EMOTIONS IN THE VOICE: HUMANISING A ROBOTIC VOICE

Tristan Bowles
Department of Theatre, Film and Television
The University of York
York, YO10 5DD, UK
thb500@york.ac.uk

Sandra Pauletto
Department of Theatre, Film and Television
The University of York, UK
York, YO10 5DD, UK
sp148@york.ac.uk

ABSTRACT

The focus of this project is the manipulation of a robotic voice signal for the purpose of adding emotional expression. In particular, the main aim was to design the emotion expressed by a robotic voice by manipulating specific acoustic parameters such as pitch, amplitude and speech rate. The three basic emotions considered were: anger, happiness and sadness.

Knowledge based on the analysis of emotional sentences recorded by actors was used to develop a program in Max/MSP to 'emotionally' manipulate neutral sentences produced by a Text-To-Speech (TTS) synthesiser. A listening test was created to verify the program success in simulating different emotions. We found that test subjects could separate the sad sentences from the others, while the discrimination between angry and happy sentences was not as clear.

1. INTRODUCTION

In many science fiction films we have become accustomed to seeing onscreen robots expressing many human-like emotions. Whether it be the fretful C-3PO (from *Star Wars*), the philosophical Jonny Five (from *Short Circuit)* or the deranged computer HAL (from *2001: A Space Odyssey*), the machines that populate these fictional realities have developed ways of expressing themselves that go beyond their basic programming. The concept of a talking machine is something that is becoming more and more a reality, with the advent of Text-To-Speech systems that allow the user to type text into a computer and have it read back to them by a synthetic voice. The nature of these voices, however, tends to be neutral in tone and often devoid of any signs of emotion and expression.

The intention of this study is to investigate the effect that emotion has on the acoustic signal of the human voice, and apply this knowledge towards replicating three fundamental emotions upon a neutral, synthetic robotic voice. The applications of this research are varied: from the enhancement of speech-based auditory displays, such as Text-To-Speech systems, to speech sound design for creative industries such as film and games.

2. HUMAN EMOTION AND SPEECH

This section introduces fundamental background knowledge on speech production and emotions to facilitate the understating of the project for the reader non-expert in these areas.

2.1 Human speech production

Human speech is produced as air is drawn into the lungs and then pushed through the vocal folds, causing them to vibrate. Depending upon the tension and position of the vocal folds, the varying pressure of air upon the glottis can create a range of different frequencies, which gives 'voice' to the sounds we utter. The sound of our breath can also produce 'unvoiced' sounds, which occur without any vibration of the vocal folds. To illustrate the difference between voiced and unvoiced sounds, Pinker [1, p.164] compares the sounds heard when one makes an unvoiced *sssssss* sound with a *zzzzzzz* sound, which is voiced. Human speech sounds are then further modified as they pass through the mouth.

Humans are also able to change the sound of their voice through intonation. This is achieved through changing the position and tension of the vocal folds, which directly affects the pitch of the voice. As Pinker states, the process of intonation is "what makes natural speech sound different from the speech of robots from old science fiction movies" [1, p.169].

2.2 Definitions of Basic Emotions

Before identifying how emotions affect the human voice it is important to consider what an emotion is. Ekman [2] identifies three important definitions in which commentators have come to understand the term 'basic' emotion. The first definition draws a distinction between 'basic' emotions such as anger, fear, disgust and sadness, believing them to be fundamentally different from one another [ibid., p.45]. The second definition suggests that a 'basic' emotion has evolved from the need to deal with fundamental life tasks, so that they are ways of dealing with situations that are thrown at us during the course of everyday life. Ekman himself believes that "the primary function of emotion is to mobilise the organism to deal quickly with important interpersonal encounters, prepared to do by what types of activity have been adaptive in the past" [ibid., p.46]. A third way of considering a 'basic' emotion is to think of it as the root of some more complex or compounded emotions [ibid., pp.46-7]. Anger, for instance, could give rise to a hot fury or a cold annoyance.

2.3 Emotional factors that affect the human voice signal

The effect of basic emotions on the human voice is something that can alter the sound of the voice. Indeed, according to Scherer [3], "even slight changes in physiological regulation will produce variations in the acoustic pattern of the speech waveform" [ibid., p.240].

