permits the presentation and analysis of voice from a convenient and accessible location. The PSTN phone system provides the perfect access tool for this type of access so an Interactive Voice Response unit connected to the telephony PSTN would provide such convenient access. The Emotion Detection engine then can be demonstrated to a multitude of interested parties as well as being “field tested” to observe performance in one of many possible applications.

The E-R Diagram below depicts one possible configuration where the IVR interacts with the caller directly and uses the emotion Detection engine to determine the speaker's emotion. The system then announces the emotion back to the caller.

18.4.3.2 E-R Diagram – as per DEMO1



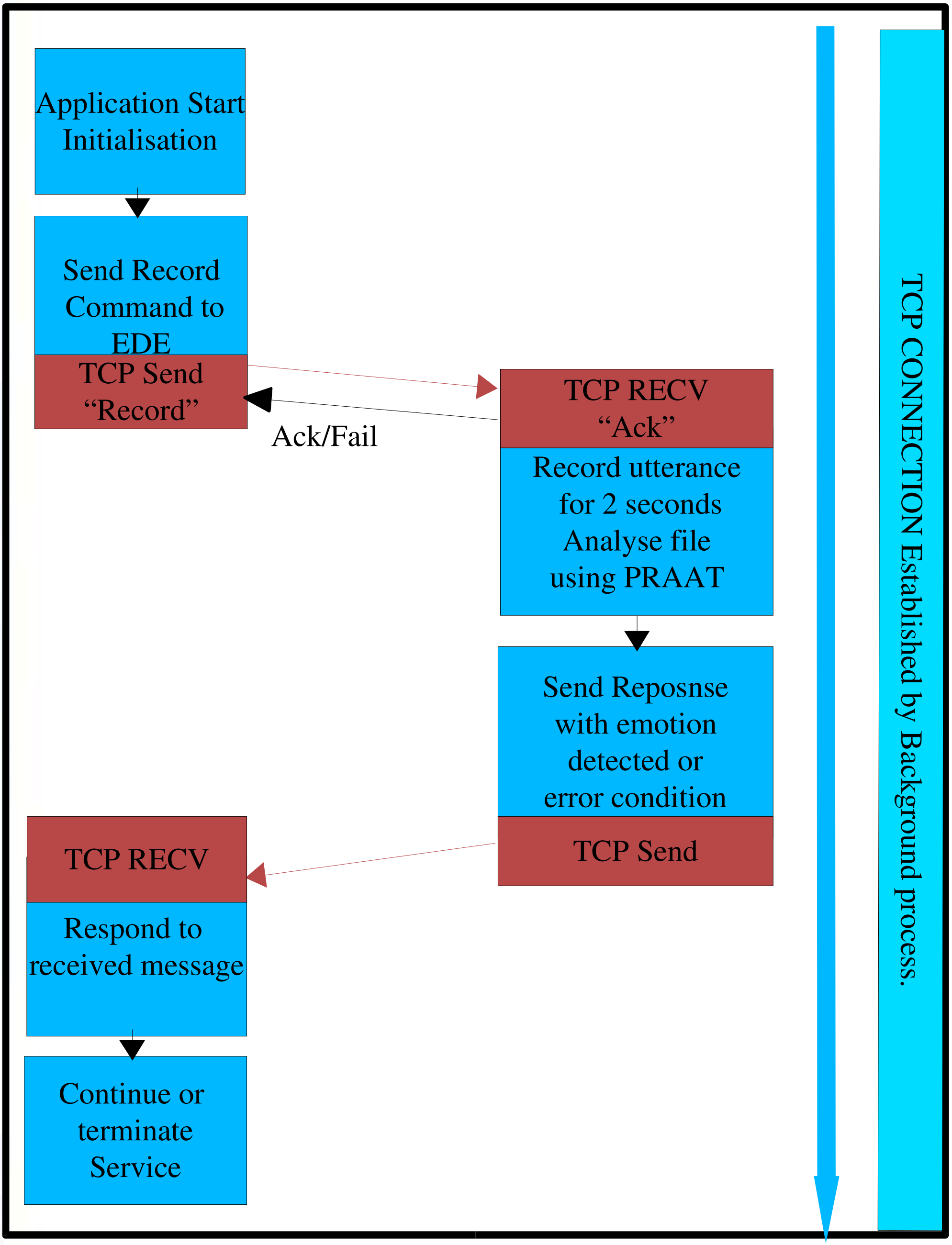
Drawing 11. ER diagram for Demo 1.

18.4.4 EDE – IVR Interface

18.4.4.1 *Messaging*

Timing Diagram

Interface between the IVR system and the EDE is via TCP/IP interface. The TCP interface ensures the sanctity of communications and provides a versatile open interface for a variety of adjunct devices and applications.


Application Start Initialisation
Send Record Command to EDE
TCP Send "Record"
TCP RECV "Ack"
Ack/Fail
Record utterance for 2 seconds Analyse file using PRAAT
Send Reposnse with emotion detected or error condition
TCP Send
TCP RECV
Respond to received message
Continue or terminate Service
TCP CONNECTION Established by Background process.

Drawing 12. Timing Diagram. IVR to Adjunct Processor IP Communications

18.4.5 Message Interface

Commands are IP based simple message strings. IP communication can be either via TCP or UDP. TCP is recommended for mission critical applications however.

Message	***String***	***Comments***
Initialise and record Command	<seq #><"GETE">	2 Byte command for Hi-Call IVR. Hicall vars @1 = 18245 (dec for 47(G),45(E)) @2 = 21573 (dec for 54(T),45(E))
Emotion analyse response	<seq #><Em Resp><Confidence>	Em Resp 01 = Happiness @X = 12337 02 = Sad @X = 12338 03 = Neutral @X = 12339 04 = Anger @X = 12340 05 = No Detection or no Activity @X = 12341 Confidence (00 – 99) @X = 12336 - 14649 ERROR 07 = internal error @X = 12343 08 = format error @X = 12344 09 = undefined error @X = 12345
Set Emotion Command	<seq #><"SETE"><bitmap>	@1 = 21317(dec for 53(S),45(E)) @2 = 21573 (dec for 54(T),45(E)) Bitmap bit 0 (1) : check for happiness bit 1 (2) : check for sad bit 2 (4) : check for neutral bit 3 (8) : check for anger
Set Emotion Acknowledge	<seq #><Resp>	Response: 00 = OK ERROR 07 = internal error @X = 12343 08 = format error @X = 12344 09 = undefined error @X = 12345

Table 16. UDP command packets and payload

18.5 Input Test Files

Data in the form of utterances is required from at least 20 individuals of both genders. An IVR based capture service has been designed to record utterances as per instructions prompted by the IVR. A number of people will be asked to follow the service instructions and record certain words with particular emotional accents. This data is necessary to analyse common elements that represent emotions in speech.

18.6 IVR Data Recorder for Emotion capture Experiments

18.6.1 Outline.

A simple IVR based service that greets the call, asks for the caller's and gender name and then, word by word and emotion by emotion, asks the caller to record. The recorded utterances are placed in individual directories for classification and later for analysis.

The service will be written in PDL and run on a Telsis Hi-Call IVR with is connected to the PSTN. Experiment 1 was conducted 20th September 2003 in the offices of Dialect Solutions Pty Ltd.

Experiment 2 will be of a similar service but this time asked callers to speak a single phrase in for different emotions.: Happiness, Sadness, Neutrality and Anger.

18.6.2 Script – Experiment 1

File Offset	*Prompt*
1	Welcome to Denis Ryan's Data acquisition experiment. Let me start off by thanking you for participating in this very important experiment. The exercise should take less than five minutes and will help Denis :- that's me!, do well in his university thesis. Beers are on the table boys and girls!.
2	OK let's get started. First I have to find out whether you are a man or a women. I think most of us can answer that. If you regard yourself as the masculine gender dial 1 or if la femme dial 2.

File Offset	***Prompt***
3	What your neither?? are you some kind of a crazy person? Come on try again surly you know what sex you are!!
4	Well it seems you have problems or maybe its just your phone. Well try from another phone and if you get this message again its time to see a shrink. See ya.
5	Righto , thanks for that. Now would you be so kind as to leave your first name just in case I have to contact you regarding some more voicing. Ready – record your name after the 2 Khz tone.
6	Brilliant, now for the fun stuff. I am going to ask you to repeat some words into the system. I will ask you to speak the same word four times with four different emotional stresses. My prompts will help you along, hopefully providing enough mental imagery so your emotions are real!! The first emotion will be neutral – that is the normal emotion you possess on about 95% of your phone conversations. So just speak normally when a neutral emotion is required. Next emotion is happiness, listen to the prompts which will supply you with a happy scenario to help bolster your happiness emotion. The next two emotions are frustration and anger. You probably haven't thought about this but you speak differently when your frustrated as compared to when you are angry – Hey take it from an expert, its true. When your frustrated you may grit your teeth and you generally will speak s l o w l y, **loudly** pronouncing EV VE RY SY LA BLE!!? When your angry you'll probably speak loud and fast, probably with a splattering of obscenities to boot. However you all have your own ways of conveying emotions , so please, do what's natural. Ready?
7	The 1st word I am going to get you to record is NO. Yep, plain old no. Nothing hard about that eh? I want you to say no in a neutral tone. Just a polite NO will suffice here. OK after the tone please record your neutral NO.
8	Perfect. Now for a happy no. This may be a bit hard but just imagine your best friend has just rang you and told you he or she has just got engaged, it and unexpected but pleasant surprise – many of us would respond with “no, really”. Of course I don't want you to say the “really”, just the no – Ready – after the tone say a happy no.

File Offset	*Prompt*
9	Yippee, your doing great. Now for a frustrated no. You've been of hold for half an hour waiting to speak to the manager regarding a complaint. The operator advises you after all that waiting only the 2nd in charge is available as the manager has just gone out to lunch. The operator asks do you want to be patched through to the 2IC. Your p'eed off but not enraged – your response will probably start off with a loud and long **no.** Think you can do it?? OK, give us a frustrated no after the tone.
10	Now time for some fun, The angry no. I guess we all have been angry a some time of our life. Anger is an easy emotion as we easily remember the emotional stress that created our anger. Just think of a time you where thoroughly angered by something and replicate into the system. Girls I'm sure an odd angry NO has been occasioned to persistent slime bags at night clubs. So ready say an angry NO after the tone.
11	Well, that's the no taken care of- now for the yes. Same pack drill. First up a neutral Yes. Once again a plain old simple polite yes will suffice. Ready – after the tone..
12	Excellent now for a happy yes. Just imagine , or remember, when the man or women of your dreams asked you to marry them. Its a question you've being dying to hear for a long time and your answer is a resounding YES. Don't mask the excitement – it helps characterise the emotion. OK a nice big happy yes after the tone.
13	OK let's try a frustrated yes. You know when you asked the same thing over and over again, your response being yes each time. Now your sick of it and the eighth time being asked ,your humor has gone, and your retort is more like as YES. Right give us a frustrated yes after the tone.
14	Well done! Now for the angry yes. Short, sharp and loud is the key. Remember those bad experiences. Give it all you've got ..an big, fat , angry YES after the tone.
15	Yeeeh, didn't that feel good? Bet you wish you could dial this service every day. Well the truth be known - you can! Now we'll experiment with some multi-syllable words. Yes I know you're all Australians but I'm sure you can do it. There even doing it on 'Neighbours' these days. The word I want you to record, first in a neutral or polite tone, is “manager”. That's right the poor old manager – the one who wears the brunt of most complaints and seldom receives complements. So lets go... a polite “manger” to start, after the tone.

16	Now what about a happy manager. Hmm a happy manager – that's a turn for the books. The scenario may read like this. Your lovely spouse has just received a promotion at work and your on the phone telling your best friend. “My Charlie has just been promoted to manager” or something of that nature. Think you can do it. Go on I know you can. A big, happy “manager” after the tone.
17	Next we have a frustrated manager – hypothetically speaking of course. We can go back to the on hold scenario where you have been on hold for an hour and told you can only speak to the assistant. You tell the operator , in no uncertain terms, No, I want to speak to the ma-na-ger. Record your frustrated version of manager after the tone.
18	Crickey where almost through! What a shame. Well our last utterance is an angry manager. Yep out Call Centre staff have probably had experience with that “I want to speak to the manager”.. You all know the score. Give it the best shot for your final recording. After the tone say “manager in an angry tone.
19	If your happy with how you recorded that dial 1, to listen to your recording dial 2 or to attempt the recording again dial 3.
20	Well that's all folks. I must thank you again for participating and sacrificing your time. Bye for now and see you all in the pub..
21	Sorry but that's invalid you can only dial 1,2 or 3

Table 17. Script for Experiment 1.

8.6.3 Script – Experiment 2

File Offset	***Prompt***
1	Welcome to Denis' second emotion acquisition experiment. Once again I thank you for participating in this very important experiment. This is more brief than the last and will take less than half the time. Let's Get cracking!
2	Again I have to find out whether you are a man or a women. I think most of us can answer that. If you are a Bruce dial 1 or if you're a Sheila dial 2.
3	What your neither?? are you some kind of a crazy person? Come on try again surly you know what sex you are!!
4	Well it seems you have problems or maybe its just your phone. Well try from another phone and if you get this message again its time to see a shrink. See ya.
5	Righto , thanks for that. Now would you be so kind as to leave your first name just in case I have to contact you regarding some more voicing. Ready – record your name after the tone.
6	Ok let's move forward. Most of you who have used this service before can dial # to skip the boring instructions. For those of you who are new to this gig – listen on.<pause>. I am going to ask you to repeat a phrase into the system. I will ask you to speak this same sentence four times with four different emotional stresses. My prompts will help you along, hopefully providing enough mental imagery so your emotions are real!! The first emotion will be neutral – that is the normal emotion you possess on about 95% of your phone conversations. So just speak normally when a neutral emotion is required. Next emotion is happiness, listen to the prompts which will supply you with a happy scenario to help bolster your happiness emotion. The next emotion is sadness and the last is anger. Sadness is generally a hard thing to do so you need to get in the right frame of mind. When your angry you'll probably speak loud and fast, probably with a splattering of obscenities to boot. However you all have your own ways of conveying emotions , so please, do what's natural. Ready?

File Offset	***Prompt***
7	Alright, like I said there is only one sentence that you have to repeat for 4 different emotions. Neutrality, Happiness, Sadness and anger. This time try to let it all out – if you think you will be distracted at the office , then try it at home. The sentence I'm Going to get you to say is " Not now I'm in the garden". Sounds strange but contextually it fits all four emotions . Don't worry If you fluff it by not saying the exact words – its the emotions that are important. . OK after the tone record the phrase "Not now I'm in the garden" in a relaxed neutral voice.
8	Perfect. To say the same thing but with a happy tone. Just think how you would say that sentence if you just found you have won the lottery – No not the piddly little $9,000,000 Aussie lotto– let's say 50,000,000 pounds on one of those Euro Jackpot thingy-a-bobs. Happy - by jingo by Crickey. OK you happy thing you , say a really happy "Not now I'm in the garden" after the tone.
9	Great Work Now for a sad sentence. I don't want to bring you down too much , so I won't give you any examples here, just to say that most of us know what it is like to be sad. Not the best of feelings I must admit. Now think of how you normally talk when you speak – generally quite slowly, poorly articulated and generally at a lower level. OK think you can pull it off – say a sad ""Not now I'm in the garden" after the sad ol' tone.
10	OK people, once again its showtime, better warn people near you in the office, else they'll think Denis is back. Anger is an easy emotion as we easily remember the emotional stress that created our anger. You may have had a falling out with someone or something very wrong has happened at work. You're in the garden for some R&R trying to recuperate in your backyard Edan, when someone calls out to you disturbing your much deserved time alone. Of course this is the last straw – you snap – yep ,go off like a gun – you scream ""Not now I'm in the garden" in amongst the red mist. OK take a deep breath say a super angry "Not now I'm in the garden" after the tone.
19	If your happy with how you recorded that dial 1, to listen to your recording dial 2 or to attempt the recording again dial 3.
20	That's all – surprised? I told you it was much shorter than before. Once again I must thank you again for participating and sacrificing your time. Bye for now and see you all in the pub..
21	Sorry but that's invalid you can only dial 1,2 or 3

18.7 Description of Outputs

The EDE must permit its analysis to be made available to adjunct equipment or processes that can perform action based on data received. The EDE will support output to an appending log file as well as a propriety messaging system. Below is a table defining the message format for the EDE Output.

