# The Generation of Affect in Synthesized Speech

Janet E. Cahn
M.I.T. Media Technology Laboratory
20 Ames Street
Cambridge, MA 02139
*e–mail: cahn@media-lab.media.mit.edu*

**Abstract**

Synthesized speech need not be expressionless. By identifying the effects of emotion on speech and choosing an appropriate representation, the generation of affect is possible and can become computational. I describe a program — the Affect Editor — which implements an acoustical model of speech and generates synthesizer instructions to produce the desired affect. The authenticity of the affect is limited by synthesizer capabilities and by incomplete descriptions of the acoustical and perceptual phenomena. However, the results of an experiment show that this approach produces synthesized speech with recognizable, and, at times, natural, affect.

## Introduction

When compared to human speech, synthesized speech is distinguished by insufficient intelligibility, inappropriate prosody and inadequate expressiveness. These are serious drawbacks for conversational computer systems. Intelligible phonemes are essential for word recognition. Prosody — intonation (melody) and rhythm — clarifies syntax and semantics and aids in discourse flow control. Expressiveness, or affect, provides information about the speaker's mental state and intent beyond that revealed by word content.

My work explores improvements to the affective component of synthesized speech. It is implemented in the Affect Editor program, which takes an abstract description of emotional speech and produces affect–generation instructions for a speech synthesizer. Its success in generating recognizable affect was confirmed by an experiment in which the affect intended was perceived as such for the majority of presentations [Cahn (1989)].

Affect is desirable in synthesized speech for reasons of naturalness, efficiency and general utility. Hearers expect affect in speech. After all, it is part of human speech. It illuminates the intentions of the speaker and is part of the context in which an utterance is interpreted. Affective information in speech is primarily non–lexical. It can therefore be transmitted concurrently with the lexical content, making fuller use of the limited speech channel bandwidth. Finally, the addition of affect to synthesized speech is useful in any application in which expressiveness is appropriate — for example, in tools for the presentation of dramatic material, in information giving systems and in synthesizers used by the speech–handicapped.

## Modeling the Effects of Emotion on Speech

The generation of affect by the Affect Editor proceeds from a model in which the effects of emotion on speech are quantified. The subjective semantic aspect of emotion is ignored, although some researchers have posited a relationship between an emotion's semantic features (e.g., pleasant or unpleasant, strong or weak) and its acoustical correlates [Davitz (1964), Scherer (1974)]. Semantic models of emotion, independent of the speech correlates, will become important when automatic control of affect is key. However, at this early stage, the main task is still the completion of the acoustical model such that it contains the right parameters in the right relation to one another.

### The speech correlates of emotion

The speech correlates of emotion have been investigated by acoustics researchers and psychologists. Acoustics researchers studied the signal characteristics of speech generated from a variety of emotional states [Fairbanks (1940), Fairbanks & Pronovost (1939), Williams & Stevens (1969)]. Psychologists studied the responses of human subjects to emotional speech [Davitz (1964), Scherer (1974)]. Although these are disparate endeavors, they share common features. First, studies in both fields are few and occur sporadically over the course of years. More importantly, the findings agree on the speech correlates that are physiologically based, and are contradictory or unclear about effects that are more intentional, that is, those effects over which the speaker has the most control.

When emotion affects physiology the corresponding effects on speech show up primarily in the fundamental frequency (F0) and timing. Thus, with the arousal of the sympathetic nervous system — as with fear, anger or joy — heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. Speech is correspondingly loud, fast and enunciated, with strong high frequency energy. With the arousal of the parasympathetic nervous system — as with boredom or sadness — heart rate and blood pressure decrease and salivation increases, producing speech that is slow, low–pitched and with little high frequency energy [Williams & Stevens (1981)].

Psychoacoustical studies found that acoustically similar but semantically different emotions — e.g., anger and enthusiasm, boredom and sadness — were often mistaken for each other. Similarities in pitch range, average pitch, speech rate, timbre (high/low frequency energy ratio) and enunciation seemed to contribute most to mistakes in identification [Davitz (1964)].

Although the findings are not entirely consistent, they do agree on the basic acoustical effects of emotion on F0 and timing. Perceptual studies indicate that these are the main conveyers of affect. This is fortuitous, since F0 and timing features can be controlled with today's synthesizers. It is possible, then, to take a description of affect in speech and reproduce its significant acoustical features in synthesized speech.

## Choosing a representation

There are two distinct possibilities for representing the effect of emotion on speech. The first is generative, proceeding from a partial description of the speaker's mental state. Such a description should include those attitudes and intentions that affect physiology or determine the syntactic and semantic content of the utterance. The second representation is descriptive, specifying the acoustic signal as perceived by the listener.

Of the two, the generative (speaker) model is theoretically the preferred approach. However, the acoustical (listener) model is better for current purposes. It is simpler and requires a less complete understanding of speech production. Moreover, since perceptual parameters are explicit and quantified in the acoustical model, their effects can be directly manipulated to test perceptual responses, thereby improving the model. Because of its relative simplicity, and because more is known about the acoustical correlates of emotion than the speaker's cognitive representations and physiological responses, the acoustical model is the one incorporated into the Affect Editor.

## The acoustical model

The acoustical model is represented by a set of parameters corresponding to the speech correlates of emotion. Each parameter varies independently. This allows direct and individual control over parameter influence and supports the investigation of relationships among the parameters. For example, we might expect correlation among parameters influenced by physiology in accord with the observation that pitch range, speech rate, loudness, timbre and enunciation often vary together as affect changes [Davitz(1964)]. However, to avoid overgeneralizing, the model incorporates few assumptions about how the parameters interact.

