Proceedings of the Institute of Acoustics

EMOTION IN THE BT LAUREATE SPEECH SYNTHESIS SYSTEM

Iain R. Murray (1) and Mike D. Edgington (2)

(1) Department of Applied Computing, The University, Dundee DD1 4HN.
(2) Speech Technology Unit, BT Labs, Ipswich IP5 3RE.

1. INTRODUCTION

Commercial speech synthesis systems have tended to be based either on formant synthesis or resynthesis by concatenation of recorded speech units, typically diphones. Formant synthesis was initially the most widespread due to the greater cost and complexity of concatenative systems, but more recently the reduction in costs and improvements in quality of concatenative systems have seen the latter rise in popularity. Another reason for this has been the more natural voice quality possible in concatenative systems (this being the voice quality of the original speaker of the speech segments), compared to the more robotic rule-based formant synthesisers.

However, although fundamentally less natural sounding, formant synthesisers offer a wider range of control over the voice during synthesis, allowing greater flexibility in the manipulation of the output voice. Thus, when attempting to add pragmatic effects such as different speaking styles and emotions into synthetic speech, formant synthesisers have typically been used in the small number of research projects in this area. These include the University of Dundee's HAMLET system [1] and M.I.T.'s Affect Editor [2] (which both use the DECtalk synthesiser), University of Essex/University of Bristol's SPRUCE system [3] and the system by K.T.H. in Sweden [4].

With increasing availability of concatenative synthesisers, it appeared timely to investigate synthesis of emotions using such a system, and this paper describes the results of a preliminary attempt to do this. The project used BT's Laureate text-to-speech (TTS) synthesis system [5], developed at BT Laboratories, Ipswich. Laureate works by concatenation of pre-recorded speech segments, mostly triphones and diphones; the concatenation is performed intelligently using techniques similar to PSOLA (Pitch Synchronous OverLap and Add) [6]. A new intonation contour (including both pitch and duration changes) is applied to the concatenated segments during production of the final audible utterance. In overall operation, the system operates in a similar fashion to other TTS systems, finally outputting a file of digital samples (in WAV or alternate format) which can be played to hear the synthesised speech.

Use of resynthesised speech in connection with emotion has tended to be for performing manipulations of speech recordings for later experimental comparison by listeners. Elson [7] performed some initial work on generating emotive speech with Laureate. For this work, Laureate pitch contours were copied from recordings of human emotional speech; results showed that the synthesised emotions were much less recognisable than the emotions in the original human speech.

As the Laureate speech is created from a database of concatenated recorded speech units, it might be possible to create some new units specifically for emotion expression. The process of recording new units and making them available to Laureate would be an involved one, although the recording segmentation can be performed largely automatically. However, it is not practical to create whole new databases for each emotion required, and there is little evidence that having an expanded database to take account of multiple emotions would be worth the effort of specifying and creating it.

Thus, the aim of the project described in this paper was to investigate the efficacy of generating emotional speech using BT's Laureate system, with the following specific objectives:

- investigation of the Laureate system with regard to emotion implementation
- development of a methodology for emotion implementation using Laureate
- use of this methodology for manual production of demonstration phrases
- recommendations for future developments of the system, including production of emotional phrases by rule

2. VOICE CONTROL PARAMETERS - DIFFERENCES AND SIMILARITIES

As a starting point to possible voice quality features in Laureate which might be altered to convey emotional affect, a review of the DECtalk parameters (especially those used by HAMLET) was undertaken. These parameters can be considered in three groups:

(a) parameters unique to DECtalk (or which worked in a totally different way in Laureate), and thus there was no useful application (e.g. sex parameter)
(b) parameters which had no direct parallel in Laureate, but whose effects could be altered via indirect manipulation of the available Laureate parameters (e.g. head size)
(c) parameters which were very similar to existing Laureate parameters (e.g. average pitch)

Of the 31 DECtalk parameters, 10 (mostly intrasegmental voice controls) were in group (a), 12 (mostly suprasegmental voice controls) were in group (b), and 9 (mostly intonational controls) were in group (c). Some of the DECtalk parameters are important for emotional implementation in HAMLET, while some parameters are not altered by HAMLET at all and thus did not need to be considered here.

Thus, despite the very different speech engines in the two synthesisers, there were a considerable number of overlaps in the control parameter set, offering a reasonable starting point for the emotion implementation using Laureate.

2.1 Standard Intonation Contours

The default intonation contour produced by Laureate is derived from the prosodic manipulation component, and implemented during the realisation phase of synthesis. This default contour and other intonational features can be controlled by altering various configuration flags (parameters) either through the synthesiser's command line interface or by editing the configuration file and re-initialising the synthesiser. The available intonation parameters are as shown in Figure 1.