O'Shaughnessy [4, pp.204-242] outlines a number of methods that have been developed for analysing the content of human speech, such as studying the time and frequency-domains and methods for estimating/detecting changes in pitch. Most of the methods have been useful for speech coders who have been trying to re-create human speech synthetically, such as in the Text-To-Speech systems, yet they can also provide analysts with a way of seeing key changes in the voice signal due to factors such as emotions. In the time-domain parameters of the voice, for instance, one can quickly interpret the intensity and rate of speech, as well as the level of the voice [ibid, p.211]. The frequency-domain, on the other hand, can give the speech analyst an idea of the frequencies contained within the voice. In a study looking into identifying personality markers in speech, Scherer [5, p.153] identifies the changes to the fundamental frequency of the voice (in terms of its pitch), the intensity of the voice (how loud certain words are spoken) and the energy distribution within the voice spectrum, which directly affects the quality of the spoken voice, as the main elements that are changed by an emotion experienced by a speaker. Added to this is the further parameter of the rate of speech. This criterion examines the speed in which words are spoken in natural speech, which may include natural pauses, silent periods, and moments where the speech flow is disrupted [ibid., p.160]. So by examining the fundamental frequency, pitch, intensity and speed of the voice, one can begin to identify the changes in the voice that occur when the speaker experiences an emotion.

Scherer [3, p.239] further alludes to studies in which key characteristics of the human voice signal have been digitally changed in an attempt to change the perception of how the voice sounds, and to influence the listener into thinking that the voice has been modified by certain emotions. The main variables of the voice signal that are adjusted in these studies are the fundamental frequency range (F0), the pitch variations or contours, the intensity, duration and accent structure of real utterances [ibid.]. Out of all these variables, Scherer reports that the F0 range had the biggest effect on the listeners, with a narrow F0 suggesting an emotional state such as sadness and a wide F0 "judged as expressing high arousal, producing attributions of strong negative emotions such as annoyance or anger, or for the presence of strongly developed speaker attitudes such as involvement, reproach, or empathic stress" [ibid.]. Similarly, high intensity in the voice signal was perceived negatively, associated with aggressive attitudes, whereas short voiced segments, uttered with a fast speech rate, were interpreted as being joyous in their nature, as opposed to slow speech rate segments with a long duration which were perceived to be akin to emotional sadness [ibid.].

A further factor that may be taken into account when investigating emotion changes in the human voice is to consider the level of emotional arousal that the voice is influenced by. In a recent study into the changes in intonation when emotions are expressed through speech, Bänzinger and Scherer [6, p.257] highlight a distinction between two levels of emotional arousal. High arousal emotions, such as hot anger, panic fear, despaired sadness and elated joy, often are associated with a raised voice, fast speech rate, and higher pitch, when compared to low arousal emotions, such as cold anger, depressed sadness, calm joy/happiness and anxious fear [ibid.]. This division of emotional states into hot and cold elements supports the view held by Ekman [2] that each emotion is not a single emotional state, but belongs to "a family of related states" [ibid., p.55], providing a variety of emotional reactions depending upon the situation in which the emotion is experienced.

2.4 Related studies and novelty of this project

One previous attempt to synthesise speech with emotion is accounted by Murray and Arnott [7], who built a speech synthesiser that modeled six primary emotions: anger, happiness, sadness, fear, disgust and grief. The speech synthesiser attempted to replicate each emotion in four stages. First of all they set some neutral pitch and duration rules that would act as the basis for the synthetic speech before any emotional effects were applied. This acted as a basis for which they developed a second set of rules which attempted to personalise the voice, to give it a certain voice quality, so that, for example, "a breathy voice would remain breathy with emotional effects added later" [ibid., p.371]. Each of the emotion dependent rules were then applied, implementing changes specific to the chosen emotion that the synthesiser was trying to replicate. This included changes such as increasing the pitch and duration of stressed vowels, altering the pitch direction of inflections on word endings, adding pauses after longer words, and eliminating abrupt changes in pitch and between phonemes [ibid., pp.376-7]. The final stage was to send the resulting phonemes and their associated pitch and duration values to the synthesiser, in order to create the speech sound. Murray and Arnott first derive the emotional rules by analysing the emotional speech performed by actors and then verify the success of the rules using a listening test in which subjects are asked to discriminate the emotion portrayed by the synthesised sentences.

