Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Iain R. Murray & John L. Arnott, *Computer Speech and Language*, Vol.22, No.2, 2008, pp.107-129. DOI: 10.1016/j.csl.2007.06.001

Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech

IAIN R. MURRAY * and JOHN L. ARNOTT

School of Computing, University of Dundee, Dundee DD1 4HN, U.K.

This is a pre-print report of research published in:

Computer Speech and Language, Vol.22, No.2, 2008, pp.107-129.

Computer Speech and Language is available online at:

http://www.elsevier.com/wps/find/journaldescription.cws_home/622808/description

ISSN: 0885-2308

URL: The DOI bookmark of the article is: http://dx.doi.org/10.1016/j.csl.2007.06.001

DOI: 10.1016/j.csl.2007.06.001

Abstract: All speech produced by humans includes information about the speaker, including conveying the emotional state of the speaker. It is thus desirable to include vocal affect in any synthetic speech where improving the naturalness of the speech produced is important. However, the speech factors which convey affect are poorly understood, and their implementation in synthetic speech systems is not yet commonplace. A prototype system for the production of emotional synthetic speech using a commercial formant synthesiser was developed based on vocal emotion descriptions given in the literature. This paper describes work to improve and augment this system, based on a detailed investigation of emotive material spoken by two actors (one amateur, one professional). The results of this analysis are summarised, and were used to enhance the existing emotion rules used in the speech synthesis system. The enhanced system was evaluated by naive listeners in a perception experiment, and the simulated emotions were found to be more realistic than in the original version of the system.

Keywords: Speech analysis; Speech perception; Emotion; Affect; Synthesis-by-rule; System evaluation.

*** Corresponding author:** Dr. Iain R. Murray, School of Computing,
University of Dundee, DUNDEE DD1 4HN, Scotland, UK.
e-mail: irmurray@computing.dundee.ac.uk
Phone: +44 1382 384155
Fax: +44 1382 385509

Preprint of: *I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129*

1 - Background And Previous Analysis Work

Many commercial speech synthesis systems are available, with varying degrees of intelligibility and naturalness, and most offer a full text-to-speech capability. Most such systems have excellent natural-sounding intonation and many also permit alteration of the "vocal identity" (i.e. the identity of the speaker, as subjectively perceived in the voice by the listener) being used - either numerically in the case of formant-based systems, or using different libraries for segment selection in the case of concatenative systems. However, no current TTS products permit naturalistic variation of emotion, partly due to the increased complexity of adding an emotion module into the text-to-speech conversion process, but mostly due to our very limited knowledge of how emotion is conveyed in the human voice.

Little early voice analysis work has been concerned specifically with emotion (Davitz (1964)), and such research has remained largely separate from the main body of speech analysis literature. However, it is apparent from the literature as reviewed by, for example, Murray and Arnott (1993) and Schröder (2001) that emotion is conveyed vocally by three components of the speech signal, namely voice quality, pitch contour (intonation) and timing. Based on this knowledge, some prototype synthetic speech-with-emotion systems have been produced, such as Murray (1989), Cahn (1990), Morton (1992), Murray and Arnott (1995) and Schröder and Trouvain (2001).

There is an increasing interest in human vocal emotion, and the closely associated topic of speaking styles (described by Eskénazi (1993)) which also affect timing pitch and voice quality, in order to increase our knowledge of emotion expression in humans and to codify it (e.g. Leinonen at al. (1997), Roach (2000)), to improve speech synthesis (e.g. current authors, Montero et al. (1999)), or to enhance human-computer interaction through affective computing (e.g. Picard (2000)).

The objective of the work described here was to conduct a detailed analysis of a small number of vocal emotions, and obtain a detailed description of those parameters of speech which are affected by the speaker's emotional state, then use these results to enhance the simulation of these emotions by a synthetic speech system – the HAMLET prototype (Murray and Arnott (1995)).

Preprint of: ***I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129***

2 - Human Voice Analysis Experiment

2.1 - Selection of Voice Sources for the Recordings

For the purposes of experimental observation and recording, real emotions can be elicited by using an artificial situation and appropriate stimulation of the subject (e.g. Ax (1953), Schachter and Singer (1962)), although imposing emotions upon subjects without their prior knowledge and consent imposes ethical difficulties; in addition, the spoken script is not controllable at all (except through guiding by accomplices interacting with the subject), and recording conditions cannot be tightly controlled. Elicitation of real emotions while avoiding ethical problems (by warning subjects of impending stimuli prior to initiating them) while also retaining control over recording conditions was achieved by Tolkmitt and Scherer (1986) by having subjects look at slides intended to stimulate varying stress levels; standard response phrases were recorded after each stimulus slide. A similar approach was taken by Raymond (1993) to elicit a range of emotions, subjects being asked to describe their feelings about the stimulus photographs. Johnstone and Scherer (1999) recorded subjects (using various methods in addition to audio recording) while playing a competitive video game, using this situation to elicit a range of emotions. A summary of experiments which have elicited emotional speech datasets is given by Douglas-Cowie et al. (2003), and guidelines for development of future databases of emotional speech are presented.

For the purposes of this study, it was felt that the most appropriate approach was to use actors to produce utterances for analysis, even though the emotions themselves remain artificial. The policy of using recordings of actors for emotional voice analysis was found to be representative of real emotions by Williams and Stevens (1972), who compared a recording of a dramatic situation (the radio announcer at the crash of the Hindenburg airship) with an actor acting out the same situation; it was found that the actor altered his voice in the same way as the person experiencing the real situation, although to a somewhat lesser degree. Banse and Scherer (1996) also found acted speech representative of genuine emotions in a detailed study. The loss of realism in the emotional expression is thus largely offset by the benefits of both being able to script the dialogue and to closely control the conditions of the recording.

2.2 - Selection of Phrases Used for the Recordings

The recorded utterances were chosen to be semantically empty phrases which can be equally valid when spoken in any of the emotions to be analysed. They are also referred to as "emotionless" or "emotionally empty" phrases as they can legitimately take on different emotions depending on the context. The following two phrases were used:

"This is not what I expected."

"You have asked me that question so many times."

The latter phrase is a modified form of the phrase used by Fairbanks and Pronovost (1939), also used by Dawes and Kramer (1966). These two emotionally neutral sentences were embedded in a series of text paragraphs which were deliberately intended to convey a particular emotion. The speakers were presented with a description of a situation and asked to speak the paragraph with the appropriate emotion for the speaker in that situation. The paragraphs and their contexts are given in full in Appendix 1.

In addition to the emotional paragraphs, the target phrases and accompanying text were recorded with neutral (i.e. unemotional) speech to enable further comparisons to be made.

2.3 - Recording Conditions

Emotional speech recordings were made in a recording studio, using a professional-quality microphone and digital audio tape recording system. The recording procedure used was similar to that employed by Vroomen, Collier and Mozziconacci (1993). Two sessions were recorded using the same material, each session employing a different speaker.

2.4 - Analysis Conditions

The speech analysis was performed using a PC-based speech workstation. The recordings were sampled at 20kHz using an input filter combination of flat + 6dB/octave. The utterance to be analysed was then isolated from the remainder of the recorded text by editing the sample. The isolated utterances thus produced were then processed to produce oscillograms, spectrograms, pitch contours and intensity contours as a graphical output for detailed manual analysis. In conducting this impressionistic analysis,

general changes in the voice (such as pitch level and range) were of interest, but particular vocal features which differed between the emotion recordings and thus might be used to convey emotion were also noted, specifically with a view to later simulation of these features in synthetic speech. These analyses were thus partly subjective, and some differences noted below (such as articulation precision) are from subjective observation only. Statistical analyses of the digital audio data were not undertaken, as such statistics would have been of no value in the synthesis stage of the project, due to the type of synthesiser being used.

"Basic emotion" theory suggests that all emotions can be made by combining a series of "basic" emotions, the number of basic emotions varying from two to eighteen, with a set of five (anger, happiness, sadness, fear and disgust) being the set most commonly adopted (see Ortony and Turner (1990) for a discussion of "basic" emotion theories and alternatives). Fónagy (1981) suggests that the "basic" emotions are primarily identifiable by differences in the prosodic content of speech, other emotions being distinguished by voice quality differences; it is likely that the situation is not as clear cut as this, Gobl and Ní Chasaide (2003) noting that voice quality alone can differentiate a range of emotions and that individual quality parameters are important discriminators for different subsets of emotions. However, as the present analysis was concerned with basic emotions, the prosodic features in particular were analysed. Prosody consists of three main elements: amplitude structure (including stress and prominence), temporal structure (pause, rhythm, and segment duration) and pitch structure (accent and intonation). All three characteristics are known to convey the expression of emotional characteristics for the five basic emotions.

