Generating Expression in Synthesized Speech

by

Janet E. Cahn

B.A., Computer Science
Mills College
Oakland, California
1983

SUBMITTED TO THE MEDIA ARTS AND SCIENCES SECTION
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

MASTER OF SCIENCE

AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1989

Signature of the Author
.....

.....
Janet E. Cahn
Media Arts and Sciences Section
May 12, 1989

Certified by
.....

.....
Chris Schmandt
Principal Research Scientist
Thesis Supervisor

Accepted by
.....

.....
Stephen A. Benton
Chairman
Departmental Committee on Graduate Students

Generating Expression in Synthesized Speech

by

Janet E. Cahn

Submitted to the Media Arts and Sciences Section on May 12, 1989 in partial fulfillment of the requirements of the degree of Master of Science at the Massachusetts Institute of Technology.

Abstract

The thesis examines the proposal that affect can be reproduced in synthesized speech by imitating the effects of emotion in human speech. A program, the Affect Editor, was constructed to allow the systematic variation of speech correlates of emotion in synthesized speech and so generate the intended affect. The task raised questions about the appropriate representation of an emotion's effect on speech, the appropriate mapping from such a representation to synthesizer parameters, and the synthesizer features needed to generate convincing affect. These questions are explored in this thesis.

The true test of synthesized affect is perceptual. An experiment was performed to test whether the intended affect was reproduced by the Affect Editor. The results confirmed that the intended affect was recognized, and furthermore, bore out predictions about areas of confusion. This work supports the conclusion that affect can be generated and systematically controlled in synthesized speech. The limits of the Affect Editor indicate that more research is needed into the determination of useful taxonomies of emotion and speech correlates of emotion. Better synthesizers are also needed, to enable more precise testing and development, and eventually, real-time generation of affect in synthesized speech.

Two tapes accompanies this thesis. *Tape 1* contains: (1) seventeen utterances from which figures in this document are generated and (2) seventeen utterance sequences — one per Affect Editor parameter — in which parameter values are cycled from the low to high. *Tape 2* contains: (1) sentences uttered with varied affects designed with the Affect Editor and (2) the stimuli used in the experiment described in Chapter 6 (30 unique utterances).

Thesis Supervisor: Chris Schmandt
Title: Principal Research Scientist

This work was supported by Nippon Telegraph and Telephone Company, as part of the Personal Computing and Telephony project, and by the U.S. Department of Air Force (Headquarters, Rome Air Development Center), as part of the Advanced Concurrent Interfaces project, contract number F30602-89-C-0022.

Contents

1 Introduction **14**

2 Foundations **16**

2.1 Terms 17

2.2 Research 19

2.2.1 Physiological correlates of emotion 19

2.2.2 Acoustical correlates of emotion 21

2.2.3 Prosodic correlates of emotion 26

2.2.4 Lexical correlates of emotion 27

2.3 Speech signal features associated with affect 28

2.3.1 The F0 contour 28

2.3.2 Duration 29

2.3.3 Phoneme articulation 29

2.3.4 Voice quality 30

2.4 Generative Intonation 30

2.4.1 Components of the theory 31

2.4.2 Intonational semantics 34

2.5 Summary 35

3 Representation and Methods 37

3.1 Representations of Emotional States 37

3.1.1 Representation in conceptual space 37

3.1.2 Representation on a speech-oriented scale 38

3.2 Parameters of the Perceptual Model 41

3.2.1 Pitch 41

3.2.2 Timing 44

3.2.3 Voice quality 46

3.2.4 Articulation 47

3.3 The Utterance 47

3.3.1 Structures and features 48

3.3.2 Maintaining semantic consistency 50

3.3.3 Sources of the linguistic analysis 50

3.4 Summary 50

4 The Affect Editor 52

4.1 Implementation Issues 53

4.1.1 A scale of parameter values 53

4.1.2 Simulation of continuous parameter change 54

4.1.3 Parameter-specific implementation issues 55

4.2 Program Flow 59

4.2.1 Synthesizer settings 60

4.2.2 Utterance composition 62

4.3 The User Interface 65

4.3.1 Emotions 65

4.3.2 Sentences 67
4.3.3 Sentence processing 67
4.3.4 Output strings 70
4.3.5 Commands 70
4.4 Summary 72

5 The Dectalk **73**

5.1 Capabilities 73
5.1.1 Side effects 73
5.1.2 Limitations 85
5.1.3 Lack of dynamic parameters 86
5.1.4 Processing capacity 87
5.2 Implementing Affect Editor Specifications 87
5.2.1 Sentence annotation 87
5.2.2 Parameters of the internal representation 89
5.3 Features of an Affect Synthesizer 104
5.3.1 Scope of control 104
5.3.2 Specification methods 109
5.4 Summary 111
5.4.1 Synthesizer limitations 112
5.4.2 Short term remedies 113
5.4.3 Recommendations 113

6 Evaluation **114**

6.1 An Experiment 114
6.1.1 The hypothesis 114

6.1.2 Stimuli 115
6.1.3 The subjects 118
6.1.4 Running the experiment 119
6.1.5 Results 123
6.2 Summary and Future Work 130

7 Summary and Future Work **131**
7.1 Contributions 131
7.2 Future Directions for the Affect Editor 133
7.2.1 Improvements to current parameters 133
7.2.2 Changes to the parameter set 134
7.2.3 Changes to the model structure 135
7.2.4 Towards the development of a theory of generative affect 136
7.3 Conclusion 137

A Structures **139**
A.1 Parameters 139
A.1.1 The param flavor 139
A.1.2 The grouped-param flavor 140
A.1.3 The synth-param flavor 140
A.2 The Emotion 141
A.2.1 The emotion structure 141
A.2.2 The correlates structure 141
A.3 The sentence 142
A.3.1 The constituent structure 142
A.3.2 The pause structure 143

A.3.3 The word structure 143

A.3.4 The phoneme structure 144

A.3.5 The phoneme-alteration structure 145

A.3.6 Interpreting for the synthesizer 146

A.3.7 The dectalk-word structure 146

B Tables **148**

B.1 Dectalk Exceptions 148

B.2 Descriptors 149

B.3 Phoneme Alteration Table 149

List of Figures

3.1 **The path from affect generation to affect perception.** 39

3.2 **The Affect Editor "accent shape" parameter.** Pitch tracks for *"I thought you really meant it."* with (a) low (b) mid-range and (c) high *accent shape* values. The F0 excursions for the pitch-accented words, *"thought"* and *"meant"* are progressively higher. 42

4.1 **The Affect Editor user interface.** The current emotion is "Afraid". The current sentence is *"You've asked me that question a thousand times."* 66

4.2 **The windows for the emotion components of the Affect Editor.** The parameters of the currrent emotion are edited to create new affect configurations. 68

4.3 **The windows for the sentence components of the Affect Editor.** The current sentence is highlighted the *SENTENCES* window, and and its recursive internal representation displayed in the *phrase structure* window. 69

4.4 **The tree structure of the discourse constituents for** *"You've asked me that question a thousand times."* 69

4.5 **The windows for the abstract and Dectalk-adapted phonologies.** . 70

4.6 **The utterance string generated for the sentence** *"You've asked me that question a thousand times."* **and for the emotions (a) fear (b) anger (c) sadness.** 71

4.7 **The** *DECTALK SETTINGS* **window, displaying the synthesizer settings for "Glad".** 72

5.1 **Effect of Dectalk phoneme mode on prosody (rhythm and pitch).** Pitch tracks for *"You've asked me that question a thousand times."* sent as (a) text (b) Arpabet phonemes. As text, *"asked"* and *"question"* receive pitch accents, but in phoneme mode they do not. Also, the duration of utterance sent as phonemes is shorter. 75

5.2 **Effect of Dectalk phoneme mode on the F0 contour.** Pitch tracks for *"You've asked me that question a thousand times."* (a) sent as plain text, showing the application of the Dectalk's stress rules (b) sent as Arpabet phonemes with word stress explicitly marked. The F0 for accented words is slightly higher for plain text. 76

5.3 **Effect of mid–sentence parameter specification on rhythm.** Pitch tracks for *"You've asked me that question a thousand times."* (a) sent as plain text (b) sent with phrase parameter changes just before "a thousand times". The second pitch track shows a longer pause between "question" and "a", and a continuation rise for "question". 78

5.4 **Effect of the Dectalk's pitch rise word stress markings on F0.** Pitch tracks for *"You've asked me that question a thousand times."* (a) sent as straight text (b) with *"question"* marked with a pitch rise (c) with *"thousand* marked with a pitch rise. In (b) and (c), the accent height of the first pitch accent is reduced because of a pitch rise that follows. Despite the different pitch rise locations, the F0 contours are pretty much the same. 80

5.5 **The effect of progressive addition of Dectalk emphatic stress on F0.** Pitch tracks for the successive addition of emphatic stress (*"*"): (a) *"You've ['}asked me that ['}question a ['}thousand ['}times."* (b) *"You've ["}asked me that ['}question a ['}thousand ['}times.* (c) *"You've ["}asked me that ["}question a ['}thousand ['}times."* (d) *"You've ["}asked me that ["}question a ["}thousand ['}times."* (e) *"You've ["}asked me that ["}question a ["}thousand ["}times."* 82

5.6 **Effect of Dectalk emphatic stress on subsequent primary stresses.** Pitch tracks for (a) *"I ['}thought you ['}really ['}meant it'.'* marked for standard Dectalk stress (b) *"I ["}thought you ['}really ['}meant it."* with initial emphatic stress. 83

5.7 **Effect of Dectalk "average pitch" on speaker identity.** Narrow band spectrograms and narrow band spectral slices for *"I ['}thought you ['}really ['}meant it."*, with an average pitch of: (a) 120 Hz (b) 160 Hz (c) 200 Hz. . 84

5.8 **Effect of final lowering on the F0 terminal contour.** Pitch tracks for *"I ['/]thought you ['/]really ['/]meant it."* with (a) minimal final lowering (b) maximal final lowering. The Affect Editor *final lowering* parameter affects the Dectalk's `assertiveness` and `baseline fall` parameters. The steepness of the terminal contour increases with an increase in *final lowering*. 91

5.9 **Effect of the Dectalk "pitch range" parameter on F0 contour.** The `pitch range` expands or contracts around the average `pitch` value. Pitch tracks for *"I ['/]thought you ['/]really ['/]meant it."* with a `pitch range` scalar of (a) 20 (b) 100 (c) and 250 . (100 is normal, i.e. 100% of a normal pitch range). 92

5.10 **Effect of the Dectalk "hat rise" parameter on F0.** Pitch tracks for *"I ['/]thought you ['/]really ['/]meant it."* with `hat rise` values of (a) 0 (b) 18 (normal) (c) 50 (d) 100. 94

5.11 **Effect of pausing and pause discontinuity on prosody.** Pitch tracks for *"And my answer has always been the same."* with: (a) no pauses (b) maximum number of fluent pauses, and smooth pause onsets (c) maximum number of fluent pauses and abrupt pause onsets (d) maximum number of hesitation pauses and smooth pause onsets (e) maximum number of hesitation pauses and abrupt pause onsets. 96

5.12 **Effect of the Dectalk "speech rate" setting on rhythm and sentence duration.** Pitch tracks for *"You've asked me that question a thousand times."* with speech rates of: (a) 122 wpm (b) 200 wpm (c) and 350 wpm. The duration of phonation and silence decreases as the speech rate increases. 97

5.13 **Effect of increased word stress frequency on the F0 contour.** Pitch tracks for *"You've asked me that question a thousand times."*: (a) minimal pitch accenting (one content word) (b) maximal accenting (all content words). The Dectalk's word stress rules are circumvented by the use of phoneme mode. Also, the greater the *stress frequency*, the more frequent the pitch accents (realized as F0 excursions upward). 98

5.14 **Effect of the Dectalk "smoothness" parameter on high frequency energy.** Energy plots for *"You've asked me that question a thousand times."* with `smoothness` values of (a) 0 (b) 50 (c) 100. The highlighted line represented energy in the 3400 to 5000 Hz band. The other line represents energy in the 120 to 440 Hz band. 100

5.15 **Effect of the Dectalk "quickness" parameter on the F0 contour.** The greater the `quickness`, the more discontinuous the F0 transitions. Pitch tracks for *"You've ['/]asked me that ['/]question a ['/]thousand ['/]times."* with `quickness` values of: (a) 0% (b) 50% (c) 100%. 101

5.16 **Effects of variations in precision of articulation.** Duration and ratios of high to low frequency energy are affected. Wide-band spectrograms for *"You've asked me that question a thousand [times]."*: (a) with minimal enunciation — vowels are reduced, consonant articulation is imprecise. (b) less imprecise — word boundaries are blurred (c) with normal enunciation (d) with some precise enunciation — enunciation is applied to the last phoneme in the word. (d) with maximum enunciation — word-initial and word-final phonemes received special emphasis and vowels and consonants are fully articulated. 105

6.1 **The twenty-eight subjects by sex, age, nationality and, for U.S. natives, dialect.** 119

6.2 **The user interface to the program that presented the stimuli and collected the data.** 120

6.3 **Plot showing the distribution of emotions in the stimuli and how they were categorized, for all subjects.** The x-axis shows the emotion stimulus, while the y-axis shows the subjects' choices. Thus, an *"Angry"* utterance was perceived as angry 61 times, disgusted 30 times, glad 13 times, sad 5 times, scared 6 times and surprised 24 times. 126

6.4 **Plot showing the distribution of sentences in the stimuli and how they were categorized, for all subjects.** The x-axis shows the sentence stimulus, while the y-axis shows the subjects' choices. Thus, *"I thought you really meant it."* was perceived as angry 40 times, disgusted 16 times, glad 4 times, sad 49 times, scared 10 times and surprised 49 times. 128

List of Tables

2.1 **Summary of Fairbanks and Hoaglin's comparison of the speech correlates of anger, contempt, fear, grief, and indifference.** 22

2.2 **Summary of Williams' and Stevens' comparison of the speech correlates of anger, fear, sorrow, and for neutral speech.** 25

4.1 **Mappings from Affect Editor parameters to Dectalk settings.** Negative values indicate settings that vary inversely with the Affect Editor parameter that controls it. 61

6.1 **Affect Editor parameter values for the six emotion stimuli.** 117

6.2 **The number of subjects presented with sentences in each the nine random presentation orders.** 121

6.3 **The number of times each intended affect was recognized, for each emotion and for all emotions, totaled across all subject responses.** . 125

6.4 **Responses counted as recognitions because of acoustic and semantic similarities between emotions.** 125

6.5 **The number of adjusted and exact recognitions, for each emotion and for all emotions, totaled across all subject responses.** 125

6.6 **The number of adjusted responses, for each emotion and for all emotions, totaled across all subject responses.** 127

6.7 **Subject tendencies to perceive emotions.** The distribution of emotions in the stimuli is even. However, the distribution of the responses per subject is not. The subject's most frequent response is presented in **boldface**, and the least frequent response is presented in *italics*. 129

B.1 **Dectalk pronunciation exceptions.** 149

B.2 Feature descriptors sets for words, contours and pauses. 150

B.3 Phoneme alterations for accomplishing precision of articulation variations, for most phonemes from "AA" to "IY". 152

B.4 Phoneme alterations for accomplishing precision of articulation variations, for most phonemes from "JH" to "ZH". 153

B.5 Phoneme alteration information for pause transitions for most phonemes from "AA" to "IX". 154

B.6 Phoneme alteration information for pause transitions for most phonemes from "JH" to "ZH". 155

Chapter 1

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. *Intelligibility* is basic — synthesis of intelligible phonemes is necessary 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.

This thesis explores improvements to the affective component of synthesized speech. Prosody and phonemic intelligibility are also in need of improvement, but are relevant to the thesis primarily for how they reveal affect.

Investigations into the perception of affect in human speech have isolated many of the acoustic and prosodic correlates of emotion. This thesis explores the hypothesis that the reproduction of these correlates in synthesized speech will result in perception of the original affect, and, that the manipulation of speech correlates of synthesized speech will produce changes in the perception of affect as predicted by the original psychoacoustic findings. To prove or disprove the hypothesis, it is necessary to extract the significant findings of psychoacoustic research, implement an affect synthesizer based on these findings and test

its output.

Chapter 2 reviews the work upon which this thesis is based. Primarily, this includes analyses of speech produced in a variety of emotional states, and studies of perceptual responses to emotional speech. A generative theory of English intonation, which has been applied to the synthesis of meaningful English intonation, is used by the affect synthesizer as a filter for the effects of emotion on speech, and so is reviewed as well.

An affect synthesizer depends upon a robust and, ideally, orthogonal set of speech correlates and for each correlate, a mapping to prosodic events and synthesizer capabilities. Chapter 3 discusses issues of abstract representation of emotional states and their acoustical correlates. Chapter 4 describes the Affect Editor, a program that directs the generation of affect in synthesized speech. The Dectalk 3 is the speech synthesizer controlled by the Affect Editor. Chapter 5 discusses how the Dectalk's capabilities and limits affected the final speech output.

Affect is a communicative and perceptual phenomenon. Thus, the true test of the hypothesis that affect can be systematically synthesized is perceptual. It was tested by an experiment in which twenty-eight subjects were asked to identify the emotional content of thirty synthesized utterances. The design and results of this experiment are described in Chapter 6.

Finally, Chapter 7 summarizes the previous chapters and contains suggestions for future work in the area of affect synthesis.

The Affect Editor is a tool with which the perception of affect can be systematically tested. It is also a tool with which to develop the rules of *automatic* generation of affect in synthesized speech. Rule-based affect generation is the long term goal of these investigations, although not of this thesis. When rules and speech correlate values have been determined, affect-generating instructions may become as straightforward as stage directions in a play. At that point, affect generation can be incorporated into any application in which expressive synthesized speech is appropriate — tools for presentation of dramatic material, information-giving systems; and synthesizers used by the speech-handicapped — to convey emotion along with the text.

Chapter 2

Foundations

This chapter discusses the foundations upon which the Affect Editor is built. It reviews terms used throughout the thesis, investigations into the objective (acoustic) and subjective (cognitive) detection of affect in speech, and a theory of intonation that is partially incorporated into the Affect Editor[1].

Affect and intonation must be considered jointly because their effects, in English, are intertwined. Both affect prosody. However, the effect of emotion upon speech, especially through physiology, is the more global effect. Intonational changes are local, tied to words, and so occur *within* the emotional context. The approach taken towards generation of affect relies upon psychoacoustics research for descriptions of the acoustical correlates of emotional states and upon the Generative Intonation theory for the linguistic component of the implementation.

[1]This theory is primarily the work of Janet Pierrehumbert. It will be referred to as the "Generative Intonation" theory in this document.

2.1 Terms

Understanding what is meant by certain speech, prosodic and discourse terms and by "emotion" is key to understanding the synthesis of affect. Terms that describe aspects of speech production are: *phoneme, allophone, fundamental frequency, F0, pitch, creaky voice* and *formant*. Terms that describe prosodic phenomena are: *intonation, pitch accent, pitch contour, F0 contour, intonation contour, rhythm* and *intensity*. Terms that describe discourse phenomena are: *phrase, utterance* and *discourse constituent*.

Phonemes are the basic and meaningful sounds of a language from which all words are constructed. These sounds are distinguished by how they are produced and how they are perceived. The replacement of one phoneme with another transforms a word into another word, as in *"bat"* versus *"vat"*, or simply into a sound sequence without meaning for the language, as in *"bit"* versus *"vit"*. **Allophones** are phonemes whose articulations differ but which may be substituted for each other without a change in word meaning (e.g., *"butter"* may be pronounced: as *"butter"* or *"budder"*).