Murray and Arnott study [7] is based on sound synthesis or *copy synthesis* [8], i.e. it is a study in which acoustic features are copied from real emotion portrayals and used to resynthesise new emotion portrayals. This type of study is relatively uncommon. Juslin and Laukka made a very comprehensive review of 104 vocal expression studies [8], searching all papers from 1900, and found that only 20% of the studies were based on *copy synthesis*. With this research we hope to add useful knowledge to a research field still relatively unexplored.

An additional novelty of this study is represented by the fact that, instead of synthesising the emotional speech, we process already synthesised speech samples and "apply" the emotions as an "effect" on the speech sample (rather like a reverb effect is applied onto a dry audio sample). The results of this study should be useful in particular for sound designers with the task of processing speech to make it sound emotional for creative applications (e.g. films and games). It is also important to note that we focus on a non-natural voice, i.e. a robotic sounding voice. For this reason the "naturalness" of the processed voice is not a concern in this study.

3. DESIGNING THE EMOTION CONTENT OF A SPEECH SIGNAL

For this project eight test phrases were chosen to be the basis for emotional analysis and the construction of the program's presets. The phrases were specifically chosen so that they were void of emotional content, i.e. they would not indicate or suggest, from their wording, the emotion being portrayed by the speaker. The 8 phrases consisted of one command statement, "put the butter on the table" (indicated in this paper as phrase 1), two descriptive statements, "the window is wide open" (indicated as phrase 2) and "the orange is round" (indicated as phrase 8), one question, "what time is it?" (indicated as phrase 3), one long statement, "one plus one equals two, two plus two equals four" (indicated as phrase 4), and three phrases containing only monosyllabic or short words: one counting up from 1 to 5 (indicated as phrase 5), one counting down from 5 to 1 (indicated as phrase 6) and one containing 5 numbers in no particular order "one, seven, nine, two, three" (indicated as phrase 7).

3.1 Voice actors

In order to investigate the changes that emotions have on the human voice, four male vocal actors were selected to perform the 8 test phrases, simulating various emotional states. The recording set-up used (equipment, recording studio, and distance between source and microphone) was the same for each performance. The actors were first asked to read out each phrase into an AKG 414 microphone in their "normal" reading voice, thus obtaining a neutral basis through which to compare the proceeding emotional states. The actors were then asked to read out each phrase again in a mildly angry voice, under the direction that they were growing increasingly annoyed at having to read out the phrases. Following this, the actors were asked to read out the phrases in an angry voice once more, but with greater intensity, in order to simulate an explosive anger. These two steps were repeated with the emotions of happiness and sadness, in order to obtain both mild and intense examples of each. It was decided that the intense emotional recordings would act as a basis for the emotional presets in the program to be built for the processing of the robotic voice, mainly because they were the most likely to make a strong impression on the listener.

Once the recordings were made, they were analysed to extract information regarding:

- the exact timing of each word and the length of any pauses or silent periods that appeared within a phrase;
- the fundamental frequency variation and the number of pitch variations or contours per phrase, and whether each pitch contour was upward or downward directed (this was obtained through the pitch analysis pane of the WaveSurfer software [9]);
- the amplitude variation within the phrase.

3.2 Analysis of the actors' voices

In order to analyse the changes that the emotions have on the actors' voices, mean averages of pitch, speech rate and amplitude parameters were taken for each phrase, based upon the data collected from the four actors' speech samples. In this way, the average duration of speech, pitch range, number of pitch contours and amplitude ranges for each of the three emotions were calculated for each phrase. The actors' mean average "neutral" voice was then used as a basis to map the changes to the duration, pitch and amplitude of each phrase.

The test phrases produced a number of common trends that related to type of phrase, the emotion spoken and its effect on the voice.

