Emotion in Speech Synthesis

Christopher Hoult

May 6, 2004

Abstract

This paper is intended to give a general overview of efforts to simulate emotion in synthetic speech in order to produce results closer to actual human speech. An introduction to the field is presented prior to a literature review covering a number of papers on the use of both voice quality and prosody in synthesizing affect. A short discussion is then presented in which the discussed papers are aggregated and conclusions on their results drawn. Of note is the indication that more positive emotions, such as *happiness* or *joy*, are not synthesized as successfully as negative emotions. Also, doubt is thrown on the usefulness and accuracy of some of the results obtained.

1 Introduction

Human communication consists of more than just words - Mehrabian determined that, on average, words only account for 7% of the meaning a listener derives from a conversation [1]. Paralinguistic information, such as prosody and voice quality, accounts for 38%, with a further 55% contributed by other, non-verbal communication. So, if the goal of speech synthesis is to mimic spoken language in its entirety, work must be performed on more than just the intelligibility of its output.

In order to improve naturalness, much research is focussed on the simulation of emotion in synthetic speech. This involves investigation into the perception of emotion and how to effectively reproduce affect in the spoken word. As an abstract concept, emotion is not an easy to measure, or even to define, but this paper aims to describe some of the efforts in this area.

Perhaps the most obvious manner of producing emotive speech would be to model the physiological effects of emotion and mental state on the vocal tract, and thus produce accurate results. However, as Rank and Pirker point out, many speech synthesizers do not model the physical attributes of the vocal tract, as in articulatory synthesis, but instead construct utterances through the selection of phonemes or other units of speech (this is known as concatenative

synthesis), or model the tract's output by constructing a signal from a number of resonators, as in formant synthesis [2]. It is, perhaps, safe to say that no definitive, effective model of the vocal tract has yet been devised in terms of naturalness and emotion control.

2 Speech Synthesis Concepts

This section is intended to provide a brief overview of some of the concepts in speech synthesis, particularly with reference to the generation of affect in synthetic speech. While it is not comprehensive, enough information should be imparted to aid the casual reader in understanding some of the terms used in this paper.

2.1 Prosody

Prosody is essentially a collection of factors that control the pitch, duration and intensity to convey non-lexical and pragmatic information in speech [3]. A number of these factors are briefly explained below.

Fundamental frequency, or ***f0***, is the frequency at which the vocal folds vibrate, and is often perceived as the pitch of speech [3]. *f0* is important in the perception of emotion as it has strong effects in conveying stressed speech, but studies have shown it to be relatively ineffectual in producing affect when altered on its own [4]. It is generally split into two smaller measures, **mean *f0*** and ***f0* range**, although several more are also in common use [2].

Segmental duration is the term used to describe the length of speech segments such as **phonemes** (the basic units of language) and syllables, as well as silences. After *f0*, this is the most important factor in emphasis of words [3].

Amplitude, perceived as intensity or loudness in speech, although not as effective as *f0* and duration for the purposes of emphasis, can also be a useful indicator of emotional state in speech. It is important to note that relative, rather than absolute, amplitude is the indicating factor in most measures - clearly, a recording taken closer to the microphone would result in a higher amplitude, yet carry exactly the same affect as an identical utterance at a greater distance.

2.2 Voice Quality

Voice quality describes the fidelity and characteristics of speech, unrelated to prosody. Essentially, it is what distinguishes one individual from another. These characteristics are created from a variety of factors in the vocal folds, and run continuously throughout a person's speech. Classification of voice quality is often performed in terms of a set of identified voice qualities: *breathy*, *whispery*,

lax-creaky, *modal*, *harsh* and *tense* [4]. Another classification method for voice quality include Wendler's RBH system [5].

2.3 Formant Synthesis

Formant synthesis is based on a model of the **formants** produced in the vocal tract. While the number of these formants is infinite, not all are important for intelligibility, so in the interest of efficiency formant synthesizers generally only use five, labelled *f1* to *f5*.