The frequency at which the vocal folds vibrate is the **fundamental frequency** of speech, or **F0**. It is this frequency that is usually referred to when **pitch** is perceived. However, pitch may also be perceived when the vocal folds are not vibrating, as in **creaky voice**. Creaky voice describes the dying off of voicing, usually at the end of an utterance, such that F0 is produced by aerodynamic means instead of vibratory motions. It sounds like a series of taps and is perceived as low-pitched [13].

The vocal tract configuration for any phoneme determines which of the fundamental frequency harmonics will be emphasized or dampened. The emphasized harmonics are the **formants** of the phoneme — the resonant frequencies of its vocal tract configuration. The first formant above F0 is called F1. Above that is F2, and so on. The formants from F1 to F5 are generally all that are needed for phoneme recognition by human hearers.

Intonation refers to changes in pitch that impart linguistic and pragmatic[2] information

[2] *Pragmatics:* the goals and purposes of the language user, with emphasis on *why* something is said rather

not discernible from words alone. Distinctive high or low pitch emphasizes words and distinguishes phrase endings. Distinctive pitch or pitch fluctuations applied to the stressed syllable of a word is called a **pitch accent** in the Generative Intonation theory. Changes in pitch at the beginnings and ends of utterances convey the role of the sentence in the discourse. For example, a high phrase–initial pitch indicates the start of a new topic, while a pitch rise at the end of phrase that is not a question indicates a continuation of the current topic. **Pitch contour, F0 contour** and **intonation contour** are used interchangeably in this thesis to describe the F0 trajectory over the course of an utterance.

Although the main carrier of linguistic and pragmatic information is *F0* [9,17], **rhythm** (the combination of phonation durations, silence durations and word stress) and *intensity* (loudness) also contribute[3]. **Prosody** refers to the control of pitch, duration and intensity to convey non–lexical linguistic and pragmatic information.

Prosody is employed in order to make a discourse more coherent. It clarifies semantic relations between utterances and signals conversational events, such as continuing, relinquishing a speaking turn, or interrupting. A **phrase** is comprised of one or more words that makes a distinct semantic contribution to the meaning of an utterance. Often a phrase is also a syntactic clause. Intonation groups words into a phrases — all phrases have intonational contours. An **utterance** is loosely defined in this thesis as the collection of phrases uttered in one speaking turn. Thus, it may be a clause, a sentence or paragraph. A **discourse constituent** is used broadly and recursively to mean as any structure that has or can have *salience* within the discourse. Thus, words, syntactic clauses, semantic phrases, sentences and the contents of one speaking turn can be discourse constituents either singly or in combination with other discourse constituents.

Emotions may be operationally defined as those mental states with identifiable effects on

---

than its actual content.

[3]The work of D. B. Fry underscores the importance of intonation to English. He found that word emphasis is accomplished mainly through changes in the F0 contour, secondarily through changes in duration and lastly through changes in intensity [9]. This replaced the belief that *intensity* was the primary conveyer of word emphasis.

physiology and behavior. This describes a continuum of mental states. For example, fear is clearly an emotion — it affects facial expression, body movements and voice. However, concentration is a mental state that is not usefully described as an emotion — it most conspicuously affects facial expression, but not necessarily or distinctly body movement or voice. In this document, the terms **emotion, speaker state** and **mental state** are used interchangeably.

2.2 Research

This section reviews the research into the speech correlates of emotion for the speaker and as perceived by the listener. The physiological, acoustic, prosodic and lexical effects of emotion are then summarized.

The earliest research into affect focused on animal vocalization and behavior for clues to the origins of human expression. It is described briefly in *The Expression of the Emotions in Man and Animal*, written by Charles Darwin and published in 1872. More recently, the speech correlates of emotion have been studied by acoustics researchers who analyzed the speech signal, by linguists investigating prosodic and lexical effects and by psychologists investigating perception of affect in speech. Through these efforts, many of the components of speech that convey affect have been identified. It is these components that the Affect Editor seeks to control in order to generate affect in synthesized speech.

2.2.1 Physiological correlates of emotion

The emotionally-driven physiological changes of broadest scope are those that increase or decrease speaker arousal. Either the *sympathetic* or *parasympathetic* subsystems of the *autonomic nervous system* are aroused. The two systems have opposing regulatory effects upon the internal organs for the purpose of maintaining physiological stability. The sympathetic nervous system tends to respond as a unit, while the parasympathetic nervous system is more specific in its influence and is responsive to some degree of voluntary control

[26].

Strong arousal of the sympathetic nervous system occurs for **anger** or **fear** and affects the organism by inducing:

- increased heart rate and blood pressure
- changes in the depth and pattern of respiratory movements
- an increase in the respiration rate, causing greater subglottal pressure [24]
- drying of the mouth
- occasional muscle tremor.

When compared with normal speech, the speech produced from this state displays:

- greater speed and loudness
- more energy in the higher frequencies relative to that in lower frequencies
- an expanded F0 range
- from increased subglottal pressure, a median F0 value that is higher than for normal speech
- disturbed rhythm, presumably from shorter durations of speech between breaths [24]
- increased fluctuations in the F0 contour
- increased precision of articulation (enunciation) [26].

Inhibition of the sympathetic nervous system, and thereby, arousal of the parasympathetic nervous system, occurs for states such as relaxation or grief. Arousal of the parasympathetic nervous system is characterized by:

- decreased heart rate
- decreased blood pressure
- increased salivation [26].

The speech produced from this state is slow and low-pitched. High frequencies are weak and articulation is imprecise [26].

The level of systemic arousal affects aspects of speaker voice quality. Subglottal pressure controls intensity. Glottal state determines the composition of the initial waveform, as well as the quality of the attack. The filtering action of the vocal tract results in loss or amplification of harmonics. Local effects are concentrated in phoneme attributes — vowel duration, completeness of closure for obstruents and the overall degree of enunciation.

2.2.2 Acoustical correlates of emotion

Fairbanks and Pronovost (1939) [8] studied the effects of five simulated emotions — anger, fear, indifference, grief and contempt — on *pitch*. They found that these emotions could be differentiated on the basis of:

- pitch range (wide or narrow)
- variations in inflection (presence or lack of, range of variations)
- average pitch (high or low)
- number of pauses and speech rate (slow or fast)
- overall slope of the F0 contour (level or downward)
- overall variability of rhythm and pitch.

Contempt exhibited a low pitch, a wide pitch range and extreme variations in inflection. **Anger** exhibited great changes in pitch and a generally downward inflection. **Fear** had the widest pitch range, a preponderance of high pitches and few pauses. **Grief** exhibited the slowest speech rate and minimal variability among features. **Indifference** exhibited the lowest pitch, the narrowest pitch range and was almost a monotone.

Fairbanks and Hoaglin (1941) studied the effects of the same five emotions on *speech rhythm*. They found that these emotions could be distinguished by:

- variations in number of pauses
- length of pauses
- ratio of pause duration to total phonation time
- speech rate.

With speaking rate measured over the total utterance duration both **grief** and **contempt** exhibited a slow speaking rate — 129 words per minute for grief and 116 words per minute for contempt. While the average speaking time for each was nearly the same, grief was characterized by long pauses, especially between phrases, such that the total pause time was nearly equal to the total phonation time. By comparison, anger, fear and indifference had relatively rapid speaking rates — 190, 202 and 209 words per minute respectively [7].

The results of both studies are summarized in Table 2.1.

	Anger	**Contempt**	**Fear**	**Grief**	**Indifference**
median F0	high	low	highest	low	lowest
F0 range	wide	wide	widest	narrow	narrowest
F0 transitions	great changes in F0	extreme variations in inflection	many F0 changes	slow F0 changes	minimal — almost a monotone
F0 contour direction	downward			slowly falling	lack of definite pattern
speech rate	190wpm	116 wpm (slow)	202 wpm	129 wpm (slow)	209 wpm 209 wpm
pauses	few	few	few	long fluent pauses	fewest
Other			preponderance of high pitches	minimal feature variability, vibrato	

Table 2.1: **Summary of Fairbanks and Hoaglin's comparison of the speech correlates of anger, contempt, fear, grief, and indifference.**

Williams and Stevens (1969) studied recordings of speakers in great distress from **fear** (pilots under stress, about to crash) or **anguish** (a radio announcer describing the crash of the Hindenburg). When contrasted with speech obtained under less drastic conditions, utterances from both sources showed increases in the median F0, fluctuation in the F0 contour, the F0 range and discontinuous F0 transitions. Also in evidence were some irregularities in the contour, possibly tremors. While they felt it was premature to assign acoustic correlates to emotional states, they suggested that a comparison of current F0 range and median F0 values with the speaker's normal values would aid in recognizing when a speaker was undergoing stress [25].

In an extension of their original study, Williams and Stevens (1972) compared the simulated emotions of **anger**, **fear**, **sorrow** and **"neutral"** and the live recording of the Hindenburg crash broadcast. They found that different emotions had distinct effects on the fundamental frequency — its median value, the average pitch range, the characteristic shape of the contour, the rate of F0 change along the contour — and the speech rate [24].

They observed that utterances spoken in **anger** exhibited the highest median F0, the widest F0 range, and the most rapid F0 changes. Despite the rapidity of F0 changes, the overall contour was relatively smooth. One or two syllables in the phrase showed greater F0 peaks, suggesting greater emphasis. The duration of the utterance was about the same as for fear, and the speech rate second only to that for neutral expression, which was the fastest (measured in syllables per second). There were some irregularities of spectral pattern in the high frequencies and greater high frequency energy overall. Consonantal closures were more defined, with abrupt intensity changes at consonant closure and release.

Fear exhibited a median F0 lower than for anger. In some cases, although the F0 range was wider, the median F0 was close to that for neutral speech. Fear exhibited rapid up and down F0 fluctuations, and sharp discontinuities from one syllable to the next, and occasional excursions into the high end of the F0 range. The speech rate was slower than for anger but still at least twice the rate of that for sorrow. Articulation of both vowels and consonants tended to be precise. The signal tended to contain relatively little energy in the lower frequencies. The correlates for fear were not however as consistent as for the

other emotions.

Of all the emotions, **sorrow** had the lowest median F0, the narrowest F0 range, and a contour that showed the fewest fluctuations. Utterances spoken in sorrow had the longest total duration consistent with a slow speech rate — less than half the rate of any other emotion. Voicing fluctuated considerably from one pulse to the next and there were voicing irregularities, sometimes to the point of whispering.

Neutral speech showed relatively slow and smooth F0 transitions. Even the more rapid F0 rises were smooth. This was the only state wherein the formant structure and glottal vibration pattern were uniform, with minimal irregularity. The speech rate was the fastest for any of the four states[4]. Articulation tended to be imprecise, especially for consonants in unstressed syllables.

These results are summarized in Table 2.2.

The acoustical correlates induced primarily by physiological changes appear consistent across languages. Subjects listening to passages in unknown languages (Kretsch [5], Beier and Zautra, 1972 [3]) recognized emotions with accuracy significantly above chance. This was particularly true for longer passages [3].

Effect of acoustical correlates on the perception of affect

Some psychoacoustics researches, notably Joel Davitz and Klaus Scherer, sought to classify emotions and their speech correlates along the axes of a semantic space, with the dimensions *activity* (presence or absence of energy or motion), *strength* (power) and *evaluation* (pleasantness or unpleasantness) [16][5]. Davitz (1964) observed that intensity (loudness), speech

[4]The high speech rate for neutral speech, presumably, the product of normal respiratory activity, can be explained by Fairbanks and Hoaglin's data, which shows indifference as having the lowest mean duration for phonations and pauses between phrases, and the second lowest mean duration for pauses within phrases [7].

[5]In studies of the internal representation of concepts, subjects were asked to rate concepts along various semantic dimensions, *activity*, *strength* and *evaluation* among them [16]. Anger, fear and joy all had high ratings along the *activity* dimension of this space; despair has a particularly low activity rating [5]. For emo-

	Anger	Fear	Sorrow	Neutral
median F0	highest	lower than **anger**	lowest	higher than **sorrow**
F0 range	widest	**anger**	narrowest	narrow
F0 transitions	most rapid; 1 or 2 peaks with greater emphasis	rapid, sharp discontinuities from 1 syllable to the next; tremors	fewest	slow smooth
speech rate	rapid, second to **neutral**	slower than **anger**; twice that of **sorrow**	slowest	fastest
high frequency energy	greatest	most energy in 300–1200 Hz range	least	similar to **sorrow**
articulation	well-defined consonantal closures	precise for vowels and consonants		somewhat imprecise
Other	higher 1st formant frequencies than for neutral speech	correlates not as consistent across speakers as for other emotions	voicing irregularities from one pulse to the next	

Table 2.2: **Summary of Williams' and Stevens' comparison of the speech correlates of anger, fear, sorrow, and for neutral speech.**

rate, median pitch, timbre and precision of articulation distinguished emotional states and additionally, varied together along the *activity* dimension [5]. Furthermore, he found that dissimilar emotions with similar *activity* ratings, such as anger and joy, were often mistaken for each other.

In the same vein, Apple and Hecht (1982) found that emotions differing in the amount of subjectively perceived energy (i.e., *activity*) levels, such as anger and boredom, were correctly identified in speech significantly more often than those, such as sadness and satisfaction, differing mainly in degree of perceived pleasantness (i.e., *evaluation*) [3].

2.2.3 Prosodic correlates of emotion

Pitch, rhythm and *intensity* are the perceptual components of speech prosody affected by changes in the speaker's emotional state. Local effects show up in the syllable[6] while global effects occur over the phrase.

Pitch

Pitch is the perceptual response to F0. Increased speaker agitation tends to expand the F0 range and increase the overall magnitude of local F0 transitions. The emotional state combines with the role of the utterance in the discourse to impart characteristic slopes to the F0 contour up to the *nuclear syllable* (the last syllable to contain a pitch accent [17]) and to the *terminal contour* (the part of the phrase bounded by the last accented word and the sentence offset). Depending upon emotional state, transitions in the F0 contour may be smooth or discontinuous [26].

---

tions, these ratings appeared to correspond to the magnitude of systemic arousal or inhibition characteristic of the emotion.

[6]When a word is stressed by perturbations of the F0 contour, these perturbations will be concentrated in the syllable carrying primary stress [17].

Rhythm

Speech rhythm derives from stress placement, and the combination of phonation and pause durations. Pauses fall into two categories. *Fluent* or *juncture* pauses normally occur between intonational (hence, semantic and syntactic) clauses. *Hesitation* pauses occur not at clause junctures but *within* clauses, typically after a function word or the first word in the clause [6]. Their introduction coincides with an increase in intensity of cognitive processing [15]. They may be *filled* with drawn–out words, or non–lexical sounds such as "um" and "uh", or *unfilled* (silent).

As the speaker's agitation increases, average syllable duration decreases and the speech rate, measured in syllables per second or words per minute, increases. Pauses are short and occur infrequently. Conversely, depression slows speech and introduces longer and more frequent hesitation pauses. Pleasant emotions have regular rhythm, as with affection, while those judged unpleasant exhibit more irregularity, as with sadness [3].

Extreme speaker agitation produces speech errors — hesitations, repetitions, corrections, omissions — and increases the proportion of nonlinguistic to linguistic sounds [21].

2.2.4 Lexical correlates of emotion

Emotional states have grammatical and lexical correlates as well. Busemann (1925, 1926) conducted some of the earliest investigations in this area. He used as a measure the *verb/adjective ratio*, which increases with increasing anxiety [3]. Osgood (1960) noted that with greater speaker arousal there is an increase in the *noun–verb/adjective–adverb* ratio [3]. Wiener and Mehrabian (1968) correlated *verbal immediacy* with the speaker's aversion or affinity to the listener, the topic or the pragmatics. The greater the aversion, the more frequent the constructions signifying temporal or spatial distance. For example, the present tense conveys greater temporal immediacy than the present perfect, while the deictics *"this"* and *"these"* convey greater spatial immediacy than *"that"* and *"those"* [3].

2.3 Speech signal features associated with affect

This section describes in more detail the speech correlates of emotion that are measurable features of the speech signal. The features are those of the F0 contour, duration, phoneme articulation and voice quality. Of these, F0 and duration effects reveal most of the influence of an emotion. Fortunately, they are also amenable to control with current synthesizer technology.

2.3.1 The F0 contour

Attributes of the F0 contour — the median F0 value, the F0 range and F0 variations — appear to be the main carriers of affective information [24]. The **median F0 value** is the average of the F0 values over the course of an utterance. It conveys the speaker's level of excitation (high median F0) or inhibition (low median F0) as it relates to the speaker's currently recognized F0 range. Its interpretation is with respect to the overall F0 range, since a small or wide range can still exhibit the same average F0 value.

The **F0 range** is the distance between the speaker's highest and lowest F0 values and normally describes the F0 activity for one or more utterances. It reveals speaker excitement or inhibition when compared to the pitch range for neutral expression. A wider than normal F0 range reflects physiological or emotional excitation; a smaller than normal range reflects inhibition.

Variations in F0 are described by how often F0 fluctuates between high and low values, the speed of the fluctuations, and whether they are smooth or abrupt. Fluctuations in F0 are both physiologically and prosodically induced. *Pitch accents* describe prosodically based fluctuations. In general, the F0 contours for emotions perceived as negative (e.g., anger, fear) are discontinuous, while those for emotions perceived as positive (e.g., happiness) are smooth.

2.3.2 Duration

Duration is the component of prosody described by *speech rate* and *stress placement*, and whose effects are perceived as *timing* and *rhythm*.

Speech rate is a function of systemic arousal and is reflected in the word or syllable rate, pausing and in the consonant to vowel duration ratio. Relative durations of vowels and consonantal closures contribute to speech rate. Williams and Stevens noted that increased durations for utterances made in anger, fear and sorrow came both from increased vowel durations and lengthened intervals of closure or constriction for consonants [24]. Relative durations are probably also a function of individual speaking style.

Notable **pause** features are *frequency of occurrence, average duration* and *location within a clause*. An agitated speaker will tend to speak quickly with fewer and shorter pauses [3], while a depressed speaker will speak slowly. Hesitation pauses, which interrupt syntactic units, reflect a disruption of cognitive functioning. Taken together, the pause features reveal the extent and type of cognitive or systemic disturbance.

Stress placement refers to the frequency of occurrence and regularity of spacing of pitch-accented words. With greater speaker agitation, word stress is frequent, while with less speaker agitation, stressed words are fewer and occur less often.

2.3.3 Phoneme articulation

Phoneme articulation varies with affect such that **precision of articulation** varies directly with speaker arousal. Increased arousal produces increased precision and decreased arousal produces decreased precision. Precise articulation is perceived as enunciated speech; imprecise articulation is perceived as slurred. Generally, precision increases with speech rate.

Intensity of release for stop consonants is a feature of consonantal articulation. It increases with the greater subglottal pressure and more complete consonantal closures that accompany increased speaker agitation [24].

2.3.4 Voice quality

Voice quality is described by characteristics that can be measured at most points in the speech signal. *Intensity, voicing irregularities, the ratio of high to low frequency energy, breathiness* and *laryngealization* contribute to perceptions of voice quality.

Intensity is observed in the signal as the amplitude of the waveform, and is correlated with the perception of *loudness*.

Voicing irregularities cover a wide range of voicing characteristics. *Vocal jitter* refers to fluctuations from one glottal pulse to the next, as in anger. The cessation of voicing is seen in emotions such as sorrow, where speech becomes simply a whisper [24]. Aperiodic vibration of the vocal cords, with F0 produced aerodynamically, is called *laryngealization*, and results in the perception of *creaky voice*.