3.2.1 Phrase duration

Figure 1 shows the average duration of each phrase. Higher bars indicate a longer duration and, when comparing the four emotional versions of the same phrase, a slower speech rate, whereas shorter bars indicate a shorter duration and, when considering different versions of the same phrase, a faster speech rate.

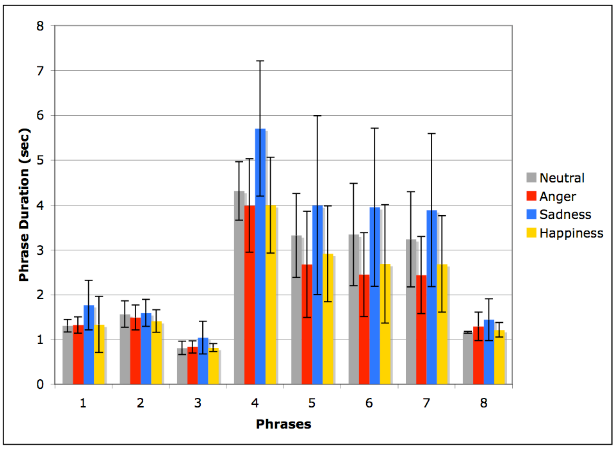

Figure 1: Phrase Duration (averages and standard deviations of the four actors' performances)

We looked at how each emotional phrase deviates from its neutral version and we noted that the phrases involving only monosyllabic or short words (phrases 5, 6 and 7) saw the greatest reduction in duration for the angry and happy phrases (-20% and below the duration of the neutral phrase). In the cases of phrases 1, 3 and 8, the average angry and happy phrases were slower than their equivalent neutral voices. Half of the sad phrases saw over a 20% increase in duration, whereas the short words phrases (phrases 5, 6 and 7) saw an increase in duration between 10-20%. The average length of pauses per phrase was also measured. The sad phrases saw the longest overall pauses with 5 out of the 8 sad phrases having the longest phrase duration. Half of the happy phrases contained pauses that were longer than the angry equivalent, whereas only angry phrases 2 and 5 contained pauses that were longer than their happy equivalent.

3.2.2 Pitch Analysis

In order to investigate the emotional changes in pitch, the maximum peak in fundamental frequency (F0), the number of pitch contours (or pitch variations) per phrase and the direction of pitch contours were examined.

Figure 2 shows the average maximum peak of F0 for each phrase.

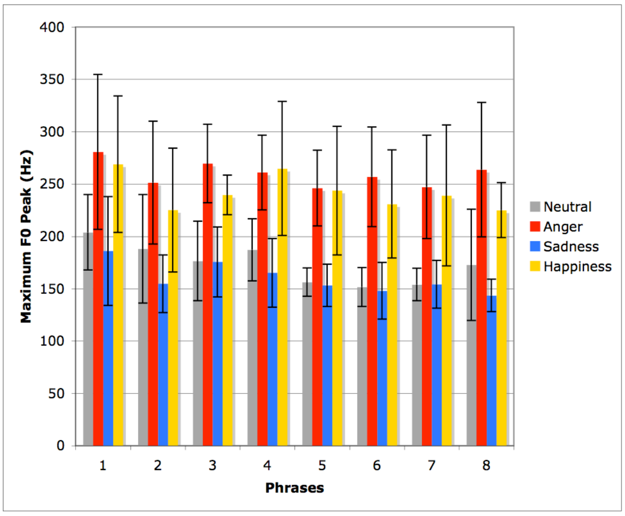

Figure 2: Maximum F0 Peak (averages and standard deviations of the four actors' performances)

For all phrases, anger and happiness have high F0 peaks, while sadness has low F0 peaks.

With the exception of phrase 4, anger phrases have the highest peak fundamental frequency of the 3 emotions. The average maximum peak frequency range of the angry phrases sits between 246Hz-281Hz, with 6 out of the 8 phrases averaging above 250Hz. The happy phrases have a range of 225Hz-269Hz, with 2 of the 8 phrases averaging above 250Hz. The sad phrases have lowest fundamental frequency peaks, operating within a range of 143Hz-186Hz.