Rules for how these **acoustic correlates** vary are applied and speech is produced without the need for the use of actual recordings. While the use of formant synthesis does not require the use of any database other than the rules to be applied, the advantages provided by its flexibility are often outweighed by the quality of its results - formant synthesis is often described by listeners as unnatural or 'robot-like' [6].

Examples of formant synthesizers include:

- **DECTalk** [7], originally developed by Dennis Klatt as **Klattalk**, later taken on by Digital Equipment Corporation, and currently owned and developed by Fonix.
- The multi-speaker formant synthesizer developed by Gutiérrez-Arriola et al. based on parameter concatenation.

2.4 Concatenative Synthesis

Concatenative speech synthesis is performed by combining small sections of recorded speech together, such as phonemes or diphones, to create words and thus phrases. These sections of speech are recorded by a human speaker and are generally monotonic to aid the combination process, which is carried out by employing an **overlap-add** (OLA) technique such as **multi-band resynthesis pitch-synchronous** OLA (MBR-PSOLA) or **time domain pitch-synchronous** OLA (TD-PSOLA) [8]. Variations in F0, segment duration and amplitude are also added in at this point.

While results from concatenative synthesis are generally impressive and relatively natural-sounding, large variations in prosody severely impair voice quality, yet again resulting in unnatural- or inconsistent-sounding results [9].

Examples of concatenative synthesizers include:

- The **Festival** system [10], developed at The Centre for Speech Technology Research in the University of Edinburgh.
- **CHATR** [11], developed at ATR Interpreting Telecommunications Research Laboratories, under the supervision of Nick Campbell.

3 Synthesis of Emotion

3.1 Voice Quality

Bulut et al. [12] set out to determine the relative importance of both voice quality and prosody in reproducing affected speech by creating a collection of eighty utterances of five declarative sentences through combinations of inventories and prosody rules for four different emotions - *angry*, *sad*, *happy* and *neutral*. Twenty actual recordings were also included, representing the five sentences spoken in each of the target emotions by a semi-professional actress. Qualitative assessment was carried out through a forced-choice test, carried out on 33 subjects, half of whom were native English speakers. Each subject was played all 100 utterances in random order, and were requested to select the most appropriate emotion for each, as well as rank on a scale of 1 to 5 the effectiveness in conveying their selected emotion.

The results from this testing showed strong results for the homogeneous pairings of voice quality and prosody rules, with all emotions bar *happy* testing at above 80% recognition rates. However, for *happy*, a recognition rate of less than 45% was recorded. Comparing this to the original recordings, *happy* tested significantly lower than the other three emotions, but was still testing at 67%, a difference of over 20 points. Other results from testing showed that utterances created with the *anger* inventory were more likely to be identified as such than other emotions, and utterances created with *sad* prosody were more likely to be identified accordingly.

Iida et al. describe the implementation of a Japanese system whereby sufferers from Amyotropic Lateral Sclerosis (ALS) who have lost the power of speech can construct emotional speech through the use of the CHATR unit-selection synthesizer with several corpora, each representing a different emotion [13]. Corpora for *joy*, *anger* and *sadness* were created by recording non-professional speakers reading a selection of texts covering the desired emotions. The user could then, by means of a GUI, enter the words to be synthesized, as well as select the affect of each section of text.

Initial testing was of a forced-choice type, carried out with 18 subjects, each given fifteen utterances (five semantically neutral sentences, produced using each corpora) - the resulting identification rates were significant improvements on chance (see Table 1). Further forced-choice tests, carried out with the target ALS sufferers, provided better results, albeit with a smaller test group: 66% for *joy*, 93.3% for *anger* and 86.6% for *sadness*.

Schröder and Grice [9] make the point that emotion is far from discrete and, as such, requires a much greater flexibility than provided by existing concatenative methods - namely, recording different inventories for each emotion, as seen

Speaker Gender	Joy	Anger	Sadness
Male	52%	51%	74%
Female	51%	60%	82%

Table 1: Perceptual test results [13]

previously ([13], [12], [2]). Instead, the approach taken is one of modelling the 'vocal correlates' of emotions. As a first step in this direction, three diphone sets, each representing a different level of vocal effort - *soft*, *modal* and *loud*, were created by recording a native male speaker of standard German, automatically labelled, hand-corrected and converted into the MBROLA format. Perceptual tests were then carried out to verify the two central hypotheses of the paper - I. all three diphone sets are identifiable as the same speaker and II. the vocal effort for each diphone set is perceived as intended.