18.8 Messaging Format

18.8.1 Sample Output Log File

STX,20030817,22:14:56,01,1,66,234,EOT

18.8.2 Log File Description

Element Number	***Description***	***Comment***
1	Start of Text STX	ASCII 02
2	Unix date stamp	Year Month Day
3	Time Stamp	Hour: minute: second
4	Message type	0 = Initialise 1 = Result 2 = Error
5	Result Speech Rec OR EME or ERROR Code if Message type = 2	RESULT 0 = No Detection 1 = Happy emotion detected 2 = Sad emotion detected 3 = Neutral emotion detected 4 = Angry emotion detected ERROR 997 = internal error 998 = format error 999 = undefined error
6	Confidence level	Certainty of result in percent
7	Sequence Number	Unique Sequence number to 0 - 9999 ring buffer
8	End Of Text EOF	ASCII 03

Table 18. Output Log File

18.9 Expected Outputs

The expected output is a binary condition coupled with a variable certainty factor. Output will be in the form of messaged based interface and an eternally appending log file. Constituents to each output iteration (based on 5 to 10 second segments of speech input):

Result: 0 = No Detection

1 = Happy emotion detected

2 = Sad emotion detected

3 = Neutral emotion detected

4 = Angry emotion detected

Certainty Level: A confidence rating in percent of the derived result.

19.1 Expected Behavior.

Below is a table of some possible scenarios and their expected results

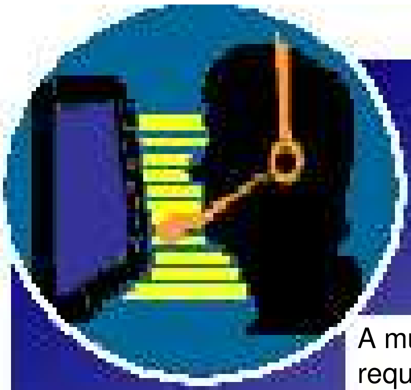
Scenarios	***Expected outcome***
Happy person speaking at a low volume	Happiness detected. Too low a volume may cause non register
Happy Person speaking at high volume	AGC on – Happiness detected AGC off – possibly anger detected
Neutral Person speaking at low volume	Neutral or non detection
Angry Person at middle volume	Anger detected
Angry person at high volume	Anger detected
Happy Person swearing at low volume (obscenities filter on)	SIR option only. Should intercept and flag anger.
Angry Person Swearing	As above SIR is in logical OR with EDE
Person banging phone on the table	SIR to flag as anger. SIR requires training on sound
Sad person talking at normal volume	Sadness detected
Neutral person at high volume	Neutrality detected
Sad person at high volume	AGC off - Possibly anger detected. AGC on sadness detected

Table 19. Expected behaviour

19.2 Scenarios Used to Verify Operation and Validate Requirement Specifications.

Scenarios	Result	Requirement ID
.wav 8KB/s @ 16bits recording supplied to simulation	Correct Detection of emotional alignment in return	A1,A2
Simulation run on supplied file	Return from simulation in console widow displayed as : emotion type 01 - 04	A3
Optional Test scenarios run on EME implemented on TI 6000 development platform	Achieves functionality and reliability of simulated system	B7
Happy person speaking at a low volume	Happy @75.00%	A4
Happy Person speaking at high volume	Happy 60.00%	A4
Neutral Person speaking at low volume	Happy 85.00%	A4
Angry Person at middle volume	Anger 75.00%	A4
Angry person at high volume	Anger 85.00%	A4
Happy Person swearing at low volume (obscenities filter on) SIR Only	anger 90.00%	A4,A6
Angry Person Swearing	anger 90.00%	A4,A6
Person banging phone on the table	Anger 90.00%	A4,A6
Above 7 scenarios tested through an IVR interfaced to the ED system	As above – however AGR may reduce certainties	C8

Table 20. Verification and Validation of Requirements



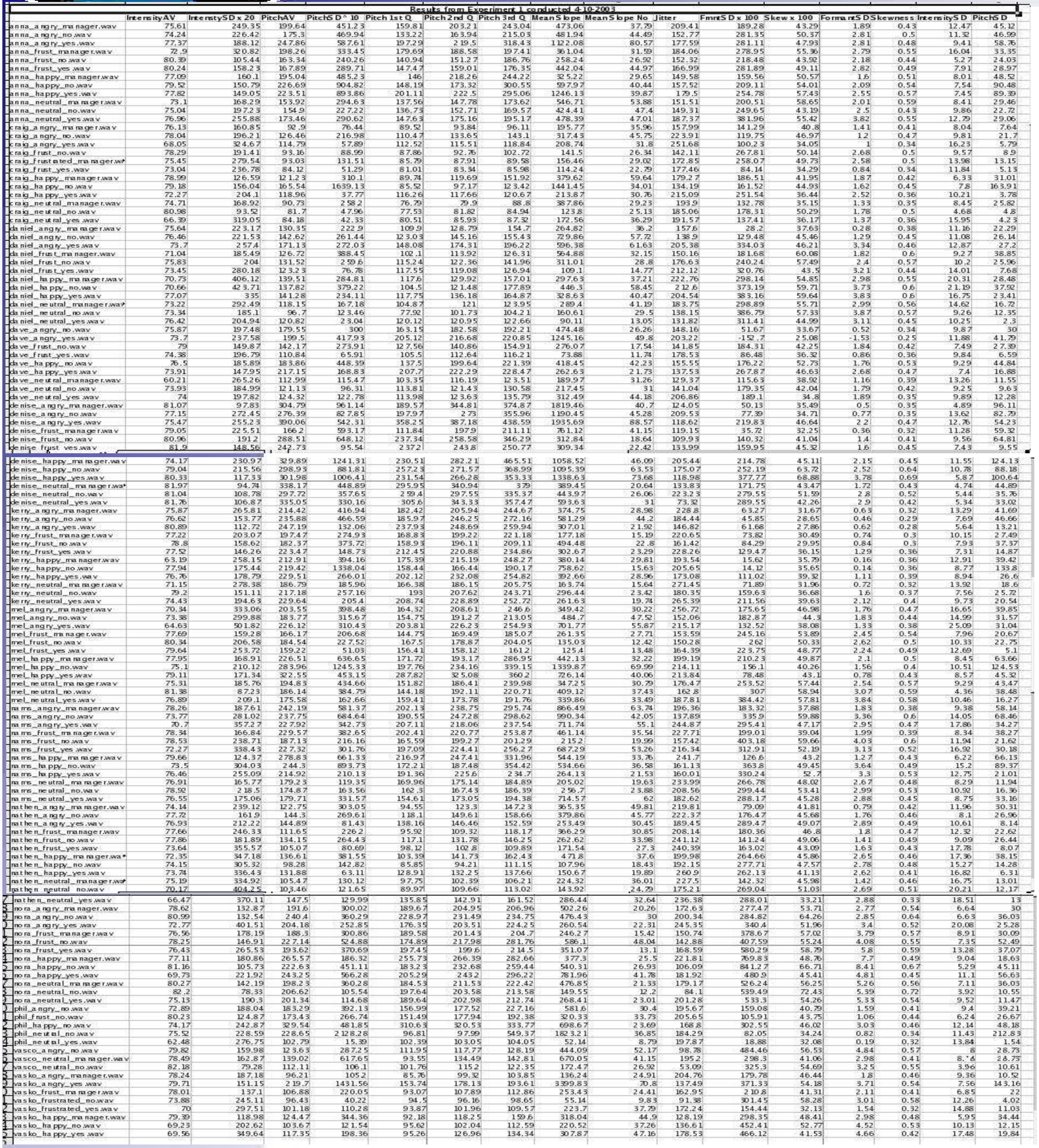
19.3 Test Plan

A multiple point regression test is used both to validate and verify design requirements of the Emotion Detection system. It also serves as a test benchmark which is performed each time an additional element or change is added or made to the system.

Test Number	Test Scenarios	Expected Result	Pass /Fail	Requirement ID
1	.wav 8KB/s @ 16bits recording supplied to simulation	Correct Detection of emotional alignment in return		A1,A2
2	Simulation run on supplied file	Return from simulation in console widow displayed as : emotion type 01 - 04		A3
3	Optional Test scenarios run on EME implemented on TI 6000 development platform	Achieves functionality and reliability of simulated system		B7
4				
5	Happy person speaking at a low volume	Happy @75.00%		A4
6	Happy Person speaking at high volume	Happy 60.00%		A4
7	Neutral Person speaking at low volume	Happy 85.00%		A4
8	Angry Person at middle volume	Anger 75.00%		A4
9	Angry person at high volume	Anger 85.00%		A4
10	Happy Person swearing at low volume (obscenities filter on) SIR Only	anger 90.00%		A4,A6
11	Angry Person Swearing	anger 90.00%		A4,A6
12	Person banging phone on the table	Anger 90.00%		A4,A6
13	Above 7 scenarios tested through an IVR interfaced to the ED system	As above – however AGR may reduce certainties		C8

Table 21. Multi point Test Plan.