Overall, the variation in the number of pitch contours was dependent on the type of phrase. Some of the angry phrases saw the greatest increase in the number of pitch contours, while the happy phrases showed greater variation between increases and decreases, from phrase to phrase. The sad phrases generally saw a decrease in the number of pitch contours, with two exceptions.

The average direction of pitch contours per phrase was calculated by counting every upward curve as a positive value (+1) and every downward directed contour as a negative value (-1). The result for each phrase was totaled and an average obtained. The majority of the neutral phrases contained downward directed pitch contours. The majority of the angry phrases contained more downward directed pitch contours, whereas the happy phrases varied between having upward directed or downward directed pitch contours. The majority of the sad phrases contained downward directed contours.

3.2.3 Amplitude analysis

Figure 3 shows the average maximum amplitude peak for each of the 8 phrases based upon the actors' performances.

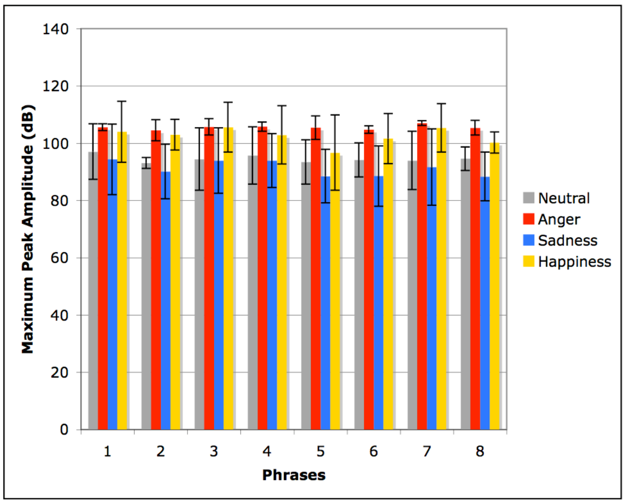

Figure 3: Maximum Amplitude Peak (averages and standard deviations of the four actors' performances)

We can see that for all the phrases, anger and happiness have the high peaks, while sadness have low peaks.

The sad phrases had the lowest maximum amplitude peaks, with all of the sad phrases peaking below 95dB. All the angry phrases and 7 out of 8 happy phrases had peaks that exceeded 100dB.

3.3 The "robotic" text-to-speech voice

In addition to the actors' recordings, the 8 phrases were also recorded by the Text-To-Speech voice "Fred" set to the "Normal" rate of speech, which can be found within the system settings for speech in a Mac computer with the OS X 10.5.8 operating system. This robotic-sounding voice was to act as the primary voice for the program and was to be manipulated to sound angry, sad or happy.

3.4 The Max/MSP program

The emotion-shaping program that was constructed used a real-time granular synthesizer built by Mattijs Kneppers for use in Max MSP 5. Kneppers' [10] patch was chosen because of the way it could smoothly alter the length and pitch of a sound.

Only a few changes were made to Kneppers' original patch in order to make it more suitable for modeling the speech rate, pitch and volume of a speech file over the course of time. Three function graphs were introduced: one controlling the speech rate, one controlling the pitch and one controlling the volume of sound files loaded into the buffer. The ranges for the pitch shift, the time stretch and the volume changes were also adapted to the needs of this project.

3.5 Constructing the Presets

A decision was made to have three presets for each phrase per emotion, so that the total number of presets for each phrase would be nine (3 x angry, 3 x sad and 3 x happy).

The first preset for each phrase was constructed using information collected from the analysis of the actors' readings of the phrases. This was achieved through plotting a graph of the average speech rate for each phrase, and transposing this onto the speech rate graph in the main interface of the Max/MSP program. Average pitch and amplitude graphs of the actors' phrases were also used as a rough guide for the pitch and volume function graphs.

The other two presets were created by tweaking the parameters of the first preset in an attempt to improve the emotional content. This would involve either increasing or decreasing each of the three elements (speech rate/pitch/volume) in order to exaggerate the phrase and make it express a strong emotion. Another method was to listen to specific actors' performances of the phrases and try and match it through the automation of points within the function graphs. In the case of the angry phrases, for example, it was helpful to listen out for the words within the phrase that were exaggerated or stressed the most. The presets that were perceived to be the most successful were then selected for testing purposes.