To test hypothesis I, two sentences that shared a minimal number of phones, yet made sense together, were designed and synthesized using the three diphone sets created for the paper as well two further sets, each at two different pitch levels. Pairs of sentences in every combination of source were compiled, and each pair played to listeners, who were asked to identify whether they believed the two sentences to have been spoken by the same person. In 99.5% of all cases, pairs of sentences from the same diphone source at the same pitch were identified as the same person; with constant pitch, but altered vocal effort, 79.9% of pairs were correctly identified. Interestingly, only 45.5% of pairings differing only in pitch were correctly identified.

Hypothesis II was tested through a continuous scale assessment of the perceived vocal effort - listeners were asked to listen to each sentence produced from new diphone sets and rate them on a scale from 0 ('Without effort') to 100 ('With great effort'). The resulting scores showed that the *soft* stimuli were rated as having less vocal effort than *modal*, with a similar result for *modal* and *loud* - confirming the hypothesis.

Gobl et al. set out to determine the impact of voice quality on emotion in synthesis when combined with variation of *f0*, compared to that of *f0* manipulation on its own [4]. In the first experiment, six different voice qualities - *breathy*, *whispery*, *creaky*, *lax-creaky*, *tense* and *harsh* - were matched, as appropriate to combinations identified in previous experiments, with six different *f0* contours as described by Mozziconacci in an earlier paper - *indignation*, *anger*, *joy*, *fear*, *boredom* and *sadness*. The second experiment saw a *modal* voice quality combined with the six emotional *f0* contours listed, as well as a *neutral* contour.

Perceptual tests were carried out by playing the 13 stimuli, described above, to listeners, who were supplied with a series of pairs of emotions and their opposites

Emotion	Experiment 1	Experiment 2
Anger	95.2%	95.7%
Happy	61.9%	65.7%
Sad	81%	84.3%
Surprise	90.5%	52.9%
Neutral	76.2%	72.9%

Table 2: Perceptual test results [14]

(e.g. *bored* and *interested*) with seven boxes between them - the central box indicating both emotions as inappropriate to the utterance, and each box to either side indicating a greater strength of affect. The results were interesting - qualitatively, the voice quality + *f0* stimuli achieved better ratings from the test subjects than those varying only in *f0* contour. There were occasions, however, where the identification of target emotions was not as desired - for example, *boredom + lax-creaky* received a higher rating for *sad* than *sadness + breathy*. This indicates that there is still a lot to be learned concerning the association voice quality and *f0* contours with particular emotions. These results do indicate that, while voice quality has a strong influence on the identification of an emotion, *f0* is still important for their perceived strength.

Montero et al. [14] performed two experiments aimed at developing an emotional Spanish concatenative synthesizer using inventories for *sadness*, *happiness*, *anger*, *surprise* and neutral. Both experiments used a forced-choice test including a 'non-identifiable' option. Experiment 1 was a diphone-based copy synthesis from inventories recorded for each target emotion. Experiment 2 was an automatic synthesis which selected appropriate sections of text from large passages of recorded speech (as appropriate to the target emotion), making use of diphones from smaller recordings in order to create more flowing sentences. As can be seen in Table 2, the results were similar to those of other papers described. However, while the majority of emotions showed an improvement from the first experiment to the second - possibly due to longer sections of each utterance being taken from the inventory - surprise shows a remarked drop in recognition rate. This is probably due to the fact that the prosodic rules for surprise used were entirely new, and had not been fully tested.