APPENDIX C – Results from Experiment 1

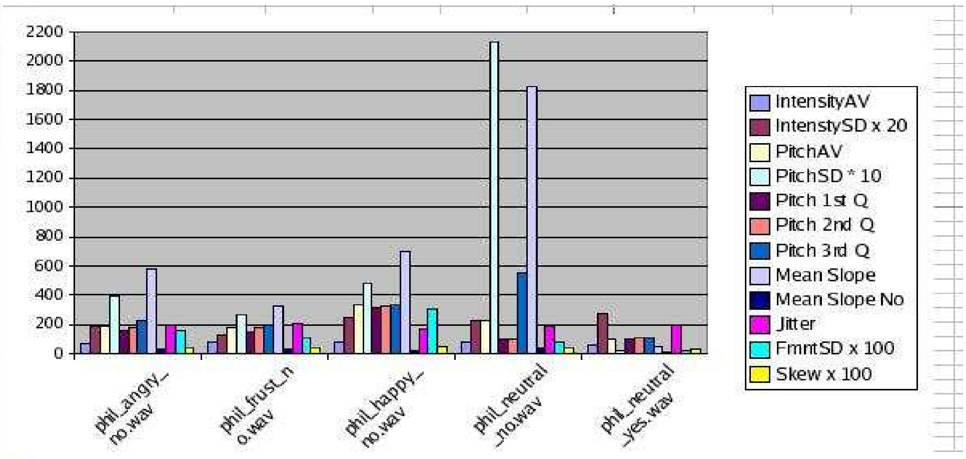


Results from Experiment 1 conducted 4-10-2003

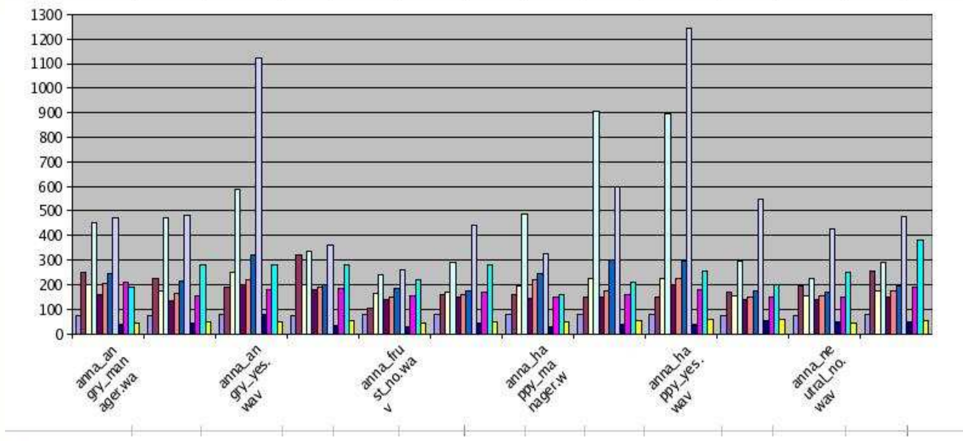
	IntensityAV	IntenstySD x 20	PitchAV	PitchSD^10	Pitch 1st Q	Pitch 2nd Q	Pitch 3rd Q	MeanSlope	MeanSlope No	Jitter	FmntSD x 100	Skew x 100	FormantSD	Skewness	IntensitySD	PitchSD
anna_angry_manager.wav	75.61	249.35	199.64	451.23	159.81	203.21	243.04	473.06	37.79	209.41	189.28	43.29	1.89	0.43	12.47	45.12
anna_angry_no.wav	74.24	226.42	175.3	469.94	133.22	163.94	215.03	481.94	44.49	152.77	281.35	50.37	2.81	0.5	11.32	46.99
anna_angry_yes.wav	77.37	188.12	247.86	587.61	197.29	219.5	318.43	1122.08	80.57	177.59	281.11	47.93	2.81	0.48	9.41	58.76
anna_frust_manager.wav	72.9	320.82	198.26	333.45	179.69	188.58	197.41	361.04	31.59	184.06	278.95	55.36	2.79	0.55	16.04	33.35
anna_frust_no.wav	80.39	105.44	163.34	240.26	140.94	151.27	186.76	258.24	26.92	152.32	218.48	43.92	2.18	0.44	5.27	24.03
anna_frust_yes.wav	80.24	158.23	167.89	289.71	147.47	159.01	176.35	442.04	44.97	166.99	281.89	49.11	2.82	0.49	7.91	28.97
anna_happy_manager.wav	77.09	160.1	195.04	485.23	146	218.26	244.22	325.22	29.65	149.58	159.56	50.57	1.6	0.51	8.01	48.52
anna_happy_no.wav	79.52	150.79	226.69	904.82	148.19	173.32	300.55	597.97	40.44	157.52	209.11	54.01	2.09	0.54	7.54	90.48
anna_happy_yes.wav	77.82	149.05	223.51	893.86	201.11	222.5	295.06	1246.13	39.87	179.5	254.78	57.43	2.55	0.57	7.45	89.39
anna_neutral_manager.wav	73.1	168.29	153.92	294.63	137.56	147.78	173.62	546.71	53.88	151.51	200.51	58.65	2.01	0.59	8.41	29.46
anna_neutral_no.wav	75.04	197.23	154.9	227.22	136.73	152.71	169.57	424.41	47.4	149.31	249.65	43.19	2.5	0.43	9.86	22.72
anna_neutral_yes.wav	76.96	255.88	173.46	290.62	147.63	175.16	195.17	478.39	47.01	187.37	381.96	55.42	3.82	0.55	12.79	29.06
craig_angry_manager.wav	76.13	160.85	92.9	76.44	89.52	93.84	96.11	195.77	35.96	157.99	141.29	40.8	1.41	0.41	8.04	7.64
craig_angry_no.wav	78.04	196.21	126.46	216.98	110.47	133.65	143.1	317.43	45.75	223.91	119.75	46.97	1.2	0.47	9.81	21.7
craig_angry_yes.wav	68.05	324.67	114.79	57.89	112.52	115.51	118.84	208.74	31.8	251.68	100.23	34.05	1	0.34	16.23	5.79
craig_frust_no.wav	78.29	191.41	93.16	88.99	87.86	92.76	102.72	141.5	26.34	142.11	267.81	50.14	2.68	0.5	9.57	8.9
craig_frustated_manager.wa	75.45	279.54	93.03	131.51	85.79	87.91	89.58	156.46	29.02	172.85	258.07	49.73	2.58	0.5	13.98	13.15
craig_frust_yes.wav	73.04	236.78	84.12	51.29	81.01	83.34	85.98	114.24	22.79	177.46	84.14	34.29	0.84	0.34	11.84	5.13
craig_happy_manager.wav	78.99	126.59	121.23	310.1	89.74	119.69	151.92	379.62	59.64	179.27	186.51	41.95	1.87	0.42	6.33	31.01
craig_happy_no.wav	79.18	156.04	165.54	1639.13	85.52	97.17	123.42	1441.45	34.01	134.19	161.52	44.93	1.62	0.45	7.8	163.91
craig_happy_yes.wav	72.27	204.1	118.96	37.77	116.26	117.66	120.67	213.87	30.76	215.09	251.54	36.44	2.52	0.36	10.21	3.78
craig_neutral_manager.wav	74.71	168.92	90.73	258.2	76.79	79.9	88.8	387.86	29.23	193.9	132.78	35.15	1.33	0.35	8.45	25.82
craig_neutral_no.wav	80.98	93.52	81.7	47.96	77.53	81.82	84.94	123.8	25.13	185.06	178.31	50.29	1.78	0.5	4.68	4.8
craig_neutral_yes.wav	66.39	319.05	84.18	42.33	80.51	85.93	87.32	172.56	36.29	191.57	137.41	36.17	1.37	0.36	15.95	4.23
daniel_angry_manager.wav	75.64	223.17	130.35	222.9	109.9	128.79	154.7	264.82	36.2	157.6	28.2	37.63	0.28	0.38	11.16	22.29
daniel_angry_no.wav	76.46	221.53	142.62	261.44	123.03	145.16	155.43	729.86	57.72	138.9	129.48	45.46	1.29	0.45	11.08	26.14
daniel_angry_yes.wav	73.7	257.4	171.13	272.03	148.08	174.31	196.22	596.38	61.63	205.38	334.03	46.21	3.34	0.46	12.87	27.2
daniel_frust_manager.wav	71.04	185.49	126.72	388.45	102.1	113.92	126.31	564.88	32.15	150.16	181.68	60.08	1.82	0.6	9.27	38.85
daniel_frust_no.wav	75.83	204	131.52	259.6	115.24	122.36	141.96	311.01	28.8	176.63	240.24	57.49	2.4	0.57	10.2	25.96
daniel_frust_yes.wav	73.45	280.18	123.23	76.78	117.55	119.08	126.94	109.1	14.77	212.12	320.76	43.5	3.21	0.44	14.01	7.68
daniel_happy_manager.wav	70.73	406.12	139.51	284.81	117.6	129.92	157.01	297.63	37.21	222.76	298.14	54.85	2.98	0.55	20.31	28.48
daniel_happy_no.wav	70.66	423.71	137.82	379.22	104.5	121.48	177.89	446.3	58.45	212.6	373.19	59.71	3.73	0.6	21.19	37.92
daniel_happy_yes.wav	77.07	335	141.28	234.11	117.75	136.18	164.87	328.63	40.47	204.54	383.16	59.64	3.83	0.6	16.75	23.41
daniel_neutral_manager.wa	73.22	292.49	118.15	167.18	104.87	121	123.95	289.4	41.19	183.75	298.89	55.71	2.99	0.56	14.62	16.72
daniel_neutral_no.wav	73.34	185.1	96.7	123.46	77.92	101.73	104.21	160.61	29.5	138.15	386.79	57.33	3.87	0.57	9.26	12.35
daniel_neutral_yes.wav	76.42	204.94	120.82	23.04	120.12	120.95	122.66	90.11	13.05	131.82	311.41	44.99	3.11	0.45	10.25	2.3
dave_angry_no.wav	75.87	197.48	179.55	300	163.15	182.58	192.21	474.48	26.26	148.16	51.67	33.67	0.52	0.34	9.87	30
dave_angry_yes.wav	73.7	237.58	199.5	417.93	205.12	216.68	220.85	1245.16	49.8	203.22	-152.7	25.08	-1.53	0.25	11.88	41.79
dave_frust_no.wav	79	149.87	142.17	273.91	127.56	140.86	154.91	276.07	17.54	141.85	184.31	42.25	1.84	0.42	7.49	27.39
dave_frust_yes.wav	74.38	196.79	110.84	65.91	105.5	112.64	116.21	73.88	11.74	178.53	86.48	36.32	0.86	0.36	9.84	6.59
dave_happy_no.wav	76.5	185.89	183.86	448.39	137.5	199.64	221.39	418.45	42.23	155.55	176.22	52.73	1.76	0.53	9.29	44.84
dave_happy_yes.wav	73.91	147.95	217.15	168.83	207.7	222.29	228.47	262.63	21.73	137.53	267.87	46.63	2.68	0.47	7.4	16.88
dave_neutral_manager.wav	60.21	265.26	112.99	115.47	103.35	116.19	123.51	189.97	31.26	129.37	115.63	38.92	1.16	0.39	13.26	11.55
dave_neutral_no.wav	73.93	184.99	121.13	96.31	113.81	121.43	130.58	217.45	31	141.04	179.35	42.04	1.79	0.42	9.25	9.63
dave_neutral_yes.wav	74	197.82	124.32	122.78	113.98	123.63	135.79	312.49	44.18	206.86	189.1	34.8	1.89	0.35	9.89	12.28
denise_angry_manager.wav	81.07	97.83	304.79	961.14	189.57	344.81	374.87	1819.46	40.7	124.05	50.13	35.49	0.5	0.35	4.89	96.11
denise_angry_no.wav	77.15	272.45	276.39	827.85	197.97	273	355.96	1190.45	45.28	209.53	77.39	34.71	0.77	0.35	13.62	82.79
denise_angry_yes.wav	75.47	255.23	390.06	542.31	358.25	387.18	438.59	1935.69	88.57	118.62	219.83	46.64	2.2	0.47	12.76	54.23
denise_frust_manager.wav	79.05	225.51	166.2	593.17	111.84	197.9	211.11	761.12	41.15	119.15	35.72	32.25	0.36	0.32	11.28	59.32
denise_frust_no.wav	80.96	191.2	288.51	648.12	237.34	258.58	346.29	312.84	18.64	109.93	140.32	41.04	1.4	0.41	9.56	64.81
denise_frust_yes.wav	81.9	148.56	242.73	95.54	237.2	243.8	250.77	309.34	22.42	133.99	159.95	45.32	1.6	0.45	7.43	9.55
denise_happy_manager.wav	74.17	230.97	329.89	1241.31	230.51	282.21	465.51	1058.52	46.09	205.44	214.78	45.11	2.15	0.45	11.55	124.13
denise_happy_no.wav	79.04	215.56	298.93	881.81	257.23	271.57	368.99	1095.39	63.53	175.07	252.19	63.72	2.52	0.64	10.78	88.18
denise_happy_yes.wav	80.33	117.33	301.98	1006.41	231.54	266.28	353.33	1338.63	73.68	118.98	377.77	68.88	3.78	0.69	5.87	100.64
denise_neutral_manager.wa	81.97	94.74	338.17	448.89	295.95	340.94	379	389.45	20.64	133.83	171.75	43.47	1.72	0.43	4.74	44.89
denise_neutral_no.wav	81.04	108.78	297.72	357.65	259.4	297.55	335.37	443.97	26.06	223.23	279.55	51.59	2.8	0.52	5.44	35.76
denise_neutral_yes.wav	81.76	106.87	335.05	330.16	305.6	343.33	357.47	593.63	31	73.32	289.55	42.26	2.9	0.42	5.34	33.02
kerry_angry_manager.wav	75.87	265.81	214.42	416.94	182.42	205.94	244.67	374.75	28.98	228.8	63.27	31.67	0.63	0.32	13.29	41.69
kerry_angry_no.wav	76.62	153.77	235.88	466.59	185.97	246.25	272.16	581.29	44.2	184.44	45.85	28.65	0.46	0.29	7.69	46.66
kerry_angry_yes.wav	80.89	112.72	247.19	132.06	237.93	248.69	259.94	307.01	21.92	146.82	61.68	27.86	0.62	0.28	5.64	13.21
kerry_frust_manager.wav	77.22	203.07	197.47	274.93	168.83	199.22	221.18	177.18	15.19	220.65	73.82	30.49	0.74	0.3	10.15	27.49
kerry_frust_no.wav	78.8	158.62	182.37	373.72	158.93	196.11	209.11	494.48	22.8	161.42	84.29	29.95	0.84	0.3	7.93	37.37
kerry_frust_yes.wav	77.52	146.26	223.47	148.73	212.45	220.88	234.86	302.67	23.29	228.26	129.47	36.15	1.29	0.36	7.31	14.87
kerry_happy_manager.wav	63.19	258.15	212.91	394.16	175.39	215.19	248.27	380.14	29.81	193.54	15.62	35.79	0.16	0.36	12.91	39.42
kerry_happy_no.wav	77.94	175.44	219.42	1338.04	158.44	166.44	190.17	758.62	15.63	205.65	14.12	35.65	0.14	0.36	8.77	133.8