2.5 - Performers

Two different actors were employed for the two recording sessions, one a male professional stage actor aged 25 (P), one an amateur actor aged 50 (A) who had performed on stage but had not had any formal voice training. Throughout all recordings, P spoke with very clear articulation (i.e. with precise and distinct pronunciation), and the overall speech rate is lower for all the utterances compared to A. However, the relative time characteristics for the five basic emotions were very similar to those obtained for A.

P's voice contained a large number of stressed accent words. The accent (sentential stress) for all emotions was much greater than for A, which is probably a result of P's dramatic training; P tried to convey the emotional affect mainly by his subjective decision to stress content words with an emphasis on their semantic meaning. P's voice quality did not change very significantly between the different emotions; this contrasted with A, whose voice quality changed noticeably and, even when taken in isolation, appeared to convey a significant part of the emotional affect. Also connected with P's individual way of speaking were the generally strong vowels with a high number of formants clearly observed in the spectrograms.

The analysis of the emotionally neutral sentences for P showed that the pitch contour did not vary within such a wide dynamic range as it did in the recordings of the same utterances from A; the pitch range was narrower for all P recordings in general. This is surprising because the pitch contour is important prosodically, providing essential information to emphasise dramatic effects.

2.6 - Voice Analysis Results

The results from the analysis of the two actors are summarised in Table 1, with further detailed comments below; comparisons made are relative to neutral speech for that actor. Terminology is from Roach (1992).

2.6.1 - Neutral

For both actors, the utterances spoken with neutral emotion can be are clearly articulated speech and show some pausing between words.

2.6.2 - Anger

A: Intensity changes corresponded to the stressed content words; the voice was breathy, with precise articulation. The spectrograms showed more high frequency energy over all the utterance, and higher first formant frequencies. The highest pitch values were placed on the first content word and, for the longer sentence, on the penultimate content word. The highest part of the pitch contour was found to correspond to the highest intensity. The end of the utterances showed strong downward inflections, and there were also sharp downward inflections at the phoneme level with few upward peaks on the stressed syllables of the content words.

P: The general level of the pitch contour was high, with a downward inflection towards the end of the sentence (as with A). High intensity was noted throughout the utterance, and on the spectrograms the distribution of the energy is in the high frequency area. Pronunciation was very clear and many distinct formant trajectories could be distinguished.

2.6.3 - Happiness

A: An increase in articulation precision was noted for content words, and the voice generally sounded breathy. The general form of the pitch contour was with a second rising part towards the end of the utterance, with a terminal fall. It was noted that the line of the pitch contour was not smooth; it had sharp small oscillations at the primary stressed syllables and local downward pitch changes which seem to be rhythmic (stressed phonemes occurring at regular intervals).

P: Happiness appeared to be a difficult emotion for P to perform, and several recordings were made. However, the results of the analysis are consistent with previous findings for happiness. The pitch was raised slightly towards the end of the sentence and remaining in a straight line; this final lifting is typical for a happy emotion. Again, many distinct formants could be distinguished, although not as many as for anger, and the formants were wider. Analysis of the emotionally loaded sentences showed that some phonemes were exaggerated, very prolonged and extended. Pauses were introduced between some words, which is more typical for a sad emotion.

2.6.4 - Sadness

A: A's sad voice exhibited an overall decrease in articulation precision. Small downward inflections at phoneme level were noted, and there were regular pauses.

P: Small downward inflections were noted at word and phoneme level. A low intensity contour was noted for all the utterances, with intensity decreasing towards the ends.

2.6.5 - Fear

A: A's spectrograms showed relatively little energy in the lower frequencies. Content words exhibited upward inflections at the beginning, and downward inflections at the end of each segment; the upward

parts reached nearly the same peak values, and the downward inflections were well pronounced. At the phoneme level, the pitch contour showed some vibrato and irregular downward inflections. The speech was interrupted by pauses between words, which formed about one quarter of the speaking time.

P: For the emotionally empty phrases, the intensity was quite loud, in the same range as for happiness and anger. Wider formant regions were noticed on some occasions, and for the emotionally loaded sentences, a high number of formant trajectories was noted. For the emotionally loaded phrases, the intensity contours were not consistent, varying between the three emotionally loaded sentences.

2.6.6 - Disgust

A: Increased articulation precision at stressed content words was noted. The pitch contour showed downward inflections at the phrase endings, and also downward pitch inflections at word endings. A rise in pitch was noted at the beginning of stressed content words. The speech rate was low, with a large number of introduced pauses, increased phonation time, and lengthening of the stressed syllables in stressed content words.

P: Intensity was quite loud, which is not a typical characteristic for disgust (Murray and Arnott (1993)), though it decreased towards the end of the utterance. For all of the disgust utterances, an increase in articulation precision was noted. Again there was an emphasis on the pitch contour changes, with much accenting and use of high intensity. Large dynamic changes within the intensity contour were noted, and variations in contour between utterances.

2.6.7 - Summary of Human Voice Analysis Experiment

A summary of the results in Table 1 are in line with the general descriptions in the existing literature (summarised by Murray and Arnott (1993)) and another investigation by Kitahara and Tohkura (1992), with no notable exceptions. However, several minor effects noted from the analysis have not previously been described including the introduction of vibrato in stressed syllables (seen on the pitch contour plots) in happy speech, and the frequent use of inflections at word and sub-word level. These suggest that pitch changes are important in emotion expression, although not necessarily able to differentiate emotions reliably.

Preprint of: ***I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129***

3 – Enhancement Of The Synthetic Speech System

3.1 - Introduction

A research group led by the current authors developed a prototype speech synthesis-with-emotion system called HAMLET (Murray and Arnott (1995)), using the limited literature (reviewed by Murray and Arnott (1993)). This system added emotion effects to synthetic speech by rule using a commercial DECtalk formant-based synthesiser (Fonix (2007)) and simulated a set of six emotions (anger, happiness, sadness, fear, disgust and grief), this set being selected as they were the emotions about which the most was known from the literature. This prototype was evaluated in a listening experiment which showed that such a system was capable of producing vocal emotions recognisable to listeners.

Recently, concatenative synthesis has become the dominant technology in commercial text-to-speech systems, due to its superior voice quality (and thus greater acceptability to many listeners). However, concatenative technology, being based on recorded speech segments, poses fundamental problems when attempting to add emotion and other low level pragmatic features. The situation is similar to comparing cartoon faces (which can be made very expressive, but are drawings) with digital photos of faces (which are more natural, but attempts to morph them into convincing facial expressions tend to be less satisfactory). Despite these difficulties, attempts at adding emotion to concatenative speech have been made (for example (using acoustic manipulation) Murray and Edgington (1998), Murray (2000), Bulut et al. (2002), Anderson (2003) and (using selection of emotional units) Hofer (2005)); however, formant synthesis which is typically controlled by rule (as here), still offers greater control within the synthesis process for the addition of pragmatic effects, and so has been selected for the current study. Burkhardt and Sendlmeier (2000) note that formant synthesis also allows systematic variation of specific acoustic features, making it an ideal tool for emotion research.

Using the analysis results described in Section 2, the emotion rules within the system were modified to produce an enhanced version of the HAMLET system, which was then evaluated in a listening experiment.

3.2 - Emotion Rulebase Implementation

In normal use, the DECtalk synthesiser is used as a text-to-speech synthesiser. However, it can also be controlled at the phoneme level, with each utterance phoneme assigned explicit pitch and duration values; the phonemes are sent to the synthesiser as a marked-up text string. The voice quality can also be controlled by setting new values for any of the 31 available voice parameters; however, voice quality changes can only be made between utterances, again via a marked-up text string. As a consequence of the way the synthesiser operates, the incorporation of emotional features from the analysis into the synthetic speech is performed in two phases:

(i) initial settings of the voice quality parameters and;

(ii) dynamic changes integrated with the phoneme string.

In each case, the rules which implement these changes operate upon an array containing initial values, each rule in turn modifying the array values before the final array is converted to the marked-up text format to be sent to the synthesiser.