### Parameters of the model

The parameters of the model are grouped into four categories — pitch, timing, voice quality and articulation. The pitch parameters describe features of F0. The timing parameters control rhythm — the combination of word stress and silence — and speech rate. Often pitch and timing parameters describe linguistic phenomena — features of words or phrases. In contrast, the voice quality parameters describe features of the speech signal as a whole.

Articulation parameters fall somewhere in between, describing features of phoneme articulation.

The distinctions among the four categories are not absolute. For example, *stress frequency*, a timing parameter, determines the number of peaks in the pitch contour. Similarly, variations in enunciation are achieved mainly by changes at the phoneme level, a feature of articulation, but also by variations in the relative strengths of high and low frequency energy, a feature of voice quality.

The parameters of each acoustical/perceptual category are discussed by category, as follows:

**Pitch parameters** *Accent shape, average pitch, contour slope, final lowering, pitch range* and *reference line* comprise the set of pitch parameters.

*Accent shape* describes the rate of F0 change for any pitch accent[1] in the utterance. Thus it describes the overall steepness or smoothness of the shape of the F0 contour at the site of a pitch accent.

*Average pitch* describes the average F0 for the utterance relative to the speaker's normal speaking pitch.

*Contour slope* describes the overall trend of the pitch range for the utterance — whether it expands, remains level or contracts.

*Final lowering* describes the terminal pitch contour — the rate and direction of F0 change at the end of an utterance. Whether it rises or falls is often a function of linguistics or pragmatics rather than of affect. For example, intent to continue speaking is typically conveyed with a rising terminal contour, regardless of affect.

*Pitch range* describes the bandwidth of the range bounded by the lowest and highest F0 for the utterance.

The *reference line* is a term borrowed from work on generative intonation [Anderson & Pierrehumbert & Liberman (1984)]. It specifies the F0 to which the pitch contour appears to return following a high or low pitch excursion.

**Timing parameters** *Exaggeration, fluent pauses, hesitation pauses, speech rate* and *stress frequency* comprise the set of timing parameters.

*Exaggeration* describes the degree to which pitch accented words receive exaggerated duration as a means of emphasis. [2]

[1] A *pitch accent* is distinctive pitch — high or low — applied to the lexically stressed syllable of a word such that the word as a whole is perceived as receiving sentential stress.

[2] The implementation of the *exaggeration* parameter introduced unwanted side effects in the speech, and

*Fluent pauses* describes the frequency of pausing between syntactic or semantic units.

*Hesitation pauses* describes the frequency of pausing within a syntactic or semantic unit. These pauses often occur after the first function word [3] in a clause [Dittmann (1974)].

*Speech rate* describes the rate of speech. It affects the number of syllables or words spoken per minute and the duration of pauses.

*Stress frequency* describes the ratio of stressed to stressable (i.e., pitch accented) words in an utterance. To the Affect Editor, stressable words may legitimately receive a pitch accent in accord with sentence semantics. Stressed words are those stressable words which actually do receive pitch accents. The greater the *stress frequency* value, the more stressable words will become stressed. However, words that are not considered stressable — usually, function words — will never receive distinctive pitch, regardless of the value of *stress frequency.*

The *stress frequency* parameter operates on words. It requires an analysis of the likelihood that a word will be stressed, as determined from syntax, semantics and pragmatics. Thus, a content word is more likely to receive stress than a function word and new information more likely to receive stress than information already mentioned.

**Voice quality parameters** *Breathiness, brilliance, loudness, pause discontinuity, pitch discontinuity* and *tremor* comprise the set of voice quality parameters. Most of these parameters except, perhaps, *brilliance* and *pitch discontinuity*, convey speaker identity as much as affect.

*Breathiness* describes the amount of frication noise that may be co–present with non–fricative phonemes (vowels, for example).

*Brilliance* describes the ratio of low to high frequency energy. A high value for this parameter indicates strong high frequency energy.

*Laryngealization* describes the creaky voice phenomena in which there is minimal subglottal pressure, a small open quotient, a narrow glottal pulse and an irregular fundamental period. Laryngealization typically correlates with speaker identity as much as with speaker emotion. The speech of older speakers is often laryngealized.

*Loudness* describes perceived loudness, a result of subglottal pressure, and therefore, the perceptual response to the amplitude of the speech signal.

*Pause discontinuity* describes the smoothness or abruptness of a pause onset. It was included

so is excluded from the current version of the Affect Editor.

[3] Function words convey primarily structural rather than semantic information. They comprise a minuscule part of most vocabularies and are rarely added to or dropped from the lexicon. Pronouns, prepositions and determiners are function words. Content words convey primarily semantic information. Their meanings may change, and they may be added to or dropped from the lexicon with relative haste. Nouns, verbs, adverbs and adjectives are usually classed as content words.

to compensate for a synthesizer introduced side effect — the abrupt cessation of phonation caused by the silence phoneme.

*Pitch discontinuity* describes the smoothness or abruptness of F0 transitions throughout the utterance, the result of more or less motor control on the part of the speaker.

*Tremor* or vocal jitter refers to irregularities between successive glottal pulses. It was observed in recordings of fearful utterances [Williams & Stevens (1972)].[4]

**Articulation parameter(s)** The sole articulation parameter is *precision*, which describes the degree of slurring or enunciation for all phoneme classes.

### Parameter values

To represent the effect of specific emotions, the parameters of the model are quantified. The amount of parameter influence on speech varies according to the emotion. Parameters are quantified on a scale centered at zero and whose values range from negative ten to ten. Zero represents the parameter influence for neutral affect, while negative ten and ten represent, respectively, the minimum and maximum influence. The effect of changing a parameter value may vary depending upon whether it is above or below zero. For example, there is little laryngealization for affectively neutral speech. Thus, the difference between the effect of no laryngealization — at negative ten — and laryngealization for neutral affect — at zero — is minimal while the difference between laryngealization at positive ten and zero is significant.