kerry_happy_yes.wav	76.76	178.79	229.51	266.01	202.12	232.08	254.82	392.66	28.96	173.08	111.02	39.32	1.11	0.39	8.94	26.6
kerry_neutral_manager.wav	71.15	278.38	186.79	185.96	166.38	186.15	205.75	163.74	15.64	271.45	71.89	31.96	0.72	0.32	13.92	18.6
kerry_neutral_no.wav	79.2	151.11	217.18	257.16	193	207.62	243.71	296.44	23.42	180.35	159.63	36.68	1.6	0.37	7.56	25.72
kerry_neutral_yes.wav	74.43	194.63	229.64	205.4	208.74	228.89	252.72	261.63	19.74	265.39	211.56	39.63	2.12	0.4	9.73	20.54
mel_angry_manager.wav	70.34	333.06	203.55	398.48	164.32	208.61	246.6	349.42	30.22	256.72	175.65	46.98	1.76	0.47	16.65	39.85
mel_angry_no.wav	73.38	299.88	183.77	315.67	154.75	191.27	213.05	484.7	47.52	152.06	182.87	44.3	1.83	0.44	14.99	31.57
mel_angry_yes.wav	64.63	501.82	226.12	310.43	203.81	226.23	254.93	701.77	55.87	215.17	132.52	38.08	1.33	0.38	25.09	31.04
mel_frust_manager.wav	77.69	159.28	166.17	206.68	144.75	169.49	185.07	261.35	27.71	153.59	245.16	53.89	2.45	0.54	7.96	20.67
mel_frust_no.wav	80.34	206.58	184.54	227.52	167.5	178.87	204.05	135.03	12.42	150.28	262	50.33	2.62	0.5	10.33	22.75
mel_frust_yes.wav	79.64	253.72	159.22	51.03	156.41	158.12	161.2	125.4	13.48	164.39	223.75	48.77	2.24	0.49	12.69	5.1
mel_happy_manager.wav	77.95	168.91	226.51	636.65	171.72	193.17	286.95	442.13	32.22	199.19	210.23	49.87	2.1	0.5	8.45	63.66
mel_happy_no.wav	75.1	210.12	283.96	1245.33	197.76	234.16	339.15	1339.87	69.99	214.11	156.1	40.26	1.56	0.4	10.51	124.53
mel_happy_yes.wav	79.11	171.34	322.55	453.15	287.82	325.08	360.2	726.14	40.06	213.84	78.48	43.1	0.78	0.43	8.57	45.32
mel_neutral_manager.wav	75.31	185.76	194.83	434.66	151.82	186.41	239.98	347.25	30.79	176.47	253.52	57.44	2.54	0.57	9.29	43.47
mel_neutral_no.wav	81.38	87.23	186.14	384.79	144.18	192.11	220.71	409.12	37.43	162.8	307	58.94	3.07	0.59	4.36	38.48
mel_neutral_yes.wav	76.89	209.1	175.58	162.66	159.41	173.78	191.76	339.86	33.49	187.81	384.42	57.81	3.84	0.58	10.46	16.27
marns_angry_manager.wav	78.26	187.61	242.19	581.37	202.13	238.75	295.74	866.49	63.74	196.36	183.32	37.88	1.83	0.38	9.38	58.14
marns_angry_no.wav	73.77	281.02	237.75	684.64	190.55	247.28	298.62	990.34	42.05	137.89	335.9	59.88	3.36	0.6	14.05	68.46
marns_angry_yes.wav	70.7	357.27	227.92	342.73	207.11	218.06	237.54	711.74	55.1	244.87	295.41	47.17	2.95	0.47	17.86	34.27
marns_frust_manager.wav	78.34	166.84	229.57	382.65	202.41	220.77	253.87	461.14	35.54	227.71	199.01	39.04	1.99	0.39	8.34	38.27
marns_frust_no.wav	78.53	238.71	187.13	216.16	165.59	199.27	201.29	215.2	19.99	157.42	403.18	59.66	4.03	0.6	11.94	21.62
marns_frust_yes.wav	72.27	338.43	227.32	301.76	197.09	224.41	256.27	687.29	53.26	216.34	312.91	52.19	3.13	0.52	16.92	30.18
marns_happy_manager.wav	79.66	124.37	278.83	661.33	216.97	247.41	331.96	544.19	33.76	241.7	126.6	43.2	1.27	0.43	6.22	66.13
marns_happy_no.wav	73.5	304.03	244.3	893.73	172.21	187.48	354.42	534.66	36.58	161.13	363.8	49.45	3.64	0.49	15.2	89.37
marns_happy_yes.wav	76.46	255.09	214.92	210.13	191.36	225.6	234.7	264.13	21.53	160.01	330.24	52.7	3.3	0.53	12.75	21.01
marns_neutral_manager.wav	76.91	165.77	179.23	119.35	169.96	175.14	184.89	205.02	19.63	233.99	266.78	48.02	2.67	0.48	8.29	11.94
marns_neutral_no.wav	78.92	218.5	174.87	163.56	162.3	167.43	186.39	256.7	23.88	208.56	299.44	53.41	2.99	0.53	10.92	16.36
marns_neutral_yes.wav	76.55	175.06	179.71	331.57	154.61	173.05	194.38	714.57	62	182.62	288.17	45.28	2.88	0.45	8.75	33.16
nathen_angry_manager.wav	74.14	239.12	122.75	303.05	94.55	123.3	147.23	365.35	49.81	219.81	79.09	41.81	0.79	0.42	11.96	30.31
nathen_angry_no.wav	77.72	161.9	144.3	269.61	118.1	149.61	158.66	379.86	45.77	222.37	176.47	45.68	1.76	0.46	8.1	26.96
nathen_angry_yes.wav	76.93	212.22	144.89	81.43	138.16	146.46	152.59	253.49	30.45	189.45	289.47	49.07	2.89	0.49	10.61	8.14
nathen_frust_manager.wav	77.66	246.33	111.65	226.2	95.92	109.32	118.17	366.29	30.85	208.14	180.36	46.8	1.8	0.47	12.32	22.62
nathen_frust_no.wav	77.86	181.89	134.15	264.43	117.1	131.78	146.25	262.62	33.98	241.12	141.24	49.06	1.41	0.49	9.09	26.44
nathen_frust_yes.wav	73.64	355.57	105.07	80.69	98.12	102.8	109.89	171.54	27.3	240.39	163.02	43.09	1.63	0.43	17.78	8.07
nathen_happy_manager.wa	72.35	347.18	136.61	381.55	103.39	141.73	162.43	471.8	37.6	199.98	264.66	45.86	2.65	0.46	17.36	38.15
nathen_happy_no.wav	74.15	305.32	98.28	142.82	85.85	94.21	111.15	107.96	18.43	192.15	277.71	47.57	2.78	0.48	15.27	14.28
nathen_happy_yes.wav	73.74	336.43	131.88	63.11	128.91	132.25	137.66	150.67	19.89	260.9	262.13	41.13	2.62	0.41	16.82	6.31
nathen_neutral_manager.wa	75.19	334.92	105.47	130.12	97.75	102.39	106.21	224.32	36.01	227.5	142.32	45.98	1.42	0.46	16.75	13.01
nathen_neutral_no.wav	70.17	404.25	103.46	121.65	89.97	109.66	113.02	143.92	24.79	175.21	269.04	51.03	2.69	0.51	20.21	12.17
nathen_neutral_yes.wav	66.47	370.11	147.5	129.99	135.85	142.91	161.52	286.44	32.64	236.38	288.01	33.21	2.88	0.33	18.51	13
nora_angry_manager.wav	78.62	132.87	191.6	300.02	189.67	204.95	206.96	502.26	20.26	172.63	277.47	53.71	2.77	0.54	6.64	30
nora_angry_no.wav	80.99	132.54	240.4	360.29	228.97	231.49	234.75	476.43	30	200.34	284.82	64.26	2.85	0.64	6.63	36.03
nora_angry_yes.wav	72.77	401.51	204.18	252.85	176.35	203.51	224.25	260.54	22.31	245.35	340.4	51.96	3.4	0.52	20.08	25.28
nora_frust_manager.wav	76.56	178.19	188.3	300.86	189.58	201.43	204.7	246.27	15.42	150.74	378.67	57.02	3.79	0.57	8.91	30.09
nora_frust_no.wav	78.25	146.91	227.14	524.88	174.89	217.98	281.76	586.1	48.04	142.88	407.59	55.24	4.08	0.55	7.35	52.49
nora_frust_yes.wav	76.43	265.53	193.62	370.69	197.45	199.6	214.5	351.07	13.1	168.59	580.29	58.79	5.8	0.59	13.28	37.07
nora_happy_manager.wav	77.11	180.86	265.57	186.32	255.73	266.39	282.66	377.3	25.5	221.81	769.83	48.76	7.7	0.49	9.04	18.63
nora_happy_no.wav	81.16	105.73	222.63	451.11	183.23	232.68	259.44	540.31	26.93	106.09	841.27	66.71	8.41	0.67	5.29	45.11
nora_happy_yes.wav	69.73	221.92	243.25	566.28	205.29	243.2	296.22	781.96	41.78	181.92	480.9	45.41	4.81	0.45	11.1	56.63
nora_neutral_manager.wav	80.27	142.19	198.23	360.28	184.53	211.53	222.42	476.85	21.33	179.17	526.24	56.25	5.26	0.56	7.11	36.03
nora_neutral_no.wav	82.2	78.33	206.62	105.54	197.64	203.58	213.58	149.55	12.2	84.1	539.49	72.43	5.39	0.72	3.92	10.55
nora_neutral_yes.wav	75.13	190.3	201.34	114.68	189.64	202.98	212.74	268.41	23.01	201.28	533.3	54.26	5.33	0.54	9.52	11.47
phil_angry_no.wav	72.89	188.04	183.29	392.13	156.99	177.52	227.16	581.6	30.4	195.67	159.08	40.79	1.59	0.41	9.4	39.21
phil_frust_no.wav	80.23	124.87	173.43	266.74	151.49	177.94	192.38	320.33	33.73	205.65	105.91	43.75	1.06	0.44	6.24	26.67
phil_happy_no.wav	74.17	242.87	329.54	481.85	310.63	320.53	333.77	698.67	23.69	168.8	302.55	46.02	3.03	0.46	12.14	48.18
phil_neutral_no.wav	75.52	228.59	228.65	2128.28	96.81	97.99	549.37	1823.21	36.85	184.29	82.05	34.24	0.82	0.34	11.43	212.83
phil_neutral_yes.wav	62.48	276.75	102.79	15.39	102.39	103.05	104.05	52.14	8.79	197.87	18.88	32.08	0.19	0.32	13.84	1.54
vasco_angry_no.wav	79.82	159.98	123.63	287.25	111.95	117.77	128.19	444.09	52.17	98.78	484.46	56.53	4.84	0.57	8	28.73
vasco_neutral_manager.wav	78.49	162.87	139.02	617.65	93.55	134.49	142.81	670.05	41.15	195.2	298.3	41.06	2.98	0.41	8.'4	28.73
vasco_neutral_no.wav	82.18	79.28	112.11	106.1	101.76	115.2	122.35	172.47	26.92	53.09	325.3	54.69	3.25	0.55	3.96	10.61
vasko_angry_manager.wav	78.24	187.18	96.21	105.2	85.76	99.32	103.85	136.24	24.91	204.76	179.78	46.44	1.8	0.46	9.36	10.52
vasko_angry_yes.wav	79.71	151.15	219.7	1431.56	153.74	178.13	193.61	3399.83	70.8	137.49	371.33	54.18	3.71	0.54	7.56	143.16
vasko_frust_manager.wav	78.01	137.1	106.88	220.05	93.07	107.89	112.86	253.43	24.41	162.95	210.8	41.31	2.11	0.41	6.85	22
vasko_frustrated_no.wav	73.88	245.11	96.43	40.22	94.5	96.16	98.65	55.14	9.83	91.38	301.45	58.28	3.01	0.58	12.26	4.02
vasko_frustrated_yes.wav	70	297.51	101.18	110.28	93.87	101.96	109.57	223.7	37.79	172.24	154.44	32.13	1.54	0.32	14.88	11.03
vasko_happy_manager.wav	79.39	118.98	124.47	344.36	92.18	118.25	159.6	318.04	44.9	128.19	298.35	48.41	2.98	0.48	5.95	34.44
vasko_happy_no.wav	69.23	202.62	103.67	121.54	95.62	102.04	112.59	220.52	37.26	136.61	452.41	52.77	4.52	0.53	10.13	12.15
vasko_happy_yes.wav	69.56	349.64	117.35	198.36	95.26	126.96	134.34	307.87	47.16	178.53	466.12	41.53	4.66	0.42	17.48	19.84