3.6 Listening test

A listening test was designed to measure the effectiveness of the presets created. The test included the neutral phrases, which were the basis for the emotional presets. This allowed us to verify whether they were really perceived as devoid of emotions.

For the test we adopted a similar approach to Bänzinger and Scherer [6 p.259] who, when testing the changes of intonation in emotional expression, provided their listeners with four visual-analogue rating scales, each representing the "intensity" (ranging from no emotion to extreme emotion) of the four emotions that were the focus of the study.

Similarly we created a test using the 8 neutral TTS phrases and the 24 presets made using the Max/MSP patch (8 x angry, 8 x sad, 8 x happy).

In the test, each participant first entered details pertaining to their date of birth, sex and whether English was their first language, then they were required to listen to each phrase and answer the following two questions:

(i.) *What is the emotional state of the voice?* …to see if their answer matched the intended emotion conveyed through phrase.

(ii.) *How intense is the emotion?* …to gauge how strongly they thought the emotion perceived was expressed.

In order to answer question (i.), the participant had to select one of the four following radio buttons: Angry, Sad, Happy, Neutral. For question (ii.) the participant had to position a horizontal slider between two points one representing weak intensity and the other strong intensity, i.e. the slider ranged from 1 (weak) to 5 (strong). Each participant was allowed to listen to each phrase as may times as they desired, but they had to answer both questions before being allowed to move on to the next phrase. The phrases were played to each participant in a random order so that they were not able to predict what emotion might come next. The test was constructed using the java based audio-visual test-builder *Skatta* [11].

4. LISTENING TEST RESULTS

20 people took part in the test, with an even split between male and female participants. 14 of the test participants were British and had learnt English as their first language, while for the other 6, who varied in ethnicity, English was not their first language.

4.1 Chi-Square Tests Results

The results of the first question of the test were analysed using the non-parametric Chi-Square test. This test can tell us if subjects attributed a particular emotion to a phrase in a random fashion (producing a non-significant result for the Chi-Square, i.e. p > 0.05) or not (producing a significant result for the Chi-Square, i.e. p < 0.05). Then we looked at which emotion had the most counts and therefore made the bigger contribution towards the significance of the Chi-Square. In the significant cases, we looked at the average score of the intensity of the emotion to gauge how strongly the emotion was portrayed. This is summarised in Table1, where A=Anger, H=Happiness, N=Neutral, S=Sadness, E=significant Emotion and I= Intensity of significant emotion.

Overall: no phrase distinction	**Chi-Square**	**Significance (p)**	**Significant Emotion**	
Anger	72.75	0.000	Happiness	
Happiness	34.35	0.000	Happiness	
Neutral	79.85	0.000	Neutral	
Sadness	222.65	0.000	Sadness	
Phrase by phrase	**Chi-Square**	**Significance (p)**	**E**	**I**
Emotion: Anger				
1-Put the butter…	7.6	0.055		
2-The window…	22.8	0.000	H	3.36
3-What time…	7.9	0.019	H	2.73
4-One plus one…	12.4	0.002	H	3
5-One, two, …	2.8	0.247		
6-Five, four, …	4.0	0.261		
7-One, seven, nine	9.2	0.027	H	2.91
8-The orange …	14.0	0.003	H	2.25
Emotion: Happiness				
1-Put the butter…	4.8	0.187		
2-The window…	6.7	0.035	H	3.17
3-What time…	2.8	0.423		
4-One plus one…	14.8	0.002	H	2.67
5-One, two, …	1.2	0.753		
6-Five, four, …	0.7	0.705		
7-One, seven, nine	2.0	0.572		
8-The orange …	16.3	0.000	H	2.87
Emotion: Neutral				
1-Put the butter…	11.2	0.011	A	3
2-The window…	7.6	0.055		