The quantification of parameter influence allows precise control over the generation and modification of affect in speech. It also supports the correlation of perceptual effects and thresholds with quantities of the model. Positioning the effects of neutral affect at the mid–range allows the straightforward implementation of descriptive quantifiers such as *more* or *less*. *More* of an affective coloration is effected by moving parameter values further away from zero (neutral affect), *less* by moving them closer. Thus, this approach provides a basis for the eventual automation of affect generation for synthesized speech.

## The Affect Editor program

The Affect Editor program implements a transfer function from an acoustical description of emotional speech to synthesized expressive speech. It is a tool for designing expressive speech and for investigating the perceptual responses to the various speech correlates of emotion. Given an emotion and an utterance, it produces output which is sent to the

[4] Tremor cannot be produced by the DECtalk3 so is not yet implemented.

synthesizer (currently a DECtalk3) to produce expressive speech. This section describes the Affect Editor input, output and flow of control.

## Input

The Affect Editor takes as input an emotion and an utterance. It represents the emotion as a set of speech correlates whose quantities guide the processing of the utterance. This representation has already been described, so the remainder of this section describes the utterance.

An Affect Editor utterance is a set of clauses, each distinguished by syntactic (thematic) or semantic role, e.g.,

> [S [[AGENT I] [ACTION saw [OBJECT your name]]] [LOCATIVE in the paper]]

where the clause divisions are justified by the Sentence, AGENT, ACTION, OBJECT and LOCATIVE classifications. The clauses are arranged in a tree structure, simulating the result of a semantic analysis in which the relations between the main and subsidiary clauses are apparent from structure. Each clause plays a particular thematic or pragmatic role as per a case frame analysis. Prosodic annotations to a clause (e.g., pitch range, speech rate) reflect its semantics and syntax and therefore its role in the utterance. Prosodic annotations to a word (e.g., pitch accents) reflect its syntactic categorization as dictated by utterance semantics. Thus, in structure and content, the utterance input required by the Affect Editor simulates the output of a text generation program in which each clause is generated to fulfill a specific informational or discourse role.

The Affect Editor performs an initial analysis of the utterance to find all possible pitch accent and pause locations. Whether these possibilities are realized depends on the parameter values for the emotion, acting as a filter on the most extreme effects. Thus, from phrase structure and syntactic category, the Affect Editor identifies all possible hesitation and fluent pause locations. The *hesitation* and *fluent pause* parameters determine at which locations pauses are actually inserted into the spoken utterance. Similarly, the *stress frequency* parameter in combination with the pitch accent probability information[5] determines how many and which words will receive pitch accents.

## Output

The Affect Editor produces instructions that enable a synthesizer to speak an utterance with the specified affect. For the DECtalk3 — the only synthesizer used so far — the Affect Editor constructs two strings. One sets the synthesizer parameters that control features

[5] The likelihood that a word will receive a pitch accent depends on how central it is to the meaning of the utterance. Changing the pitch accenting probabilities may well change the interpretation of an utterance.

of prosody and voice quality. The second string is the utterance itself, a combination of English text, ARPAbet[6] phonemes, phoneme durations, pauses and intonation markings. Depending on the affect, this string may include pauses and modifications to word intonation and pronunciation.

### Program flow of control

The Affect Editor interprets the acoustical parameter values to produce lexical and non–lexical effects. It first sends the synthesizer instructions for producing non–lexically based effects (e.g., phrase features such as pitch range or speech rate, and voice quality effect) and then processes the utterance. As described previously, the initial utterance is an arrangement of one or more clauses, simulating the tree structured output of a text generation program. During the processing, the utterance becomes a linear phonology whose words and intonational phrases are marked with acoustical features as per parameter values. The acoustical features are then interpreted for the synthesizer, producing a synthesizer phonology in which all possible acoustical features are expressed as synthesizer specific instructions. These instructions (for the DECtalk3, a combination of text, phonemes, diacritics) are assembled in the appropriate order to form the utterance spoken by the synthesizer. Figure 1 illustrates the program flow of control and Figure 2 its expression in the Affect Editor interface.

### Synthesizer considerations and effects

Because the acoustical representation is synthesizer independent its parameters must be interpreted for each synthesizer it drives. The mapping of Affect Editor parameters to DECtalk3 capabilities involves both one–to–many and many–to–one mappings from the acoustical parameters to the synthesizer settings. The parameters not represented in the DECtalk's own parameter set are implemented in software where possible. Thus, a rising or falling contour slope is approximated by assigning a high F0 to the word at the end or beginning of the utterance; pauses are added by inserting a silence character; the quality of pause onset — smooth or abrupt — is effected by inserting phonemes prior to the silence; and precision of articulation is achieved by phoneme substitutions or additions.

The DECtalk3 was chosen for the scope and variety of its prosodic and voice quality controls. However, its limitations made it hard to determine whether an emotion had been poorly specified or correctly specified but poorly reproduced. The limitations are of two kinds — side effects and limited capabilities. For example, a side effect of specifying a word with phonemes instead of English text is that it is spoken with a lower F0. Another side effect is produced by word stress markings, which should cause F0 perturbations only for the word they mark. However, they sometimes affect the pitch contour for the entire utterance

[6] ARPAbet is a phonemic alphabet for English, developed as an ASCII approximation of the International Phonetic Alphabet symbols.