20.1 GRAPHS OF RESULTS

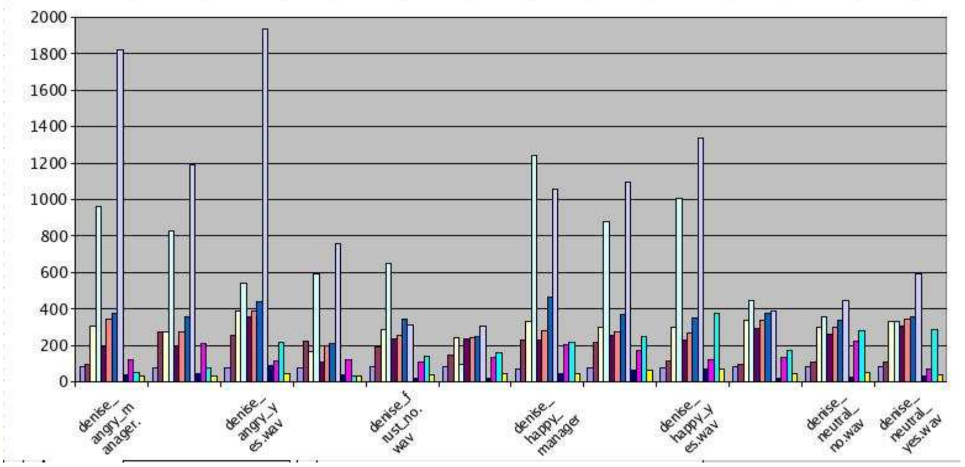
Phil



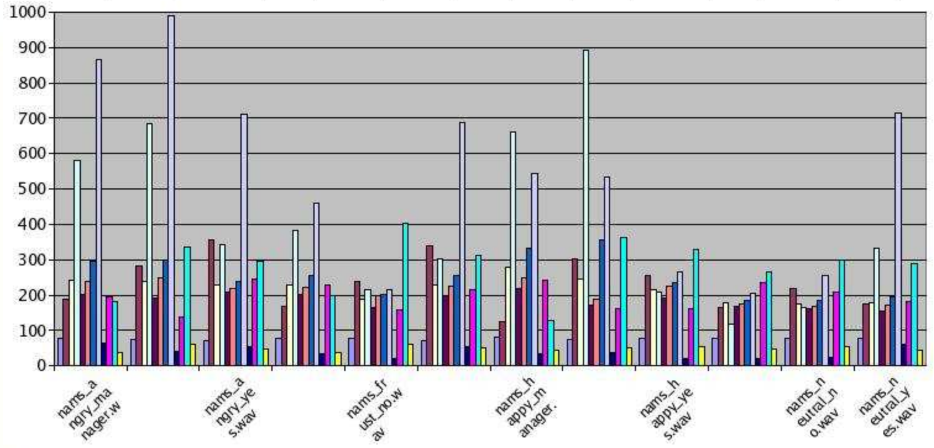
Anna



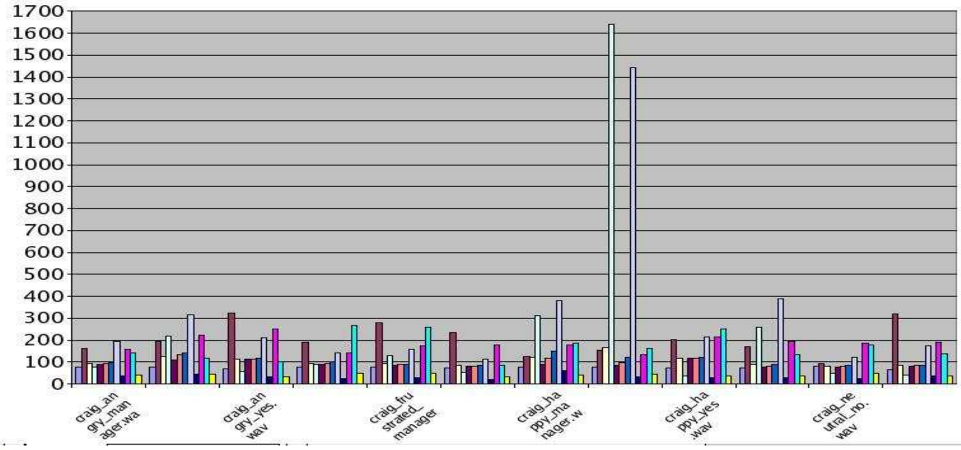
Denise



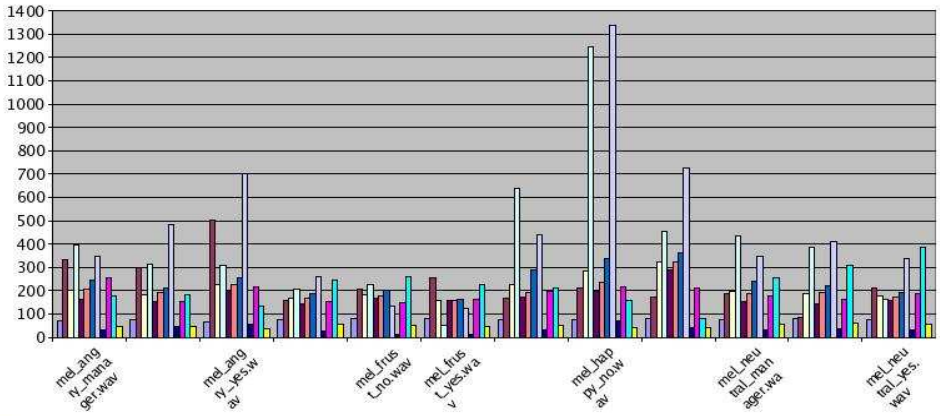
Naomi



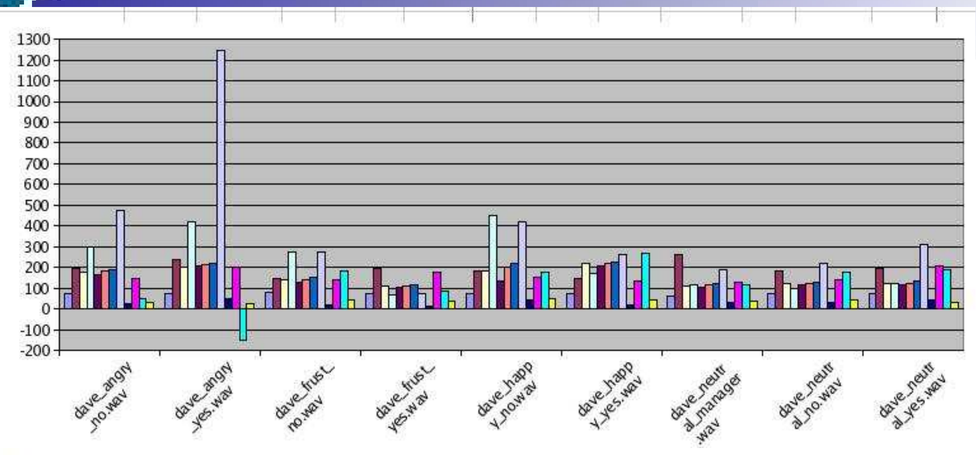
Craig



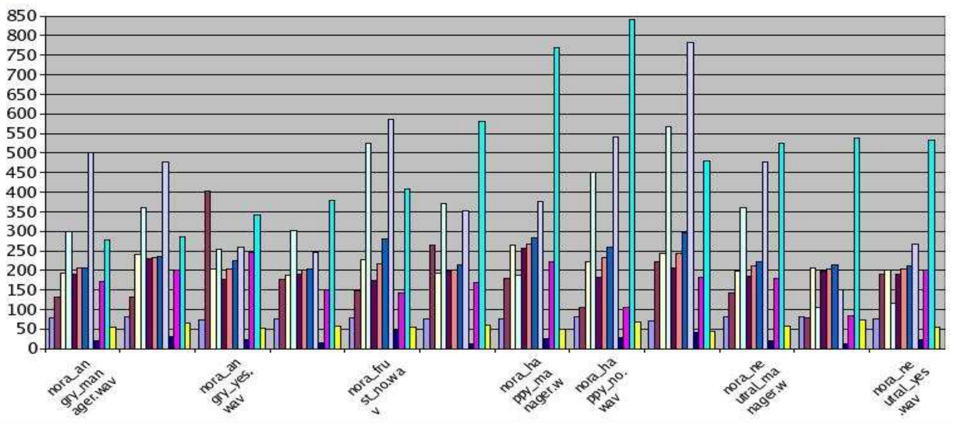
Melody



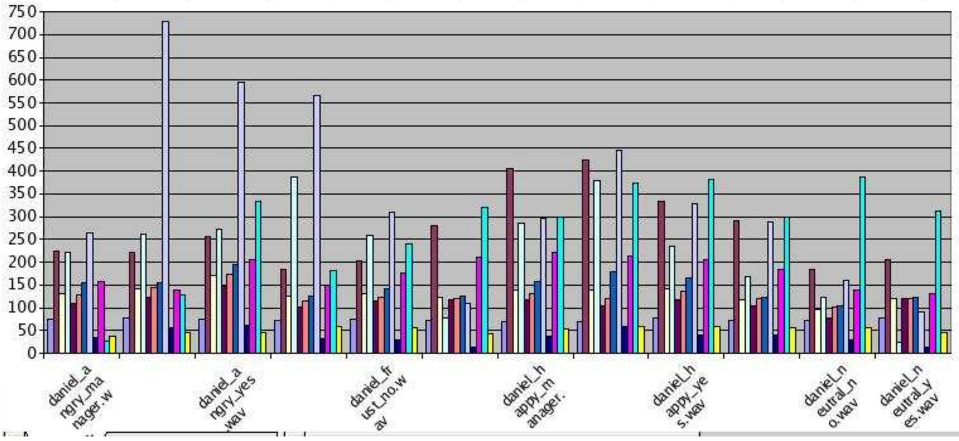
Dave



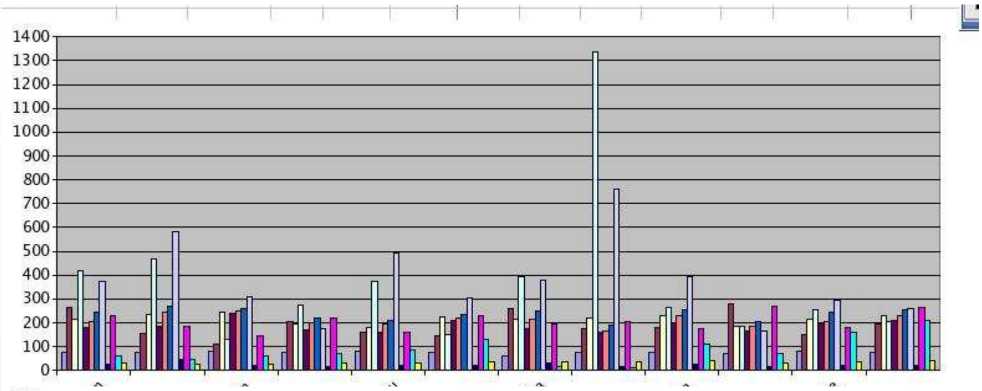
Nora



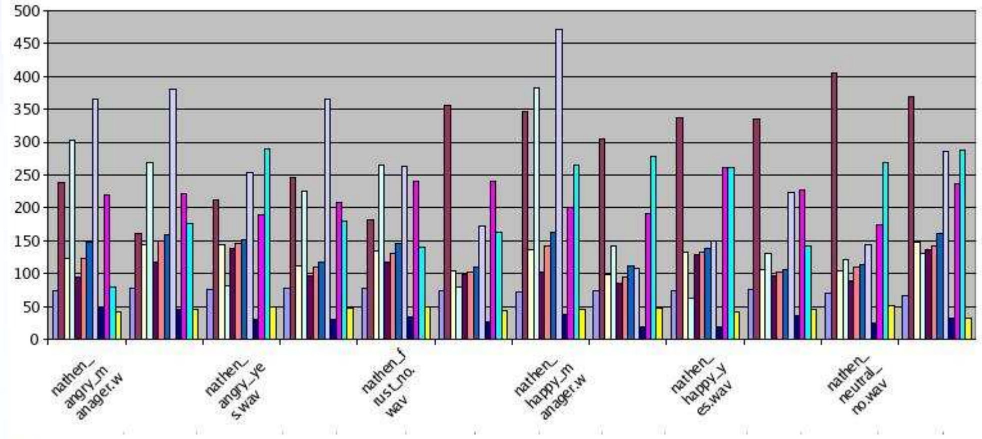
Daniel



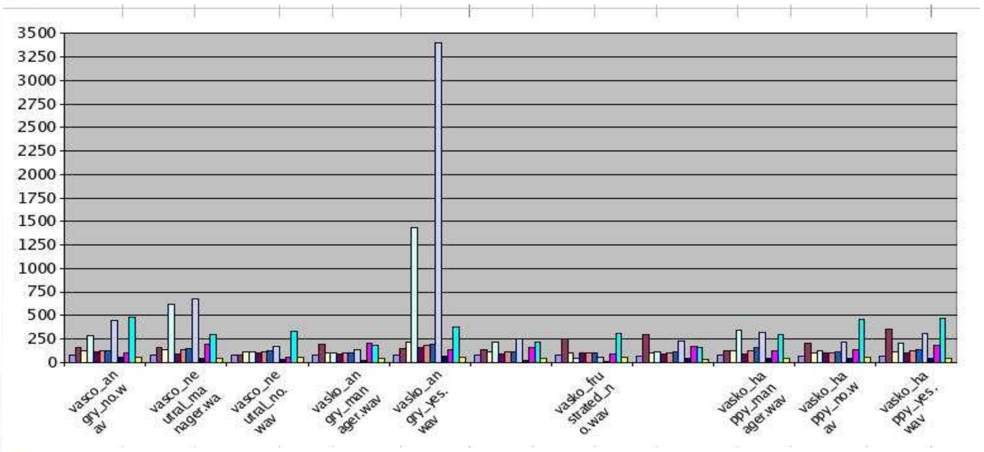
Kerry



Nathan



Vasko



Sample File	***Result***	***Pass/Fail***
anna_angry_manager	Angry	Pass
anna_angry_no	Angry	Pass
anna_angry_yes	Angry	Pass
anna_frustrated_manager	Not checked by analyser	
anna_frustrated_no	Not checked by analyser	
anna_frustrated_yes	Not checked by analyser	
anna_happy_manager	Sad	Fail
anna_happy_no	Happy	Pass
anna_happy_yes	Happy	Pass
anna_neutral_manager	Sad	Fail
anna_ neutral _no	Neutral	Pass
anna_ neutral _yes	Neutral	Pass
craig_angry_manager	Neutral	Fail
craig_angry_no	Neutral	Fail
craig_angry_yes	Neutral	Fail
craig_frustrated_manager	Not checked by analyser	
craig_frustrated_no	Not checked by analyser	
craig_frustrated_yes	Not checked by analyser	
craig_happy_manager	Sad	Fail
craig_happy_no	Happy	Pass
craig_happy_yes	Neutral	Fail
craig_ neutral _manager	Neutral	Pass
craig_ neutral_no	Neutral	Pass
craig_ neutral _yes	Neutral	Pass
daniel_angry_manager	Neutral	Fail
daniel_angry_no	Neutral	Fail
daniel_angry_yes	Neutral	Fail
daniel_frustrated_manager	Not checked by analyser	
daniel_frustrated_no	Not checked by analyser	
daniel_frustrated_yes	Not checked by analyser	
daniel_happy_manager	Neutral	Fail
daniel_happy_no	Angry	

Sample File	***Result***	***Pass/Fail***
daniel_happy_yes	Neutral	Fail
daniel_ neutral _manager	Neutral	Fail
daniel_ neutral _no	Neutral	Pass
daniel_ neutral _yes	Neutral	Pass
dave_ neutral _manager	Neutral	Pass
dave_ neutral _no	Neutral	Pass
dave_ neutral _yes	Neutral	Pass
dave_angry_no	Angry	Pass
dave_angry_yes	Angry	Pass
dave_frustrated_no	Not checked by analyser	
dave_frustrated_yes	Not checked by analyser	
dave_happy_no	Angry	Fail
dave_happy_yes	Happy	Pass
denise_angry_manager	Happy	Fail
denise_angry_no	Happy	Fail
denise_angry_yes	Angry	Pass
denise_frustrated_manager	Not checked by analyser	
denise_frustrated_no	Not checked by analyser	
denise_frustrated_yes	Not checked by analyser	
denise_happy_manager	Happy	Pass
denise_happy_no	Happy	Pass
denise_happy_yes	Happy	Pass
denise_ neutral _manager	Sad	Fail
denise_ neutral_no	Sad	Fail
denise_ neutral _yes	Sad	Fail
kerry_angry_manager	Angry	Pass
kerry_angry_no	Sad	Fail
kerry_angry_yes	Neutral	Fail
kerry_frustrated_manager	Not checked by analyser	
kerry_frustrated_no	Not checked by analyser	
kerry_frustrated_yes	Not checked by analyser	
kerry_happy_manager	Angry	Fail

Sample File	Result	Pass/Fail
kerry_happy_no	Happy	Pass
kerry_happy_yes	Neutral	Fail
kerry_ neutral _manager	Neutral	Pass
kerry_ neutral_no	Neutral	Pass
kerry_ neutral _yes	Neutral	Pass
mel_angry_manager	Angry	Pass
mel_angry_no	Angry	Pass
mel_angry_yes	Angry	Pass
mel_frustrated_manager	Not checked by analyser	
mel_frustrated_no	Not checked by analyser	
mel_frustrated_yes	Not checked by analyser	
mel_happy_manager	Sad	Fail
mel_happy_no	Happy	Pass
mel_happy_yes	Sad	Fail
mel_ neutral _manager	Angry	Fail
mel_ neutral_no	Sad	Fail
mel_ neutral _yes	Neutral	Pass
naomi_angry_manager	Angry	Pass
naomi_angry_no	Angry	Pass
naomi_angry_yes	Angry	Pass
naomi_frustrated_manager	Not checked by analyser	
naomi_frustrated_no	Not checked by analyser	
naomi_frustrated_yes	Not checked by analyser	
naomi_happy_manager	Sad	Fail
naomi_happy_no	Happy	Pass
naomi_happy_yes	Neutral	Fail
naomi_ neutral _manager	Neutral	Pass
naomi_ neutral_no	Neutral	Pass
naomi_ neutral _yes	Sad	Fail
nathen_angry_manager	Angry	Pass
nathen_angry_no	Neutral	Fail
nathen_angry_yes	Neutral	Fail

Sample File	Result	Pass/Fail
nathen_frustrated_manager	Not checked by analyser	
nathen_frustrated_no	Not checked by analyser	
nathen_frustrated_yes	Not checked by analyser	
nathen_happy_manager	Angry	Fail
nathen_happy_no	Neutral	Fail
nathen_happy_yes	Neutral	Fail
nathen_ neutral _manager	Neutral	Pass
nathen_ neutral_no	Neutral	Pass
nathen_ neutral _yes	Neutral	Pass
nora_angry_manager	Sad	Fail
nora_angry_no	Sad	Fail
nora_angry_yes	Neutral	Fail
nora_frustrated_manager	Not checked by analyser	
nora_frustrated_no	Not checked by analyser	
nora_frustrated_yes	Not checked by analyser	
nora_happy_manager	Neutral	Fail
nora_happy_no	Sad	Fail
nora_happy_yes	Angry	Fail
nora_ neutral _manager	Sad	Fail
nora_ neutral_no	Neutral	Pass
nora_ neutral _yes	Neutral	Pass
vasko_angry_manager	Neutral	Fail
vasko_angry_no	Neutral	Fail
vasko_angry_yes	Not Recorded	Fail
vasko_frustrated_manager	Not checked by analyser	
vasko_frustrated_no	Not checked by analyser	
vasko_frustrated_yes	Not checked by analyser	
vasko_happy_manager	Sad	Fail
vasko_happy_no	Neutral	Fail
vasko_happy_yes	Neutral	Fail
vasko_ neutral _manager	Neutral	Pass
vasko_ neutral_no	Neutral	Pass
vasko_ neutral _yes	Neutral	Pass

Table 22. Experiment 1 results

20.1.1 Cumulative Results

Subject	Score	Gender	Comments
Anna	7/9	Female	Trained Actor. Two SAD false Positives
Craig	4/9	Male	Did not detect ANGER. Two SAD false positives
Daniel	2/9	Male	Nothing but NEUTRAL – except False +ve ANGER
Dave	5/7	Male	Trained actor. Two failures detecting HAPPY
Denise	4/9	Male/ Female	Put on voice. Researcher imitating female. ANGER analysed as HAPPY
Kerry	4/9	Female	Soft voice – No ANGER detected
Melanie	5/9	Female	NEUTRAL & HAPPY misses
Naomi	6/9	Female	Trained Actor. False +ves with SAD
Nathan	4/9	Male	NEUTRAL False +ves
Nora	2/9	Female	Strong Hispanic accent. Only NEUTRAL Hits
Vasko	3/9	Male	Only NEUTRAL hits.
Denis	9/9	Male (Researcher)	Four featured analyser modeled on myself
Emina	8/9	Female (Researcher's wife)	Not trained – very expressive voice

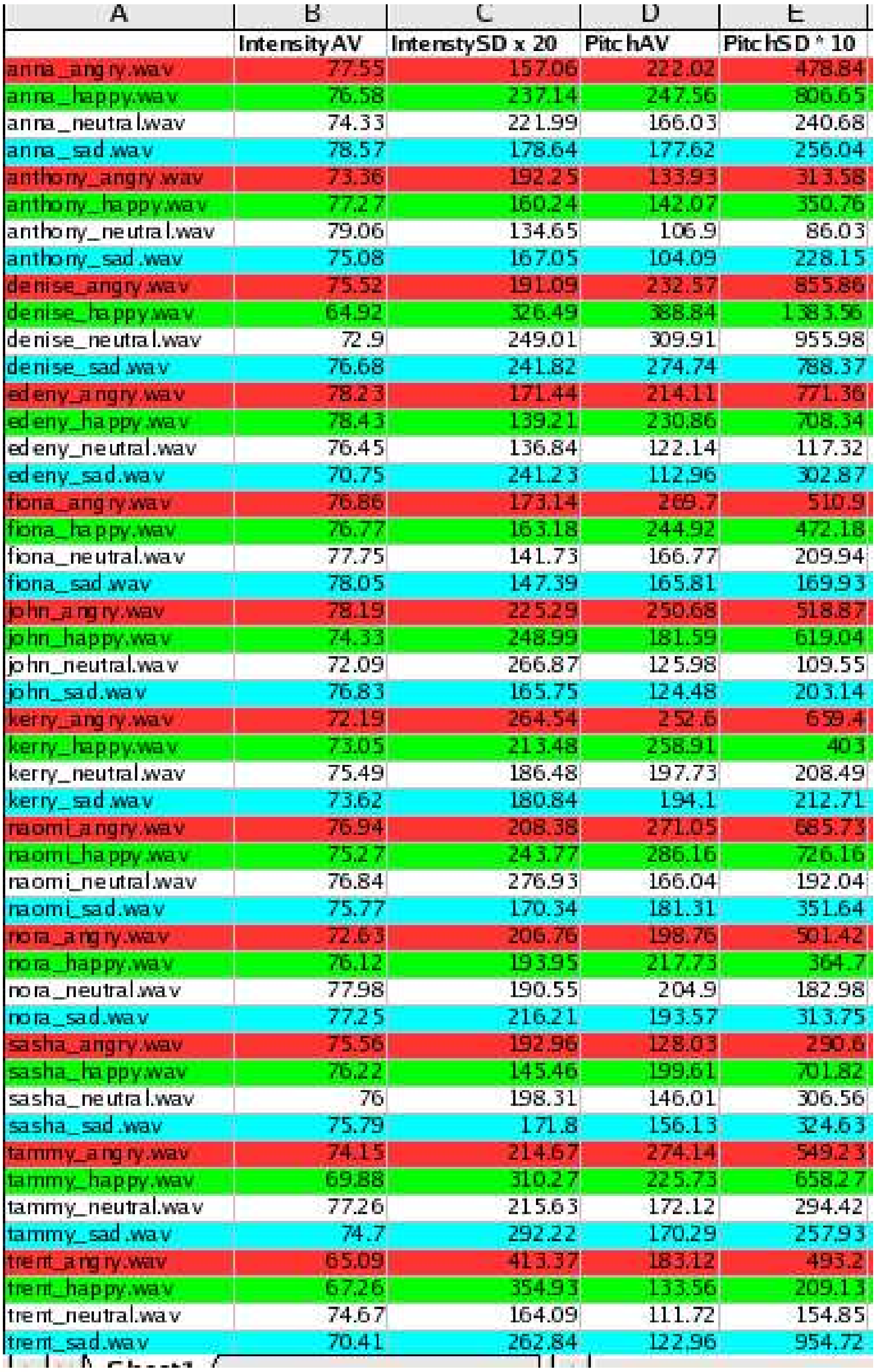
Table 23. Experiment 1 - Cumulative results