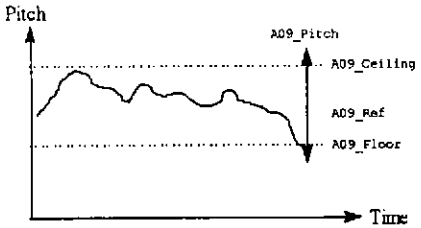

FIGURE 1 - General Laureate pitch contour and control variables

The Ref parameter is in effect an average pitch value, and the Ceiling controls the pitch range of the utterance. The main Pitch parameter shifts the entire contour up or down by a number of semitones. It

was found that by using extreme values of Pitch, certain secondary voice quality changes could also be produced. The Speak_Rate parameter also produced some secondary voice quality changes at low values. Laureate also provides some input flags for features such as accent and focus, but these are not of direct relevance to emotion implementation (most of the features controlled would actually be emotion independent and therefore specified elsewhere in the TTS process) and their use was not investigated.

2.2 Precise Contour Manipulation

The standard version of Laureate does not allow user manipulation of the pitch contour within an utterance, other than through the general parameters described above. However, a special version which did allow the pitch of individual segments to be specified by the user was produced to support the work reported here.

In order to implement emotional contours, Elson [7] recorded actors speaking the same phrases using a range of emotions, and by analysing the contours used, obtained values for copy synthesis using Laureate. These emotion contours, though designed to be realistic, did not appear so from the user testing of the synthesised phrases. These findings, and other aspects of emotion related to Laureate, were outlined by Edgington [8].

3. LAUREATE EXPERIMENTATION

A subset of four emotions was selected for prototype development: anger, happiness, sadness and fear. These were selected as they utilise different changes in vocal parameters; anger and happiness show increased activity in the pitch contour (both in level and variation), anger also showing noticeable voice quality changes, sadness shows decreased activity in the pitch contour (both in level and variation), and fear shows little change in the pitch contour (other than level) being more dependent on voice quality changes. Two other emotions simulated by the HAMLET system were not selected for implementation here - grief is generally similar to sadness, and disgust is principally dependent on detailed prosodic changes which were not easy to implement manually.

For the experimentation process, a PC workstation (running Windows 95) was used for system control, file storage and manipulation; the Laureate speech engine was running on a Unix server accessed remotely from the PC.

The process of incorporating emotion control into Laureate speech was carried out in three stages:

(a) **Standard parameters:** investigating manipulations of the existing Laureate system (using the HAMLET parameters to guide the initial Laureate parameter settings); parameter settings and corresponding WAV outputs to be recorded;

(b) **Post-processing:** performing manipulations on the WAV files produced in (a) to investigate if the emotional affect can be improved in this way, and;

(c) **Modifying Laureate:** altering the code within Laureate itself to allow additional effects to be added; further WAV manipulations to also be applied to speech produced by the modified system, if appropriate.

During the experimentation process, any utterances created which were of particular interest or which demonstrated particular voice features were recorded as WAV files and the corresponding parameters which produced them were noted.

3.1 Standard Parameters

Using HAMLET, default parameters (voice quality and intonation) were noted for the default set of six emotions. These were translated into approximately corresponding values of the five main Laureate parameters; these values were used as the starting parameters for the Laureate prototype set of four emotions (additional relevant information about vocal emotion taken from [9] was also incorporated). From these starting values, each parameter was varied systematically, and interesting new voices derived were noted. Some Laureate parameter settings appeared to give rise indirectly to voice quality-type changes, and this was exploited; for example, a low pitch value was found to produce a voice which sounds slightly creaky.

During experimentation, Sound Recorder windows were used to store and replay neutral speech and the "latest best" version, allowing these to be easily compared with the results of ongoing parameter modifications (in the Unix Laureate window). The parameters were modified by editing the Laureate configuration file, rather than being altered directly in the Laureate window.

In general, the initial speed and contour shapes (derived from HAMLET) produced reasonable effects in Laureate, although further heuristic changes to the main contour led to more satisfactory contours for the four emotions. The main changes implemented were as follows:

- Anger has a raised pitch contour and pitch range. Ref and Ceiling have been made artificially high followed by downward correction by Pitch to introduce some "laryngealisation". Speech rate is increased.
- Happiness also has a raised pitch contour. This is further raised using Pitch to give a high-pitched exciting-sounding voice, though this is restrained to avoid sounding "squeaky". Speech rate is slightly increased.
- Sadness has a lowered pitch contour, further lowered using Pitch again to give a bit of "laryngealisation" in the voice. Speech rate is slowed down, though not by too much.
- Fear is similar to happiness, with an even more raised pitch contour verging on "squeaky" to give a "panicky" voice. Speech rate is very much increased.

The emotion voices produced at this stage probably represent the closest approach to these emotions that could easily be achieved by manipulation of the small Laureate parameter set.