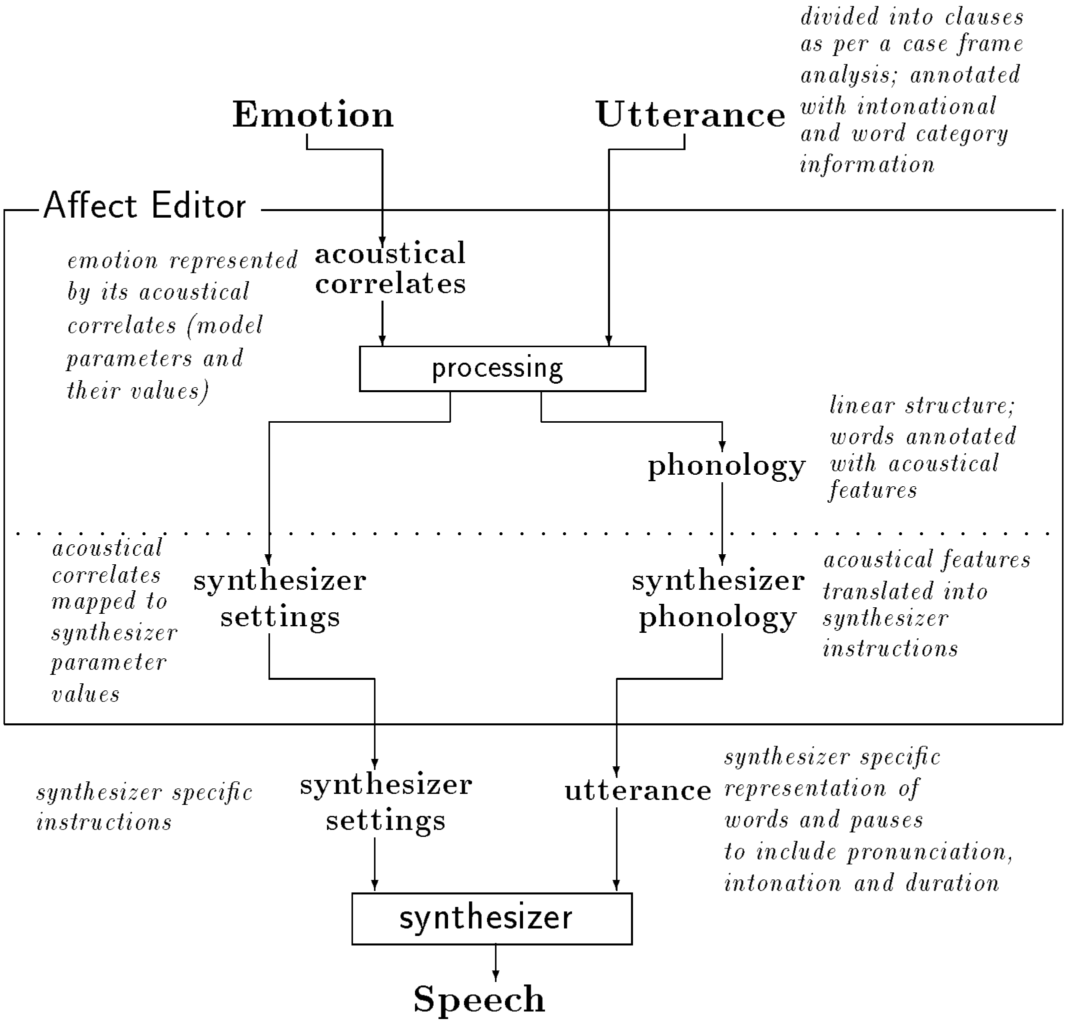
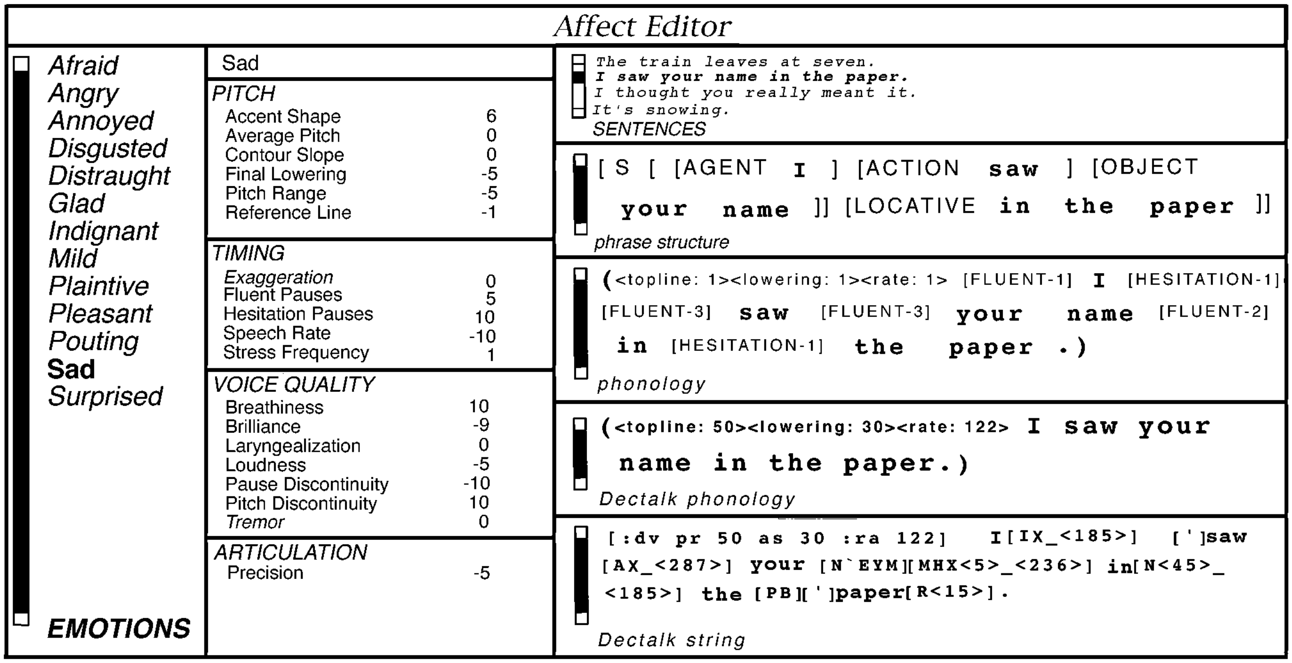