3.2.1 - Initial Settings (Static Rules)

It is apparent (from the literature and the preceding voice analysis) that each emotion results in particular changes in voice quality which remain largely constant throughout the utterance, and which convey a significant part of the emotional affect. For synthesis with DECtalk, these static features are implemented by a string of the "voice design" (i.e. voice quality) parameters which is sent in advance of the utterance to be spoken. Large changes in the voice design parameters result in changes in the "voice identity" of the synthesiser i.e. it sounds like a different person speaking. For the purposes of this project, smaller changes to certain of the parameters were required to suggest the same voice speaking with different emotions. Maximal use of the synthesiser's feature space can be achieved if the combinations for the initial settings are designed well apart from each other adjusted according to the number of emotions implemented. The selection of which parameters to use (see Appendix 2 for a list of DECtalk voice parameters, including those manipulated by HAMLET) for emotion expression was made heuristically; it will be appreciated from the list that while some parameters (e.g. resonator gains) control synthesiser

parameters directly, some others (e.g. assertiveness) are more esoteric and do not map directly to simple human speech variables.

The static rules take as their starting values the current voice design parameters, rather than starting with absolute values each time. Thus emotional changes can be implemented on any of DECtalk's "voice identities" and not just on the default voice.

3.2.2 - Dynamic Changes (Procedural Rules)

The dynamic changes are associated with the pitch and duration of individual phonemes. These changes are each implemented by a different procedure operating on the current utterance phonemes – each rule to be run (determined by the emotion selected) passes through the phoneme data and causes changes to be made to the phoneme, its duration and/or its pitch; the set of phonemes ultimately produced, together with timing and duration values, is then formatted into marked-up text and sent for synthesis.

The voice analysis showed that the voice quality tends not to vary appreciably during the utterance, except for intensity which for anger and happiness increases with the first content word and for sadness and fear decreases towards the end of the utterance. These changes could be implemented by inserting a further voice design string into the phoneme string at the required location; however, voice design changes within an utterance cause DECtalk to pause slightly, which would be undesirable for some emotions, so such changes were not implemented for this reason.

Dynamic changes in the pitch contour play an important role in conveying emotional affect (Murray and Arnott, op. cit.). They affect the general form for the whole sentence, as well as introducing fluctuations at word and phoneme levels. One major modification to the HAMLET speech dynamics was the generation of emotion-specific default pitch contours, rather than the simple linear declination contour used in the original system. Several contour shapes were observed from the voice analysis, and these were simulated by HAMLET for the five basic emotions; all were based on a triangular contour segment specified by end points and peak points, which could thus be implemented at the utterance, clause, or word level. The full contour in each case was built up from a series of segments - an utterance-level contour with clause-level segments and inflections added on top as appropriate. The utterance-level pitch contours implemented were as follows:

NEUTRAL, ANGER, SADNESS, DISGUST: Rise-fall contour, with pitch peak on the first stressed syllable.

HAPPINESS: Fall-rise contour, with lowest pitch on the last stressed syllable.

FEAR: Rise-fall-rise contour, with pitch peaks on the first and last stressed syllables.

Dynamic changes in the time characteristics include:

(i) introducing pauses between words and clauses; the pause durations varied according to the selected emotion;

(ii) modelling the increased phonation time in some cases and;

(iii) adjusting the rhythm characteristics by changing the phoneme durations.

The overall speech rate of the utterance is governed by a scaling operation on all utterance durations, based Allen et al. (1987).

The dynamic changes have been implemented as a set of procedural rules, a different subset of which is activated by each emotion. The rules set previously used was as follows (summarised from Murray and Arnott (1995)):

(1) increase pitch of stressed vowels

(2) increase duration of stressed vowels

(3) increase articulation precision

(4) decrease articulation precision

(5) increase rate of pitch declination

(6) eliminate abrupt changes in pitch between phonemes

(7) add downward pitch inflections at word endings

(8) reduce degree of pitch drop at end of utterance

(9) replace downward inflections with upward inflections

(10) modify durations for regular stressing

(11) add pauses after long words (words with more than 3 phonemes)

Following the analysis, all of the above rules had their parameters altered, and three additional rules were added to implement the following effects:

(12) add downward inflections at word level

(13) increase phonation time of some final consonants

(14) add vibrato to vowels

3.2.3 Emotion Implementation Example - Anger

As an example of the emotion implementation process, anger is executed as follows (it is assumed that text has been entered and converted to phonemes); the actual values implemented were determined heuristically (based on the literature and acted speech analysis) and the units correspond to the synthesiser's units for that parameter:

(i) Static rules - of the 20 parameters manipulated by HAMLET rules, 11 are manipulated for anger, as follows:

Average pitch is increased by 6Hz to 136Hz

Pitch range is increased by 10% to 110%

Smoothness is increased by 3% to 6%

Laryngealisation is increased by 60% to 60%

Cascade formant 4 frequency is decreased by 165Hz to 3135Hz

Frication gain is increased by 1% to 71%

Aspiration gain is increased by 3% to 73%

Nasal gain is decreased by 1% to 73%

Begin baseline fall is increased by 2Hz to 20Hz

Quickness is increased by 2Hz to 20Hz

Speech rate is increased by 30wpm to 210wpm

The phoneme array is then processed to set initial values for phoneme duration and phoneme pitch; these values will vary depending on some of the parameters set above, such as the speech rate and average pitch.

(ii) Dynamic rules - of the 14 dynamic rules, only 2 are utilised for anger:

(1) increase pitch of stressed vowels - the phoneme array is processed and stressed vowels have their pitch increased by a factor dependent on the type of stress and pitch range

(4) decrease articulation precision - any strong vowels in the utterance are replaced by their reduced forms

The final voice design parameter values and phoneme array are then formatted into a textual mark-up and sent to the synthesiser for speech production.

4 - Synthetic Speech Perception Experiment

4.1 - Introduction

In order to evaluate the modified HAMLET system and confirm that the emotions produced were recognisable and realistic, a listening experiment was designed and conducted. In the experiment, subjects listened to the emotive synthetic speech and were asked to note the vocal emotion which they perceived in the speech. The procedure for this experiment was based on previous listening experiments used to evaluate the Affect Editor (Cahn (1990)) and HAMLET (Murray (1989) and Murray and Arnott (1995)) and the same test phrases were used to allow direct comparison with evaluation results from the latter experiment.

The objectives for the experiment can be summarised as follows:

- to estimate the degree of realism of each of the vocal emotions produced by HAMLET;
- to produce recognisability rankings for all emotions simulated;
- to estimate the degree to which voice quality alone contributes to emotion perception;
- to evaluate the perception of the human emotional speech recordings used for the analysis.

4.2 - Selection of Subjects

Thirty five subjects took part in the experiment, a number sufficiently large to ensure statistical significance of the results. The group consisted of 28 males and 7 females; all were volunteers recruited by sign-up sheets from undergraduate courses in various University faculties. All subjects were unfamiliar with the system being evaluated, and were paid for their participation in the experiment.

4.3 - Text Phrases for Experimentation

The test phrases used in the main experiment were of two main types, those **empty of semantic emotion clues** and those **with text intended to indicate a particular emotion** (in the latter case, experimental results showed that the expected emotion or neutral was the main emotion perceived from the text in the majority of cases). These phrases were synthesised both **with** and **without vocal emotion effects**

(appropriate to the text emotion in the case of the second group). The emotionally empty phrases were expected to produce a high "no emotion" score when synthesised with no vocal emotion effects, but when synthesised with a particular vocal emotion, to produce a high score corresponding to that emotion. For the phrases with textual emotion, it was expected that the emotion would be more easily identified when vocal emotion clues were present in the synthesised utterance. Use of the same texts in these utterance pairs (i.e. the same phrase texts synthesised with and without vocal emotion effects) allowed differences between the results of the pairs to be calculated; these differences therefore represented the net effect of the HAMLET vocal emotion rules, any perception effects caused by the text itself being cancelled out by the comparison.

A total of 39 test phrases were used, as used for the previous HAMLET evaluation experiment (Murray and Arnott (1995)). These phrases were each synthesised **without** emotion effects, then with **all HAMLET effects**, and then with **HAMLET voice quality effects only**. The actual phrase texts used are given in Appendix 3. All synthesis used DECtalk's standard male voice (Paul).