3.2 Prosody

Cahn describes the development of the Affect Editor, a preprocessor that automatically marks up text for synthesis by the DECtalk formant synthesizer according to the desired emotions of the user [3]. Cahn lists a series of 17 factors, relating both to prosody and voice quality - although the changes in voice quality are in terms of type of voice (such as *creaky* or *breathy*), rather than related to a specific emotion, owing to the formant nature of DECtalk. An experiment was designed to test the hypothesis that the speech correlates

Emotion	Recognition rate
Angry	43.9%
Disgusted	42.1%
Glad	48.2%
Sad	91%
Scared	51.8%
Surprised	43.9%
All emotions	53.5%

Table 3: Emotion recognition test results [3]

Emotion	Recognition rate
Angry	65.5%
Disgusted	80.7%
Glad	83.2%
Sad	97.1%
Scared	72.7%
Surprised	72.7%
All emotions	78.7%

Table 4: Results after adjustment for closeness [3]

of emotion could be synthesized to such a degree that the emotion would be recognised by human listeners. Five sentences were each synthesized with the six chosen emotions - *angry*, *disgusted*, *glad*, *sad*, *scared* and *surprised* - with the settings for the Affect Editor chosen according to previous studies on the speech correlates of emotion. A forced-choice test was then conducted for each of the thirty stimuli, and listeners were also requested to mark, on a scale of 1 to 10, how much of the emotion was in each utterance, and how sure they were of their choices. Listeners were also given leave to add comments further to the forced-choice tests.

The results of the experiment are shown in Table 3, with all emotions being recognized at a much higher rate than chance. Of interest is the strength of recognition of *sad* as predicted by Cahn, due to its being 'among the most distinct in the set' of target emotions. *Anger* and *disgust* were often mistaken for each other; similarly for *gladness* and *surprise*. Initial results obtained from testing were adjusted to account for close matches due to strong semantic similarities between certain of the target emotions, as well as the consistent matching of certain emotions. Thus, a second set of results was created, and are presented in 4

Tartter proposes the hypothesis that, not only can a listener perceive whether a speaker is smiling or frowning in normal registers of speech (as proposed in a

	Normal Register	**Whisper Register**
Frown	62.9%	60.7%
Happy	57.6%	52.2%

Table 5: Recognition rates [14]

previous paper), but the same effect can be achieved in the whisper register [15]. This is relevant to the area of emotion in speech as the ability to imply smiling or frowning could conceivably go a long way to increasing the effectiveness and believability of synthesized emotion. Six native American English speakers were recorded effecting a neutral tone of voice whilst smiling and frowning and speaking in the normal and whisper registers. Pairs of syllables, both of neutral expression and frowning, and neutral and smiling, were then played to each of six listeners (all native speakers) who were asked which of the two syllables sounded happier, although in half of the frown samples, the listeners were asked which syllable of each pair sounded like it was frowned.

The results of the tests showed better-than-chance recognition rates of both smiled and frowned speech in both the normal and whisper registers - see Table 5 - and, as such, confirms the central hypothesis. Further analysis of the recorded syllables was performed, and it was determined that frowning lowers *f2* and increases syllable duration, whereas smiling raises *f2*.

Mozziconacci and Hermes [16] explore the role of intonation patterns in conveying emotion initially by labelling recordings of three native Dutch speakers in terms of a Dutch intonation grammar (created by 't Hart, Collier and Cohen). The distribution of these patterns across emotions was then analysed, before the results of this analysis were applied to synthesize speech with target emotions. Analysis of the distribution of patterns proved unremarkable, in that no specific pattern was closely associated with any one emotion - in fact, one pattern was frequently used in most emotional contexts. In the second part, a forced-choice perception test was carried out on two synthesized utterances produced using some of the more frequent combinations of patterns identified - the target emotions being *indignation*, *neutral*, *fear*, *sadness*, *anger*, *boredom* and *joy*. The results of these tests, though beyond the scope of this paper, indicate that, while there is no 'clear-cut' one-to-one relationship between particular patterns and emotions, some patterns are more effective in conveying some emotions than others.