divided into clauses as per a case frame analysis; annotated with intonational and word category information
Emotion
Utterance
Affect Editor
emotion represented by its acoustical correlates (model parameters and their values)
acoustical correlates
processing
linear structure; words annotated with acoustical features
phonology
acoustical correlates mapped to synthesizer parameter values
synthesizer settings
synthesizer phonology
acoustical features translated into synthesizer instructions
synthesizer specific instructions
synthesizer settings
utterance
synthesizer specific representation of words and pauses to include pronunciation, intonation and duration
synthesizer
Speech

Figure 1: **The Affect Editor program flow of control, illustrated primarily by data structure creation and transformation.** The input is an emotion and an utterance. The emotion is represented internally by a set of quantified acoustical correlates. These determine the synthesizer settings and control prosodic and phonemic modifications to the original utterance. The synthesizer independent activity occurs above the dotted line and synthesizer dependent activity below.



Affect Editor

Afraid
Angry
Annoyed
Disgusted
Distraught
Glad
Indignant
Mild
Plaintive
Pleasant
Pouting
**Sad**
Surprised

**EMOTIONS**

Sad

| PITCH | |
|---|---|
| Accent Shape | 6 |
| Average Pitch | 0 |
| Contour Slope | 0 |
| Final Lowering | -5 |
| Pitch Range | -5 |
| Reference Line | -1 |

| TIMING | |
|---|---|
| *Exaggeration* | 0 |
| Fluent Pauses | 5 |
| Hesitation Pauses | 10 |
| Speech Rate | -10 |
| Stress Frequency | 1 |

| VOICE QUALITY | |
|---|---|
| Breathiness | 10 |
| Brilliance | -9 |
| Laryngealization | 0 |
| Loudness | -5 |
| Pause Discontinuity | -10 |
| Pitch Discontinuity | 10 |
| *Tremor* | 0 |

| ARTICULATION | |
|---|---|
| Precision | -5 |

*The train leaves at seven.*
***I saw your name in the paper.***
*I thought you really meant it.*
*It's snowing.*
*SENTENCES*

```
[ S [ [AGENT I ] [ACTION saw ] [OBJECT
 your name ]] [LOCATIVE in the paper ]]
```
*phrase structure*

```
(<topline: 1><lowering: 1><rate: 1> [FLUENT-1] I [HESITATION-1]
[FLUENT-3] saw [FLUENT-3] your name [FLUENT-2]
 in [HESITATION-1] the paper .)
```
*phonology*

```
(<topline: 50><lowering: 30><rate: 122> I saw your
 name in the paper.)
```
*Dectalk phonology*

```
[:dv pr 50 as 30 :ra 122]  I[IX_<185>]  [']saw
[AX_<287>] your [N`EYM][MHX<5>_<236>] in[N<45>_
<185>] the [PB][']paper[R<15>].
```
*Dectalk string*

Figure 2: **The Affect Editor user interface**. In this example, the sentence, *"I saw your name in the paper."* will be spoken with a *Sad* affect. The parameters of the model and their quantities appear in the middle column, labeled with the current emotion (*Sad*).

by preventing other stressed words from receiving stress. Lastly, changes to average pitch automatically affect the pitch range as well, such that one perceives a change of speaker rather than affect.

The *tremor* parameter could not be implemented because the synthesizer could not produce tremors in its output. The most limiting feature, however, was the synthesizer's inability to handle an instruction in which many ASCII characters specified a short–lived event. Too many pitch and duration instructions, for example, caused it to temporarily stop speaking. This, in part, prevented the implementation of the *exaggeration* parameter. It also prevented the implementation of precise word–by–word pitch contour control.

Some of the unwanted word related side effects can be overcome by implementing an intonational description system with primarily local effects, such as the two tone annotation developed by Pierrehumbert and colleagues [Pierrehumbert (1980), Liberman & Pierrehumbert (1981), Anderson & Pierrehumbert & Liberman (1984)]. The separation of pitch range and average pitch effects would allow greater F0 variation without affecting the perception of speaker identity. Synthesizer capabilities should be expanded, perhaps with the addition of features currently implemented in the Affect Editor software, particularly the ability to specify precision of articulation and overall pitch contour slope.

### Summary

The Affect Editor incorporates an acoustical/perceptual model of the effect of emotion on speech for the purposes of generating affect in synthesized speech and investigating how to generate better affect. Because work in this area is just beginning, the Affect Editor incorporates few assumptions about interrelations among parameters. However, it provides a foundation for exploring parameter influence, and thus, for automating the infusion of affect into synthesized speech.

## Experimental Verification

An experiment was performed to verify that the Affect Editor could produce recognizable affect. Subjects heard utterances produced by the Affect Editor and were asked to choose from six adjectives the one that best described the utterance.