4.4 - Subjects' Input

Based on previous experience (Murray and Arnott (1995)), a forced response test (where the subject is forced to pick one emotion from a presented list of choices) was retained for the current experiment. The emotion choice list presented was the same as used previously, and included seven additional "distracter" emotions intended to disguise the number of actual emotions being tested (the distracters being the next most common emotions drawn from the literature (Murray and Arnott (op. cit.)). At each showing, the order of the emotion list was randomised to eliminate any ordering effects and further disguise the actual set of target emotions. The emotion list was as follows:

* Anger	Loving
* Sadness	Worry
* Happiness	Amusement
* Disgust	Embarrassment
* Fear	Surprise
* Grief	Jealousy
* No Emotion	Sarcasm
	Other

(* indicates one of the emotions under test)

"Other" was included to allow subjects to indicate a perceived emotion which was not on the list.

An additional stage was added to the subjects' input after each stimulus utterance to gauge (on a 5-point scale) how satisfied the subjects were with the response they had just given. This procedure was used by Cahn (1990) and allows a more detailed weighted profile of the results to be produced.

4.5 - Experimental Procedure

The experiment was conducted in a quiet windowless office, with the subject sitting at a PC; the subject followed instructions given on the computer's monitor, and listened to the synthetic speech produced via headphones (the text of the utterance was not presented on the screen, the stimuli being presented acoustically only). The program conducting the experiment was designed to be operated by simple controls and give clear on-screen instructions as to what was expected of the user at each stage.

The subjects were told that it was purely the voice upon which their emotion judgements were to be made, and not the actual words spoken. After hearing each stimulus utterance once only, the subject was asked to select the appropriate emotion from the presented list using the mouse. They were then asked to rate how well the chosen emotion was conveyed in the voice by choosing one of the options from the rating list (very well / well / moderately / poorly / very poorly), again using the mouse.

The 117 stimulus utterances were presented to each subject in random order, and their responses were given after each one; this part of the experiment generally lasted around 35 minutes. At the conclusion of the perception experiment, each subject was asked to comment on the experimental procedure and the

vocal emotion implementation, and to make any general remarks. Following completion of this procedure, each subject undertook the further listening experiment to evaluate the human speech samples.

4.6 - Synthetic Speech Perception Results

From the results of the forced response test, two complete confusion matrices were produced for further analysis of the weighted and unweighted results. From these matrices, partial confusion matrices were constructed for each of the six utterance groups. Sections of these matrices corresponding to the emotions under test are shown within Tables 2-8 below. Note that scores for the distracter emotions have been omitted for clarity, and that throughout this section (including Tables), all values shown refer to the *number of subjects* making that selection.

4.6.1 - Neutral texts with no vocal emotion

A summary confusion plot for utterances with neutral text spoken with no vocal emotion is shown in Table 2; the results for the eighteen utterances in this category have been averaged. The following observation can be made from the full results:

- 78% of the 18 row maxima (i.e. the most frequent emotion attribution for the utterance) occurred as expected in the "no emotion" column (compared to 28% in the original HAMLET evaluation) with 11% also occurring under anger, 11% under sadness and 6% under surprise, indicating general perception of the utterances as neutral.

4.6.2 - Emotive texts with no vocal emotion

A summary confusion plot for the utterances with emotive text and no vocal emotion effects are shown in Table 3, averaged over twenty-one test utterances; the results are also shown in Table 4, averaged for each stimulus emotion (three different phrase texts). The following observations can be made from the full results:

- 71% of the 21 row maxima occurred as expected in the "no emotion" column. This indicated that the subjects were generally reacting as instructed to the vocal stimuli, rather than to the semantic effect of the text.

- The remaining 29% of the row maxima all occurred under the column corresponding to the emotion in the text (two each of anger, sadness and happiness) indicating that some subjects did default to the emotion in the text rather than in the voice on some occasions.

4.6.3 - Neutral texts with vocal emotion

The confusion plots for the utterances with neutral text synthesised with vocal emotion effects are shown in Table 5. The results for each stimulus emotion (three different phrase texts) in this category have been averaged. The following observations can be made from the full results:

- 40% of the 18 row maxima occurred in the expected column, compared to 28% from the evaluation of the previous version of HAMLET. These maxima were distributed as follows: 11% for anger (no improvement from the previous version of HAMLET), 17% for sadness (no improvement), 6% for happiness (improved result), and 6% for fear (improved result).

- All three grief utterances were perceived as sadness, probably due to the similar ways in which these emotions are expressed.

- All three disgust utterances were perceived as sadness, confirming the finding from the original HAMLET experiment which indicated an underlying sadness in the voice.

- Happiness was not clearly recognised; one of the happy utterances was perceived as surprise (perhaps resulting from the chosen form of the pitch contour model (terminal rise)), the other happy utterance being perceived as worry.

- The emotions produced by the modified system appeared to be less easily mistaken for one another (in most cases, confusion occurred between emotions which are difficult to differentiate, e.g. sadness for grief, happiness for surprise); notable confusions are disgust being strongly recognised as sadness, and fear being strongly recognised as grief and less strongly as sadness.

4.6.4 - Emotive texts with vocal emotion

The confusion plot for the utterances with emotive text synthesised with appropriate vocal emotion effects is shown in Table 6; the results for each stimulus emotion (three different texts) in this category have been averaged. The following observations can be made from the full results:

- 76% of the 21 row maxima occurred in the expected columns, indicating that the emotions were being recognised reliably by the subjects. The same result was obtained for the original version of HAMLET.

- The implementation of fear has been improved compared to the original version; all three fear phrases were recognised as expected compared to only one phrase for the original version.

- The implementation of disgust is slightly better than the original version, with all three phrases being recognised as expected by 3, 11 (row maxima) and 4 subjects respectively, compared with two disgust phrases not recognised by any of the subjects for the original version.

4.6.5 - Neutral texts with vocal emotion implemented by voice quality only

These are the phrase utterances in which there was no emotion expressed in the text, but the phrase was synthesised with a particular vocal emotion implemented by using only the voice quality rules, without the effect of the procedural rules. Table 7 shows the confusion matrix for each emotion averaged over three utterances. The full results indicate:

- 44% of the 18 row maxima appeared as expected. Surprisingly, one of the fear phrases was better recognised when implemented only by the initial setting of the voice quality parameters; all other row maxima appeared as in section 4.6.3. The implementation of the procedural rules (timing characteristics and pitch contour changes) does not seem to have introduced additional improvement in the emotion modelling. The reason for this effect could be the basic level of control available for those features; a better low level integration of the prosodic changes with the internal structure of the speech-producing algorithm might result in a better quality of the emotion modelling.

- It was noted that all three grief utterances and all three disgust utterances were perceived as sadness in both full emotional implementation and voice quality implementation only.

4.6.6 - Emotive texts with vocal emotion implemented by voice quality only

These are the phrase utterances in which there was a particular emotion in the text, and the phrase was synthesised with the corresponding vocal emotion implemented by using only the voice quality rules, without the effect of the procedural rules. Table 8 shows the confusion matrix for each emotion averaged over three utterances. The full results indicate:

- 76% of the 21 row maxima occurred in the expected columns. Similar results were obtained for the full emotion implementation (section 4.6.3). The voice quality only condition provided better perception for one of the happy phrases and poorer perception for one of the disgust phrases.

- When the total number of subjects who correctly identified each emotion when fully implemented by the HAMLET rules were compared with the results obtained from the voice quality implementation only, only small differences in perception were noted. Adding the procedural rules to the voice quality did not substantially improve the perception of some emotions as had been expected.

4.7 - Differences Caused by the HAMLET Rules

Difference matrix plots were produced from the data, by examining the figures for the same texts synthesised both with and without vocal emotion effects, and subtracting one figure from the corresponding one; this thus produced the net difference in vocal emotion perception caused by the effects produced by the modified HAMLET rules. The difference data was analysed using McNemar's test (Sprent (1989)), a nonparametric statistical test which does not consider the overall number of subjects who passed or failed altered or unaltered tests, but rather the relative numbers of individual subjects whose performance *improved* or *deteriorated* between the altered stimuli. For this experiment, this meant comparing only the subjects who had identified the emotion in each phrase with vocal emotion but not without, and vice versa. The test was applied only to the confusion matrix entries where the stimulus and perceived vocal emotions were the same. Table 9 shows this subset of the difference confusion matrices (other elements of the full confusion matrices being omitted for clarity).