The ratio of high and low frequency energy changes with speaker agitation. Great high frequency energy correlates with agitation; minimal high frequency energy correlates with depression or calm. Thus, there is greatest energy in the higher frequencies for anger, and least for sorrow [24].

Breathiness and *laryngealization* reflect characteristic individual vocal tract "settings" and so are predominantly correlated with speaker identity rather than affect [13]. Their influence on affect is barely mentioned in the literature. **Breathiness** describes the generation of breath noise along with voicing. Female voices tend to be breathy. An excess of breathiness produces a whisper [4]. Increases in **laryngealization** produce *creaky voice*, wherein voicing is turned off at the sentence beginnings and ends [4].

2.4 Generative Intonation

Both the physiological and prosodic effects of emotion show up in the speech signal. The key prosodic element is intonation. In the Affect Editor, intonational phenomena is described

in Generative Intonation terminology, because this theory directly addresses the *synthesis* of intonation. This section reviews the components of the theory and their semantic interpretation.

The synthesis of English intonation is addressed in part by Pierrehumbert and Liberman [17,18,14]. Pierrehumbert's Ph.D. thesis [17] describes a *grammar* of English intonation, a major step towards the systematic generation of correct and meaningful intonation. Pierrehumbert and Hirschberg's recent account of the *semantics* of intonation opens the possibility of generating by rule intonation appropriate to the propositional and pragmatic function of the utterance [11].

2.4.1 Components of the theory

The significance of Generative Intonation is that it replaces a literal pitch contour description, in terms of points on a frequency/time axis, with a symbolic one described by combinations of high (**H**) and low (**L**) tones. Word emphasis and terminal contours are specified as annotations to text, from which intonation (the F0 contour) is automatically generated. The role of pitch accents, the composition of phrases, the speaker's contribution and the role of prominence are key components of the theory and are described in the following sections.

Pitch accents

Pierrehumbert describes English intonation in terms of high and low tone levels, scaled higher or lower relative to a *reference line* — an F0 value — for the phrase [2]. Six combinations of high and low tones form the *pitch accents*, denoted by **H***, **L***, **H*+L**, **H+L***, **L*+H** and **L+H*** [18,17]. They describe F0 changes that occur within an emphasized word, to the *syllable* bearing primary lexical stress. **H*** and **L*** are the *simple* pitch accents. **H*+L**, **H+L***, **L*+H** and **L+H*** are *complex* pitch accents and denote pitch movement within one syllable. The tone followed by an asterisk is aligned in time with the

start of the syllable. Tones not followed by asterisks are "leading" or "trailing" tones and are of generally shorter duration [17].

Phrases

A sequence of *pitch accents* followed by a *phrase accent*[7] is the minimal grouping in Generative Intonation. It defines an *intermediate phrase*, alternatively referred to as a "breath group" [24] or a "phonemic clause" [6]. A sequence of intermediate phrases followed by a *boundary tone* — the final pitch for the phrase — constitutes an *intonational phrase*. Pitch range has scope over intermediate and intonational phrases. Coverage by the theory stops here, although it should be noted that sequences of intonational phrases make up the various discourse units — sentence, paragraph — and intonation has meaning at these levels as well [23].

Terminal contour

The terminal F0 contour of an intermediate phrase is shaped by the nuclear (final) pitch accent, followed by a *phrase accent* (**H** or **L**). The addition of an **H%** or **L%** *boundary tone* to the final intermediate phrase completes the terminal contour for an intonational phrase.

Reference Line

The *reference line* is defined as that F0 value from which all **H** accents are scaled upward, and all **L** accents are scaled downward. It seems to be the linguistic equivalent of *average pitch*. Perhaps the biggest difference is that changes in *average pitch* derive from physiological events, while changes in the *reference line* are under more conscious control. The reference line is deliberately and intentionally employed to convey meaning. For example, a raised *reference line* (and possibly an attendant expanded pitch range) signals the start

[7] A *phrase accent* is a tone that is not time–aligned with a word. It occurs after the last pitch accent (nuclear accent) and shapes the remainder of the contour.

of a new topic. Its movement upwards is correlated with "speaking up" [2].

The Speaker A speaker's lowest speaking F0 is called *baseline* in Generative Intonation. It is physically constrained by the speaker's vocal apparatus (for example, by head size and vocal tract length) [19]. A speaker's pitch range is bounded by the baseline, which is fixed, and a *topline*, which varies. The topline is the highest F0 in the utterance. It increases with sentence length [23], with the introduction of a new topic, perhaps, with a raised reference line [2] as well as with speaker agitation. A speaker's highest pitch (a shriek) is rarely reached. However, the baseline pitch is often reached as proximity to the baseline is the measure of final lowering and, therefore, the finality of the statement.

Prominence

In its earliest incarnation (1980), Generative Intonation posited the existence of *prominence* — a percent applied to the pitch range — as a measure of accent height or depth [17]. The addition of the *reference line* in 1984 elegantly redefined prominence as a scalar applicable in either direction away from the reference line such that the calculation of an **H** tone frequency is:

$$ReferenceLine + (prominence \times (topline - ReferenceLine))$$

and for an **L** tone:

$$ReferenceLine - (prominence \times (ReferenceLine - baseline))$$

However, prominence is not mentioned in more recent publications. In particular, *catathesis* or *downstepping*, the systematically decreasing F0 in list contours, is explained without the use of prominence [19]. It is triggered by the presence of a bitonal accent which compresses the pitch range for the words that follow, and thereby reduces the height of subsequent pitch accents. A contour displaying catathesis can be calculated from the pitch range, accent types and accent positions. Despite this, prominence appears to quantify meaning such that the most prominent word is the most salient for the sentence. Prominences plays

a crucial role in the Affect Editor. Words with the highest prominence are the first or last to display the effects of emotion in speech, depending on the effect.

Currently, prominence is only a feature of words. However, it could easily and consistently extend to any of the higher level structures — intermediate phrases, intonational phrases, and paragraphs — to quantify the relative salience of each unit. For phrases and larger units, this quantification would scale the more global features of speech — pitch range, final lowering and speech rate.

2.4.2 Intonational semantics

Word meaning is communicated by **pitch accents**. Clause meaning and its relation to the rest of the discourse is communicated by pitch range and by the shape of the **terminal F0 contour**.

Pitch accent semantics

The semantic theory of pitch accent and terminal contour usage is still under development. However, variations in both appear to correspond systematically to variations in sentence meaning in a pragmatic context. It becomes important not only that a word is emphasized, but why it is emphasized. Pierrehumbert and Hirschberg claim that pitch accents indicate salience, and their type indicates the type of salience. For example when the starred tone is **H**, as in **H***, **H***+**L** or **L**+**H***, the speaker intends the accented information to be added to the hearer's beliefs. The converse is true when the starred tone is **L**. The speaker is emphasizing information assumed to already reside in the hearer's belief space. Complex pitch accents indicate assumptions about the mutual belief space content, and additionally, convey a relationship between belief space contents and the current proposition. In general, **L**+**H** tones invoke a sense of scale upon which the emphasized item should be placed, while **H**+**L** tones invoke an inference path supporting a predication[11].

Terminal contour semantics

The phrase accents and boundary tones in the terminal contour express the relation of a phrase to other phrases in the discourse. An **H** phrase accent indicates that the current intermediate phrase it ends is part of a larger unit that includes the following phrase, while an **L** phrase accent separates the current intermediate phrase from any that follow. Similarly, an **H%** boundary tone signals interpretation of the current intonational phrase with respect to succeeding utterances, while an **L%** boundary tone signals interpretation with respect to previous utterances [11].

The combination of *nuclear pitch accent*, *phrase accent* and *boundary tone* describes fully the terminal contour. For example, an **L* L H%** contour has a steep rise at the end. It tends to be used mostly in yes-no questions where the answer is uncertain or to convey doubt or incredulity. The high-rise contour, **H* H H%**, when part of a statement, conveys doubt about the relevance of the new information (emphasized with an **H***) being added to the mutual belief space. Used to end a yes–no question, this same contour implies that the speaker expects a positive response [11]. **H%** tones may signal the speaker's intent to continue, as in the *continuation rise*, **L H%**, and also establish satisfaction–precedence relationships [10] within a sequence of statements. **L%** tones, especially in **L L%** phrase endings, convey finality and completion [11].

2.5 Summary

Acoustic, linguistic and cognitive inquiries have shown that vocalization is influenced by emotion — the cognitive and physiological responses of an organism to its environment. In human vocalization, emotion affects the acoustic features of the speech signal and its linguistic — paralinguistic (non–lexical) and lexical — content. The acoustic features affected by emotion are mostly the F0 and durational correlates — the key components of *prosody*. Emotion affects the median F0, F0 range, and the number, speed and smoothness of F0 transitions. Its influence on duration can be measured in the speech rate and in

stress frequency. Other affected areas are voice quality — intensity, voicing irregularities, ratio of high and low frequency energy, breathiness and laryngealization – and precision of articulation.

Linguistic affects of emotion may be divided into paralinguistic and lexical effects. Prosodic events are represented at this level. They include F0 changes — intonation — and duration — pause frequency and pause location. Lexical effects include the presence or absence of speech errors, and frequency distributions of parts of speech over the utterance.

This supports the goal of automatic synthesis of affect in speech depends ultimately on developing a theory of affect in speech, with which to generate affect appropriate to the linguistic content, and the speaker's mental state (intention and belief). It must be consistent with other theories of speech events. Toward this end, the Generative Intonation theory is included in the Affect Editor representation of intonational events.

Chapter 3

Representation and Methods

3.1 Representations of Emotional States

The path from emotion to speech signal requires at least one intermediate representation whose function is to quantify the effects of emotion on speech. This is an abstract representation, independent of the current technology. Variations in its quantities are mapped to linguistic and synthesizer parameters to produce the corresponding variations in the speech signal and finally, in perceived affect.

This section discusses possible candidates for the intermediate representation. One represents an emotion in terms of its semantic features. Others represents features of speech, encoded in either a production (speaker state) or perceptual (listener perception) model.

3.1.1 Representation in conceptual space

Psychoacoustics researchers have adopted the *activity*, *evaluation* and *strength* dimensions of Osgood, *et al's* [16] semantic space to quantify variations of affect in speech[1]. Davitz

[1]These are dimensions originally employed to quantify the meaning of concepts. They are described in Section 2.2.2. The *activity* measures the presence or absence of energy. The *evaluation* (or *valence*) dimension

and Scherer observed separately that changes along the *activity* and *evaluation* dimensions appeared to correspond to changes in speech phenomena, particularly those linked to physiology. Thus, intensity, pitch, timbre, speech rate and precision of articulation tended to vary with changes along the *activity* dimension [5], while prosodic phenomena -- regularity of changes, rhythm, inflection — tended to vary with changes in *evaluation* [22,20].

Although this tripartite representation has been useful in psychoacoustical research, its dimensions are not specific to speech. It is appropriate to a higher level representation of emotion. But because there is no generative theory of affect, it is not, at this stage, appropriate for the task of driving the *generation* of affect in speech. Instead, a useful representation is one that directly measures the effect of emotion on the speaker or the hearer.

3.1.2 Representation on a speech–oriented scale

Emotions may be operationally described as those mental states for which there are identifiable physiological, linguistic, and cognitive correlates. These together act upon speech production. The emotion is identified by listeners from its characteristic grammatical and acoustical correlates. The path from emotion to perceived affect proceeds generally as follows:

$$emotion \Rightarrow effect\ on\ speaker \Rightarrow effect\ on\ listener$$

Figure 3.1 describes this formulation in more detail. The schema it presents contains two levels of useful abstraction. One describes the effect of emotion upon the speaker, the other the features of speech that contribute to the affect perceived by the listener.

---

measures the perceived pleasantness or unpleasantness. The *strength* dimension measures perceived power or "weight" of a concept.

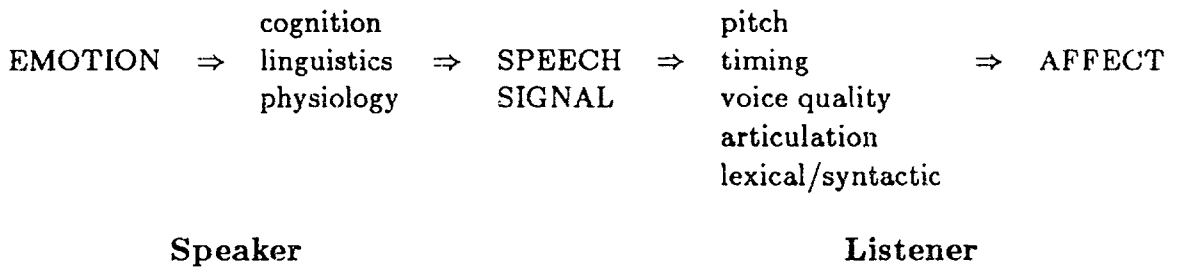


Figure 3.1: **The path from affect generation to affect perception.**

The production or speaker model

A model of the speaker depicts the state of the speech production apparatus. This is interpreted in the broadest sense to include the cognitive, linguistic and physiological aspects of speech production. Manipulations to the model control the process of producing affect in speech. If the simulation is adequate, it will generate speech in which a listener naturally perceives the intended affect. Cognitive processes, linguistic and paralinguistic considerations, and physiology are all employed in the production of speech and affect. Their contributions are summarized below.

Cognition At the cognitive level, the *meaning* of the utterance is formulated. Its semantics and its relation to the discourse expresses the speaker's intentions. Cognitive considerations include the semantics within and across utterances, deep structure syntax and the processes that control the retrieval of concepts and words. For example, difficulty in locating relevant concepts, and consequently, in retrieving the appropriate words springs from this level. It will show up in speech as hesitation pauses, repetitions or speech errors. The combination of careful sentence construction and conceptual clarity may show up as fluent pauses. In faster speech, the fluent pauses may disappear but phrases will be well-marked intonationally according to their meaning and their function in the discourse as represented at this level.

Linguistics The objective at the linguistic level is the *packaging* of the information. The meaning represented at the cognitive level is manifest linguistically and paralinguistically at this level. Words are grouped into structures that convey discourse constituency and syntactic and intonational information. Especially, intonation is assigned as a commentary on the strictly textual component of speech. It clarifies the packaging of the information such that discourse, semantic and/or syntactic roles and structures are revealed. It indicates the relationships — for example, similarity, dependency and reference — among the discourse constituents.

Physiology Activity at the physiological level affects the overall *presentation* of the information. The effects of emotion upon the respiratory system, larynx, vocal tract, muscular system (especially as it affects motor control), heart rate and blood pressure are modeled here. The physiological effects of emotion are the filter through which is passed the more intentional information of words, phrases and concepts [2].

The perceptual or listener model

The alternative to a speaker state and process model is a *perceptual* model, comprised of abstractions about what a listener hears. Reverse engineering reconstructs the speech signal from its perceptually based feature descriptions.

There are several advantages to using this model. First, it is the simpler model. It requires minimal theorizing about the mind and eschews complex explanations of intonational and discourse events. Its descriptors describe only speech events. Secondly, because the perceptual parameters are explicitly declared, this model is well suited for testing the perceptual effects of affect in speech. Finally, current technology supports such a model. Most commercial synthesizers allow variations of the *acoustical* features of the speech signal or their

[2] A physiological description is most directly useful with a vocal tract model synthesizer. However, the construction of a functional vocal tract model synthesizer is not yet feasible. Not enough is known about what is intrinsic and necessary, and the computational costs of incorporating what *is* known are currently too high [12].

corresponding *perceptual* interpretations.

The linguistic and cognitive information necessary to speech production is intrinsic only to the speaker model. It is acquired by the perceptual model through the sentences it speaks. The incorporation of linguistic and discourse analyses into the sentences completes the information needed by the perceptual model to produce the appropriate output.

Of the three representations of affect in speech, the perceptual model is currently the most tractable. Therefore, it is the model incorporated into the Affect Editor. Its parts and parameters are described in the following sections.

3.2 Parameters of the Perceptual Model

Speech parameters may be grouped by the areas from which they originate — cognition, linguistics or physiology — or by the areas in which their influence is primarily felt. The first approach is closer to a speaker model, discussed previously. In the second, parameters are grouped by their influence on *pitch*, *timing*, *voice quality* or *articulation*. The features in each category are described in this section.

3.2.1 Pitch

Pitch effects are described by *accent shape*, *average pitch*, *pitch contour slope*, *final lowering*, *pitch range* and *reference line*.

Accent shape describes the rate of F0 change for any pitch accented word, and thereby, the *shape* of the F0 contour for any pitch accent. It is interpreted as a smoothing function applied to the F0 contour at the site of a pitch accent. Although *accent shape* increases and decreases directly with speaker arousal, it is revealed intonationally as a scalar that applies to all pitch accent prominences. A high *accent shape* value indicates maximum agitation, visible as a steep rise and fall in the F0 contour with a high F0 at the peak. Thus, the large and high F0 excursions for anger, noted by Williams and Stevens [24], can be specified with

a high *accent shape* value. Perceptually, the value for *accent shape* corresponds to how noticeable each accent is. Whereas the pitch range magnifies or diminishes all F0 values in the contour, *accent shape* scales F0 values relative to the pitch range and only for accented words. The effect of progressive increases in *accent shape* on pitch accents is illustrated in Figure 3.2.

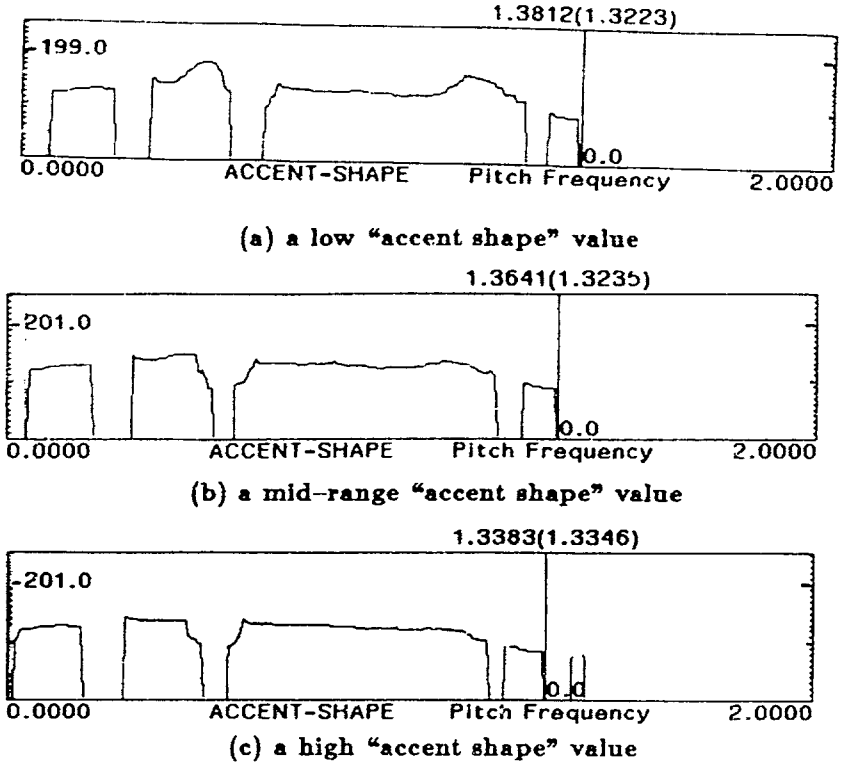


Figure 3.2: **The Affect Editor "accent shape" parameter.** Pitch tracks for *"I thought you really meant it."* with (a) low (b) mid-range and (c) high *accent shape* values. The F0 excursions for the pitch-accented words, *"thought"* and *"meant"* are progressively higher.