20.1.2 Overall Results

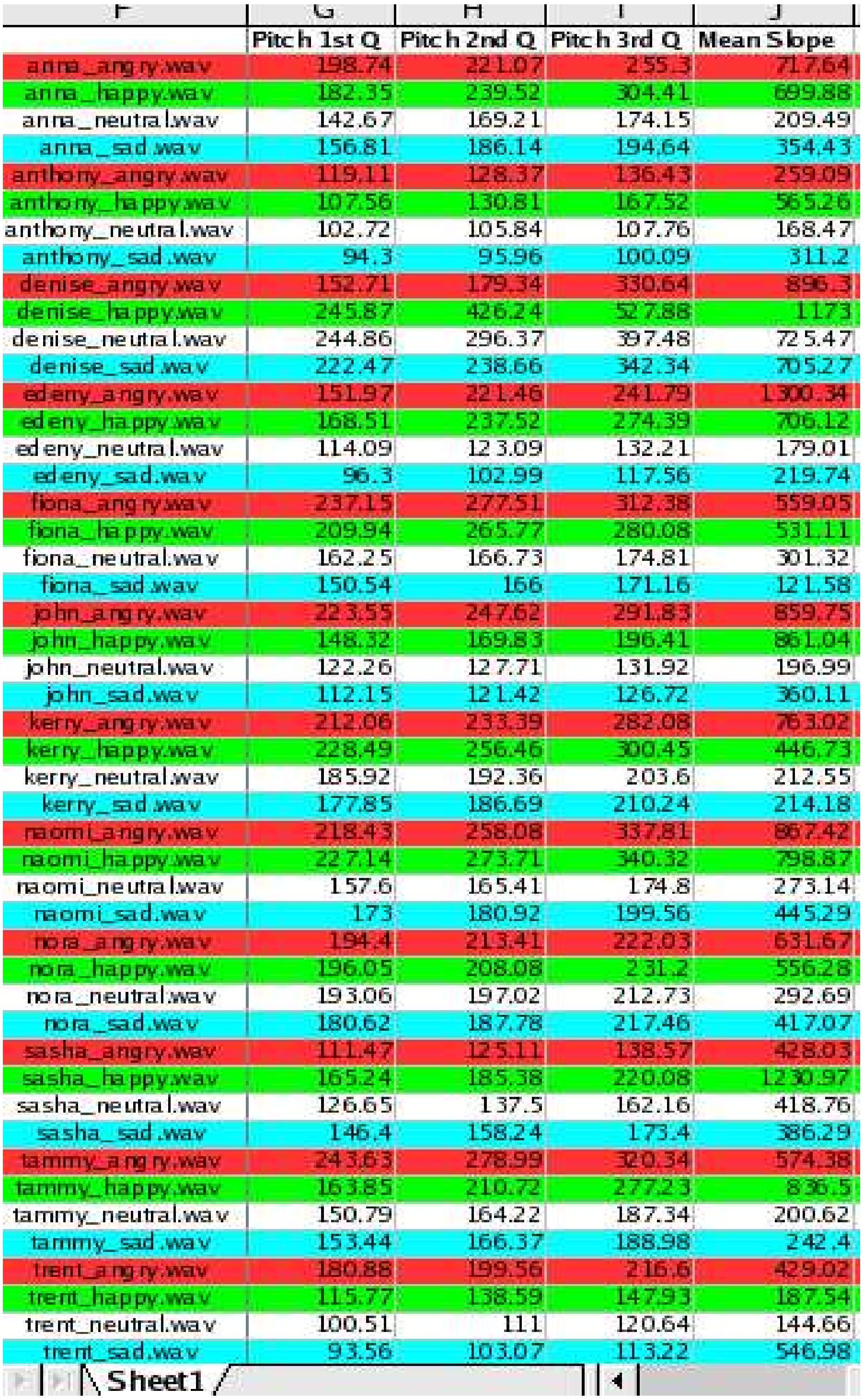
EMOTION	SCORE	COMMENTS
Happiness	10/32	Most prone to error. Difficult to detect in male subjects
Anger	13/32	Great results with trained or expressive voices
Neutral	24/33	Most Reliable result

Table 24. Experiment 1 - Overall Results

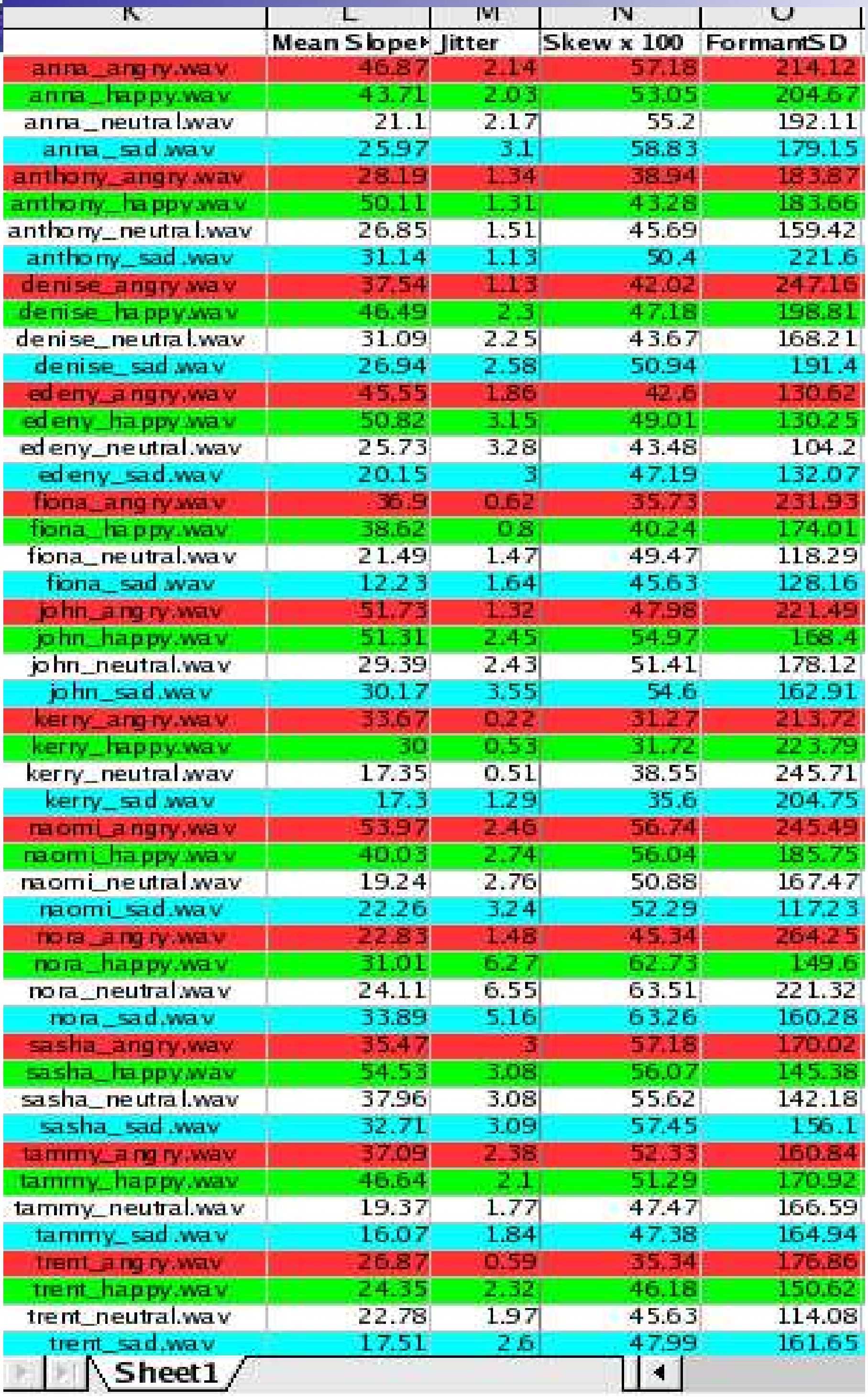
APPENDIX D - Results from Experiment 2



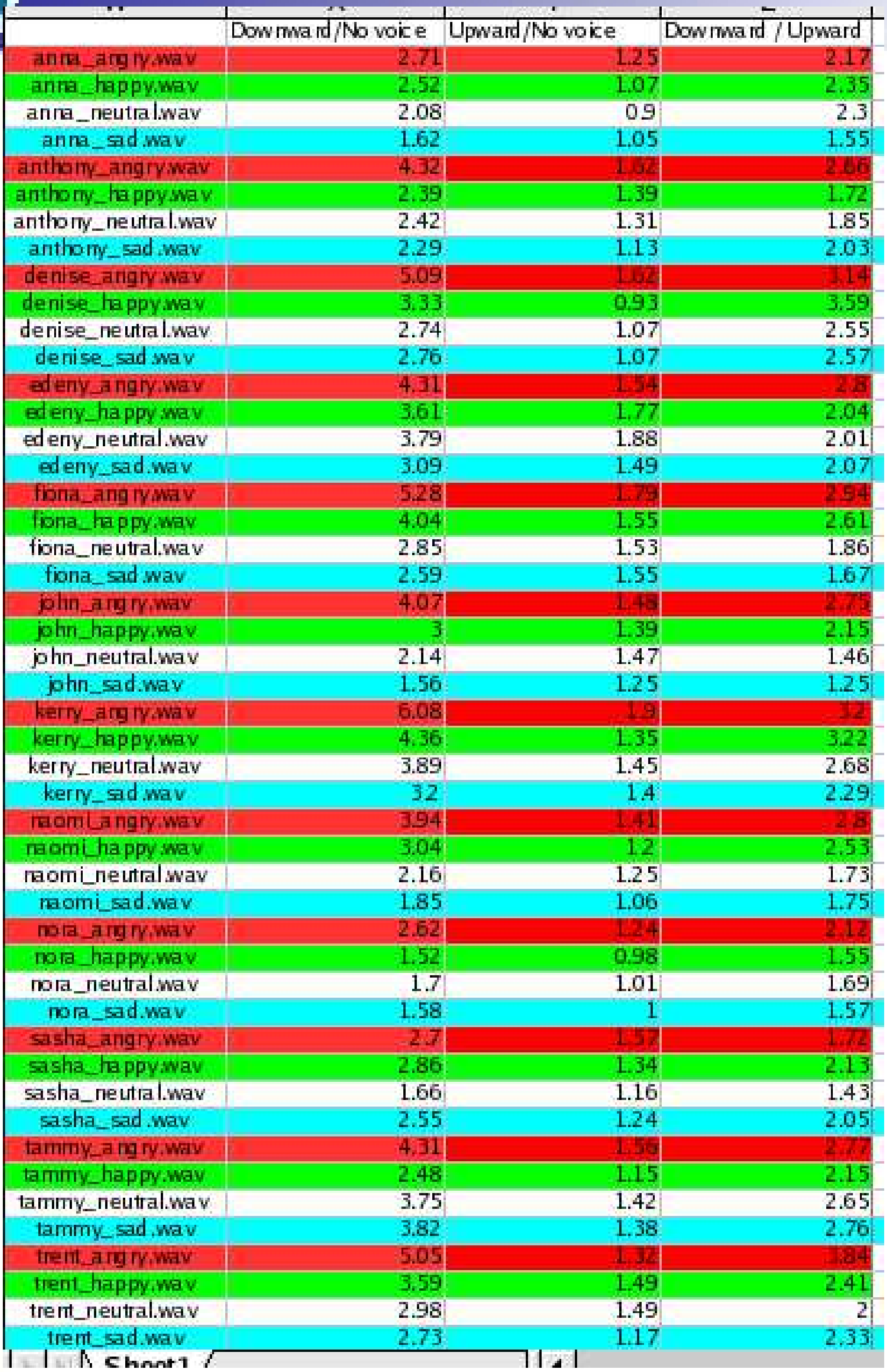
A	B	C	D	E
	IntensityAV	IntenstySD x 20	PitchAV	PitchSD * 10
anna_angry.wav	77.55	157.06	222.02	478.84
anna_happy.wav	76.58	237.14	247.56	806.65
anna_neutral.wav	74.33	221.99	166.03	240.68
anna_sad.wav	78.57	178.64	177.62	256.04
anthony_angry.wav	73.36	192.25	133.93	313.58
anthony_happy.wav	77.27	160.24	142.07	350.76
anthony_neutral.wav	79.06	134.65	106.9	86.03
anthony_sad.wav	75.08	167.05	104.09	228.15
denise_angry.wav	75.52	191.09	232.57	855.86
denise_happy.wav	64.92	326.49	388.84	1383.56
denise_neutral.wav	72.9	249.01	309.91	955.98
denise_sad.wav	76.68	241.82	274.74	788.37
edeny_angry.wav	78.23	171.44	214.11	771.36
edeny_happy.wav	78.43	139.21	230.86	708.34
edeny_neutral.wav	76.45	136.84	122.14	117.32
edeny_sad.wav	70.75	241.23	112.96	302.87
fiona_angry.wav	76.86	173.14	269.7	510.9
fiona_happy.wav	76.77	163.18	244.92	472.18
fiona_neutral.wav	77.75	141.73	166.77	209.94
fiona_sad.wav	78.05	147.39	165.81	169.93
john_angry.wav	78.19	225.29	250.68	518.87
john_happy.wav	74.33	248.99	181.59	619.04
john_neutral.wav	72.09	266.87	125.98	109.55
john_sad.wav	76.83	165.75	124.48	203.14
kerry_angry.wav	72.19	264.54	252.6	659.4
kerry_happy.wav	73.05	213.48	258.91	403
kerry_neutral.wav	75.49	186.48	197.73	208.49
kerry_sad.wav	73.62	180.84	194.1	212.71
naomi_angry.wav	76.94	208.38	271.05	685.73
naomi_happy.wav	75.27	243.77	286.16	726.16
naomi_neutral.wav	76.84	276.93	166.04	192.04
naomi_sad.wav	75.77	170.34	181.31	351.64
nora_angry.wav	72.63	206.76	198.76	501.42
nora_happy.wav	76.12	193.95	217.73	364.7
nora_neutral.wav	77.98	190.55	204.9	182.98
nora_sad.wav	77.25	216.21	193.57	313.75
sasha_angry.wav	75.56	192.96	128.03	290.6
sasha_happy.wav	76.22	145.46	199.61	701.82
sasha_neutral.wav	76	198.31	146.01	306.56
sasha_sad.wav	75.79	171.8	156.13	324.63
tammy_angry.wav	74.15	214.67	274.14	549.23
tammy_happy.wav	69.88	310.27	225.73	658.27
tammy_neutral.wav	77.26	215.63	172.12	294.42
tammy_sad.wav	74.7	292.22	170.29	257.93
trent_angry.wav	65.09	413.37	183.12	493.2
trent_happy.wav	67.26	354.93	133.56	209.13
trent_neutral.wav	74.67	164.09	111.72	154.85
trent_sad.wav	70.41	262.84	122.96	954.72