4.7.1 - The effect of adding vocal emotion to phrases with neutral text

The first column of Table 9 shows the differences between phrases synthesised from texts with no textual emotion. The full results indicate:

- Of the eighteen utterance pairs, improvements (which are significant at the 5% level) in emotion perception caused by adding vocal emotion occurred for one of the anger phrases, one of the happiness phrases, one of the sadness phrases, one of the fear phrases and one of the grief phrases; improvements significant at the 1% level occurred for one of the anger phrases, two for the sadness phrases and one for the fear phrases. This total of nine statistically significant improvements in perception compares to a total of five for the original version. Disgust was poorly recognised in all conditions.

- It was concluded from the evaluation of the original HAMLET system that the spread of results for some emotions indicated that the "high variability (of the results was) dependent on the actual phrase used" in some cases; this was rejected by the present perception experiment, as there were persistent improvements for all emotions, excepting a slight degradation for one of the disgust phrases.

4.7.2 - The effect of adding vocal emotion to phrases with emotive text

The second column of Table 9 shows the differences between phrases synthesised from texts with textual emotion). The full results indicate:

- The first three phrase utterance pairs were identical (neutral being the vocal emotion used in both), and only minor differences in the perception totals occur as expected. This result is consistent with the finding from the evaluation of the original version.

- For the other eighteen phrase pairs, improvements in emotion perception caused by adding vocal emotion occurred for nine of the phrases, all significant at the 1% level. For the original version there was also the total of nine phrases with significant improvements, though with only seven at the 1% level.

- Recognition of all three grief phrases was significantly improved at the 1% level, a result also obtained for the original version, confirming a high reliance on the context for this emotion.

- A significant perception improvement was noted for fear, with all three phrases correctly identified at 1% level. This is a better result than observed for the original version, indicating that the implementation of fear has been improved.

- The following rankings for emotion perception improvements using all rules were found:

    fear (best)
    grief
    anger
    sadness
    happiness
    disgust (worst)

4.7.3 - The effect of adding vocal emotion implemented by voice quality only to neutral text

The third column of Table 9 shows the differences between phrases synthesised using voice quality rules only from texts with no emotion. The full results indicate:

- Perception has been significantly improved for six out of eighteen utterances, as follows: three of the sadness phrases at 1% level, two of the fear phrases at the 1% level and one of the grief phrases at the 5% level. These results are slightly worse than the case when the vocal emotion implementation included procedural rules as well as vocal quality effects.

- The results indicate that the voice quality implementation of sadness and fear contributes significantly to the emotion perception and conveys a significant part of the emotion's vocal characteristics.

4.7.4 - The effect of adding vocal emotion implemented by voice quality only to emotive text

The fourth column of Table 9 shows the differences between phrases synthesised using voice quality rules only from texts with textual emotion. The full results indicate:

- Of the eighteen phrase pairs, improvements significant at the 5% level occurred for four phrases: two of the anger phrases and two of the sadness phrases. Improvements significant at the 1% level occurred for six phrases: one of anger, one of sadness, three of fear and one of disgust. These results indicate very good implementation of the emotional characteristics when using only the voice quality parameters, other than for happiness and disgust which are poorly perceived (disgust is the poorest perceived in human speech also (Fairbanks and Pronovost (1939)).

- This part of the experiment resulted in the greatest number of statistically significant results; they are better than the results from the original version and, remarkably, are slightly better than the case when the emotion effects included intonational effects produced using the procedural rules.

- The following rankings for emotion perception improvements using voice quality rules only were found:

fear (best)

sadness

anger

grief

happiness

disgust (worst)

4.8 - Perception of the Human Speech Recordings

In order to confirm that the emotions produced by the actors were perceived as expected, a formal experiment was conducted as part of the evaluation procedure for the enhanced synthetic speech with emotion system reported above. Four sets of recordings were produced from the original actor recordings, each comprising twenty-four utterances (two phrases with six emotions each from each actor) in randomised order. Following the conclusion of the synthetic speech evaluation, each subject was played one set of utterances chosen at random and asked to choose the emotion conveyed by each utterance in a forced response test. A confusion plot for the twenty-four utterances is shown in Table 10; each emotion was conveyed in four utterances, and averaged results are shown.

It was observed from the results that only anger and neutral were correctly identified by the majority of subjects, with happiness, sadness and disgust all regularly identified as anger. It is surprising to note that in informal discussion following the listening experiment, most subjects commented that identification of emotion in the human speech recordings was easier than for the synthetic speech-with-emotion; despite this, they identified the emotions as expected more often in the synthesised speech than in the human speech.

4.9 - Summary of Speech Synthesis Perception Experiment

Subjects were able to identify neutral vocal emotion more consistently than subjects in the previous evaluation (Murray & Arnott (1995)).

The effect of the semantic content of the text appeared less significant than in the results obtained from the testing of the original version (op. cit.). This conclusion is based on the more consistent pattern of the improved scores spread over the phrases for a given emotion.

The vocal emotion implementation of anger and happiness using only the voice quality rules was perceived to a similar degree as the full emotion implementation including the intonation rules (in one case even slightly better). This concurs with results of Gobl and Ní Chasaide (2003) and the study by O'Sullivan et al. (1985) comparing voice quality, speech content, face and body channels on speaker judgements, which indicated that voice quality only is used in judging the speaker if nothing else was available.

5 - Conclusion

This paper has described a human vocal emotion analysis experiment whose results were used to enhance a synthetic speech-with-emotion system. The enhanced system was then evaluated in a listening experiment, and the results compared with the previous evaluation. The following notable improvements in the perception of the enhanced HAMLET speech over the original were identified by this experiment:

- a significant improvement for fear;
- some improvement for happiness;
- better general scores for anger and sadness and;
- a slight improvement for disgust.

Carried out at the same time as the synthetic speech evaluation, a listening experiment on the human speech recordings showed that subjects were better able to perceive the intended emotion from the synthetic speech than from the human speech (although subjectively they thought that the reverse was the case). Despite this finding, the analysis procedure described has resulted in a synthetic speech system with improved realism.

Preprint of: ***I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129***

Acknowledgements

The work reported in this paper was developed from research originally carried out under a U.K. SERC/EPSRC Research Grant and a BT Short-term Research Fellowship. The authors wish to express their thanks to Dr. Elizabeth Rohwer who contributed to this work.

Preprint of: *I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129*

APPENDIX 1 - EMOTIVE TEXTS FOR ANALYSIS EXPERIMENTS

The following texts were used to elicit the vocal emotions from the actors during the recording sessions. The actors were given the context information, and asked to read each paragraph using the emotion indicated. The texts were available to the actors prior to the recording session. The neutral phrases were not included in a context-setting paragraph, but were recorded both at the beginning and at the end of the recording session. The paragraphs were informally judged to indicate the emotion intended by over half of ten subjects presented with the texts (without the context information).

NEUTRAL

> "This is not what I expected. I thought things would be different."

> "You've asked me that question so many times. You know what the answer is."

ANGER

You have purchased an item of hi-fi equipment which has turned out to be faulty. You return to the store to get it exchanged, but the salesman you speak too tries to fob you off with excuses:

> "It's just not good enough. I want to see the manager right now. I'm not spending all this money on something that isn't going to work. Your store advertises that the customer comes first, but that is plainly untrue. This is not what I expected. I thought things would be different. I thought I'd get the prompt, courteous service you advertise, but all you've given me is excuses, and I'm not having it. I'm not leaving here until I get satisfaction."

Your friend relies on you to fix his car, but it has been particularly unreliable lately, and you are fed up spending all your time fixing it for him:

> "Why don't you take the stupid thing to a garage and get it serviced properly. I'm fed up spending all my spare time under the bonnet of your car. I wouldn't really mind, because I enjoy fiddling with engines, but you just expect me to drop everything every time it breaks down. And don't ask me what's wrong with it now. You've asked me that question so many times. You know what the

answer is. There's always something new gone wrong with it every time you come. The thing's a wreck, and I wish you'd stop bothering me with it."

SADNESS

You have received a quote from your garage for repairs to your car. The quote is so high, that you won't be able to afford to go on holiday. You have to tell your spouse about the letter:

"Well, they want five hundred pounds to fix the car. We'll have to get the work done, and that means we can't afford to go on holiday at the end of the month now. I was so looking forward to getting the car fixed and going away, too. This is not what I expected. I thought things would be different. I thought it would only be a hundred pounds or so, and that would have been no problem. I really was looking forward to the holiday so much. I guess we'll just have to wait a little while longer."