Average pitch quantifies how high or low the speaker appears to be speaking relative to normal functioning. The listener's ability to establish a norm suggests that there are clues to normal *average pitch* (and other speech features) which aid in the detection of abnormalities. Possibly a speaker's normal *average pitch* is a fixed distance from the *baseline*, or is indicated by voice quality parameters, for example, the ratio of high to low frequency energies or the relative strengths of formant frequencies. *Average pitch* is the average F0 value of the

contour. However, its value tells us nothing about the variance among the values from which it is calculated. That is the function of the *pitch range* and *reference line*.

Contour slope describes the general direction of the pitch contour, whether the pattern exhibited by successive pitch accents heights and by F0 values for unaccented words is rising, level or falling. It may be described by variations in descending utterance baseline, a pattern of rising or falling pitch accent prominences or an expanding or contracting pitch range[3].

Final lowering refers to the steepness and depth of the pitch fall that occurs at the end of a declarative phrase and some wh-questions. It correlates primarily with the meaning and discourse role of a phrase, indicating whether the phrase is to be interpreted according to what came before or what will come after. A phrase spoken with great final lowering signifies the end of a topic and possibly the end of a speaking turn. Also, the greater the *final lowering*, the more emphatic the affect. A phrase with minimal *final lowering* or even a continuation rise at the end, often indicates the speaker's intention to continue talking. When not used as a turn-taking cue, it conveys affect that is perceived as tentative.

Pitch range measures the bandwidth of the range bounded by the lowest and highest F0 of the utterance. The simplest of specifications uses the speaker baseline[4] to anchor the pitch range. *Pitch range* has scope over a phrase.

In its role as the F0 value from which all pitch accent **H** or **L** tones are scaled, the **reference line** obscures or exaggerates differences among pitch accent prominences. For example, significantly different prominences will yield similar F0 values for **H** tones if the distance between the *reference line* and the topline is small, and similar F0 values for **L** tones if the distance between the *reference line* and speaker baseline is small. In particular, a raised *reference line* reduces the distinctions among pitch accent prominences for **H** tones. Characteristic contours, such as catathesis, may instead sound like a series of undistinguished

[3]The Affect Editor uses pitch accent prominence patterns within a phrase, and pitch range differences between phrases, to approximate rising or falling F0 contours.

[4]Pierrehumbert defines "speaker baseline" as the lowest F0 produced for all utterances [19]. Its value is a function of the speech apparatus.

high–pitched accents.

Physiology determines **average pitch** while paralinguistic and linguistic considerations determine the **reference line**. However, they might possibly describe the same phenomenon. Raising the average pitch to indicate speaker arousal seems analogous to raising the reference line for "speaking up" [2]. Yet, for a sentence with only **H*** accents, most F0 values are above the reference line, yielding an average pitch that is distinct from and higher than the reference line. Perhaps *average pitch* has more of an effect upon voice quality, for example, the relative energies of high and low frequencies. Since the difference and relationship between the two is not clear, both parameters are included in the in the set of pitch parameters.

3.2.2 Timing

Timing parameters contribute to speech rhythm. They describe speed (*speech rate*), syncopation (*exaggeration, stress frequency*) and location of silences (*fluent pauses, hesitation pauses*). Timing effects are primarily products of cognitive processes or linguistic necessity. They are usually under conscious control, since, like pitch effects, they are used to consciously convey meaning. However, physiological occurrences may induce extremes of timing effects. For example, an increased respiration rate creates shorter units of speech [24] and, thereby, changes speech rhythm. Timing effects are described by *exaggeration, fluent pauses, hesitation pauses, speech rate* and *stress frequency*.

Exaggeration affects rhythm and represents the extent to which the speech rate slows for stressed words. An exaggerated word is given increased duration and the fullest articulation, such that pitch transitions within the word are slowed as well.

Pauses may be filled with sound (e.g., the paraverbals "um" or "uh") or silence. A pause that is vocalized is called a *"filled"* pause, while one that is silent is an *"unfilled"* pause. Hesitations may be expressed in either form, or additionally, by false starts or repetitions.

Fluent pauses occur between intonational units[5]. They highlight semantic units which are also distinguished by intonation. Their duration is a function of the discourse role of the clause and perhaps also the difficulty of clause construction. The smallest intonational unit that can be preceded by a fluent pause is the intermediate phrase.

Hesitation pauses occur *within* an intonational clause, interrupting its semantic coherency. Dittmann locates them following the first function word in a clause [6]. Since function words provide the structure in which content words are arranged but otherwise have minimal intrinsic content, an alternate view observes that hesitation pauses occur *before* high content items (a word or subordinated clause) as a function of the difficulty of retrieving words that express a concept, or even the concept itself.

Speech rate is physiologically based – an excited speaker will breathe and move quickly, and speech is correspondingly fast. *Speech rate* is measured in words or syllables per minute. Its components include how quickly and individual word or syllable is uttered, the duration of sound to silence within an utterance and the relative durations of phoneme classes.

Stress frequency quantifies the frequency of word stress occurrence (pitch accents). The more agitated the speaker, the more emphatic the speech and the more frequently words are stressed. *Stress frequency* is the discontinuity parameter primarily responsible for the smoothness or abruptness of prosodically motivated F0 transitions. The greater the *stress frequency*, the more frequent the pitch accents, and the more frequent and sharp the intervening F0 transitions. If the speech rate is slowed to accommodate the greater number of pitch accents, F0 transitions, though numerous, will remain smooth.

Stress frequency is a key parameter since many effects of emotional state show up first or most noticeably in stressed words. The effects of word–based features — precision of articulation, accent shape and exaggeration — are filtered through *stress frequency*.

[5]Dittmann uses the term *phonemic clauses* to describe intonational units [6].

3.2.3 Voice quality

Voice quality parameters quantify phenomena observed throughout the whole utterance, and which are primarily of physiological origin. The voice quality parameters of the model are *breathiness, brilliance, laryngealization, loudness, pause discontinuity, pitch discontinuity* and *tremor.*

Breathiness is not of great importance in conveying affect. It refers to a burst of frication produced in the speech signal. Its increase adds a tentative and weak quality to the voice, so is greatest when the speaker is minimally excited.

Brilliance describes the perceptual effect of the relative energies of the low and high frequencies. When the speaker is agitated, higher frequencies predominate and the voice sounds harsh or "brilliant". When the speaker is relaxed or depressed, lower frequencies predominate, and the voice sounds soothing and warm. Brilliant voices tend to carry and stand out above background noise [4].

Laryngealization describes the creaky voice phenomena. It is the result of minimal subglottal pressure such that F0 is produced by aerodynamic means instead of by vocal fold vibrations.

Large waveform amplitude is perceived as **loudness**. The more excited the speaker, the greater the subglottal pressure and signal amplitude and the louder the voice. Although people perceive stressed words as being louder they are actually responding to increased pitch [9]. The loudness of speech is much more a result of speaker physiology than consciously applied prosody. Thus, *loudness* primarily conveys affective and not linguistic information.

Pause discontinuity describes the smoothness of the transitions from sound to silence for unfilled pauses. The Affect Editor recognizes four categories of pause onset — *smooth, firm, firmer* and *abrupt.* The type of pause onset expresses the speaker's relative difficulty of retrieving words or concepts, or of formulating the phrase. The greater the cognitive difficulty or the emotional upset, the more abrupt the silences that reflect this.

Pitch discontinuity describes the effect on the F0 contour of more or less motor control on the part of the speaker. It controls the smoothness or abruptness of F0 transitions throughout the contour, and the degree to which intended F0 targets are reached. With less motor control, F0 transitions are abrupt and quickly fluctuating. With more control, F0 transitions are smoother.

Tremor, also called *jitter*, describes voicing irregularities that occur between successive waveforms. It has been observed in the expression of fear [24].

3.2.4 Articulation

Articulation effects are currently described only by their **precision**. *Precision of articulation* covers a range of articulation styles from slurring to enunciation. An imprecisely enunciated (slurred) consonant has minimal frication noise, and, if originally unvoiced, may be realized as voiced. An imprecisely enunciated vowel is reduced, usually to a schwa. In the case of diphthongs, its component vowels are reduced. For consonants, enunciation is the result of increased frication and aspiration and of correctly applied voicing. For vowels, it is the result of fully articulating vowels that are normally unstressed or reduced.

Changes in *precision of articulation* have some affect on rhythm. For example, an increase of precision causes the fuller articulation of more syllables. Relative syllable durations change such that words are longer and the time between pitch accents increases. Consequently, speech rhythm is altered.

3.3 The Utterance

Speaker intentions are present via the utterances, through an analysis of the linguistic and pragmatic structures. The analysis is independent of any particular emotion. It is cursory, and consists mostly of marking all *possible* intonational clauses and discourse constituents of a sentence. Whether some or all of these divisions are manifest in the synthesized

speech depends upon the emotion. This section describes the structure and contents of the utterances that accompany the perceptual model. Issues of semantic consistency and the source of the analysis that is encoded into the utterance are also discussed.

3.3.1 Structures and features

The Affect Editor takes as input a sentence whose structure is a synthesis of discourse and intonational groupings. As a discourse constituent, it is a recursive structure whose parts are grouped by syntactic and semantic informational content. As a speech component, it is also marked for intonational features. It is not, however, an intonational phrase — an intonational phrase is a linear structure. It may be considered a *proto–intonational* phrase, an intonationally marked discourse constituent. A sentence, to the Affect Editor, is a discourse constituent composed of one or more of these proto–intonational phrases. It applies all its operations to one proto–intonational phrase at a time.

Intonational information

Each utterance must be marked with intonational features. Intonational information is carried by intonational clauses and individual words. Each intonational clause is an *intonational phrase* or an *intermediate phrase*, as defined by Generative Intonation. They are annotated as follows:

- An intonational or intermediate phrase has a phrase accent, a pitch range and a speech rate.
- In addition, an intonational phrase is labeled with final lowering and a final boundary tone.
- Every possible location for a fluent or hesitation pause is labeled with the type of pause it receives and the relative likelihood of pause occurrence. (Pause likelihood of occurrence, or pause ranking, is described in section 4.1.3.)

All content words and some function words are candidates for pitch accenting. Each of these words is marked with the type of accent it would receive under maximum *stress frequency* and for accent prominence. The accent prominence reflects the word's salience with regard to the central concept for the sentence. All words are marked grammatically with their syntactic category for the sentence.

Discourse information

The division of an utterance into intonational parts occurs as an overlay to the basic *discourse constituent* structure. As a discourse constituent, might carry information about an event or action, the agent or object of an event or action, or time. Usually a discourse constituent consists of one or more syntactic units, and therefore, one or more semantic units. For example, in the following analysis —

[S [EVENT You've asked me [OBJ that question]] [TIME a thousand times.]]

— *"You've asked me that question"* is a discourse constituent that describes an event. It contains the discourse constituent, *"that question"*, which describes an object of the event. Note, however, that this division is not exactly parallel to the syntactic parse:

[S [NP You] [AUX 've] [VP asked [NP me] [NP that question]] [NP a thousand times.]]

This classification by information type is one that might be used by a text generator in top–down generation of text. The constituent labels are included for completeness, as they are part of a [primitive] discourse analysis. However, the Affect Editor needs to know only where an informational unit begins and ends and nothing about its contents. It uses the division into discourse constituents so that fluent pauses may be inserted at constituent boundaries. Thus, the only information currently of significance to the Affect Editor is whether an utterance component is a constituent (which may be intonationally distinguished) or a word.

3.3.2 Maintaining semantic consistency

Sentence meaning is kept the same for all emotions so that the factors contributing to the perception of affect are few and independent. There is only one analysis for any sentence — word groupings into discourse constituents do not vary with emotion. Similarly, prominence values for accented words also do not vary with emotion. The word most salient to the meaning of the sentence receives the highest prominence and so reflects a constant meaning across emotional variations. (The issue of whether emotions have propositional content which might affect word groupings and prominences is outside the scope of the thesis.)

3.3.3 Sources of the linguistic analysis

The addition of programs that perform automatic parsing, discourse analysis and intonational marking would complete a system for automatically generating expressive speech from text. For the Affect Editor, the source of the syntactic/semantic, discourse and intonational analyses is irrelevant, and need not be automated.

3.4 Summary

The path from emotion to speech output requires an intermediate representation. A semantic abstraction of emotion is simple but not so useful because it is not tailored to speech. A speaker state model is perhaps the most desirable but the information needed to make it complete and usable is lacking. A perceptual model, one which describes features of the speech output, is the simplest and most useful of the three. It is the model around which the Affect Editor is built. Speaker intentions are represented in the minimal intonational and semantic analysis of the input utterance.

The internal representation of an emotion describes its effects on the perception of pitch, timing, voice quality and articulation. To adequately apply these parameters, an utterance is annotated for intonational, syntactic, semantic and discourse features. The intonational

description comes from Generative Intonation. The syntactic information consists of parts of speech. The semantic information is contained in word prominence assignments and discourse constituent groupings and labels. The annotations represent a small but sufficient piece of the speaker's *possible* linguistic and discourse intentions.

The model is responsible for conveying the perceptual response to the effects of the speaker's intentions and physiology on speech, and, when applied to the utterance, functions as a filter that allows some or all of the annotated features to manifest. Its nineteen[6] parameters and minimal analysis combine to make the synthesis of affect a tractable problem.

[6]Seventeen of the nineteen parameters are currently implemented because of hardware constraints, described in detail in Chapter 5.

Chapter 4

The Affect Editor

This chapter discusses the design and implementation issues raised by the incorporation of the speech correlates and sentence analyses discussed in Chapter 3 into the Affect Editor. It focuses on general programmatic issues[1]. General representations — parameter scales, and the simulation of a continuum of parameter change — are discussed first. This is followed by a discussion of the speech parameters — *stress frequency, contour slope, exaggeration* and *pauses* — whose implementation was especially problematic or complex. Following this, the program flow is described. The chapter closes with a brief description of the Affect Editor's user interface.

The Affect Editor is a program that applies speech parameter values for an emotion to a pre–analyzed sentence in order to synthesize the sentence with the correct affect. Its output is two strings which are interpreted by the synthesizer (in this case, the Dectalk3). The first string specifies the synthesizer settings. These affect acoustic and prosodic events. The second string is the utterance itself, specified with plain text, phonemes and prosodic markings. The synthesizer settings are adjusted first, and then the utterance is spoken.

[1]Specific implementation issues are detailed in Chapter 5, since parts of the implementation was shaped by Dectalk requirements.

4.1 Implementation Issues

This section describes aspects of parameter implementation — parameter scaling, the simulation of a continuum of parameter change for parameters for which there is no explicit mapping to synthesizer settings and parameters whose implementation is problematic or complex.

4.1.1 A scale of parameter values

The Affect Editor's parameters are measured on a scale of values that range from -10 to 10. A value of -10 designates the minimal effect of a parameter on speech, while 10 designates its maximum effect. For parameters whose main effects are on synthesizer settings — *accent shape, average pitch, final lowering, pitch range, reference line, speech rate, breathiness, brilliance, laryngealization* and *loudness* — a value of 0 represents its influence at the speaker's "neutral" setting. Thus change in these values shows up as non-linear change in the synthesizer settings, occurring at different rates above and below the norm. Since neutral affect may be characterized by the complete absence of some features (e.g., breathiness and laryngealization) the effects of a value of -10 and 0 are often the same.

Parameters describing prosodic effects that show up mostly as lexical features — *contour slope, stress frequency, exaggeration, fluent pauses, hesitation pauses, pause discontinuity (pause onset quality)* and *precision of articulation* — are not represented in the synthesizer settings. Values for these parameters are not interpreted around a zero norm. Instead, changes in these parameters effect linear changes in the speech output. This is an artifact of the current implementation, of the first pass at implementing parameters not explicitly part of the synthesizer repertoire. It is entirely possible to normalize all speech parameters around values around zero. Functionality would not change but the internal representation would become more consistent.

Interpretation of extreme values for both types of parameters it the same. The highest value represents the parameter's maximal influence and the lowest value, the minimal.

4.1.2 Simulation of continuous parameter change

For Affect Editor parameters that map to synthesizer settings, continuous change is straightforward. A change in its numerical value is reflected by a change in the numerical value of the Dectalk parameters it controls. The synthesizer settings for which this holds tend to describe features of speech that derive primarily from *physiological* events — *breathiness, brilliance, loudness, laryngealization* and *speech rate.* They affect general aspects of the speech itself rather than the more local and intentional events such as words, phrases and intonation. The continuum over which their values change is a numerical scale understood by the synthesizer.

Mapping from Affect Editor parameters to speech features not represented in the synthesizer settings is more complex. There is no numerical value to send to the synthesizer to effect a change. Instead, the changes are included in the utterance string. This is consistent, since the unrepresented speech features — *contour slope, fluent pauses, hesitation pauses, stress frequency, pause discontinuity,* and *precision of articulation* — tend to describe prosodic and articulation of linguistic units — words and phrases.

Changes for these parameters show up as increases or decreases in phoneme substitutions, phoneme additions, silences and intonational markings. These binary and discrete specifications (the presence or absence of phonemic and prosodic representations) must be handled such that changes appear *continuous.* The orchestration of increase or decrease is effected via an ordered set of distinct characteristics which define the path from minimal to maximum parameter influence. However, the characteristics describe what occurs for only a few discrete levels of parameter influence[2].

The simplest use of a descriptor set is one that produces identical output for any parameter value that maps into a region bounded by any two descriptors. In this scheme, there are only as many variations in the speech output as there are descriptors.

More and incremental variation can be introduced if the influence of effects at a region's

[2]Table B.2 shows the Affect Editor descriptors for variations in parameter influence.

boundaries are mediated through word prominence, such that for intermediate values, words with the highest prominence exhibit the characteristics of one boundary descriptor, while words with lower prominence exhibit the characteristics of the other. For a parameter whose effects show up first in pitch–accented words, increases in its value will be reflected by the increasing number of pitch–accented words exhibiting its characteristics. Words of highest prominence will exhibit the effect first, followed by words of successively lower prominence. When parameter value reaches the upper edge of the region, *all* the pitch–accented words — that can exhibit characteristic of the upper boundary will do so. This is how Affect Editor interprets word feature descriptors. It guarantees that all words in an utterance are not always enunciated or accented in the same fashion and thereby simulates continuous parameter variation. It is hoped that this variation makes for more natural output as well.

4.1.3 Parameter–specific implementation issues

Some voice quality and prosodic parameters can directly manipulate Dectalk settings. These are: *accent shape, average pitch, final lowering, pitch range, reference line, speech rate, breathiness, brilliance, laryngealization* and *loudness.* Their influence is expressed in the synthesizer setting string produced by the Affect Editor. This section, however, describes the implementation of parameter effects that are *not* so straightforward. These parameters — *contour slope, stress frequency, exaggeration, fluent pauses, hesitation pauses, pause discontinuity* (pause onset quality) and *precision of articulation* — filter or modify features included in the original sentence analysis. Their influence is expressed in the Affect Editor's utterance string.

Stress frequency

The Affect Editor implementation of *stress frequency* rests on the assumption that accented words, which carry most of the linguistic and paralinguistic information [3], will also carry

[3]This is at least true for enunciation. Stressed (pitch accented) words are fully articulated, while the vowels in words with lesser or no stress are often reduced

most of the affective information. Through *stress frequency*, words are selected (or rejected) for pitch accenting. *Stress frequency* is a key parameter — its influence determines whether or not a word exhibits an affective feature.