### Equipment

The program that presented the stimuli and collected the responses ran on a Symbolics 3650 Lisp Machine. The synthesized speech was produced by a DECtalk3 and sent to a Sansui AU3900 amplifier. Subjects heard the speech through Koss KC–180 headphones or NEC

RS-500-R speakers, at their preference. The amplifier settings — bass, treble and balance — were set to their mid–points for all subjects.

## Stimuli

The subjects heard thirty utterances, combinations of five sentences and six affects — angry, disgusted, glad, sad, scared or surprised. The sentences were intended to be affectively neutral so that subjects would draw their conclusions from the non–lexical features of the utterance. However, in trial runs, subjects perceived as incongruous or meaningless affectively neutral utterances spoken with affect. The criteria were relaxed to require mainly that the sentences be plausible in each affective context. The subjects heard these sentences:

```
I'm almost finished.
I saw your name in the paper.
I thought you really meant it.
I'm going to the city.
Look at that picture.
```

The six emotions were selected because they were semantically distinct and often reflected acoustical or semantic extremes as well. Thus, the subjects' judgments were more likely to reflect the Affect Editor's performance than their own internal representations of emotion semantics, whereas judgments of semantically or acoustically indistinct emotions would more likely reflect individual biases.

The acoustical correlates for the emotions were culled from the research upon which the Affect Editor is based — the acoustical and perceptual descriptions of the effect of emotion on human speech [Fairbanks (1940), Fairbanks & Pronovost (1939), Williams & Stevens (1969), Davitz (1964), Scherer (1974)]. These were interpreted for the Affect Editor parameters (see Table 1) such that, for example, frequent pitch contour fluctuations were effected by a high value for the *stress frequency* parameter, a rising pitch contour by a high value for *contour slope* and slurred speech by a low value for *precision of articulation.*

## Subjects

Twenty–eight subjects participated in the experiment. Most were MIT students whose ages ranged from nineteen to thirty–five. There were nine women and nineteen men. The first language of twenty–four of the subjects was some form of General American English; the first language of the other four subjects was not English. Of the American English speakers, four were from New England, eight from the Mid-Atlantic states, six from the Midwest and three from the South and Southwest. The subjects were not paid.

## Method

Subjects heard synthesized speech and were asked to choose from among six adjectives the one best describing the affective quality of the speech. To compensate for the limitations of forced choice responses, subjects could optionally qualify their answers by answering the questions "How much?" (magnitude) and "How sure are you?" (certainty). Subjects could also type in comments to more fully explain their choices or describe their perceptions.

**Affect Editor parameter values**

| | *Angry* | *Disgusted* | *Glad* | *Sad* | *Scared* | *Surprised* |
|---|---|---|---|---|---|---|
| Accent shape | 10 | 0 | 10 | 6 | 10 | 5 |
| Average pitch | -5 | 0 | -3 | 0 | 10 | 0 |
| Contour slope | 0 | 0 | 5 | 0 | 10 | 10 |
| Final lowering | 10 | 0 | -4 | -5 | -10 | 0 |
| Pitch range | 10 | 3 | 10 | -5 | 10 | 8 |
| Reference line | -3 | 0 | -8 | -1 | 10 | -8 |
| Fluent pauses | -5 | 0 | -5 | 5 | -10 | -5 |
| Hesitation pauses | -7 | -10 | -8 | 10 | 10 | -10 |
| Speech rate | 8 | -3 | 2 | -10 | 10 | 4 |
| Stress frequency | 0 | 0 | 5 | 1 | 10 | 0 |
| Breathiness | -5 | 0 | -5 | 10 | 0 | 0 |
| Brilliance | 10 | 5 | -2 | -9 | 10 | -3 |
| Laryngealization | 0 | 0 | 0 | 0 | -10 | 0 |
| Loudness | 10 | 0 | 0 | -5 | 10 | 5 |
| Pause discontinuity | 10 | 0 | -10 | -10 | 10 | -10 |
| Pitch discontinuity | 3 | 10 | -10 | 10 | 10 | 5 |
| Precision of articulation | 5 | 7 | -3 | -5 | 0 | 0 |

Table 1: **The Affect Editor parameter values used to synthesize the affect stimuli in the experiment.** The descriptions in the acoustic and psychoacoustic literature were adapted for the Affect Editor.

The experiment took place in a large office. It proceeded as follows:

- The experimenter explained that the subject would hear synthesized utterances spoken with different emotional qualities and that the subject was to choose the emotion that best described the emotional quality with which the utterance was spoken.

- The experimenter explained the three judgment scales (affect descriptors, magnitude, certainty) and the comment facility.

- The experimenter explained the program interface and commands.

- The experimenter left the room.

- Using a mouse, the subject clicked on **START** to begin the experiment.
- To accustom the subjects to synthesized speech, the DECtalk3 spoke this paragraph:

  > *Hello. This is a perceptual experiment. There are no right or wrong answers. Just go by what you hear. I'll speak some sentences with varying emotional qualities. Click on the word that best describes the quality or emotion you hear. OK! Here is the first sentence.*

  with [relatively] neutral affect.

- Thirty synthesized utterances, unique combinations of six emotions and five sentences, were presented in one of nine random orders. The subjects could replay each utterance as many times as necessary before entering their judgment. The runs varied in duration from eight to twenty–eight minutes.