3.2 Post-Processing

To add further affective effects, the output WAV files produced from the parameter manipulation phase were modified using a waveform editing package (GoldWave 3.24). The main effects incorporated by editing the WAV files were:

- **silence**: periods of silence could be removed (to reduce utterance duration) or inserted (to simulate pauses)
- **gain control**: the amplitude could be varied, either over the whole utterance, or over particular sounds; this simulated intensity changes
- **equalisation**: the "bass" and "treble" frequencies were filtered or boosting; this simulated changes in smoothness and richness
- **vibrato and tremolo**: rapid pitch variations (vibrato) and rapid amplitude variations (tremolo) could be added using preset or user-defined effects functions in GoldWave
- **breath noise**: recorded (real) breath noise was mixed in to simulate breathy voice quality, but this was not successful
- **others**: other effects options were experimented with, but none gave useful affective changes

These effects are added partly to simulate effects which could be added at earlier stages within the synthesis process, but which are not possible without making major modifications to the speech engine. Other effects which contribute usefully to the affect could be implemented in due course by additional processing either within Laureate or by performing automatic batch post-processing on the Laureate output. All of the effects are quantifiable, and could have their values altered if required, including the magnitude of the vibrato and tremolo effects. The range of effects implemented was again guided by experience with HAMLET and observations from [9].

The outcome of this stage was as follows:

- For anger, the main perceptible change is the slightly perturbed voice quality, similar to added laryngealisation.
- For happiness, all changes made are to the amplitude, and there is little overall perceptible change in the voice.
- For sadness, the reduced amplitude and added pauses are clearly perceptible.
- For fear, the vibrato and tremolo have created a very disturbed voice.

The post-processing effects implemented have added a number of useful affective features to the speech, although some were clearly more affective than others. Sadness and fear both benefit substantially from the post-processing, and both are reasonably convincing. There is a perceptible change in the voice quality of anger, and a reasonable affect results. Happiness is the least satisfactory, with the main perception being a raised pitch overall.

4. IMPROVED LAUREATE

4.1 Adding Phoneme-level Variations

All work conducted up to this stage had been performed using the standard Laureate system. However, to assist with the incorporation of emotion features, a custom version was produced, with the following additional features:

- ability to specify the duration of phonemes individually, to allow non-linear changes in the speaking rate
- ability to specify the pitch of phonemes individually, to allow detailed pitch contours to be created
- ability to specify pauses to be added into the speech

To permit these changes, Laureate was altered to accept as input a list specifying phonemes and associated parameters, rather than normal text. The list was a standard ASCII text file, an example of which is shown in Figure 2.

```
#:    50    10    148
j    149    10    148
u    148    10    173
h     20    10    157
{     20    10    157
v     43    10    155
A     79    10    182
```

FIGURE 2 – Example of a PHO file

On each line of the file is included the required SAMPA phoneme, its duration (in ms), the point at which the target pitch is to be reached (as a % of the phoneme length), and the target pitch in Hz. This utterance specification file format is similar to that used by the MBROLA synthesiser [10], and the same file extension (.PHO) was adopted for the Laureate files.

The use of the specification files meant that every phoneme could be controlled precisely by specifying its required pitch and duration by editing the PHO file. If required, phonemes could also be changed, inserted (including silences) or deleted by editing the file.

To test the operation of Laureate with the PHO file format and that the specified parameters were being produced correctly by the synthesiser, several song PHO files were written. Though of limited musical excellence, these at least demonstrated (by virtue of being "in tune") that pitch targets were indeed being met correctly.

4.2 Emotion Implementation

The initial procedure employed in modifying the Laureate output was to create emotional speech using HAMLET, note the phoneme, pitch and duration values, then enter these directly into the Laureate PHO file. This process was performed for one of the test phrases for all four emotions. However, the results were unsatisfactory; this was believed to be due to the difference between the default pitch contours used by HAMLET and Laureate.

However, as the nature of the intonational emotion changes implemented by HAMLET were derived from the vocal emotion literature, it was clear that the same effects were still appropriate to implement in Laureate. However, it was clear that it would be necessary to apply the changes to the Laureate default phoneme values rather than to those used by HAMLET (which were derived from the DECtalk default phoneme values).

Thus, the technique adopted was to characterise the emotional changes made by HAMLET, and then implement similar changes in the default Laureate phoneme values. For each emotion, each of the relevant HAMLET rules was then taken in turn and its effects manually calculated and this data was added into the PHO file derived from unemotional Laureate speech, with some heuristic adjustments applied also. It was noted that as each rule was implemented, the subjective perceptual evaluation of the emotion improved (i.e. none of the rules caused undesirable effects), suggesting that the rules are indeed universal, or at least synthesiser independent. An example of the pitch contours is shown in Figure 3.