Your dog has not been seen for several days. You try to comfort your spouse:

"I wish we knew something about him, but no one's seen him for days now. It's so unlike him to stay away for so long. I know he's gone missing before, but then he's usually only been out overnight. I don't know what can have become of him this time. You've asked me that question so many times. You know what the answer is. I've wracked my brain for anywhere he might have gone, but I've looked everywhere and come up with nothing. I guess we'll just have to wait and hope that he will come back to us this time."

HAPPINESS

You have just received a letter telling you that you have won £10,000 in a competition. You rush to tell your neighbour the news:

"Hey, listen, I've got some great news! I've won ten thousand pounds in that competition. There were fifty prizes, and I've got the top one! It's incredible - I've never won anything before in my life. This is not what I expected. I thought things would be different. I was due some big bills and

> some loan repayments, but I needn't bother about them now. And I can get that new car I've been wanting, and a holiday too. I'm going to spend and spend."

You have won first prize in a competition - a cheque for £10,000. After the presentation, you realise that you've really got the money, and tell your friend what you're going to do with it:

> "I won. I never won anything before in my life, not even a plastic piggy bank at a fun fair. Ten thousand pounds – wow! Look at this cheque. I don't even see the numbers. You know what I see? A trip round the world, a car, no debts, some decent clothes, maybe a house. I'm going to spend and spend. That's all I want to do. You've asked me that question so many times. You know what the answer is. This can only happen once, so I'm going to spend, spend, spend."

DISGUST

You have been clearing out your Uncle and Aunt's garden for them while they're out for the day. However, while cleaning the pond, you've slipped and got yourself covered in horrible slimy mud. You relate your experiences to your Uncle on his return:

> "I started on the shed - what a mess! There were cobwebs everywhere, and the dust! In one corner there were loads of tins with yucky liquids in them - don't ask me what they were, I just got rid of them. Then I drained out the pond water, and when I went in to clean the bottom, I slipped and landed right in the mud. What a smell! And it was so slimy and sticky, it took me ages to wash it all off. This is not what I expected. I thought things would be different. I thought I'd get it finished off quickly and spend the rest of the day enjoying the sun. No such luck."

Your friends are having their dog put down because there will be no room for it in their new flat. You express your disgust to one of your friends at their decision:

"I don't know how you can face doing such a thing. He's been your friend for so many years, and now you're having him put down just for your own convenience. What an incredibly heartless decision to make. Of course I think it's wrong. You've asked me that question so many times. You know what the answer is. I believe you shouldn't destroy+ animals unless there is no other choice on health grounds. You have many alternatives, and this is surely the worst of them."

Preprint of: ***I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129***

FEAR

You have borrowed your boss's new car to go on an errand for him. Against his instructions, you have also gone on an errand of your own, and on the way there, you have brushed against a parked car and scratched the paint down one side. You telephone your spouse for advice:

> "What'll I do? The boss will kill me when he finds out about this. The car was only a week old, and it's his pride and joy. He told me not to go off anywhere in it, but it was only a small detour to the shops. This is not what I expected. I thought things would be different. I thought it would be a good chance to do the boss a favour and get out of the office for a while, but now I've really messed things up. He'll be furious. He might even sack me for going against his instructions. What am I going to do?"

You have been sentenced to die for murder. As your cell door is opened at dawn on the day of execution, you are seized with a fit of uncontrollable terror:

> "Please don't let them take me away now. Let me have one more day ... one more hour ... to live. I don't deserve hanging for the thing I did. I didn't know then that a man's life meant so much. But I know now, I know, and please forgive me. I don't know how it happened. Honest, I don't. One minute I was standing there, and the next minute there was a smoking gun in my hand. I don't know how it got there. You've got to believe me this time, even if you never did before. You've got to believe it in time to keep them from hanging me. Every night you ask me how it happened. But I don't know! I don't know! I can't remember. There is no other answer. You've asked me that question so many times. You know what the answer is. You can't figure out things like that. They just happen. And afterwards you're sorry. I'm that way now. I'm sorry. Oh, please, stop them ... quick ... before it's too late!"

APPENDIX 2 - SUMMARY OF DECtalk VOICE DESIGN PARAMETERS

PARAMETER	UNIT	DEFAULT	USED
Speaker sex	M or F	M	No
Head size	%	100	Yes
Average pitch	Hz	130	Yes
Pitch range	%	100	Yes
Richness	%	70	Yes
Smoothness	%	3	Yes
Breathiness	dB	0	Yes
Laryngealisation	%	0	Yes
Lax breath	%	0	Yes
Assertiveness	%	100	Yes
Open period sample		0	Yes
Cascade formant 4 frequency	Hz	3300	Yes
Cascade formant 4 bandwidth	Hz	260	Yes
Cascade formant 5 frequency	Hz	3650	No
Cascade formant 5 bandwidth	Hz	330	No
Hat rise	Hz	18	No
Stress rise	Hz	32	No
Frication gain	dB	70	Yes
Aspiration gain	dB	73	Yes
Voicing gain	dB	65	Yes
Nasal gain	dB	74	Yes
Resonator 1 gain	dB	68	No
Resonator 2 gain	dB	60	No
Resonator 3 gain	dB	48	No
Resonator 4 gain	dB	64	No
Resonator 5 gain	dB	86	Yes
Begin baseline fall	Hz	18	Yes
Quickness	%	40	Yes
Speech rate	wpm	180	Yes
Period pause duration	ms	0	No
Comma pause duration	ms	0	No

Parameter names are taken from the DECtalk V4.2 documentation (wherein the exact function of some is not made clear); default values are for standard male voice (Paul), which was used throughout this project. Selection of which parameters were manipulated by HAMLET for emotion expression was done heuristically. Head size, laryngealisation, smoothness and richness have the greatest audible effect on the DECtalk speech (other than speaker sex, which was not used as it has so extreme an effect). Increasing the smoothness parameter results in a decrease in the voicing energy at higher frequencies; its operation is thus analogous to a the bass control on an audio system. Richness is analogous to the treble control, giving more emphasis to higher frequencies.

The overall intensity of the speech can be adjusted using the two groups of gain parameters; resonator gains (connected with the structure of the formant synthesiser) and a group phonetically affecting the articulation (frication gain, aspiration gain, voicing gain and nasal gain). Changes to some parameters can result in clipping in the speech output, which has to be corrected by changes in the gain parameters used.

Preprint of: *I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129*

APPENDIX 3 - STIMULUS TEXTS USED FOR HAMLET PERCEPTION EXPERIMENT

The test phrases used in the HAMLET perception experiment were divided into two main groups, those with text indicating a particular emotion, and those empty of semantic emotion clues. These phrases were synthesised both with and without vocal emotion effects (appropriate to the text emotion in the case of the first group). Use of the same texts in these utterance pairs (i.e. the same phrase texts synthesised with and without vocal emotion effects) allowed differences between the results of the pairs to be calculated; these differences therefore represented the net effect of the HAMLET vocal emotion rules, any perception effects caused by the text itself being cancelled out by the comparison.

A total of 39 test phrases were used; these were the same phrases used for the previous HAMLET evaluation experiment (Murray (1989)), allowing a direct comparison with the previous results to be made. The 39 phrases were each synthesised in three ways:

- without emotion effects;
- with all HAMLET effects implemented;
- with HAMLET voice quality effects only implemented

This gave a total of one hundred and seventeen synthesised stimulus utterances, as follows:
(i) 18 neutral phrases (3 for each of the six HAMLET emotions), designed to be applicable to any emotion, all synthesised in a neutral voice (Table 11a);
(ii) 21 emotionally loaded phrases (3 for each of the six HAMLET emotions plus neutral), all synthesised in a neutral voice (Table 11b);
(iii) The same phrases as (i) above, synthesised with one of the six emotions (Table 11a);
(iv) The same phrases as (ii) above, synthesised with the appropriate vocal emotion (Table 11b);
(v) The same phrases as (i) above, synthesised with one of the six emotion implemented by using only the voice quality effects (Table 11a);
(vi) The same phrases as (ii) above, synthesised with the appropriate vocal emotion implemented by using only the voice quality effects (Table 11b).
In Table 11, the number in the first column indicates the number of the phrase with the given text synthesised without vocal emotion effects, the number in the second column indicates the number of the phrase with the given text synthesised with full vocal emotion effects and the number in the third column

indicates the number of the phrase with the given text synthesised with vocal emotion effects (voice quality only).