Rank and Pirker [2] built on the work of Cahn [3] and applies it to an Austrian German concatenative synthesizer, varying prosodic information (such as *f0* contour, segmental duration of phonemes and spectral energy) as well as voice quality (by introducing noise to the signal). A demi-syllable inventory was prepared and stored as linear predictive coding (LPC) coefficients. Simple

Emotion	Recognition Rate
Anger	40%
Fear	17.8%
Sadness	68.9%
Disgust	22.2%

Table 6: Perceptual test results [2]

residual excited linear prediction (SRELP) synthesis was then used to synthesize five sentences each with four of the target emotions - *anger*, *sadness*, *fear* and *disgust*. These stimuli were then used for two perceptual tests. In the first of these tests, listeners were presented with a series of pairs of sentences, each pair consisting of the same sentence, but uttered in different target emotions, with the listener asked to rate, on a scale of 1 to 5, how well the emotion of the two utterances could be distinguished. In the second test, listeners were presented with 20 stimuli and were asked to identify the emotion of each in a forced-choice manner, and then to rate, again on a scale of 1 to 5, the effectiveness of the utterance in conveying the chosen emotion.

The results of the testing provides some interesting results regarding the distinctness of the chosen emotions. It was found that *disgust*, as defined by their source data on the emotion, was actually perceived as sadness in the majority of cases. Further to this, both *anger* and *fear* were frequently interpreted as disgust. In fact, the only notable success arising from testing was in *sadness* - see 6 for a breakdown of perception results. One likely factor in this lack of success (compared to other, similar experiments described in this paper) is not the method of synthesis, nor testing, but the choice of emotion to synthesize - the emotions chosen can all be classified as negative, whereas other comparable examples use a mixture of negative and positive with forced-choice testing, and as such allow the subjects a more clear-cut choice.

Murray et al. [17] implemented a proof-of-concept using a modified version of BT's Laureate synthesizer to show that rule-based concatenative methods could produce effective results in the synthesis of affect. The LAERTES (Language And Evaluation Research Tool for Emotional Speech) tool was used to provide initial parameters for the Laureate system, and the waveforms produced as a result were then edited by hand using a commercial waveform editor. Pilot testing showed that the hand-edited waveforms performed significantly better than both LAERTES and the HAMLET formant rule-based synthesizer, indicating that rule-based post-processing of concatenated speech would be effective as well as possible.

4 Discussion

When reading the works covered in this paper, a number of trends and shortcomings can be identified. The majority of perceptual experiments described use forced-choice testing for recognition rates for emotions, coupled with a selection of generally distinct emotions - the result being that the recognition rates reported are inflated due to the lack of choice and the 'black and white' nature of the emotions chosen. Indeed, it is notable that Rank and Pirker [2] chose to use only negative emotions, and, perhaps as a consequence, reports significantly lower recognition rates than others. This criticism of forced-choice testing is one also pointed out by Schröder [6], although the paper in question also points out the ease with which such tests may be conducted, compared with other forms of assessment.

Assessment methods aside, there is also an intriguing trend in results that sees positive emotions - such as *joy*, *happiness* and *surprise* - score significantly lower than negative ones. This might point to a number of factors - perhaps there are more subtle differences in the positive emotions, such as hinted at by Tartter [15], and so the understanding of the speech correlates with these emotions is less than that for the correlates of negative emotions; or perhaps there are fewer distinct characteristics of positive emotions compared to negative. Whatever the case, it certainly lends its weight to the proposition by Murray et al. that '... synthetic speech sounds depressing by default ...' [17].

Future work in the sphere of emotive speech should include investigating the problems regularly faced when synthesizing emotions such as *happiness* and *joy* and, as Schröder also proposes [6], investigation should be carried out into the effectiveness of forced-choice selection in testing where no emotion presented to subjects can be said to be similar to another.