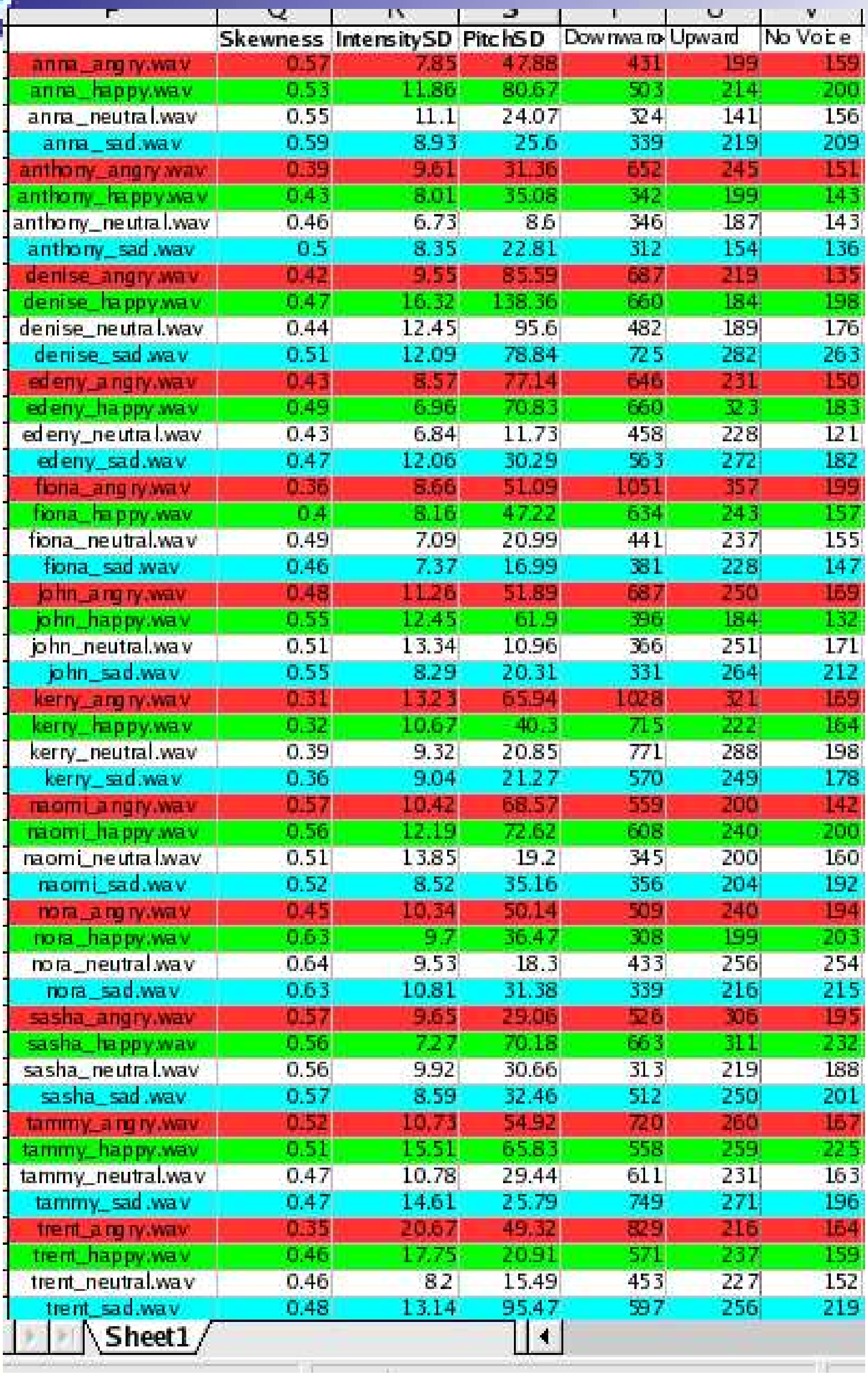
F	G	H	I	J
	Pitch 1st Q	Pitch 2nd Q	Pitch 3rd Q	Mean Slope
anna_angry.wav	198.74	221.07	255.3	717.64
anna_happy.wav	182.35	239.52	304.41	699.88
anna_neutral.wav	142.67	169.21	174.15	209.49
anna_sad.wav	156.81	186.14	194.64	354.43
anthony_angry.wav	119.11	128.37	136.43	259.09
anthony_happy.wav	107.56	130.81	167.52	565.26
anthony_neutral.wav	102.72	105.84	107.76	168.47
anthony_sad.wav	94.3	95.96	100.09	311.2
denise_angry.wav	152.71	179.34	330.64	896.3
denise_happy.wav	245.87	426.24	527.88	1173
denise_neutral.wav	244.86	296.37	397.48	725.47
denise_sad.wav	222.47	238.66	342.34	705.27
edeny_angry.wav	151.97	221.46	241.79	1300.34
edeny_happy.wav	168.51	237.52	274.39	706.12
edeny_neutral.wav	114.09	123.09	132.21	179.01
edeny_sad.wav	96.3	102.99	117.56	219.74
fiona_angry.wav	237.15	277.51	312.38	559.05
fiona_happy.wav	209.94	265.77	280.08	531.11
fiona_neutral.wav	162.25	166.73	174.81	301.32
fiona_sad.wav	150.54	166	171.16	121.58
john_angry.wav	223.55	247.62	291.83	859.75
john_happy.wav	148.32	169.83	196.41	861.04
john_neutral.wav	122.26	127.71	131.92	196.99
john_sad.wav	112.15	121.42	126.72	360.11
kerry_angry.wav	212.06	233.39	282.08	763.02
kerry_happy.wav	228.49	256.46	300.45	446.73
kerry_neutral.wav	185.92	192.36	203.6	212.55
kerry_sad.wav	177.85	186.69	210.24	214.18
naomi_angry.wav	218.43	258.08	337.81	867.42
naomi_happy.wav	227.14	273.71	340.32	798.87
naomi_neutral.wav	157.6	165.41	174.8	273.14
naomi_sad.wav	173	180.92	199.56	445.29
nora_angry.wav	194.4	213.41	222.03	631.67
nora_happy.wav	196.05	208.08	231.2	556.28
nora_neutral.wav	193.06	197.02	212.73	292.69
nora_sad.wav	180.62	187.78	217.46	417.07
sasha_angry.wav	111.47	125.11	138.57	428.03
sasha_happy.wav	165.24	185.38	220.08	1230.97
sasha_neutral.wav	126.65	137.5	162.16	418.76
sasha_sad.wav	146.4	158.24	173.4	386.29
tammy_angry.wav	243.63	278.99	320.34	574.38
tammy_happy.wav	163.85	210.72	277.23	836.5
tammy_neutral.wav	150.79	164.22	187.34	200.62
tammy_sad.wav	153.44	166.37	188.98	242.4
trent_angry.wav	180.88	199.56	216.6	429.02
trent_happy.wav	115.77	138.59	147.93	187.54
trent_neutral.wav	100.51	111	120.64	144.66
trent_sad.wav	93.56	103.07	113.22	546.98



K	L	M	N	O
	Mean Slope▸	Jitter	Skew x 100	FormantSD
anna_angry.wav	46.87	2.14	57.18	214.12
anna_happy.wav	43.71	2.03	53.05	204.67
anna_neutral.wav	21.1	2.17	55.2	192.11
anna_sad.wav	25.97	3.1	58.83	179.15
anthony_angry.wav	28.19	1.34	38.94	183.87
anthony_happy.wav	50.11	1.31	43.28	183.66
anthony_neutral.wav	26.85	1.51	45.69	159.42
anthony_sad.wav	31.14	1.13	50.4	221.6
denise_angry.wav	37.54	1.13	42.02	247.16
denise_happy.wav	46.49	2.3	47.18	198.81
denise_neutral.wav	31.09	2.25	43.67	168.21
denise_sad.wav	26.94	2.58	50.94	191.4
edeny_angry.wav	45.55	1.86	42.6	130.62
edeny_happy.wav	50.82	3.15	49.01	130.25
edeny_neutral.wav	25.73	3.28	43.48	104.2
edeny_sad.wav	20.15	3	47.19	132.07
fiona_angry.wav	36.9	0.62	35.73	231.93
fiona_happy.wav	38.62	0.8	40.24	174.01
fiona_neutral.wav	21.49	1.47	49.47	118.29
fiona_sad.wav	12.23	1.64	45.63	128.16
john_angry.wav	51.73	1.32	47.98	221.49
john_happy.wav	51.31	2.45	54.97	168.4
john_neutral.wav	29.39	2.43	51.41	178.12
john_sad.wav	30.17	3.55	54.6	162.91
kerry_angry.wav	33.67	0.22	31.27	213.72
kerry_happy.wav	30	0.53	31.72	223.79
kerry_neutral.wav	17.35	0.51	38.55	245.71
kerry_sad.wav	17.3	1.29	35.6	204.75
naomi_angry.wav	53.97	2.46	56.74	245.49
naomi_happy.wav	40.03	2.74	56.04	185.75
naomi_neutral.wav	19.24	2.76	50.88	167.47
naomi_sad.wav	22.26	3.24	52.29	117.23
nora_angry.wav	22.83	1.48	45.34	264.25
nora_happy.wav	31.01	6.27	62.73	149.6
nora_neutral.wav	24.11	6.55	63.51	221.32
nora_sad.wav	33.89	5.16	63.26	160.28
sasha_angry.wav	35.47	3	57.18	170.02
sasha_happy.wav	54.53	3.08	56.07	145.38
sasha_neutral.wav	37.96	3.08	55.62	142.18
sasha_sad.wav	32.71	3.09	57.45	156.1
tammy_angry.wav	37.09	2.38	52.33	160.84
tammy_happy.wav	46.64	2.1	51.29	170.92
tammy_neutral.wav	19.37	1.77	47.47	166.59
tammy_sad.wav	16.07	1.84	47.38	164.94
trent_angry.wav	26.87	0.59	35.34	176.86
trent_happy.wav	24.35	2.32	46.18	150.62
trent_neutral.wav	22.78	1.97	45.63	114.08
trent_sad.wav	17.51	2.6	47.99	161.65



	Downward/No voice	Upward/No voice	Downward / Upward
anna_angry.wav	2.71	1.25	2.17
anna_happy.wav	2.52	1.07	2.35
anna_neutral.wav	2.08	0.9	2.3
anna_sad.wav	1.62	1.05	1.55
anthony_angry.wav	4.32	1.82	2.66
anthony_happy.wav	2.39	1.39	1.72
anthony_neutral.wav	2.42	1.31	1.85
anthony_sad.wav	2.29	1.13	2.03
denise_angry.wav	5.09	1.62	3.14
denise_happy.wav	3.33	0.93	3.59
denise_neutral.wav	2.74	1.07	2.55
denise_sad.wav	2.76	1.07	2.57
edeny_angry.wav	4.31	1.54	2.8
edeny_happy.wav	3.61	1.77	2.04
edeny_neutral.wav	3.79	1.88	2.01
edeny_sad.wav	3.09	1.49	2.07
fiona_angry.wav	5.28	1.79	2.94
fiona_happy.wav	4.04	1.55	2.61
fiona_neutral.wav	2.85	1.53	1.86
fiona_sad.wav	2.59	1.55	1.67
john_angry.wav	4.07	1.48	2.75
john_happy.wav	3	1.39	2.15
john_neutral.wav	2.14	1.47	1.46
john_sad.wav	1.56	1.25	1.25
kerry_angry.wav	6.08	1.9	3.2
kerry_happy.wav	4.36	1.35	3.22
kerry_neutral.wav	3.89	1.45	2.68
kerry_sad.wav	3.2	1.4	2.29
naomi_angry.wav	3.94	1.41	2.8
naomi_happy.wav	3.04	1.2	2.53
naomi_neutral.wav	2.16	1.25	1.73
naomi_sad.wav	1.85	1.06	1.75
nora_angry.wav	2.62	1.24	2.12
nora_happy.wav	1.52	0.98	1.55
nora_neutral.wav	1.7	1.01	1.69
nora_sad.wav	1.58	1	1.57
sasha_angry.wav	2.7	1.57	1.72
sasha_happy.wav	2.86	1.34	2.13
sasha_neutral.wav	1.66	1.16	1.43
sasha_sad.wav	2.55	1.24	2.05
tammy_angry.wav	4.31	1.56	2.77
tammy_happy.wav	2.48	1.15	2.15
tammy_neutral.wav	3.75	1.42	2.65
tammy_sad.wav	3.82	1.38	2.76
trent_angry.wav	5.05	1.32	3.84
trent_happy.wav	3.59	1.49	2.41
trent_neutral.wav	2.98	1.49	2
trent_sad.wav	2.73	1.17	2.33



P	Q	R	S	T	U	V
	Skewness	IntensitySD	PitchSD	Downward	Upward	No Voice
anna_angry.wav	0.57	7.85	47.88	431	199	159
anna_happy.wav	0.53	11.86	80.67	503	214	200
anna_neutral.wav	0.55	11.1	24.07	324	141	156
anna_sad.wav	0.59	8.93	25.6	339	219	209
anthony_angry.wav	0.39	9.61	31.36	652	245	151
anthony_happy.wav	0.43	8.01	35.08	342	199	143
anthony_neutral.wav	0.46	6.73	8.6	346	187	143
anthony_sad.wav	0.5	8.35	22.81	312	154	136
denise_angry.wav	0.42	9.55	85.59	687	219	135
denise_happy.wav	0.47	16.32	138.36	660	184	198
denise_neutral.wav	0.44	12.45	95.6	482	189	176
denise_sad.wav	0.51	12.09	78.84	725	282	263
edeny_angry.wav	0.43	8.57	77.14	646	231	150
edeny_happy.wav	0.49	6.96	70.83	660	323	183
edeny_neutral.wav	0.43	6.84	11.73	458	228	121
edeny_sad.wav	0.47	12.06	30.29	563	272	182
fiona_angry.wav	0.36	8.66	51.09	1051	357	199
fiona_happy.wav	0.4	8.16	47.22	634	243	157
fiona_neutral.wav	0.49	7.09	20.99	441	237	155
fiona_sad.wav	0.46	7.37	16.99	381	228	147
john_angry.wav	0.48	11.26	51.89	687	250	169
john_happy.wav	0.55	12.45	61.9	396	184	132
john_neutral.wav	0.51	13.34	10.96	366	251	171
john_sad.wav	0.55	8.29	20.31	331	264	212
kerry_angry.wav	0.31	13.23	65.94	1028	321	169
kerry_happy.wav	0.32	10.67	40.3	715	222	164
kerry_neutral.wav	0.39	9.32	20.85	771	288	198
kerry_sad.wav	0.36	9.04	21.27	570	249	178
naomi_angry.wav	0.57	10.42	68.57	559	200	142
naomi_happy.wav	0.56	12.19	72.62	608	240	200
naomi_neutral.wav	0.51	13.85	19.2	345	200	160
naomi_sad.wav	0.52	8.52	35.16	356	204	192
nora_angry.wav	0.45	10.34	50.14	509	240	194
nora_happy.wav	0.63	9.7	36.47	308	199	203
nora_neutral.wav	0.64	9.53	18.3	433	256	254
nora_sad.wav	0.63	10.81	31.38	339	216	215
sasha_angry.wav	0.57	9.65	29.06	526	306	195
sasha_happy.wav	0.56	7.27	70.18	663	311	232
sasha_neutral.wav	0.56	9.92	30.66	313	219	188
sasha_sad.wav	0.57	8.59	32.46	512	250	201
tammy_angry.wav	0.52	10.73	54.92	720	260	167
tammy_happy.wav	0.51	15.51	65.83	558	259	225
tammy_neutral.wav	0.47	10.78	29.44	611	231	163
tammy_sad.wav	0.47	14.61	25.79	749	271	196
trent_angry.wav	0.35	20.67	49.32	829	216	164
trent_happy.wav	0.46	17.75	20.91	571	237	159
trent_neutral.wav	0.46	8.2	15.49	453	227	152
trent_sad.wav	0.48	13.14	95.47	597	256	219

Subject	***Gender***	***Emotion***	***Result with original 4 feature analyser***	***Result with 2nd generation Analyser***	
Anna	F	Calm (neutral)	**Neutral**	**neutral**	
		Happy	**Happy**	sad	
		Sad	neutral	**sad**	
		Angry	sad	Can't decide between Anger and Happy	
Anthony	M	Calm (neutral)	**neutral**	**neutral**	
		Happy	sad	sad	
		Sad	neutral	**sad**	
		Angry	**angry**	Can't decide between Anger and Happy	
Denise	F (pseudo)	Calm (neutral)	happy	happy	
		Happy	**happy**	**happy**	
		Sad	happy	angry	
		Angry	happy	**angry**	
E deny	M	Calm (neutral)	**neutral**	**neutral**	
		Happy	sad	Can't decide between Anger and Happy	
		Sad	angry	neutral	
		Angry	happy	**angry**	
Fiona	F	Calm (neutral)	**neutral**	sad	
		Happy	sad	Can't decide between Anger and Happy	
		Sad	neutral	neutral	
		Angry	sad	**angry**	
Helen	F	Calm (neutral)	**neutral**	**neutral**	
		Happy	angry	neutral	
		Sad	neutral	neutral	
		Angry	happy	**angry**	
John	M	Calm (neutral)	**neutral**	**neutral**	
		Happy	angry	sad	
		Sad	neutral	**sad**	