Preprint of: *I.R. Murray, J.L. Arnott / Computer Speech and Language 22 (2008) 107–129*

		Speaker A only	Common Features	Speaker P only
Neutral			Intensity decreasing over utterance, clear, some pauses	
Anger	Pitch		**Wide range**, downward inflections, terminal fall	**Much higher**
	Intensity		**Higher**	
	Speed	**Slightly higher**		
	Other	Breathy, **tense**		
Happiness	Pitch	**Rhythmic stresses**	**Higher, wide range**	Raised at end
	Intensity		**Higher**	
	Speed		Higher	
	Other	Breathy, increased precision		Some pauses
Sadness	Pitch	**Lower**	**Narrow range**	Slightly higher
	Intensity		**Lower**	Decreasing over utterance
	Speed	**Lower**		
	Other	**Decreased precision**, pauses		
Fear	Pitch	Slightly **higher**, slightly **wider range**, vibrato		
	Intensity	Lower		Variable, tending to be high
	Speed	Slightly **higher**		
	Other	Many pauses		
Disgust	Pitch	**Lower**		**Wide range**
	Intensity	**Lower**		Higher, decreasing over utterance
	Speed	**Lower**		
	Other	Many pauses	Increased precision	

TABLE 1 - SUMMARY OF NOTABLE EFFECTS FROM ACTOR RECORDING ANALYSIS

Features in **boldface** agree with the literature (summarised in Murray and Arnott (1993)), features given relative to neutral speech

Perceived vocal emotion						
No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
9.6	3.9	3.8	1.8	2.3	0.3	0.7

TABLE 2

Partial confusion matrix for **neutral** texts synthesised **without** vocal emotion effects; figures are the number of subjects making that selection, averaged over 18 test utterances

Perceived vocal emotion						
No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
11.8	2.7	4.0	3.1	2.2	0.9	0.9

TABLE 3

Partial confusion matrix for **emotive** texts synthesised **without** vocal emotion effects; figures are the number of subjects making that selection, averaged over 21 test utterances

		Perceived vocal emotion						
		No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
Stimulus textual emotion	**No emotion**	18.6	0.7	3.0	2.0	0.7	0.0	1.0
	Anger	5.0	10.3	1.0	1.6	4.3	0.7	0.7
	Sadness	8.6	1.0	10.3	1.0	2.0	1.0	2.0
	Happiness	9.0	0.3	2.0	8.7	0.7	0.0	1.3
	Disgust	13.0	0.7	5.3	2.0	3.7	2.0	0.0
	Fear	14.0	1.7	1.0	1.3	0.7	4.0	0.7
	Grief	10.3	1.3	4.0	4.0	1.0	0.7	0.7

TABLE 4

Partial confusion matrix for **emotive** texts synthesised **without** vocal emotion effects; figures are the number of subjects making that selection, averaged over 3 test utterances

		Perceived vocal emotion						
		No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
Stimulus vocal emotion	**No emotion**	18.0	1.3	1.3	1.3	1.3	0.3	0.0
	Anger	3.0	14.3	0.3	3.3	3.3	0.3	0.3
	Sadness	4.0	0.3	16.0	0.0	0.7	0.3	4.7
	Happiness	3.7	0.7	3.3	3.0	0.3	1.0	0.7
	Disgust	6.7	0.0	16.3	0.7	1.3	0.0	2.3
	Fear	0.3	1.0	5.3	0.7	0.3	7.7	11.0
	Grief	0.7	2.0	12.0	0.3	2.3	0.0	6.0

TABLE 5

Partial confusion matrix for **neutral** texts synthesised **with** vocal emotion effects; figures are the number of subjects making that selection, averaged over 3 test utterances

		Perceived vocal emotion						
		No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
Stimulus textual & vocal emotion	**No emotion**	18.0	1.3	1.3	1.3	1.3	0.3	0.0
	Anger	1.0	22.0	0.0	1.0	5.0	0.7	0.0
	Sadness	3.3	0.33	17.0	0.0	1.3	0.0	7.0
	Happiness	2.0	0.7	3.0	8.7	0.0	1.3	1.3
	Disgust	5.7	1.0	9.3	0.3	6.0	0.0	0.7
	Fear	0.3	0.0	2.0	0.0	0.0	19.3	4.3
	Grief	1.3	0.0	11.7	0.0	1.7	0.3	11.3

TABLE 6

Partial confusion matrix for **emotive** texts synthesised **with** vocal emotion effects; figures are the number of subjects making that selection, averaged over 3 test utterances

		Perceived vocal emotion						
		No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
Stimulus vocal emotion	**No emotion**	20.3	0.0	2.0	2.7	1.3	0.0	0.0
	Anger	1.3	11.0	0.7	1.7	6.3	2.0	0.3
	Sadness	4.0	0.3	14.0	0.0	1.0	0.3	4.7
	Happiness	3.3	0.0	4.7	3.7	0.0	2.7	1.0
	Disgust	6.0	0.0	18.7	0.3	1.0	0.0	0.7
	Fear	0.0	0.3	5.3	1.3	0.3	7.7	6.7
	Grief	3.7	0.7	14.0	0.0	1.3	0.3	4.7

TABLE 7

Partial confusion matrix for **neutral** texts synthesised **with** vocal emotion effects (**voice quality only**) ; figures are the number of subjects making that selection, averaged over 3 test utterances

		Perceived vocal emotion						
		No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
Stimulus textual & vocal emotion	**No emotion**	20.3	0.0	2.0	2.7	1.3	0.0	0.0
	Anger	0.7	25.7	0.0	0.0	4.0	0.7	0.0
	Sadness	3.3	0.0	19.3	9.3	0.0	0.7	4.7
	Happiness	3.0	0.0	2.7	9.3	0.0	2.7	2.7
	Disgust	5.3	0.3	14.3	0.3	3.3	0.0	2.3
	Fear	1.0	0.3	0.3	0.7	0.0	19.0	3.0
	Grief	3.0	0.0	12.0	0.0	0.0	0.0	8.7

TABLE 8

Partial confusion matrix for **emotive** texts synthesised **with** vocal emotion effects (**voice quality only**) ; figures are the number of subjects making that selection, averaged over 3 test utterances

		ALL EMOTION RULES		VOICE QUALITY RULES ONLY	
		Neutral texts	**Emotive texts**	**Neutral texts**	**Emotive texts**
STIMULUS VOCAL EMOTION	**No emotion**	-	5	-	-1
	No emotion	-	0	-	-1
	No emotion	-	7	-	0
	Anger	13**	15	6	20*
	Anger	19*	24**	6	21*
	Anger	1	10**	1	19**
	Happiness	6*	3	1	4
	Happiness	0	2	6	3
	Happiness	0	-5	0	-5
	Sadness	12**	4	10**	10**
	Sadness	10*	10**	10**	10*
	Sadness	15**	6	13**	7*
	Fear	7	12**	4	11**
	Fear	6*	21**	10**	19**
	Fear	10**	13**	9**	15**
	Grief	5	10**	4	9**
	Grief	8*	14**	4	8
	Grief	4	8**	5*	7
	Disgust	0	2	0	1
	Disgust	-2	5	-3	-1
	Disgust	0	0	0	-1

TABLE 9 - PARTIAL DIFFERENCE CONFUSION MATRIX

Figures indicate perception total for utterances **with** vocal emotion effects (3 for each emotion) **minus** perception total for *same* utterance **without** vocal emotion effects; only results where perceived vocal emotion was the same as the stimulus are shown for clarity – this corresponds to the data analysed with McNemar's test, and significant results are indicated:

* result significant at 5% level

** result significant at 1% level

		Perceived vocal emotion						
		No emotion	**Anger**	**Sadness**	**Happiness**	**Disgust**	**Fear**	**Grief**
Stimulus human vocal emotion	**No emotion**	12.0	0.0	5.5	0.0	2.8	0.5	0.0
	Anger	3.3	24.5	6.5	12.0	8.3	4.25	0.0
	Sadness	6.3	0.8	5.5	1.3	2.3	3.8	1.0
	Happiness	0.0	0.3	0.0	0.5	0.5	1.0	0.3
	Disgust	3.3	4.3	2.8	3.3	5.3	1.8	0.3
	Fear	0.3	0.5	0.5	1.5	0.4	4.5	0.6

TABLE 10

Partial confusion matrix for **human** emotional speech; figures are the number of subjects making that selection, averaged over 4 test utterances

ANGER	He has made a spectacle of himself.
ANGER	He said that he would be here by ten o'clock.
ANGER	I can see some people over there.
HAPPINESS	I am expecting to hear soon.
HAPPINESS	It is a very odd-looking thing.
HAPPINESS	It isn't here anymore.
SADNESS	It was in there the last tune I looked.
SADNESS	It's certain to be the way they choose.
SADNESS	She cannot remember what I said.
FEAR	There is no other answer.
FEAR	They gave it to my neighbour.
FEAR	They told me that they didn't have one.
GRIEF	They will be there the next time.
GRIEF	This is not what I expected.
GRIEF	You have asked me that question so many times.
DISGUST	She has changed my appointment.
DISGUST	The telephone has not rung today at all.
DISGUST	The shops are all closed this week.