Its effect on the utterance is regulated by word *prominence*. For its lowest value there is minimal accenting — only the word with highest prominence is accented. As the value increases, words with successively lower prominences are permitted to carry stress. This keeps sentence meaning consistent in different emotional contexts.

Sometimes changes in emotions will affect prominence such that the key stress is moved and sentence meaning is changed. (From empirical observation, it seems that there are melodies that are characteristic of certain emotions or attitudes.) To systematically reproduce this requires a theory that relates the propositional content of emotional states to prosodic events. Lacking such a theory, the Affect Editor uses prominence and stress to *preserve* the sentence meaning throughout changing emotional states.

Contour slope

Contour slope values are interpreted as follows: 0 denotes a level F0 contour, 10 denotes a steeply rising F0 contour and -10 denotes a steeply falling F0 contour. If the synthesizer allowed, steep *contour slopes* could be effected by increasing or decreasing pitch accent prominences, or perhaps by expanding or contracting the pitch range over the course of the sentence. However, the Dectalk has no mechanism for approximating pitch accents, and mid-sentence pitch range changes introduce unwanted pauses. So steep slopes are approximated by maximizing the difference between the F0 values of the first and last accentable words. For example, a falling contour, the first accentable word receives the highest F0 for the sentence. This word need not be currently assigned stress by the *stress frequency* parameter. Assigning a pitch accent outside of the effects of *stress frequency* is the only concession to possibility that changing emotions could alter the prominence relationships and thereby alter sentence meaning.

Exaggeration

Exaggeration refers to word emphasis achieved by increasing word duration. A syllable's duration is increased by increasing the duration of its vowel, or by inserting an extra vowel into the syllable. This extra vowel, usually a schwa, represents an intermediate vocal tract configuration that exists as the vocal tract moves slowly from its initial to its target configuration. *Exaggeration* is a feature that applies first to the most prominent pitch–accented words. The word is emphasized by duration and pronunciation as well as pitch.

The *exaggeration* feature is unimplemented in the current version of the Affect Editor because its implementation requires the use of the Dectalk's phoneme mode. Phoneme mode has side effects (see Chapter 5) that make it desirable to retain pitch-accented words as text for as long as possible.

Pauses

This section describes the features of the Affect Editor's implementation of pausing — the likelihood of fluent or hesitation pause occurrence, and variations in the quality of pause onset. To indicate hesitation or fluent pauses, the Affect Editor inserts only unfilled pauses. Simplicity motivates this constraint — the relation of all the varieties of pausing to affect is not yet clear.

Pause type by likelihood of occurrence: In the sentence analysis, possible pause locations are marked for pause type and pause likelihood. This analysis is independent of any affect. However, whether the pauses are expressed in the final output is a function of the particular affect.

The Affect Editor distinguishes among pauses according to the semantic or intonational unit they precede. *Fluent pauses* are intonationally (therefore syntactically and semantically) governed and so are ranked in the initial sentence analysis according to the type of discourse constituent they precede. This ranking is meant to reflect the relative difficulty of

formulating the constituent. FLUENT−1 pauses precede an intonational phrase, FLUENT−2 pauses precede an intermediate phrase and FLUENT−3 pauses precede a constituent within an intermediate phrase. FLUENT−1 pauses have the longest duration, since they precede a segment of the most conceptual or discoursal weight. FLUENT−3 pauses are short since they precede units that are not normally intonationally distinguished[4].

The ranking of *hesitation pauses* is also intended to represent the relative difficulty of formulating the concept they precede. Dittmann[6] observed that most hesitation pauses follow the first function word in an intonational phrase. These locations are marked with HESITATION−1 pauses, and are given the highest rank. HESITATION−2 and HESITATION−3 designate hesitation pauses that occur at other locations. HESITATION−2 pauses follow a negative element ("not", "never") or an adverb when it precedes another content word. HESITATION−3 pauses follow an adjective. Pauses preceding adverbs are ranked higher than those before adjectives because adverbs have the greater scope — they can occur before more word categories.

Pause rank is the result of its location in the utterance structure and controls the order in which pauses are added to the utterance string. When the affect indicates little disruption of mental processes, pauses are infrequent and occur only between phrases (fluent pauses). As disruption increases, pauses are more frequent. More phrases are *separated* by pauses, and more phrases are *disrupted* by pausing from within.

Pause onset Pause onset characteristics vary with affect. They are regulated by the *pause discontinuity* parameter and describe the transition from sound to silence when a pause follows a word. The quality of pause onset varies from *smooth* to *abrupt*. A *smooth* transition to silence is accomplished by either lengthening the final phoneme of the word, or adding a phoneme which represents the vocal tract configuration in transition. An *abrupt* pause onset requires no intervening transition into silence. Phonation simply stops.

[4]Alternatively, only two fluent pause types are needed: one to mark an intonational unit (an intonational or intermediate phrase) and one to mark a normally non-intonationally distinguished unit (a constituent within an intermediate phrase). An intonational phrase is then preceded by two adjacent pauses of the first type — the second belongs to the first intermediate phrase — and is expressed as one long silence.

Precision of articulation

Precision of articulation refers to the precision or lack of precision with which phonemes are articulated. Slurring is achieved by increasing the incompleteness of closure for consonants and reducing vowels. The Affect Editor simulates incomplete consonantal closure by replacing normally unvoiced consonants with their voiced equivalents. Reduced vowels are usually represented by schwas. Conversely, precise articulation is achieved by substituting unvoiced consonants for voiced, and fully articulating vowels, even those normally represented by schwas.

The location of increased or decreased precision effects is ordered so that the effects show up first at word boundaries — at the word end and then at the word beginning — and then within the word. Enunciation at the word boundaries is increased by duplicating the initial or final phoneme so that attacks and decays are sharper. Following a word with a short abrupt silence also has a similar effect. Slurring at word boundaries is represented by adding a phoneme that represents a smooth transition to silence or to the next word. Coarticulation effects are increased. The Dectalk implementation of phoneme substitution and word boundary effects are illustrated in Tables B.3 through B.6.

Words with the highest prominence retain precise articulation for as long as possible. They are enunciated first, and slurred last.

4.2 Program Flow

This section describes the processes by which the Affect Editor generates the strings that specify synthesizer settings and the utterance composition. The calculation of synthesizer settings is described first, followed by a description of how the utterance contents are determined.

4.2.1 Synthesizer settings

Although both Affect Editor parameters and synthesizer settings are represented numerically, the mapping between them is not necessarily one–to–one, nor do linear changes in parameter values produce linear changes in synthesizer settings.

A synthesizer setting may be affected by more than one Affect Editor parameter. Its value is the sum of the effects of its controlling parameters. Each controlling parameter affects a percent of the Dectalk setting's final value. The total of the *absolute values* of these percents must be 1 (100%). A synthesizer setting may vary inversely or directly with an Affect Editor parameter. The assortment, amount and direction of Affect Editor parameter effects on the Dectalk are displayed in Table 4.1.

The mapping occurs in two stages. First, the results of Affect Editor parameter multiplication of the synthesizer's normal values are summed. The value for each Affect Editor parameter is a percent — *ParamPercent* — expressing its relation to its total range — (*ParamMax - ParamMin*):

$$ParamPercent = (ParamValue \,/\, (ParamMax - ParamMin))$$

Once calculated, *ParamPercent* is multiplied by the percent, from the mapping table, by which it affects the synthesizer parameter. Together, these two percents scale the default synthesizer value (denoting its value for neutral affect) for the setting, as follows:

$$SynthSettingValue = (ParamPercent \times MappingPercent \times SynthSettingDefault)$$

This occurs for each Affect Editor parameter that affects a synthesizer parameter. The final value from the synthesizer setting is the sum of all the effects of its controlling parameters.

In the second stage, this value (*SynthSettingValue*) is interpreted around the *norm* for the synthesizer setting to achieve a final synthesizer value for two conditions:

- If the ratio of the *SynthParamValue* to its synthesizer's range is greater than or equal

Name	Dectalk symbol	Controlling Parameter	Percent of control
average pitch	ap	*average pitch*	1
assertiveness	as	*final lowering*	.8
		contour direction	.2
baseline fall	bf	*contour direction*	-.5
		final lowering	.5
breathiness	br	*breathiness*	1
4th formant bandwidth	b4	*loudness*	-1
5th formant bandwidth	b5	*loudness*	-1
comma pause	:cp	*speech rate*	-1
gain of frication	gf	*precision of articulation*	1
gain of aspiration	gh	*precision of articulation*	1
gain of voicing	gv	*loudness*	.6
		precision of articulation	.4
hat rise	hr	*reference line*	1
laryngealization	la	*laryngealization*	1
loudness	lo	*loudness*	1
lax breathiness	lx	*breathiness*	1
period pause	:pp	*speech rate*	-1
pitch range	pr	*pitch range*	1
quickness	qu	*pitch discontinuity*	1
speech rate	:ra	*speech rate*	1
richness	ri	*brilliance*	1
smoothness	sm	*brilliance*	-1
speech rate	:ra	*speech rate*	1
stress rise	sr	*accent shape*	.8
		pitch discontinuity	.2

Table 4.1: **Mappings from Affect Editor parameters to Dectalk settings.** Negative values indicate settings that vary inversely with the Affect Editor parameter that controls it.

to .5, subtract .5 from the ratio to obtain the percent above the **norm**, and then calculate:

$$DefaultValue + (2 \times PercentAboveNorm \times (\max - DefaultValue))$$

- If the ratio of the *SynthParamValue* to its range is less than .5, calculate as follows:

$$2 \times ValueRangeRatio \times DefaultValue$$

Appendix A describes the structures involved in these calculations.

4.2.2 Utterance composition

The utterance that the Affect Editor initially sees is composed of words grouped into discourse constituents and annotated for *possible* pitch accents and pause locations. In the course of applying the speech correlate model, word and phrase features are added or blocked via further annotation. The annotations are qualitative feature descriptions. They become quantitative as the the Affect Editor interprets for the Dectalk. The utterance is finally represented by a string which combines prosodic markings, phonemes and straight text. The path that turns the initial recursive utterance structure into an ASCII string is as follows:

1. Phrases and words are marked with feature descriptors, as per current affect specifications. This completes the abstract phonological description.
2. Feature descriptors are interpreted for the Dectalk such that new phrase structures and words are created. A Dectalk word is composed of the characters that precede it, the text or phonemes of the word itself, and characters that follow it (see Figure A.3.7 in Appendix B).
3. The Dectalk string is assembled from phrase header information, and for each word, from the combination of prosodic annotation, the phonemes that describe word boundary activity, the pronunciation of the word itself — represented by phonemes or text — and the pause that follows the word.

4. After the synthesizer settings are adjusted, the utterance string is sent to the Dectalk and spoken.

The parts of this process — assigning phrase and word features, interpreting of each feature descriptor and the construction of the Dectalk string — are described in following sections.

Assignment of phrase and word features

Phrase and word features assignments result from the following steps:

1. **Filter out excess features.** Adjust the sentence representation, originally marked for maximum word stress and pause occurrence, for the current affect.
   - Use the *stress frequency* value to determine which of the accentable (content) words receive pitch accents.
   - Use the *fluent pauses* and *hesitation pauses* values determine where pauses will occur.
2. **Assign phrase and word features.** Use the values of Affect Editor parameters to select the appropriate phrase and word feature descriptors[5].
   - Use the *accent shape* value to select the *prominence* descriptor that will apply globally to all pitch–accented words.
   - Use the *contour slope* value to select the contour slope descriptor.
   - Use the *exaggeration* value to select the exaggeration descriptors.
   - Use the *pause discontinuity* value to select the pause onset descriptors.
   - Use the *enunciation* value to select the enunciation descriptors.
3. **Mark the pitch contour peak.** This is the first accentable for a downward slope or the last for an upward slope.
4. **Assign word features.**

[5]The *exaggeration* parameter is included even though it is not currently implemented for theAffect Editor.

- Assign enunciation descriptors to all words.
- Assign *prominence* descriptors to pitch–accented words.
- Assign *exaggeration* descriptors to pitch–accented words.

5. **Assign pause features.** Mark fluent and hesitation pauses with a pause onset descriptor.

Appendix B explains in detail the treatment of specific of phrase and word feature descriptors.

Once features based on syntactic or semantic groupings (e.g., pauses) have been assigned, the recursive utterance structure can be discarded. A depth–first tree traversal recovers the surface structure of the utterance. At this point, the utterance is represented by a simple list, as displayed in the *phonology* window of the user interface (see Figure 4.1).

Interpretation for the Dectalk

Once feature descriptors are assigned, they are interpreted for the Dectalk. For a word, this may involve:

- placing it in phoneme mode
- preceding it with a word stress marking
- preceding or following it with a phoneme string to effect the specified *precision of articulation* value
- following it with a phoneme string that includes the silence phoneme for pausing.

. For a phrase, this may involve:

- altering the Dectalk parameters that are controlled by *speech rate, final lowering* and *pitch range.*
- changing the default final punctuation from a period to a comma, for minimal *final lowering*, to effect a continuation rise.

At this point, the utterance is represented as a simple list that contains phrase feature information and Dectalk approximations of the word features specified in the abstract phonology. To obtain the final utterance string, the list is traversed and the separately stored phrase and word features it contains are extracted and combined into a string. This string mixes straight text, phonemes and prosodic symbols.

4.3 The User Interface

This section describes the Affect Editor user interface. The Affect Editor is implemented on a Symbolics 3650 Lisp machine and makes use of the Symbolics graphics and window packages. Editing and command execution is possible with either the keyboard or the mouse. Figure 4.1 shows the Affect Editor interface. The screen is divided into areas for emotions, sentences, program output or commands. The emotion windows display a list of emotions, the current emotion and its speech correlates. The sentence windows display a list of pre-analyzed sentences, the current sentence and the results of applying the current emotion to the current sentence. The output windows display the utterance string and the Dectalk settings. The command menu contains frequently used commands that are accessible with the mouse. The other command window accepts keyboard input to the Affect Editor or the Lisp interpreter.

4.3.1 Emotions

The *EMOTIONS* window contains a list of emotions. Choosing one will cause the current sentence to be spoken with the qualities for that emotion. The emotion list is dynamic. New emotions can be added and old emotions deleted. The column immediately to the right displays the current emotion and its speech correlates. The correlates (parameters) are organized by the area they affect in the *PITCH*, *TIMING*, *VOICE QUALITY*, and *ARTICULATION* windows. The bulk of the editing occurs in these four windows. Parameter values can be changed and used to generate a new version of the current utterance. The

Figure 4.1: The Affect Editor user interface. The current emotion is "Afraid". The current sentence is *"You've asked me that question a thousand times."*

current configuration can be further changed, abandoned or saved as a new emotion in the *EMOTIONS* window. Figure 4.2 shows the emotion list and parameter values for "Angry", the current emotion.

4.3.2 Sentences

The *SENTENCES* window contains a list of pre-analyzed sentences. Currently, this list is not dynamic. Sentences must be added *prior* to running the Affect Editor program and may not be deleted. The current sentence is highlighted in boldface. Figure 4.3 shows an example of the sentence windows. The internal representation for the sentence — a tree composed of discourse constituents — is displayed in the *phrase structure* window. The information stored in the internal representation and all its components (discourse constituents and individual words) is accessible by the via the mouse. Figure 4.4 shows more clearly the tree structure of the utterance in the *phrase structure* window.

4.3.3 Sentence processing

The sentence processing windows show how, the initial tree structure of the sentence becomes linear as the original constituent structure is discarded. The *phrase structure* window shows the initial tree structure. The *phonology* window displays the abstract [linear] phonology derived from this structure. It shows the results of feature annotation and pause marking. The *Dectalk phonology* shows the subsequent interpretation of the phonological representation for the Dectalk. Pauses that follow a word in the abstract phonology are incorporated into words in the Dectalk phonology. They are not separate objects in this window (see section A.3.7 in Appendix A). Figure 4.5 shows an example of the contents of the phonology windows.

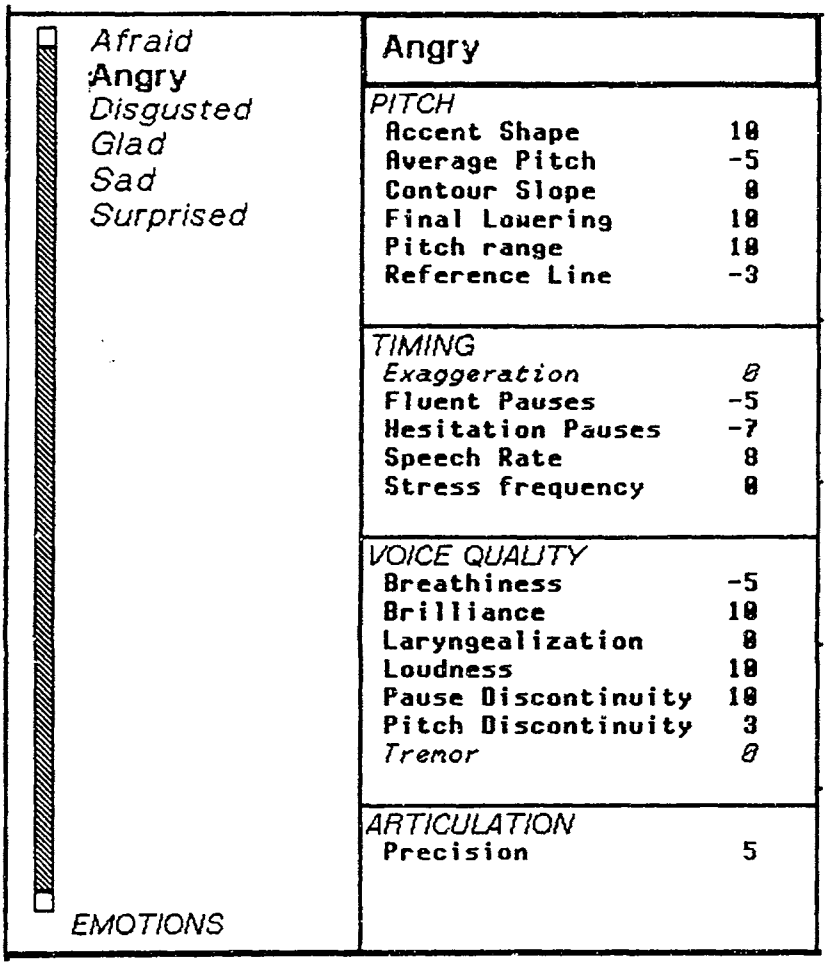


Figure 4.2: **The windows for the emotion components of the Affect Editor.** The parameters of the currrent emotion are edited to create new affect configurations.