Subject	*Gender*	*Emotion*	*Result with 4 feature analyser*	*Result with 4 feature plus contour*	*Pass / Fail*
		Angry	angry	Can't decide between Anger and Happy	
Kerry	F	Calm (neutral)	neutral	neutral	
		Happy	angry	angry	
		Sad	neutral	neutral	
		Angry	**angry**	**angry**	
Naomi	F	Calm (neutral)	**neutral**	**neutral**	
		Happy	angry	Can't decide between Anger and Happy	
		Sad	**sad**	**sad**	
		Angry	**angry**	Can't decide between Anger and Neutral	
Nora	F	Calm (neutral)	**neutral**	**neutral**	
		Happy	angry	sad	
		Sad	angry	**sad**	
		Angry	**angry**	neutral	
Richard	M	Calm (neutral)	**neutral**	**neutral**	
		Happy	neutral	neutral	
		Sad	neutral	neutral	
		Angry	neutral	neutral	
Rick	M	Calm (neutral)	sad	sad	
		Happy	angry	sad	
		Sad	happy	happy	
		Angry	neutral	neutral	
Ross	M	Calm (neutral)	**neutral**	**neutral**	
		Happy	angry	Can't decide between Anger and sadness	
		Sad	neutral	neutral	
		Angry	neutral	neutral	
Sasha	M	Calm (neutral)	angry	sad	
		Happy	sad	angry	

Subject	***Gender***	***Emotion***	***Result with 4 feature analyser***	***Result with 4 feature plus contour***	***Pass / Fail***
		Sad	**sad**	**sad**	
		Angry	neutral	sad	
		Happy	**happy**	**happy**	
		Sad	angry	**sad**	
		Angry	**angry**	sad	
Suzanne	F	Calm (neutral)	angry	happy	
		Happy	neutral	Can't decide between Anger and Happy	
		Sad	happy	neutral	
		Angry	neutral	**angry**	
Tammy	F	Calm (neutral)	**neutral**	**neutral**	
		Happy	angry	sad	
		Sad	neutral	neutral	
		Angry	**angry**	**angry**	
Trent	M	Calm (neutral)	**neutral**	**neutral**	
		Happy	neutral	neutral	
		Sad	happy	happy	
		Angry	**angry**	**angry**	
PERFORMANCE			24/72	29/72	

Table 25. Experiment 2 results

APPENDIX E – Emotion Detection Engine Source Code

22.1 1st Generation analyser PRAAT script

```
form Enter Filename excluding the .wav extension
   text  Filename
endform

ip_type$ = "UDP"
# UDP or TCP

pitchSD = 999

Read from file... /home/denis/Documents/UNI/Thesis/PRAAT/'filename$'.wav
 select Sound 'filename$'
 # Play
 To Intensity... 100 0
 intensitySD = Get standard deviation... 0 0
 select Sound 'filename$'
 To Pitch... 0 75 600
 pitchSD = Get standard deviation... 0 0 Hertz
 pitchQuant2nd = Get quantile... 0 0 0.5 Hertz
echo pitchSD = 'pitchSD'

if pitchSD = 999
   emotion = 9
else
 if pitchSD < 30
                        emotion = 1
 elsif pitchSD > 75
                        emotion = 2
 elsif intensitySD <9
                        emotion = 3
 else
                        emotion = 4
 endif
endif
 if emotion = 1
                echo "Neutral"
 elsif emotion = 2
                echo "Happy"
 elsif emotion = 3
                echo "Sad"
 elsif emotion = 4
                echo "Angry"
 else
     echo "No detection"
 endif
 if emotion = 1
                echo "Neutral"
                system cd /home/denis/Documents/uni/thesis/PRAAT; perl TCP_neutral.pl
 elsif emotion = 2
                echo "Happy"
                system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl TCP_happy.pl
 elsif emotion = 3
                echo "Sad"
                system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl
'ip_type$'_sad.pl

 elsif emotion = 4
                echo "Angry"
                system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl TCP_angry.pl
 else
     system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl TCP_no_activity.pl
                echo "No activity"
 endif
#system rm 'filename$'.wav
exit
```

22.2 2nd Generation analyser PRAAT script.

```
form Enter Filename excluding the .wav extension
   text  Filename
   integer Downward 0
   integer No_voice 0
endform
down_voice = downward / no_voice

ip_type$ = "UDP"
# UDP or TCP

# Do not test intensity if AGC on
agc_on$ = "yes"
# 'yes' or 'no'

Read from file... /home/denis/Documents/UNI/Thesis/PRAAT/'filename$'.wav

select Sound 'filename$'
 # Play
 To Intensity... 100 0
 intensitySD = Get standard deviation... 0 0
 select Sound 'filename$'
 To Pitch... 0 75 600
 pitchSD = Get standard deviation... 0 0 Hertz
 pitchQuant1st = Get quantile... 0 0 0.25 Hertz

 select Sound 'filename$'
 To Pitch... 0 75 600
 pitchAV = Get mean... 0 0 Hertz
```

```
# Try and determine sex
if pitchAV < 150
                gender$ = "male"
else
                gender$ = "female"
endif

select Sound 'filename$'
To PointProcess (extrema)... yes no Sinc70
jit = Get jitter (local)... 0 0 0.0001 0.02

select Sound 'filename$'
To Pitch... 0 75 600
mean_slope = Get mean absolute slope... Hertz

#echo filename = 'filename$'
#echo pitchSD = 'pitchSD'

# Initialise Emotion point scores  **************
sadness = 0
happiness = 0
anger = 0
neutral = 0
# *************************************************
emotion = 0
 if pitchSD < 20
                neutral = 9
 elsif pitchSD < 30
                neutral  = 5
                sadness  = 4
 endif
 if pitchSD > 85
                happiness = 9
 elsif pitchSD > 70
    happiness = happiness + 5
            anger = anger + 5
 endif
if agc_on$ = "no"
 if intensitySD <9
            sadness = sadness + 5
  endif
endif
if jit > 3
            sadness = sadness + 7
elsif jit > 2.5
            sadness = sadness + 4
endif
if mean_slope < 250
            neutral = neutral + 5
elsif mean_slope > 1000
            anger = anger + 8
elsif mean_slope > 700
            anger = anger + 5
            happiness = happiness + 5
endif
#printline Pitch 1st Quantile = 'pitchQuant1st''newline$'
#printline Subject is a 'gender$''newline$''newline$'
print 'filename$''tab$'
#print 'pitchQuant1st''tab$'
print 'gender$''tab$'
if gender$ = "male"
 if pitchQuant1st < 100
            sadness = sadness + 8
 elsif pitchQuant1st < 115
            sadness = sadness + 6
 endif
```

```
else
 if pitchQuant1st < 160
                        sadness = sadness + 8
 elsif pitchQuant1st < 190
                        sadness = sadness + 6
 endif
endif
if down_voice > 5
                        anger = anger + 9
elsif down_voice > 4
                        anger = anger + 5
                        happiness = happiness + 5
elsif down_voice < 2
                        sadness = sadness + 4
else
                        neutral = neutral + 4
endif

if downward  > 700
                        anger = anger + 8
endif
#print 'down_voice"tab$'
print 'sadness"tab$'
print 'happiness"tab$'
print 'anger"tab$'
print 'neutral"tab$'
#print 'mean_slope"tab$'
#  **** Derive emotion from pointscore  *******

emotion = 0
```

```
if sadness > neutral
        if sadness > happiness
                if sadness > anger
                        emotion = 3
                endif
        endif
endif
if emotion = 0
 if happiness > neutral
        if happiness > sadness
                if happiness > anger
                        emotion = 2
                endif
        endif
 endif
endif
if emotion = 0
 if anger > neutral
        if anger > happiness
                if anger > sadness
                        emotion = 4
                endif
        endif
 else
        emotion = 1
 endif
endif
if emotion = 1
        print Emotion is Neutral 'newline$'
#       #system cd /home/denis/Documents/uni/thesis/PRAAT; perl
'ip_type$'_neutral.pl
```

```
 elsif emotion = 2
                        print Emotion is Happy 'newline$'
#                       #system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl
'ip_type$'_happy.pl


 elsif emotion = 3
                        print Emotion is Sad 'newline$'
#                       #system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl
'ip_type$'_sad.pl


 elsif emotion = 4
                        print Emotion is Angry 'newline$'
#                       #system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl
'ip_type$'_angry.pl
else
 #      system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl 'ip_type$'_no_activity.pl
                        print No Voice Detected 'newline$'
 endif
#system rm 'filename$'.wav
select Sound 'filename$'
plus Intensity 'filename$'
plus Pitch 'filename$'
Remove


exit
```

22.3 Unix application script

```
#!/bin/bash
#getsound1.sh

cd /home/denis/Documents/UNI/Thesis/PRAAT
    socketID = $1  #load socket info for Comms

                        echo $! " = Rec task number"
                        echo
                        /bin/sleep 1

                        pid=$!
                        ((pid = $pid + 4))
                        #/bin/kill -INT $pid # sends SIGTERM (CRTL-C) to last background job +4
                        echo $pid " = killed task"
                        #/bin/kill -9 $pid # sends SIGTERM (CRTL-C) to eradicate pid
                        /bin/sleep 2

                        ./praat check_emotion      #PRAAT Script
                        echo
                        echo "PRAAT task number = "$!
                        echo

 /home/denis/Documents/UNI/Thesis/PRAAT/test_sound.wav
                        #echo "removed /
home/denis/Documents/UNI/Thesis/PRAAT/test_sound.wav"
#INT = CNTL C
#TERM = CNTL Z
```

22.4 Updated PRAAT script for 2nd Order Analyser

```
form Enter file without extension
                    text File_name
endform


Read from file... /home/denis/Documents/UNI/Thesis/PRAAT/'file_name$'.wav



select Sound 'file_name$'
To Pitch... 0 75 600
Write to short text file... /
home/denis/Documents/UNI/Thesis/PRAAT/object_data.Pitch


#  Call Perl Process to perform 2nd order analysis.  Pass file name


printline 'file_name$''tab$'


system cd /home/denis/Documents/UNI/Thesis/PRAAT; perl
pitch_analize_RT.pl 'file_name$'



exit


##  FINIS  ##
```

22.5 Perl Script to Extract Pitch Contour and zero voiced samples

```
#!/usr/bin/perl
   use IO::Socket;
# this programme opens a PRAAT pitch object file and analyses the pitch contour
# as well as deriving the pitch breaks
#
# Returns
#
#  Numb_Breaks (Integer 8 bit)
#  Pitch_slope (string:- "up","down","flat")
#  Denis Ryan 2003
#  Version 2.00  Used real time for analysis for Emotion Detector

$port = 2000;                  #UDP Port
$server='192.168.0.2';         #peer Address
$line_num = 0;
$file_type = qq("Pitch"\n);   # File identifier
$candidate_pos_count = 12;    # Initially candidate count 13 lines from TOF
$frame_count = 0;              # Forthcoming frame counts
$toggle = 0;                   # used to get every second line
$pos_num = 0;                  #initialise array pointer
$first_freq = 13;              #1st freq at line 13
$totalHz1st = 0;               #initialise freq counter
$totalHz2nd = 0;               #initialise 2nd 1/2 freq counte
$rec_num = 0;                  #total number of records
$weight_1 = 0;                 #weighting for 1st half
$weight_2 = 0;                 #weighting for 2nd half
$lower_count = 0;              #number of downward pitch changes
$upper_count = 0;              #number of upward pitch changes
$stationary_count=0;           #number of valid non movements
$zero_count=0;                 #number of unvoiced frames
$freq_cnt = 0;                 #pointer to valid frequencies
```

```
open (SOUNDFILE,"object_data.Pitch") || die "Cannot open file 'pitch_data.Pitch'";
while(<SOUNDFILE>)
 {
  $line_num++;
  $objectdata_line[$line_num] = $_;

 # print "\n$line_num\n";
  if ($line_num == 2){     # check if File OK
   if ($_ ne $file_type){
             die "\nNot a pitch object file\n";  # Check Script
   }
  }
  if ($line_num == 5){
   $utterance_length = $_;  #load the total file time in seconds
  }
 }
$rec_num = $line_num;          #preserve record count
#print "\n$utterance_length\n";
#print "\n Number of Records = $rec_num\n";


$line_num = $first_freq;              #reset for analysis
$frame_count = $objectdata_line[$candidate_pos_count];

#Get first half of utterance  ########################

for( ; $line_num < $rec_num/2 ; ){
 for ($x=0 ; $x<$frame_count ; $x++){
  if ($objectdata_line[$line_num] != 0){     #weight a positive recording
   $weight_1++;
   $freq_array[$freq_cnt++] = $objectdata_line[$line_num];
   if ( $freq_array[$freq_cnt-2] >  $freq_array[$freq_cnt-1] ){ #log trend
             $lower_count++;
   }
      elsif ( $freq_array[$freq_cnt-2] <  $freq_array[$freq_cnt-1]){
                  $upper_count++;
      }
      else {
                  $stationary_count++;
                  }
    }
    else {
      $zero_count++;
      }
    $totalHz1st = $totalHz1st + $objectdata_line[$line_num++];
    $line_num++;                          #ignor strength
  }
  $line_num++;                            #ignor intensity at end of frame
  $frame_count = $objectdata_line[$line_num++];   #Located next frame count
}
#print "\nTotal Hertz for 1st half  = $totalHz1st\n";
$_ = $totalHz1st * $rec_num /($weight_1* 2);
#print "\n Weighted Total for 1st half = $_\n";

#Now Get 2nd half of utterance   ######################

for( ; $line_num  < $rec_num ; ){
  for ($x=0 ; $x<$frame_count ; $x++){
    if ($objectdata_line[$line_num] != 0){     #weight a positive recording
      $weight_2++;
      $freq_array[$freq_cnt++] = $objectdata_line[$line_num];
         if ( $freq_array[$freq_cnt-2] >  $freq_array[$freq_cnt-1]){ #log trend
                  $lower_count++;
                  # print "\n freq_array[$freq_cnt-2] = $freq_array[$freq_cnt-2]\n freq_array
[$freq_cnt-1] = $freq_array[$freq_cnt-1]\n";
      }
    elsif ( $freq_array[$freq_cnt-2] <  $freq_array[$freq_cnt-1]){
                        $upper_count++;
      }
      else {
                        $stationary_count++;
                        }
  }
   else {
      $zero_count++;
    }
    $totalHz2nd = $totalHz2nd + $objectdata_line[$line_num++];
    $line_num++;                          #ignor strength
   # print "\nTotal Hertz = $totalHz\n";
  }
  $line_num++;                              #ignor intensity at end of frame
  $frame_count = $objectdata_line[$line_num++];   #Located next frame count
}
#print "\nTotal Hertz for 2nd Half  = $totalHz2nd\n";
$_ = $totalHz2nd * $rec_num /($weight_2* 2);
#print "\n Weighted Total for 2nd Half = $_\n";
#print "\t $lower_count";
#print "\t $upper_count";
#print "\n Number of non movements = $stationary_count\n";
#print "\t $zero_count";

#   $socket = IO::Socket::INET->new
#   (
#                       PeerPort=> $port,
#       PeerAddr => $server,
#                       Type     => SOCK_DGRAM,
#       Domain   => PF_INET,
#       Proto    => 'udp',
                        #Listen              => 1,
                #Reuse    => 1,
#   ) || die "Bind failed\n";

#print $socket "\n Number of lower movements = $lower_count\n";
#print $socket "\n Number of upper movements = $upper_count\n";
#print $socket "\n Number of non movements = $stationary_count\n";
#print $socket  "\n Number of zeros  = $zero_count\n";

# Call PRRAT Emotion Checker Script *****************

$file_name = shift @ARGV;
system ("./praat check_emotion_RT.praat $file_name $lower_count $zero_count" );
```