TABLE 11a - SEMANTICALLY NEUTRAL / EMOTIONALLY "EMPTY" PHRASES

NEUTRAL	The green book is lying on the table.
NEUTRAL	There are ten cars in the car park.
NEUTRAL	The picture is hanging on the wall.
ANGER	How dare you speak to me like that.
ANGER	You have just wasted an hour of my time.
ANGER	That man must have stolen my wallet.
HAPPINESS	I've won a fortune on the football pools.
HAPPINESS	Today I bought the car that I've always wanted.
HAPPINESS	That meal was really delicious.
SADNESS	I have left my umbrella on the train.
SADNESS	Someone has stolen by new bicycle.
SADNESS	I cannot come to your party tomorrow.
FEAR	Someone is creeping about downstairs.
FEAR	Please put that knife down.
FEAR	A police car is following us.
GRIEF	My wife has been in a serious accident.
GRIEF	A lorry ran over my dog yesterday.
GRIEF	My brother has become very ill.
DISGUST	The weather is really awful today.
DISGUST	This coffee tastes terrible.
DISGUST	He gave a very poor performance.

TABLE 11b - SEMANTICALLY LOADED PHRASES

References

Ax, A.F., "Physiological Differentiation of Emotional States", Psychosomatic Medicine, 15, 1953, pp. 433-442.

Allen, J., Hunnicutt, M.S., Klatt, D., Armstrong, R.C. and Pisoni, D., "From Text to Speech: The MITalk System", Cambridge University Press, Cambridge, UK, 1987.

Anderson, S., "Emotion Rules for the Festival Speech Synthesiser", Project Report, Applied Computing, University of Dundee, 2003.

Banse, R. and Scherer, K.R., "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, 70(3), 1996, pp. 614-636.

Bulut, M., Narayanan, S.S. and Syrdal, A.K., "Expressive Speech Synthesis using a Concatenative Synthesiser", Proceedings of ICSLP '02, Denver, CO, USA, 2002.

Burkhardt, F. and Sendlmeier, W.F., "Verification of acoustical correlates of emotional speech using formant synthesis", Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, 5-7 September 2000, pp. 151-156.

Cahn, J.E., "Generating Expression in Synthesised Speech", MIT Media Laboratory Technical Report, 1990.

Davitz, J.R., "The Communication Of Emotional Meaning", MacGraw-Hill, New York, 1964.

Dawes, R.M. and Kramer, E., "A Proximity Analysis of Vocally Expressed Emotions", Perceptual And Motor Skills, 22, 1966, pp. 571-574.

Douglas-Cowie, E., Campbell, N., Cowie, R. and Roach, P., "Emotional speech: towards a new generation of databases", Speech Communication, 40, 2003, pp. 33-60.

Eskénazi, M., "Trends in speaking styles research", Proceedings of Eurospeech '93, Berlin, Germany, 1993, pp. 501-509.

Fairbanks, G. and Pronovost, W., "An Experimental Study of the Pitch Characteristics of the Voice during the Expression of Emotion", Speech Monographs, 6, 1939, pp. 87-104.

Fónagy, I., "Emotions, Voice and Music", In Sundberg, J. (Ed), "Research Aspects On Singing: Proceedings From A Seminar Organised By The Committee For The Acoustics Of Music", Royal Swedish Academy Of Music, 33, 1981, pp. 51-79.

Fonix, "About Fonix Speech", 2007. Retrieved Jan 8 2007 from http://www.fonixspeech.com/pages/index.php

Gobl, C. and Ní Chasaide, A., "The role of voice quality in communicating emotion, mood and attitude", Speech Communication, 40, 2003, pp. 189-212.

Hofer, G.O., Richmond, K. and Clark, R.A.J., "Informed Blending of Databases for Emotional Speech Synthesis", Proceedings of Interspeech 2005, Lisbon, Portugal, Paper #1836.

Johnstone, T. and Scherer, K.R., "The effects of emotions on voice quality", Geneva Studies in Emotion and Communication, 13(3), 1999. Retrieved Jan 8 2007 from http://www.unige.ch/fapse/emotion/ publications/geneva_studies.html

Kitahara, Y. and Tohkura, Y., "Prosodic Control to Express Emotions for Man-Machine Speech Interaction", IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (Japan), E75-A(2), February 1992, pp. 155-163.

Leinonen, L., Hiltunen, T., Linnankoski, I. and Laakso, M., "Expression of emotional-motivational connotations with a one-word utterance", Journal of the Acoustical Society of America, 102(3), 1997, pp. 1853-1863.

Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Macías-Guarasa, J., Enríquez, E. and Pardo, J.M., "Development of an Emotional Speech Synthesiser in Spanish", Proceedings of Eurospeech '99, Budapest, Hungary, September 1999.

Morton, K., "PALM: PsychoAcoustic Language Modelling", Proceedings of the Institute of Acoustics, 14(6), 1992, pp. 189-197.

Murray, I.R., "Simulating Emotion In Synthetic Speech", PhD Thesis, University Of Dundee, 1989.

Murray, I.R., "Rule-based Emotion Synthesis using Concatenated Speech", Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, 5-7 September 2000, pp. 173-177.

Murray, I.R. and Arnott, J.L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", Journal of the Acoustical Society of America, 93(2), 1993, pp. 1097-1108.

Murray, I.R. and Arnott, J.L., "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", Speech Communication, 16, 1995, pp. 369-390.

Murray, I.R. and Edgington, M.D., "Emotion in the BT Laureate speech synthesis system", Proceedings of the Institute of Acoustics, 20(6), pp. 263-270, 1998.

Ortony, A. and Turner, T.J., "What's Basic about Basic Emotions?", Psychological Review, 97(3), pp. 315-331, 1990.

O'Sullivan, M., Ekman, P., Friesen, W. and Scherer, K, "What You Say and How You Say It: The Contribution of Speech Content and Voice Quality to Judgements of Others", Journal of Personality and Social Psychology, 48, 1985, pp. 54-62.

Picard, R.W., "Affective Computing", MIT Press, 2000.

Raymond, N., "Emotional Speech In Men And Women", BSc dissertation, University of Leeds, 1993.

Roach, P., "Introducing Phonetics", Penguin, 1992.

Roach, P., "Techniques for the phonetic description of emotional speech", Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, September 2000, pp. 53-59.

Russell, J. A., "A Circumplex Model of Affect", Journal of Personality and Social Psychology, 39, 1980, pp. 1161-1178.

Schachter, S. and Singer, J.E., "Cognitive, Social and Physiological Determinants of Emotional State", Psychological Review, 69(5), 1962, pp. 379-399.

Schröder, M., "Emotional Speech Synthesis: A Review", Proceedings of Eurospeech 2001, Aalborg, Denmark, 2001.

Schröder, M. and Trouvain, J., "The German Text-to-Speech System MARY: A Tool for Research, Development and Testing", Proceedings of the 4th ISCA Workshop on Speech Synthesis, Blair Atholl, UK, 2001.

P. Sprent, Applied Nonparametric Statistical Methods. London: Chapman and Hall, 1989.

Tolkmitt, F.J. and Scherer, K.R., "Effect of Experimentally Induced Stress on Vocal Parameters", Journal Of Experimental Psychology: Human Perception And Performance, 12(3), 1986, pp. 302-313.

van Bezooijen, R. A. M. G., Characteristics and Recognizability of Vocal Expressions of Emotion, Foris Publications, Dordrecht, 1984.

Vroomen, J., Collier, R. and Mozziconacci, S., "Duration and intonation in emotional speech", Proceedings of Eurospeech '93, Berlin, Germany, 1993, pp. 577-580.

Williams, C.E. and Stevens, K.N., "Emotions and Speech: Some Acoustic Correlates", Journal Of The Acoustical Society Of America, 52(4(2)), 1972, pp. 1238-1250.