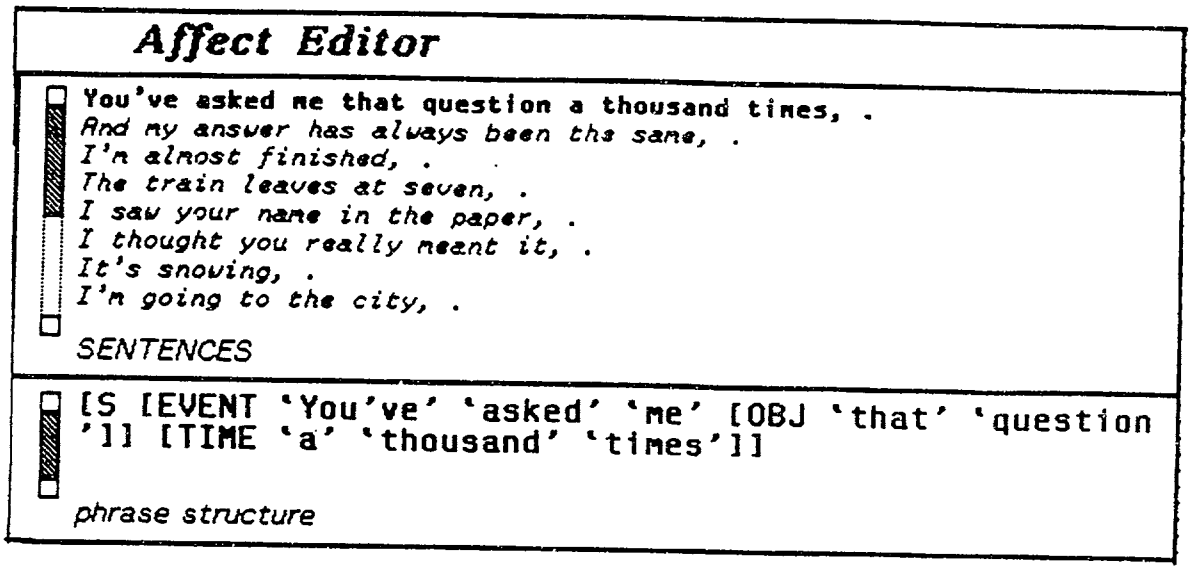


Figure 4.3: The windows for the sentence components of the Affect Editor. The current sentence is highlighted the *SENTENCES* window, and and its recursive internal representation displayed in the *phrase structure* window.

[S [EVENT You've asked me [OBJ that question]] [TIME a thousand times]]

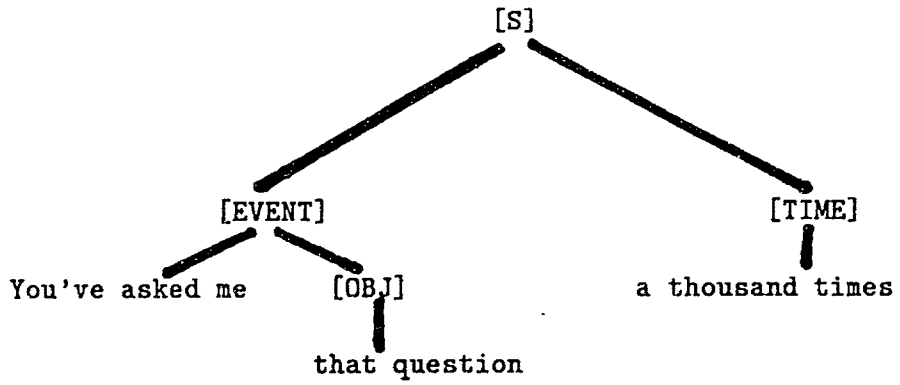


Figure 4.4: The tree structure of the discourse constituents for *"You've asked me that question a thousand times."*

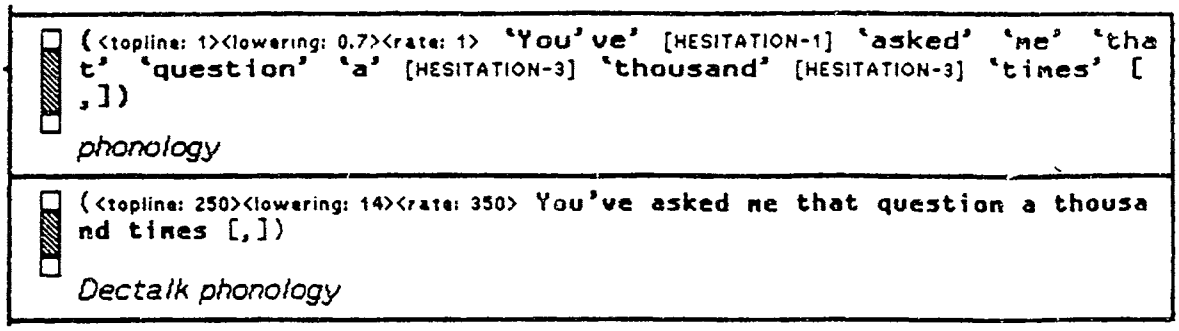


Figure 4.5: **The windows for the abstract and Dectalk–adapted phonologies.**

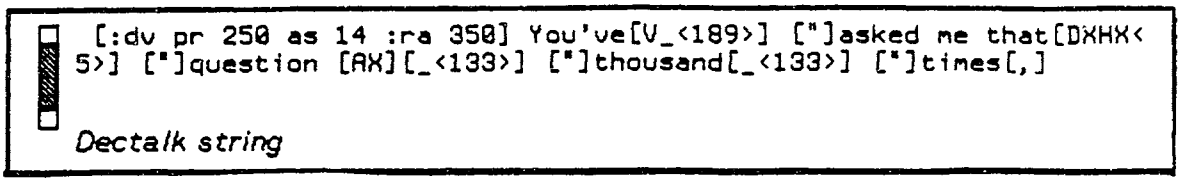
4.3.4 Output strings

The result of sentence processing is the string displayed in the *Dectalk string* window. Figure 4.6 shows an example of the Dectalk strings generated for the sentence *"You've asked me that question a thousand times."* for fear, anger and sadness.

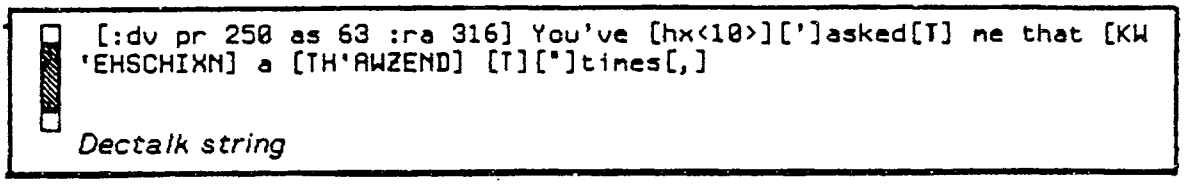
The other string generated by the Affect Editor controls the Dectalk settings. The Dectalk setting symbols and values are displayed in the *DECTALK SETTINGS* window, as shown in Figure 4.7.

4.3.5 Commands

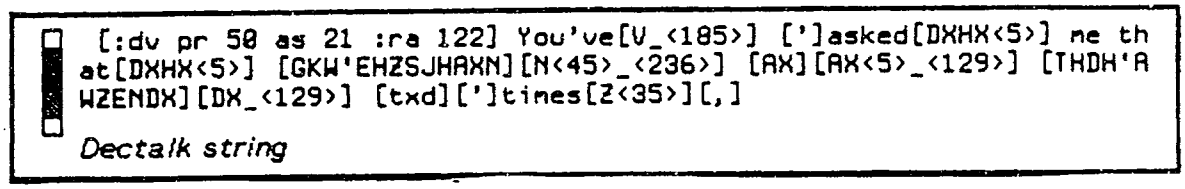
Commands may be selected from the Affect Editor command menu (see Figure 4.1) with the mouse, or entered in the Affect Editor command window from the keyboard. Most of the commands fulfill generic editing functions such as *selecting* an affect configuration to edit, *changing* the current configuration, *saving* it or *deleting* it. Additionally there are *playback* functions with which to test the current configuration, the effects of systematically increasing or decreasing the parameter values or to compare the applications of several configurations to the sentences.



(a) Fear



(b) Anger



(c) Sadness

Figure 4.6: **The utterance string generated for the sentence** *"You've asked me that question a thousand times."* **and for the emotions (a) fear (b) anger (c) sadness.**

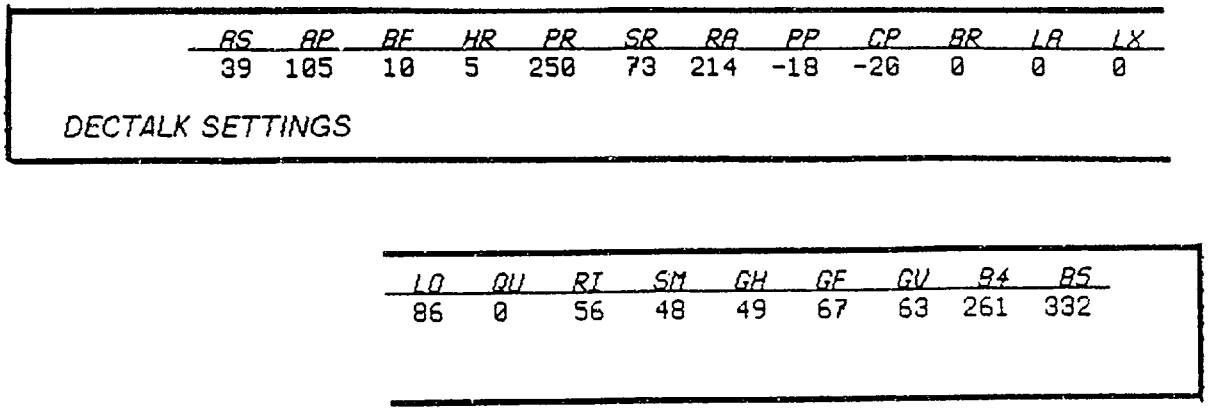


Figure 4.7: **The *DECTALK SETTINGS* window, displaying the synthesizer settings for "Glad".**

4.4 Summary

The Affect Editor is a tool with which to produce and vary affect in synthesized speech. Its parameters define different aspects of the acoustical and linguistic correlates of speech. The parameter values and a representation of an utterance, minimally marked for syntax and intonation, are interpreted to produce two strings — one specifying synthesizer settings and the other specifying the word pronunciation and prosody for the utterance.

The most easily controlled speech correlates are those explicitly represented in the synthesizer's own parameter set. These mainly control aspects of voice quality, and secondarily, prosody. Control of other features, mainly those of prosody and articulation, must be simulated. This is accomplished using the *stress frequency* parameter to designate which words will exhibit these features, marking these words qualitative description of the feature and then interpreting them for the Dectalk. The simulated continua of change for some parameters points out features that are currently missing in the Dectalk (and other synthesizers) and indicates approaches for expanding its capabilities. Synthesizer capabilities and limitations, and their influence on the Affect Editor, are explored in greater detail in the Chapter 5.

Chapter 5

The Dectalk

5.1 Capabilities

The Dectalk 3 was selected for the control it allows over vocal tract settings and intonation. However, this chapter begins with a discussion of the Dectalk's deficiencies. It focuses on two: *side effects*, where changes to one speech attribute produce unwanted changes in another, and *limits* of the Dectalk's capabilities. The specifics of the Dectalk's strengths and weaknesses will become apparent in the section in which the implementation of each Affect Editor parameter is described. The chapter closes with a discussion of the features of a speech synthesizer that would make it easier to achieve convincing affect.

5.1.1 Side effects

Unwanted side effects are produced in five instances:

- when words are sent to the Dectalk as [Arpabet] phonemes instead of text
- when Dectalk parameter specifications are embedded in the middle of an utterance
- when some types of word stress markings are applied
- for average pitch specifications outside of the normal pitch range

- for extremes of brilliance specifications.

The side effects of phoneme, phrase parameter specifications, word stress markings, average pitch changes and extremes of brilliance are discussed in the following sections.

Side effects of phoneme mode

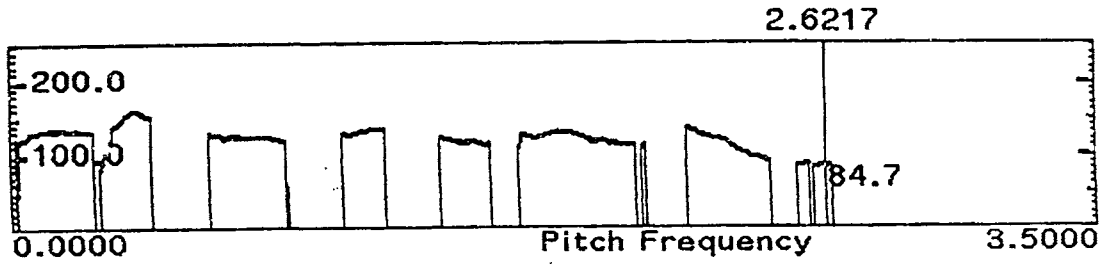
When words are expressed as phoneme strings instead of English text, side effects may occur for features of F0, rhythm and pronunciation. These effects and their corresponding compensatory measures are discussed in this section.

Prosodic effects

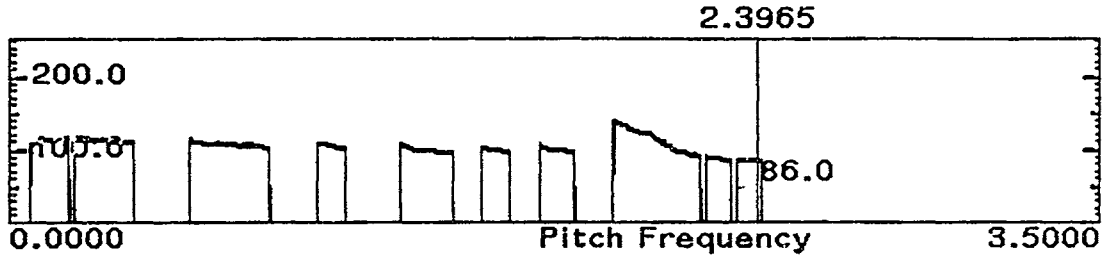
When a word is in phoneme mode, its syntactic function — minimally, whether it is a function or content word — is hidden from the Dectalk. Unfortunately, the Dectalk relies on this information to shape the F0 contour, assign relative duration and to locate pauses. When it is lacking, the Dectalk cannot apply the full force of its syntactically driven prosodic rules. Figure 5.1 contrasts the pitch tracks of an utterance sent first as text and then as phonemes.

Unless the F0 value is explicitly specified, phoneme mode tends to produce, F0 values lower than what is normally generated for straight text. This is illustrated by the pitch tracks in Figure 5.2.

Phoneme mode may adversely affect the rhythmic component of prosody. This is mostly a result of the requirement that pitch and duration instructions be embedded within a phoneme string. Pitch in Hz and duration in milliseconds must be specified for a specific phoneme. There is no higher level representation that would allow, for example, pitch specification for the stressed *syllable*. If exact pitch control is desired, a whole word must be sent as phonemes. This creates long ASCII strings that specify only momentary events. Unfortunately, the Dectalk appears to be unable to handle many specifications when they

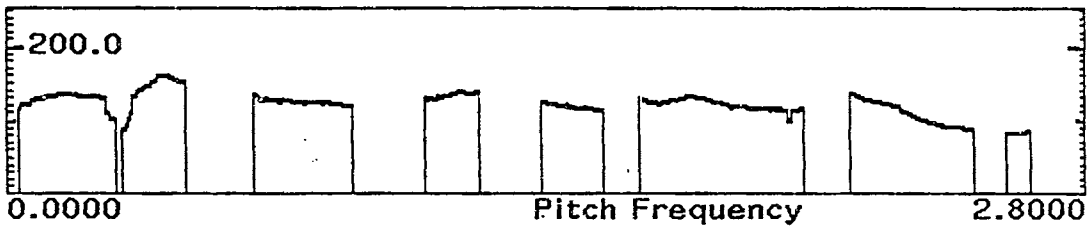


(a) plain text

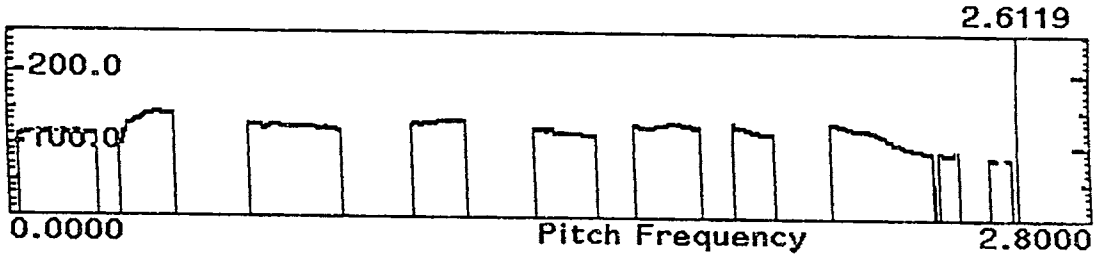


(b) phoneme mode

Figure 5.1: **Effect of Dectalk phoneme mode on prosody (rhythm and pitch).** Pitch tracks for *"You've asked me that question a thousand times."* sent as (a) text (b) Arpabet phonemes. As text, *"asked"* and *"question"* receive pitch accents, but in phoneme mode they do not. Also, the duration of utterance sent as phonemes is shorter.



(a) plain text



(b) phoneme mode

Figure 5.2: **Effect of Dectalk phoneme mode on the F0 contour.** Pitch tracks for *"You've asked me that question a thousand times."* (a) sent as plain text, showing the application of the Dectalk's stress rules (b) sent as Arpabet phonemes with word stress explicitly marked. The F0 for accented words is slightly higher for plain text.

map to a short instant of time. It may stop speaking for a moment, or speak with an altered voice quality.

The cessation of output is the most unpredictable side effect of phoneme mode. Because of this, a pitch contour cannot be specified with explicit F0 and duration values. And Generative Intonation, whose principles drive the calculation of such a pitch contour, cannot be fully implemented for the Affect Editor.

Pronunciation effects

A word in phoneme mode is almost always spoken as if it were being recited singly, and not as part of an utterance. This is called its *citation* pronunciation, where the word is enunciated more clearly than for normal speech. Articulation is more precise, such that the vowels in unstressed or reduced syllables tend *not* to be reduced. This fuller articulation is modified slightly by phoneme coarticulation rules. However, since fully articulated syllables are of longer duration, altered pronunciation produces altered rhythm.

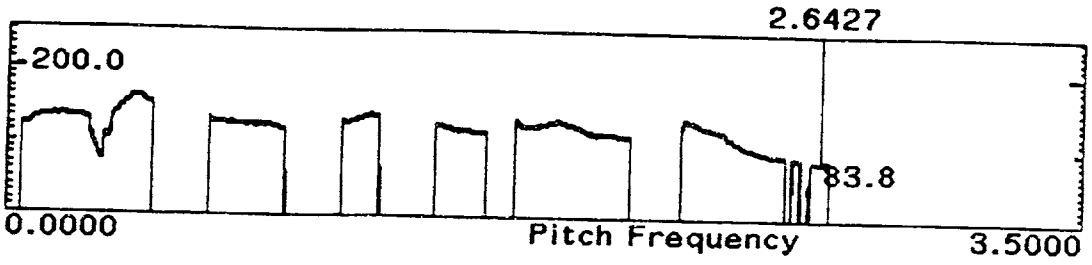
Conclusion

In light of the many side effects of phoneme mode, the Affect Editor strives to retain words as text. Phoneme mode is deliberately employed for only two purposes — to disguise the grammatical function of a content word and thereby prevent the Dectalk from applying a pitch accent to it, and to effect extreme slurring. For extreme slurring, reduced phonemes are substituted for the relatively more precise enunciation of the default pronunciation.

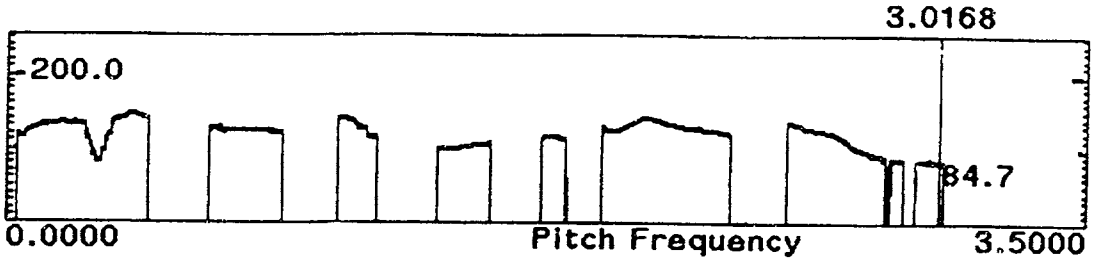
Side effects of phrase parameter specification

It is often necessary to change the pitch range, final lowering or speech rate parameters of an intermediate phrase to reflect its role in the discourse. However, the insertion of such instructions in the middle of an utterance invariably causes a pause in the output at the point where the parameter change is specified (see Figure 5.3) and introduces phrase final intonation as well. Indeed, any reset of a Dectalk vocal tract or prosodic parameter within

an utterance introduces unwanted pausing and phrase final contours.