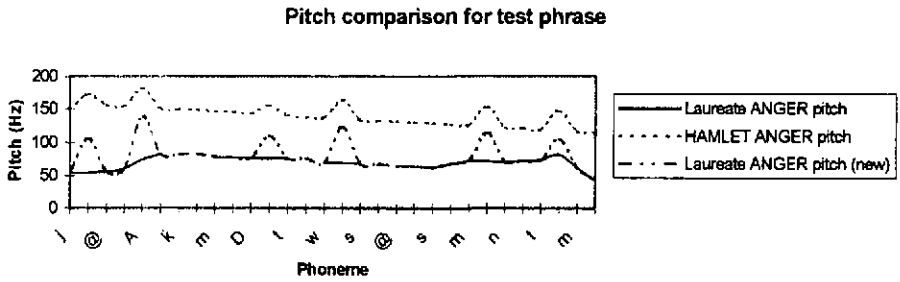

FIGURE 3 – HAMLET / Laureate pitch contour example for Anger

From the final versions of the PHO files, WAV files of the Laureate speech were created. These were then subject to the same wave editing regime which had been developed previously, thus producing final versions of the four emotive phrases.

Using the manual procedure outlined here, it was estimated that the total time to implement an emotional utterance from a new phrase was approximately 2½ - 3 hours.

5. DISCUSSION

5.1 Assessment Of Sample Utterances

The final WAV files produced appear to be affective, and do (subjectively) sound recognisably like the emotions which they are intended to portray, and were clearly differentiable. The addition of pitch and duration effects has a noticeable effect in improving the affective content. The wave editing performed also contributes to the overall affect, though possibly not as noticeably as when no pitch and duration effects were included.

The affective content of the demonstration utterances is probably of a similar order of emotional realism as the emotions produced by the HAMLET system, and both systems may offer about the best emotional affects possible with their respective synthesis engines.

Due to the time constraints on the project, and the fact that the demonstration utterances were produced manually, only ten utterances (two phrases in each of five emotions, including neutral) were produced during the project, and no formal subjective testing of the affective phrases was possible.

5.2 Further Work

This project has shown that it is possible to alter Laureate speech to produce affect. However, the examples have been produced with a large amount of manual processing, and it would be highly desirable to automate as much as possible of the procedure. This could be done via a suitable scripting language, program coding, or based on spreadsheeting the emotion rules. This might even be further extended to give other stress and speaking style controls.

To this end, further work on the Laureate-with-emotion system has been conducted [11] in order to examine automatic implementation of emotion rules. This work has demonstrated some automation of the process, altering some of the Laureate parameters automatically, and further work is needed to fully automate the process to the level already demonstrated by manual means.

6. CONCLUSION

This project had the aim of demonstrating whether the Laureate concatenative synthesis system could produce speech with affective content. Subjective appraisal of the various emotional speech utterances produced suggest that reasonably convincing examples of the four emotions have been produced. Further work is being undertaken to automate the emotion implementation process for Laureate speech.

7. ACKNOWLEDGMENTS

The work described in this paper was carried out under a BT Short-term Research Fellowship in 1997.

8. USEFUL LINKS

Audio samples of the phrases developed in the project are available at:
`http://alpha.mic.dundee.ac.uk/~irmurray/hamlet.html`
Information about Laureate is available at:
`http://innovate.bt.com/showcase/laureate/index.htm`
Information about BT's Short-term Research Fellowship scheme is available at:
`http://www.labs.bt.com/recruitment/fellow/index.htm`

9. REFERENCES

[1] I.R. Murray and J.L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", *Speech Communication*, **16(4)**, 1995, pp. 369-390.

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

[3] E. Lewis and M. Tatham, "High Specification Text-to-Speech Synthesis for Information Systems", *Digest of the Second Language Engineering Convention*, London, October 1995, pp. 145-152.

[4] R. Carlson, "Synthesis: Modelling Variability and Constraints", *Speech Communication*, **11(2-3)**, 1992, pp. 159-166.

[5] J.H. Page and A.P. Breen, "The Laureate text-to-speech system - architecture and applications", *BT Technology Journal*, **14(1)**, 1996, 11 pp.

[6] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech", *Speech Communication*, **16(2)**, 1995, pp. 175-205.

[7] B.P. Elson, "Synthesis of vocal emotion using the Laureate TTS system", *4th Year MEng report*, University of York, 1997.

[8] M.D. Edgington, "Investigating the limitations of concatenative synthesis", *Proceedings of Eurospeech '97*, Rhodes, Greece, September 1997.

[9] I.R. Murray and J.L. Arnott, "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.

[10] `http://tcts.fpms.ac.be/synthesis/mbrola.html`

[11] Campion, D., "Emotional Speech Synthesis", *MSc Thesis*, University of Dundee, 1998.