3-What time…	1.3	0.522		
4-One plus one…	9.2	0.027	N	3.1
5-One, two, …	22.8	0.000	N	3.14
6-Five, four, …	7.2	0.007	N	3.19
7-One, seven, nine	3.7	0.157		
8-The orange …	21.6	0.000	N	2.43
Emotion: Sadness				
1-Put the butter…	11.2	0.011	S	3.09
2-The window…	5.0	0.025	S	2.2
3-What time…	28.9	0.000	S	3.24
4-One plus one…	5.0	0.025	S	3.21
5-One, two, …	7.2	0.007	S	3.63
6-Five, four, …	12.4	0.002	S	3.07
7-One, seven, nine	24.1	0.000	S	4.06
8-The orange …	26.8	0.000	S	3.2

Table 1: Chi-Square results

From these results, we can see that the sad emotion is consistently the best recognised across all phrases. 4 out of the 8 neutral phrases have been correctly recognised, whereas 1 is wrongly recognised as angry. 3 out of 8 happy phrases have been correctly recognised. The angry phrase type failed to be recognised correctly, attaining results of non-significance or significantly happy (in 5 cases). In terms of successful matches per phrase, phrases 4 and 8 have the most number of correctly identified emotions, each sharing an overall success rate of 75%. Phrases 1, 3 and 7 generated the least matches, with only 1 out of the 4 emotions identified correctly.

4.2 Intensity Results

Seven out of eight sad phrases score an intensity higher than 3. The highest intensity score for sadness is 4.06 for phrase 7. Phrase 5 follows with 3.63, then phrases 3, 4, 8, 1, 6, and 2. The highest intensity score for the significantly happy phrases is phrase 2 with a score of 3.17. Phrases 4 and 8 score 2.67 and 2.87 respectively.

Five out of eight angry phrases (2,3,4,7,8) have been incorrectly recognised as happy. Phrases 2 and 4 have scores of 3.63 and 3 respectively, while the rest of the sentences have scores between 2 and 3. It is interesting to note that for these phrases, which were manipulated to portray anger but were recognised as happy, the next emotion that has high scores of intensity is anger indicating that subjects seemed confused between these two emotions.

Four phrases (4,5,6,8) in the neutral case score as significantly neutral, with 3 of these (4,5,6) having intensity scores higher than 3. The highest intensity score is 3.19 for phrase 6. Phrase 1 results as significantly angry with an intensity score of 3. However, as with the confusion between idenification of anger and happy outlined above, the second most chosen emotion is the intended emotion, neutral, with an intensity score of 3.

4.3 Overall confusion matrix

Table 2 shows the confusion matrix for the overall experiment. The numbers represent the percentage of subjects that selected a particular perceived emotion for all the phrases with a specific intended emotion. The Chi-Square is significant for all the emotions.

Intended emotion → Perceived emotion ↓	**Anger**	**Happiness**	**Neutral**	**Sadness**
Anger	30	11	26	10
Happiness	**49**	**43**	14	2
Neutral	17	25	**53**	13
Sadness	4	21	7	**75**

Table 2: Confusion matrix without distinction of phrase

This overall result shows that sad, neutral and happy phrases are relatively well recognised, while angry phrases are confused as being happy.

5. DISCUSSION

Based upon the results of this study, it is possible to assert that specific changes to the speech rate, pitch and amplitude of the human voice do affect the way the voice is perceived emotionally. However, not all manipulations have been successful in this project. The most evident unsuccessful result is the manipulation to obtain angry phrases. In some cases these have been mistaken for happy.

Less than half of the happy phrases were recognised as happy showing that the approach used is possibly going in the right direction, but requires further refinement.

From the analysis of the actors' voices we can see that anger and happiness have similar trends in terms of speech rate, pitch and amplitude variation. Although anger usually exceeded happiness for the majority of these parameters, there were a few phrases where happiness exceeded anger. This seems to indicate a close relationship between the emotions of anger and happiness that makes it a lot more difficult to distinguish between the two when attempting to simulate them onto a robotic-sounding voice. One possible reason why the angry and happy phrases obtain poor results could be that further stimuli, such as a facial expression or an emotionally-loaded phrase, is needed for a participant to be able to distinguish between an angry voice and a happy voice. Alternatively, maybe further alterations to the voice, that go beyond simply adjusting the speech rate, pitch and amplitude, need to occur before anger becomes more distinct from happiness.

The successful recognition rate for the sad phrases confirms that if the pitch, duration and amplitude of the robotic neutral voice all decrease, then the voice will sound sad.