## Hypotheses

The null hypothesis predicted that each of the six affects would be recognizable only at the level of chance, at 17%. A recognition rate significantly above this would disprove the null hypothesis and prove its inverse, namely, that the intended affect was recognized at an incidence significantly greater than chance, and therefore that recognizable affect could be added to synthesized speech. More specifically, significant recognition rates would support the approach embodied in the Affect Editor — that the result of modeling the acoustical correlates of speech, allowing manipulation of the model parameters and mapping their effects to speech synthesizer capabilities parameters would be a system which produced recognizable affect in synthesized speech for a wide range of emotions. Based on the results of studies of the perception of human emotional speech [Davitz (1964)], I predicted that when the intended affect was not perceived, subjects would perceive instead an emotion with similar acoustical or semantic correlates.

## Results

Only the results of the forced choice data were analyzed. A chi–squared was computed from the data and found to be extremely significant. The obtained value, with five degrees of freedom, was 823.21. This is extremely significant at $p = .01$.

Each emotion was perceived for approximately 50% of its presentations — far above chance. Errors were not random, but followed the pattern of errors made in the identification of affect in human speech. Thus, sadness, with the most acoustically distinct features — soft, slow, halting speech with minimal high frequency energy — was the most recognizable. Emotions with similar acoustical features, such as gladness and surprise or anger and surprise, were often confused. Even more consistent and frequent substitution occurred for emotions with similar semantics, for example, between anger and disgust or gladness and surprise. The stimuli and the identifications they elicited are presented in Figure 3.

Except for sadness, with a 91% recognition rate, the intended emotions were recognized in approximately 50% of the presentations and were mistaken for similar emotions in an additional 20%. The responses, exact and adjusted (allowing as correct the most frequently substituted descriptor) are summarized in Table 2.

Implicit in the predictions was the hypothesis that the intended affect would be recognized regardless of the utterance semantics. In fact, utterance semantics colored some of the judgments. For example, *"I thought you really meant it."* was rarely perceived as glad, scared or disgusted, while *I'm almost finished."* was most often perceived as glad.

**Exact and adjusted recognition, per emotion**

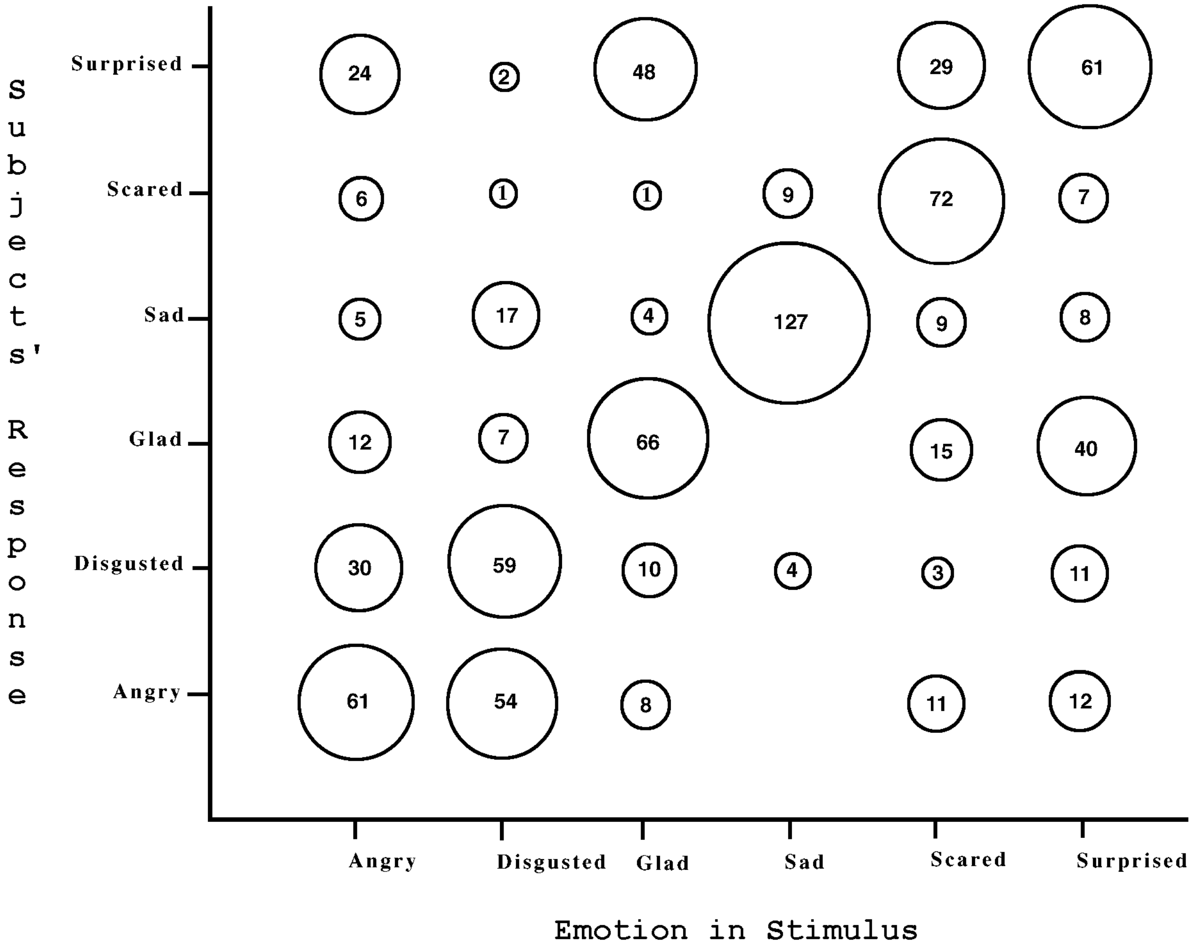
| stimulus ⇒ | *Angry* | *Disgusted* | *Glad* | *Sad* | *Scared* | *Surprised* | For All Emotions |
|---|---|---|---|---|---|---|---|
| Total presentations | 139 | 140 | 137 | 140 | 139 | 139 | 834 |
| Total recognized | 61 | 59 | 66 | 127 | 72 | 61 | 446 |
| Percent recognized | 43.9 | 42.1 | 48.2 | 91 | 51.8 | 43.9 | 53.5 |
| Total recognized (adjusted) | 91 | 113 | 114 | 136 | 101 | 101 | 656 |
| Percent recognized (adjusted) | 65.5 | 80.7 | 83.2 | 97.1 | 72.7 | 72.7 | 78.7 |

Table 2: **The number of exact and adjusted recognitions, for each emotion and for all emotions, totaled across all subject responses.**

Individual subjects tended to favor some emotions over others, especially emotions with similar semantics or acoustics. These biases were individual, however, and not characteristic for any of the age, sex, regional or national subgroupings.