(a) plain text



(b) with mid–sentence phrase parameter

Figure 5.3: **Effect of mid–sentence parameter specification on rhythm.** Pitch tracks for *"You've asked me that question a thousand times."* (a) sent as plain text (b) sent with phrase parameter changes just before "a thousand times". The second pitch track shows a longer pause between "question" and "a", and a continuation rise for "question".

To avoid unspecified pauses from parameter resets, Dectalk parameters are set once, at the beginning of the utterance. Although pitch range, final lowering and speech rate are part of the phrase description for all subsequent phrases, they are never actually sent to the Dectalk. Simulation of attributes that might change dynamically over the course of an utterance — for example, a change from voicing to whispering — is not even attempted because it would introduce unwanted silence.

Side effects of word stress markings

The main unwanted side of effect of word stress markings is on the F0 contour. The Dectalk recognizes six symbols for effecting lexical stress:

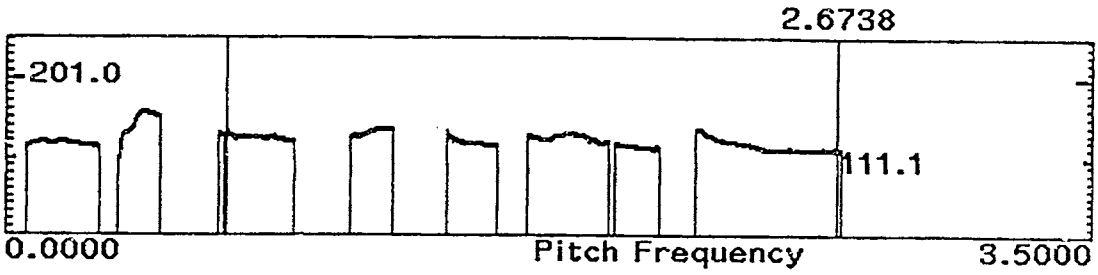
- "ˈ" for primary stress
- "ˌ" for secondary stress
- "ˈˈ" for emphatic primary stress
- "\" for pitch fall
- "/" for pitch rise
- "/\ " for pitch rise and fall [4].

Pitch accents apply only to the stressed syllable of a pitch accented word. Any of the primary stress markings should do the same. Or at least, the influence of a word stress marking should only affect the contour for duration of the word. This, however, is not the case with the rise/fall stresses nor, at times, for emphatic primary stress.

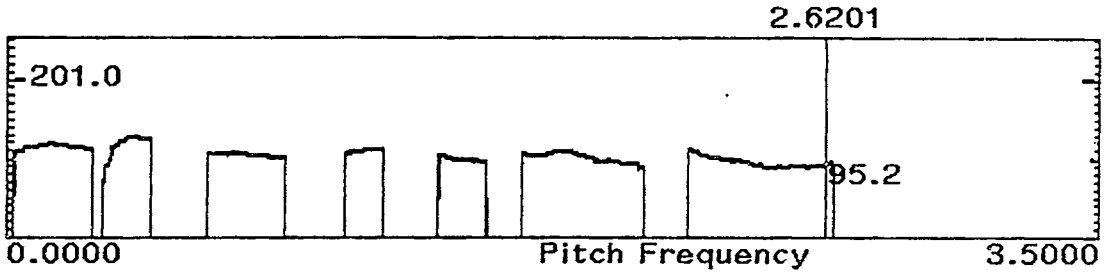
The influence of the rise/fall stresses tends to extend inappropriately beyond the stressed word, even to other accented words. In some instances, F0 remains low after a pitch fall even to the point of overriding the effect of adjacent stress markings. In Figure 5.4b the pitch accent height of *"asked"* is reduced because of a pitch rise on *"question"*. The contour Figure 5.4c is similar, even though the pitch rise has been moved to *"thousand"*.

Additionally, the application of the rise/fall stresses is constrained such that a pitch rise must be the first stress in a phrase, and pitch rises and falls must alternate. Were these operating as true pitch accents, their use would be constrained by the type and strength of word salience (i.e., pitch accent type and pitch accent prominence) but *not* by the type of preceding stress.

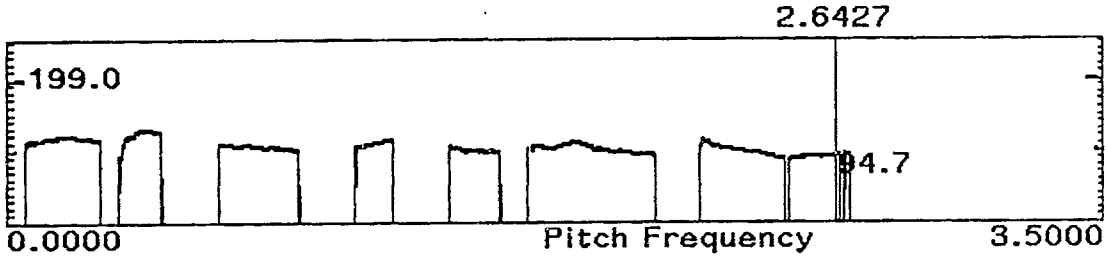
In some instances a sequence of emphatic ("ˈˈ") stresses lowers the F0 for succeeding unaccented words (see Figure 5.5). For other instances, a "ˈˈ" stress lowers the F0 such that



(a) straight text



(b) pitch rise on "question"



(c) pitch rise on "thousand"

Figure 5.4: **Effect of the Dectalk's pitch rise word stress markings on F0.** Pitch tracks for *"You've asked me that question a thousand times."* (a) sent as straight text (b) with *"question"* marked with a pitch rise (c) with *"thousand* marked with a pitch rise. In (b) and (c), the accent height of the first pitch accent is reduced because of a pitch rise that follows. Despite the different pitch rise locations, the F0 contours are pretty much the same.

a succeeding "1" stress is without effect on the F0 contour (see Figure 5.6). This Dectalk artifact disturbed the otherwise progressive changes in accenting frequency and style implemented in the Affect Editor.

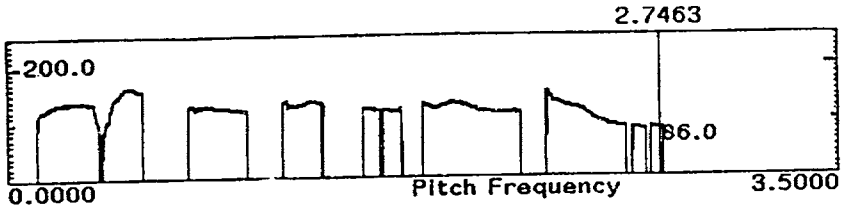
To avoid F0 side effects from word stress markings, rise/fall markings are not used, and emphatic stress is used sparingly.

Side effects of average pitch changes

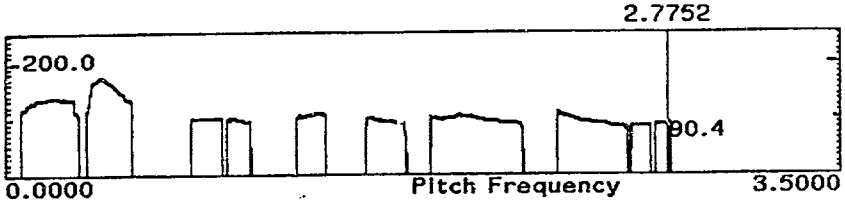
As the Dectalk's `average pitch` value changes, the impression is eventually that of a different speaker instead of one whose emotional state has changed. This follows from the implementation, whereby the Dectalk's `pitch range` boundaries are raised or lowered in parallel, proportional[1] to corresponding changes in `average pitch` (see Figure 5.7). In consequence when average pitch changes so does the speaker baseline. As `average pitch` is raised, the voice quickly becomes a falsetto. This may be valid for *male* voices, but it should only happen for very high `average pitch` values.

Speaker baseline, as defined by Generative Intonation, is a function of the speech apparatus, and therefore part of a speaker's identifying characteristics. When the baseline (the bottom of the pitch range) changes, it may well convey a change of speaker instead of a change in affect. It is likely that we detect changes in average pitch by gauging its perceived relation to the overall pitch range, and especially to its distance from the speaker baseline. From this comparison, we derive discourse information (e.g., whether the speaker is speaking up, and therefore expanding the pitch range to signal the start of a new topic), affective information (e.g., a higher than normal average pitch indicates agitation, a lower than normal average pitch indicates calm or depression). However, because the Dectalk implementation of average pitch has side effects, the Affect Editor restricts the range of acceptable `average pitch` values to just part of what is possible. Additionally, the *reference line* parameter can be used to vary the perception of average pitch. It controls the `hat rise` setting, which

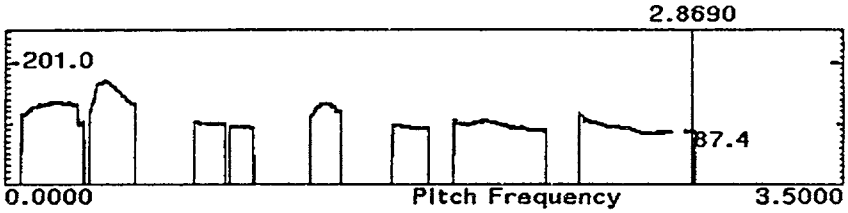
[1] about one-quarter of the distance between the bottom and the top of the pitch range



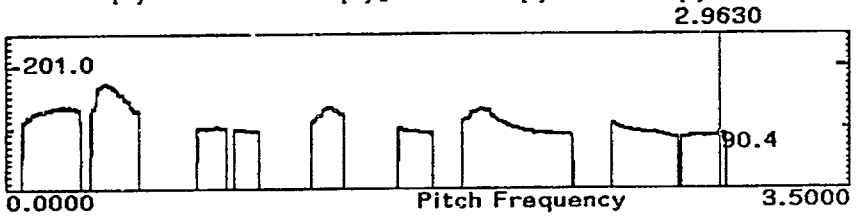
(a) *"You've [']asked me that [']question a [']thousand [']times."*



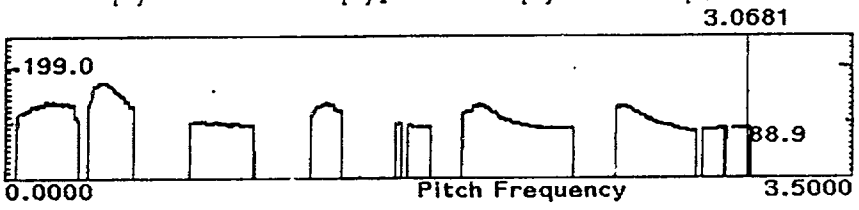
(b) *"You've ["]asked me that [']question a [']thousand [']times.*



(c) *"You've ["]asked me that ["]question a [']thousand [']times."*

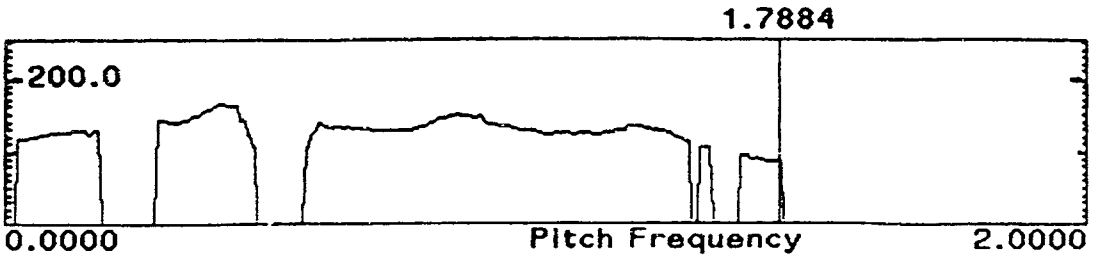


(d) *"You've ["]asked me that ["]question a ["]thousand [']times."*

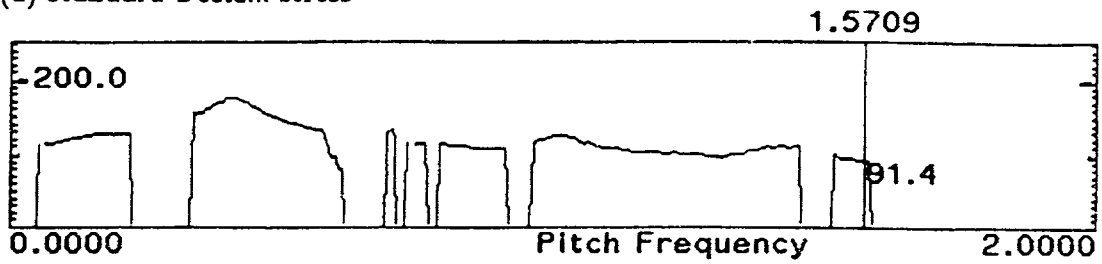


(e) *"You've ["]asked me that ["]question a ["]thousand ["]times."*

Figure 5.5: **The effect of progressive addition of Dectalk emphatic stress on F0.** Pitch tracks for the successive addition of emphatic stress ("¨"): (a) *"You've [']asked me that [']question a [']thousand [']times."* (b) *"You've ["]asked me that [']question a [']thousand [']times.* (c) *"You've ["]asked me that ["]question a [']thousand [']times."* (d) *"You've ["]asked me that ["]question a ["]thousand [']times."* (e) *"You've ["]asked me that ["]question a ["]thousand ["]times."*



(a) standard Dectalk stress



(b) initial emphatic stress

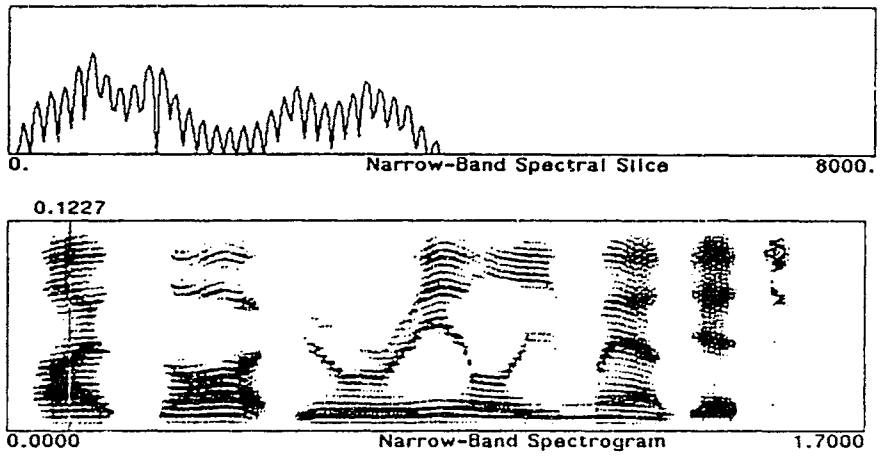
Figure 5.6: **Effect of Dectalk emphatic stress on subsequent primary stresses.** Pitch tracks for (a) *"I [']thought you [']really [']meant it'."* marked for standard Dectalk stress (b) *"I [*]thought you [']really [']meant it."* with initial emphatic stress.

does not as drastically affect pitch range or voice quality.

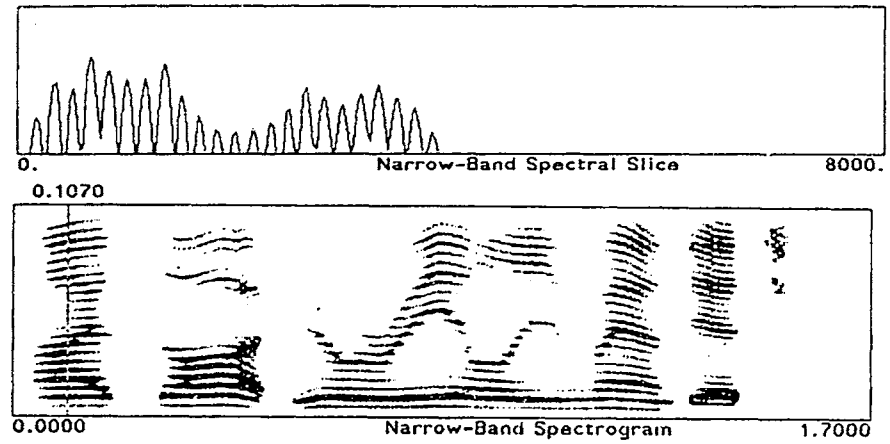
Another corrective measure is simply to avoid high values for the Affect Editor *average pitch* parameter, except when high pitch is intrinsic to the expression of an emotion. Thus, for fear, *average pitch* is set to 10, the maximum on the -10 to 10 scale, but for most other emotions it is set much lower. Anger and gladness, for example, have an *average pitch* value of only -3.

Side effects of changes in brilliance

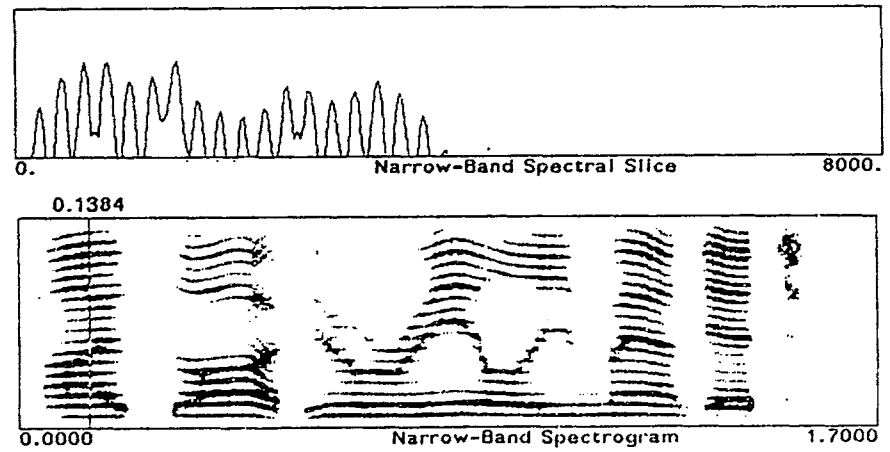
The Affect Editor's *brilliance* parameter controls the ratio of high to low frequency energy. These changes also eventually convey, although less strongly than for `average pitch`, an impression of a change of speaker instead in affect. This seems incorrect, but since it is a less drastic effect, no attempt is made to compensate.



(a) average pitch set to 120 Hz



(b) average pitch set to 160 Hz



(c) average pitch set to 200 Hz

Figure 5.7: **Effect of Dectalk "average pitch" on speaker identity.** Narrow band spectrograms and narrow band spectral slices for *"I ['/]thought you ['/]really ['/]meant it."*, with an average pitch of: (a) 120 Hz (b) 160 Hz (c) 200 Hz.

5.1.2 Limitations

The section describes the three most apparent Dectalk limitations — the way intonation is specified and controlled, its lack of dynamic parameters and the limits of its processing capacity.

Intonation specification

The propositional content of intonation is not adequately represented by the set of Dectalk intonational markings. Precise F0 contour specification cannot be achieved through these symbolic markings, but rather, as discussed in section 5.1.1, only through F0 and duration specifications embedded in a phonemic representation. The Dectalk's default intonation, and the means for changing it, are described below.