While this paper cannot claim comprehensive coverage of emotion in speech synthesis, it can, at least, offer some insight into a number of efforts towards more natural-sounding speech synthesis and their relative success. What is apparent from the papers covered in previous sections is that there is still a long way to go until emotional speech synthesis approaches the capabilities of human speech. Apart from the topics covered in this paper, there are significant, unresolved issues being tackled, such as the continuous nature of emotion and its representation therein (for example, the concept of 'emotional dimensions' briefly discussed by Schröder [9]), or the inference of emotional context from text. These are just some of the barriers that stand in the way of an all-powerful, emotive speech synthesizer, and will probably not be overcome for a long time yet.

References

[1] A. Mehrabian, "Communication without words", *Psychology Today*, vol. 2, pp. 53–56, 1968.

[2] E. Rank and H. Pirker, "Generating emotional speech with a concatenative synthesizer", in *Proceedings, ICSLP '98, Sydney, Australia, Vol. 3*, November 1998, pp. 671–674.

[3] J. Cahn, "Generating expression in synthesized speech", Master's thesis, Massachusetts Institute of Technology, May 1989.

[4] C. Gobl, E. Bennett, and A. Ní Chasaide, "Expressive synthesis: How crucial is voice quality?", in *Proceedings, IEEE Workshop on Speech Synthesis, Santa Monica*, September 2002.

[5] T. Nawka, L.C. Anders, and Wendler J., "Die auditive beurteilung heiserer stimmen nach dem rbh-system", *Sprache Stimme Gehör*, vol. 18, pp. 130–133, 1994.

[6] M. Schröder, "Emotional speech synthesis", in *Proceedings, Eurospeech 2001, Aalborg, Denmark, Vol. 1*, September 2001, pp. 561–564.

[7] D. Klatt, *DecTalk user's manual*, Digital Equipment Corporation, 1990.

[8] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*, Taylor & Francis, 2nd edition, 2001.

[9] M. Schröoder and M. Grice, "Expressing vocal effort in concatenative synthesis", in *Proceedings, 15th International Congress of Phonetic Sciences, Barcelona, Spain*, June 2003.

[10] A. Black and P. Taylor, *The Festival Speech Synthesis System: system documentation*, Human Communications Research Centre, University of Edinburgh, Scotland, January 1997.

[11] A. Black, *CHATR, Version 0.8, A Generic Speech Synthesis*, ATR Interpreting Telecommunications Laboratories, Kyoto, Japan, March 1996.

[12] M. Bulut, S. Narayanan, and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer", in *Proceedings, ICSLP 2002, Denver, Colorado, USA, Vol. 2*, September 2002, pp. 1265–1269.

[13] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A speech synthesis system with emotion for assisting communication", in *Proceedings, ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland*, September 2000, pp. 167–172.

[14] J.M. Montero, J. Gutirrez-Arriola, J. Cols, J. Macas-Guarasa, E. Enrquez, and J.M. Pardo, "Development of an emotional speech synthesiser in spanish", in *Proceedings, 6th European Conference on Speech Communication and Technology*, 1999, pp. 2099–2102.

[15] V. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers", *Journal of the Acoustical Society of America*, vol. 96, pp. 2101–2107, October 1994.

[16] S. Mozziconacci and D. J. Hermes, "Role of intonation patterns in conveying emotion in speech", in *Proceedings, International Conference of Phonetic Sciences, San Francisco*, August 1999, pp. 2001–2004.

[17] I. Murray, M. Edgington, D. Campion, and J. Lynn, "Rule-based emotion synthesis using concatenated speech", in *Proceedings, ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland*, September 2000, pp. 167–172.

[18] J. Pierrehumbert, "Synthesizing intonation", *Journal of the Acoustical Society of America*, vol. 70, pp. 985–995, October 1981.

[19] S. Mozziconacci, "The expression of emotion considered in the frameword of an intonational model", in *Proceedings, ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland*, September 2000, pp. 45–52.

[20] J. M. Gutiérrez-Arriola, J.M. Montero, J.A. Vallejo, Córdoba R., R. San-Segundo, and J.M. Pardo, "A new multi-speaker formant synthesizer that applies voice conversion techniques", in *Proceedings, Eurospeech 2001, Aalborg, Denmark*, September 2001.

Typeset in LaTeX