Four out of eight neutral phrases were recognised as neutral. For three neutral phrases no clear emotional state could be assigned therefore the "neutralness" of these phrases is unclear in this test. It is possible that with a larger number of subjects the perceived emotional state of these sentences could emerge more clearly.

In the case of neutral phrase 1, the participants rated the emotion as being angry. A possible reason for this may be the words within the phrase. Whilst the phrase itself doesn't contain any emotion-specific words, the nature of the sentence is a command. Participants may have thought this sounded forceful and angry in this instance. Murray and Arnott [7] allude to a similar occurrence when testing the phrases from their HAMLET system, where the words within the phrase itself caused participants to imagine "a stereotypical situation where the phrase might be used" [ibid., p.387].

Overall, this project is successful in giving more insight into methods that could be employed to design emotional expression in a robotic voice. There are, however, a number of improvements that can be made to the work.

Difficulties were encountered when constructing the presets. This related to how the pitch and amplitude data, based on the actors' voices, could be used accurately in the program. As the pitch and volume of the actors' voices vary significantly over the course of a phrase, it was particularly difficult to judge on which words to make specific pitch and amplitude adjustments. Indeed, one might benefit from trying to model the pitch and amplitude of a robotic voice based on one actor's performance, rather than an average of several.

Currently, the function graphs in the user interface of the Max/MSP patch used to manipulate the phrases can be used to sketch out the changes in speech rate/pitch/amplitude over the course of a phrase with relative ease, but at the cost of accuracy.

Furthermore, the presets used for testing could be improved by gaining feedback from a panel of experts (e.g. theatre directors, actors, etc.).

A further, larger study that includes the improvements mentioned above has been planned and is under development at present. We plan to report the results of this new experiment in the near future.

6. CONCLUSION

A program was constructed during this study that "applies" three emotional states onto a robotic-sounding Text-To-Speech voice. This study has examined how changes in speech rate, pitch and amplitude affect the human voice and has applied this knowledge towards creating a number of emotional presets that allow the user to make a robotic speech signal sound angry, sad or happy. The resulting presets were tested on a wider audience and the results from this test indicated that decreases in speech rate, pitch and amplitude create the impression of sadness. The test also showed that the angry and happiness presets were very closely related, indicating that speech rate, pitch and amplitude adjustments alone, or without a greater degree of accuracy, are not enough to make anger distinct from happiness. From these results, a number of enhancements and further work was suggested, which should help to clarify how we can improve emotion expression in a robotic voice.

7. REFERENCES

[1] S. Pinker: *The Language Instinct*, St Ives: Penguin Books, pp.158-191, 1995.

[2] P. Ekman: "Basic Emotions", pp. 45-60, in T. Dalgleish and M. Power: *Handbook of Cognition and Emotion*, Chichester: John Wiley & Sons, 1999.

[3] K. R. Scherer: "Expression of emotion in voice and music", *Journal of Voice* 9(3): 235-248, 1995.

[4] D. O'Shaughnessy: *Speech Communication: Human and Machine*, United States: Addison-Wesley Publishing Company, pp. 204-242, 1987.

[5] K. Scherer: "Personality markers in speech", in K. Scherer and H. Giles (eds.): *Social Markers in Speech* London: Cambridge University Press, pp. 147-210, 1979.

[6] T. Bänzinger and K. Scherer: "The role of intonation in emotional expressions", *Speech Communication*, vol 46, pp.252-267, 2005.

[7] I. Murray and J. Arnott: "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", *Speech Communication*, vol.16, issue 4, pp.369-390, 1995.

[8] P.N. Juslin and P. Laukka: "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?", *Psychological Bulletin*, Vol. 129, No. 5, 770 – 814, 2003.

[9] Wavesurfer Software: available from <http://www.speech.kth.se/wavesurfer/> [Last Accessed 14 June 2010].

[10] M. Kneppers: "Real-time, natural sounding granular time stretcher/ pitch shifter", available from <http://www.cycling74.com/share.html> [Last Accessed 14 June 2010].

[11] Skatta Software: available from <http://sourceforge.net/projects/skatta/> [Last Accessed 14 June 2010]