The magnitude, certainty and comment input facilities were primarily devices for minimizing the frustration that often arises with forced choice. They allowed the subjects to qualify their answers and to feel that their responses accurately conveyed their perceptions. Although unused in the tabulations, these data are instructive in pointing out issues that await exploration, e.g., the difference between recognizability and naturalness, how the perception of speaker identity affects the perception of affect. Some of the comments are presented in Table 3.

As evidenced by the tabulated and informal responses, the results support the hypothesis that recognizable affect can be generated in synthesized speech.



**Emotion in stimulus**

Figure 3: **Plot showing how the intended emotions in the stimuli were perceived, for each emotion, over all subjects.** The x-axis shows the emotion stimuli and the y-axis the subjects' responses. The numbers along the ascending right–to–left diagonal show exact matches between the intended affect and subject perceptions. For example, an angry utterance was perceived as angry in sixty–one presentations, disgusted in thirty, glad in thirteen, sad in five, scared in six and surprised in twenty–four.

| *Subject* | *Intended* | *Perceived* | *Comment* |
|---|---|---|---|
| #20 | Scared | Scared | Depends a lot on what assumption I have of the speaker; e.g. whether this is a young boy, old lady, or adult man. Could be the normal speech of a cartoon character.... |
| #21 | Sad | Sad | Can't get the sense: I don't understand what DECtalk is saying |
| | Disgusted | Angry | Ooh, that's a good one |
| #22 | Glad | Surprised | hard2get |
| | Sad | Sad | hard2get |
| | Scared | Surprised | hard2get |
| #23 | Disgusted | Angry | Barely controlled anger |
| | Disgusted | Angry | and, again, disgusted as well. |
| | Angry | Surprised | or possibly angry |
| #25 | Angry | Angry | Sound more impatient than anything else |

Table 3: **Subject comments showing the intended affect, the perceived affect and the subject's comment.** These comments were optional but were encouraged to capture feedback obscured by forced choice.

## Conclusions and Future Work

The Affect Editor program demonstrates that recognizable and even natural–sounding affect can be produced by imitating in synthesized speech the effects of emotion in human speech. It also serves as a tool for exploring what is needed in an affect generating system.

Its effectiveness would be enhanced with better hardware and with improvements to the model that reflect a better understanding of perception of affect in speech. Hardware improvements include: more synthesizer parameters; synthesizer parameters that vary independently such that side effects are minimized; and an increase in the overall processing abilities of the synthesizer.

The significant software improvements will be driven by a better understanding of the perception and production of affect in speech. Thus, the current model may see the incorporation of parameter dependencies, the addition of new parameters and the merging or removal of existing parameters. With better synthesizers and better models, the mappings between levels (from the emotion to its acoustical representation to its synthesizer specific expression) can be tested and improved.

Ultimately, the automatic generation of affect in synthesized speech will be best served with a generative model, most likely a representation of the speaker's mental and physiological states. The construction of such a model depends upon the identification of the relevant descriptive parameters and, more fundamentally, upon the development of a theory of the use and interpretation of affect in speech.

## References

Anderson, M. and Pierrehumbert, J. B. and Liberman, M. Y. (1984). Synthesis by rule of English intonation patterns. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*, page 2.8.1 to 2.8.4.

Cahn, J. E. (1989). Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology.

Davitz, J. (1964). *The Communication of Emotional Meaning*, pages 57–68,105–154. McGraw-Hill.

Dittmann, A. T. (1974). The body movement–speech rhythm relationship as a cue to speech encoding. In Weitz, editor, *Nonverbal Communication*, pages 168–177. Oxford University.

Fairbanks, G. (1940). Recent experimental investigations of vocal pitch in speech. *Journal of the Acoustical Society of America*, (11):457–466.

Fairbanks, G. and Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech Monographs*, 6:87–104.

Liberman, M. Y. and Pierrehumbert, J. B. (1981). Intonational invariance under changes in pitch range and length. In *Language Sound Structure*, chapter 10. M.I.T. Press.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, M.I.T., Dept. of Linguistics.

Scherer, K. R. (1974). Acoustic concomitants of emotional dimensions: Judging affects from synthesized tone sequences. In Weitz, editor, *Nonverbal Communication*, pages 105–111. Oxford University.

Williams, C. E. and Stevens, K. N. (1969). On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine*, 40(12):1369–1372, Dec.

Williams, C. E. and Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52(4 (Part 2)):1238–1250.

Williams, C. E. and Stevens, K. N. (1981). Vocal correlates of emotional states. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 189–220. Grune and Stratton, Inc.

## Biography

Janet Cahn is currently a doctoral student in the Speech Research Group at the M.I.T. Media Laboratory. She also works with the Natural Language Group at Hewlett Packard Laboratories in Palo Alto, California. Her work on affect was the subject of her master's thesis at the Media Laboratory.