Dectalk intonation is based upon an F0 contour shape called a "hat rise" [4]. This describes the tendency for F0 to rise on the first stressed syllable and remain high until the end of the utterance, when there is either a dramatic fall or a fall–rise pattern [12]. The default dramatic fall at the end is modified if the Dectalk's `assertiveness` value is low, or if the final punctuation is "!" or "?" in which case alternate terminal contours are applied. Word stress (pitch accents) takes the form of a local rise above the hat–shaped intonation pattern [12]. The shape of the default pitch contour may be altered with:

- lexical markings for standard stress ("'"), emphatic stress ("") or for pitch rise and fall
- the ")" clause marking to indicate the start of a verb phrase
- a comma, to indicate the end of a clause
- end of a sentence punctuation – "." for a declarative, "!" for an exclamation and "?" for a question
- the Dectalk `stress rise` setting, which adjusts the height of the F0 peak on stressed syllables

- the Dectalk `hat rise` setting, which adjusts the height of the initial rise for the hat shaped F0 contour
- the Dectalk `assertiveness` setting, which affects final lowering at the end of an sentence
- the Dectalk `baseline fall` setting, affecting slightly the contour slope.

The combinatorics of mixing the Dectalk's semantically determined prosodic parameters — word stress markings, terminal contour parameters — with those that are physiologically determined — speech rate, pitch range — are powerful enough to create a variety of melodies. However, these parameters are not sufficiently local, atomic, or orthogonal so full and precise control is not possible.

The set of word stress markings cannot be expanded for more precision. However, precision can be achieved in the opposite direction, by excluding markings with unpredictable effects on F0. Thus, in the Affect Editor, finality and uncertainty are conveyed only via the `assertiveness` and `baseline fall` settings and by punctuation — a period for finality or a comma for tentativeness. Exclamation points, which emphatically convey finality, or question marks, which can convey uncertainty, are excluded because they are redundant. Also, they affect not only the sentence ending, but the whole F0 contour that precedes it.

5.1.3 Lack of dynamic parameters

The lack of dynamic parameters means that parameters, such as loudness, voicing, and the ratio of high and low frequencies, cannot vary over the course of an utterance without being explicitly reset. However, resets introduce pauses into the speech output, so no attempt is made to simulate dynamic changes for these parameters.

5.1.4 Processing capacity

As previously noted, the Dectalk's output is interrupted when it processes complex instructions that must be realized over a short period of time (measured in milliseconds). This precludes the extensive use of phoneme mode, and the introduction of mid–sentence changes.

5.2 Implementing Affect Editor Specifications

This section describes how the Affect Editor's sentence annotations are interpreted for the Dectalk, and how each parameter of the Affect Editor's internal representation is implemented. The Dectalk's capabilities and limitations should be clearer as a result.

5.2.1 Sentence annotation

In this section the contents of the sentence annotation are reviewed, and the implementation of pitch accents, phrase parameters and discourse information are discussed.

A brief review

To understand how the sentence annotation is adapted for the Dectalk, it is useful to review the contents of the annotation. Basically, a sentence is annotated with discourse and intonational information. Its words are grouped into intonational phrases and discourse constituents prior to any further processing by the Affect Editor. The annotation describes:

- **the type of information in the discourse constituent**: This is something that a text generator, speaking through the Affect Editor, might use to indicate the relevance of the utterance to the discourse. However, the sentences the Affect Editor utters are not part of any discourse. so the information in their discourse constituents is not obviously domain dependent (there is no discourse, so no discourse domain). Instead,

the information is described by its *generic* content —- e.g., time, state, event —- or role — e.g., sentence, subject, or object. The information descriptions are included to justify the division of the sentence into constituents. Figures 4.3 and 4.4 show examples of this annotation.

- **phrase parameters**: For intonational and intermediate phrases these are *pitch range*, *final lowering* and *speech rate*. Unless the otherwise specified, the Affect Editor defaults are an L phrase accent and an L% final boundary tone. Together they describe a falling terminal contour. The Dectalk doesn't interpret Generative Intonation annotation, so an **L L%** terminal contour is effected with a phrase final period, while the **H H%** or **L H%** contour is effected with a phrase final comma.
- **pitch accents**: Pitch accents are described by type and prominence. Pitch accent types are those of Generative Intonation. Pitch accents indicate the word's propositional content, but, because the Dectalk cannot reliably reproduce them, they serve only to indicate that a word is accentable. Prominence, however, is directly useful to the Affect Editor. It determines the order in which word effects will manifest in a sentence.

Discourse information

The Dectalk makes little direct use of discourse information other than to faithfully carry out the pausing as previously determined from the division of the sentence into discourse constituents.

Phrase parameters

Phrase parameters are those that affect the whole phrase — pitch range, reference line, speech rate, etc — and those describing the terminal contour. Because the Dectalk pauses whenever its vocal tract or prosodic parameters are reset, parameters affecting the whole phrase are set only at the beginning of an utterance. Phrase parameters for subordinate clauses are not sent to the Dectalk.

There is no direct Dectalk translation for terminal contours. Phrase accent and boundary tones can be loosely approximated by varying the `baseline fall` in combination with a comma to effect a continuation rise (**H H%** or **L H%**) or a period to simulate finality (**L L%**).

Pitch accents

There is no direct Dectalk translation for the Generative Intonation pitch accents. The stresses recognized by Klattalk, and therefore by the Dectalk, are always excursions upwards from the original hat rise contour [12]. Thus only approximations of **H*** accents — **H***, **H*+L** and **L+H*** — are possible. The rise and fall markings may be a rough approximation of **L*** accents (**L***, **L*+H**, **H+L***) but their effects are too unpredictable to be of use. Consequently, pitch accent notation is significant only insofar as it reveals an intonationally stressed word. Currently only variations of **H*** accents annotate words.

5.2.2 Parameters of the internal representation

This section describes the implementation of each Affect Editor parameter. The parameters are classified according to their effects on **pitch, timing, voice quality** or **articulation**. Two parameters, *exaggeration* and *tremors*, are never implemented because of side effects (*exaggeration*) or synthesizer limitations (*tremors*).

Pitch parameters

This section discusses the effect of the Dectalk on the implementation of the Affect Editor's pitch parameters — *accent shape, average pitch, contour slope, final lowering, pitch range* and *reference line.*

Accent shape As with most speech effects, the effects of physiologically–based discontinuities are magnified in accented words. Most effort goes into the articulation of these words, as per sentence meaning. The Dectalk `stress rise` setting regulates the F0 height of accented words such that the greater the distance of the peak F0 from the average pitch, the greater the rate of F0 change within the word and the greater the perceived discontinuity of F0 transitions. `Stress rise` varies directly with accent *shape*.

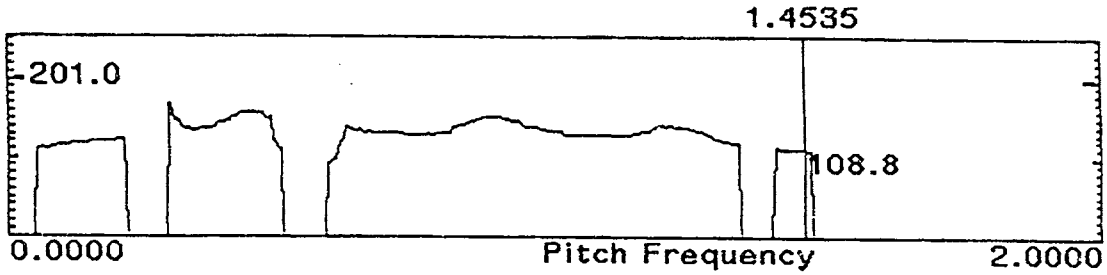
The *accent shape* parameter affects the kind of accenting received by a word. As it increases, the F0 transitions around an accent become progressively steeper and the contour increasingly perturbed. To effect this with the Dectalk, the accent marking changes eventually from primary to emphatic stress. In Generative Intonation parlance, this suggests that the accent type changes from **H*** to a bitonal (**H*+L** or **L+H***). There is nothing in the theory that relates accent type to affect, so this is not necessarily a principled extension of the theory. It is justified on two counts – the mapping of Dectalk stress types to pitch accents is not exact, and ultimately, increased F0 fluctuations in accented words are perceptually correlated with changes in affect.

Average pitch The Dectalk `average pitch` setting varies directly with changes in the Affect Editor's *average pitch* parameter. However, the range over which it varies is restricted by the Affect Editor. Because of previously mentioned side effects, high *average pitch* values are used sparingly.

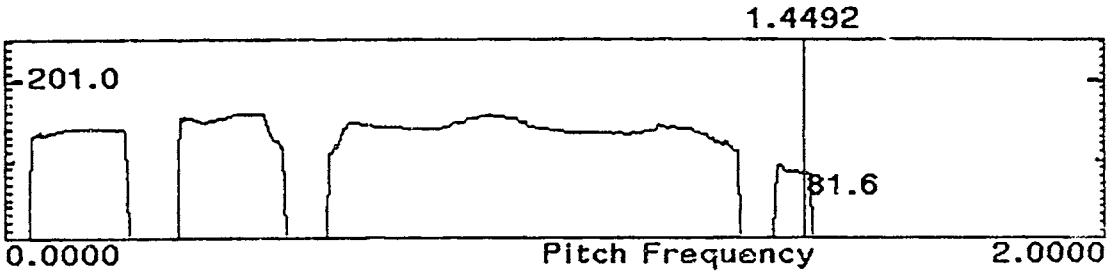
Contour slope The Dectalk's `baseline fall` setting varies with the value of *contour slope* as well as with *final lowering*.

Contour slope is also approximated by raising the F0 of the first accentable word for a falling contour, or the last accentable word for a rising contour. These constitute the peak F0 values for the pitch contour. Once the slope peak is identified, the F0 is raised by applying Dectalk emphatic stress to the word. The accent height at the location of the stress is determined by settings, e.g., `stress rise` and `pitch range`, that affect the whole pitch range.

Final lowering This is approximated by varying the Dectalk's `baseline fall` (slightly) and `assertiveness` settings. A high `assertiveness` value produces dramatic final lowering at the end of the utterance; a low value produces a tentative, questioning effect. With extremely low *final lowering* values, a comma is substituted for a period at the phrase end, since the Dectalk's characteristic terminal contour for a comma is a continuation rise. This underscores tentativeness. Figure 5.8 illustrates the difference in the steepness of the terminal contour for minimal and maximal *final lowering*.



(a) minimal final lowering



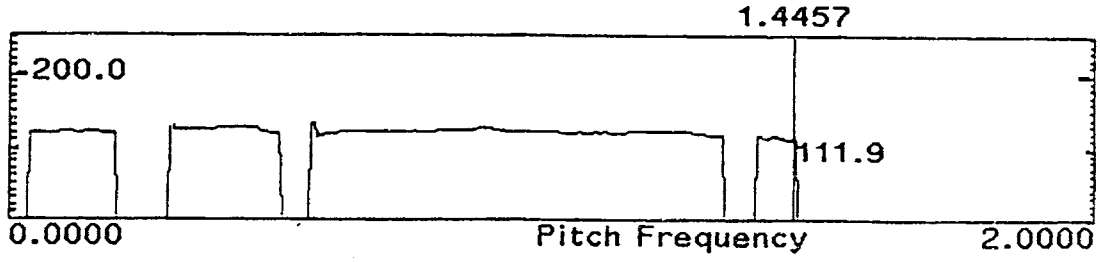
(b) minimum final lowering

Figure 5.8: **Effect of final lowering on the F0 terminal contour.** Pitch tracks for *"I [']thought you [']really [']meant it."* with (a) minimal final lowering (b) maximal final lowering. The Affect Editor *final lowering* parameter affects the Dectalk's `assertiveness` and `baseline fall` parameters. The steepness of the terminal contour increases with an increase in *final lowering*.

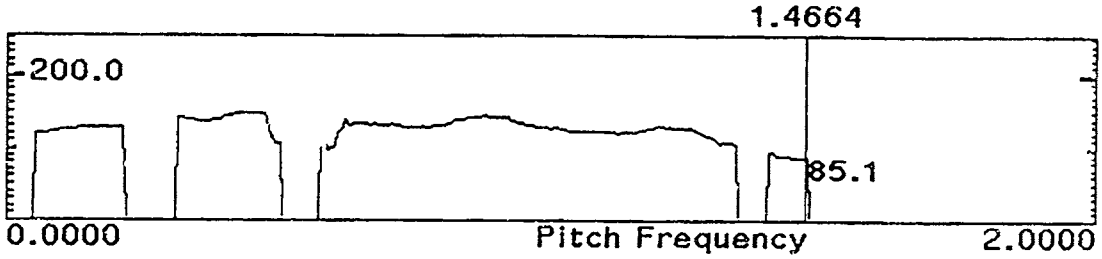
Pitch range There is no straightforward way to impose a pitch range specified as a bandwidth. This is because the Dectalk uses its `pitch range` value to scale a previously calculated F0 value, as follows:

$$average\ pitch\ +\ ((pitch\ range\ (original\ F0\ -\ 120))\ /\ 100)$$

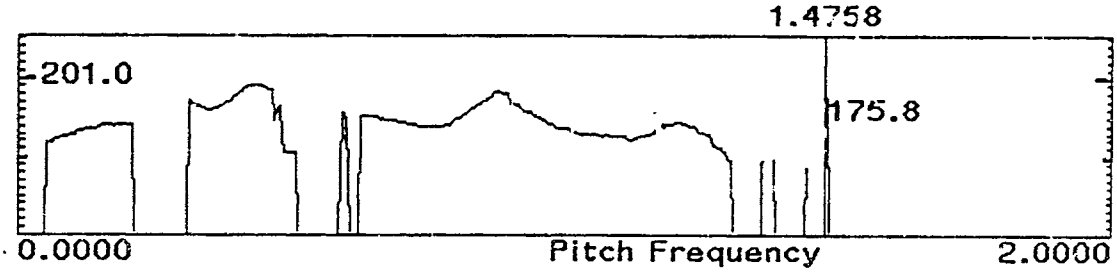
There is no way to know what the original F0 value was, or which other parameters are applied before or after the pitch range scaling. Ideally, the pitch range would expand from or contract towards the speaker baseline. Instead pitch range growth or shrinkage centers around the average pitch and fluctuates both above and below (see Figure 5.2.2).



(a) pitch range set to 20



(b) pitch range set to 100



(c) pitch range set to 250

Figure 5.9: **Effect of the Dectalk "pitch range" parameter on F0 contour.** The pitch range expands or contracts around the average pitch value. Pitch tracks for *"I ['}thought you ['}really ['}meant it."* with a pitch range scalar of (a) 20 (b) 100 (c) and 250 . (100 is normal, i.e. 100% of a normal pitch range).

Reference line This intonational parameter has no real correlate in the Dectalk scheme. The closest approximation is the hat `rise` setting. Hat `rise` describes an F0 contour that is shaped like a hat. It starts low, rises, is relatively flat throughout most of the utterance, and then returns to a low pitch at the end [4] (see Figure 5.10). Raising the hat `rise` value

has the advantage over `average pitch` of not changing speaker identity quite as drastically.

Timing

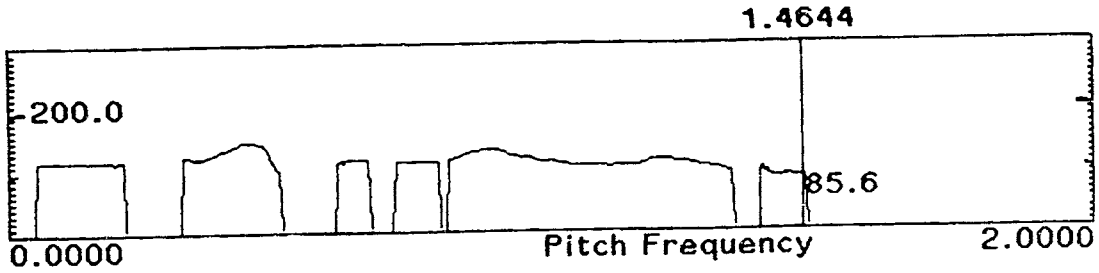
This section discusses the effect of the Dectalk on the implementation of the Affect Editor's timing parameters — *exaggeration*, *fluent pauses*, *hesitation pauses*, *speech rate* and *stress frequency*.

Exaggeration *Exaggeration* can be effected by either: lengthening phoneme durations, especially for the vowel of a lexically stressed syllable; or by adding extra phonemes. These can be either duplicates of the exaggerated phoneme, or a phoneme that represents an intermediate articulatory configuration as the vocal tract slowly moves from one phoneme to the next. Usually this intermediate configuration is a schwa. Either approach places the word in phoneme mode. This is unfortunate, because *exaggeration* applies to stressed words. These are the words that should remain in text form to take advantage of the Dectalk's rhythm and pronunciation rules. Therefore, the effects of the *exaggeration* descriptor are never actually translated for the Dectalk.

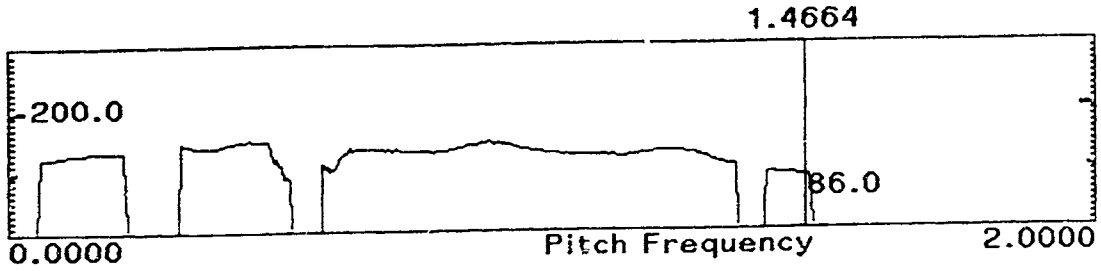
Pauses The implementation of *fluent* and *hesitation pauses* are discussed together. *Pause discontinuity*, an Affect Editor voice quality parameter, is also briefly discussed.

The Dectalk's own rules insert pauses before a verb phrase, after a comma, after a period and after a paragraph [4]. A slight pause is automatically inserted by the Dectalk before a verb phrase. Neither the comma or period are useful for fluent or hesitation pause markers because they affect intonation as well as duration. A comma effects a continuation rise right before the pause, while a period induces phrase–final intonation. So the silence phoneme is the only available effector of controlled pausing.

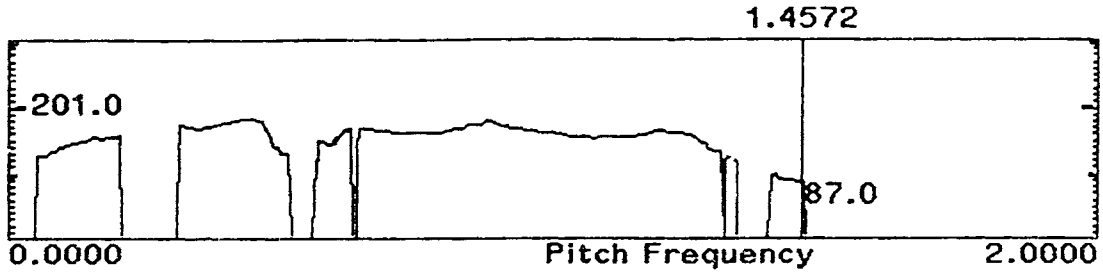
Other than the difference in location and duration (hesitation pauses are shorter) fluent and hesitation pauses are implemented identically. Their key features are duration and the



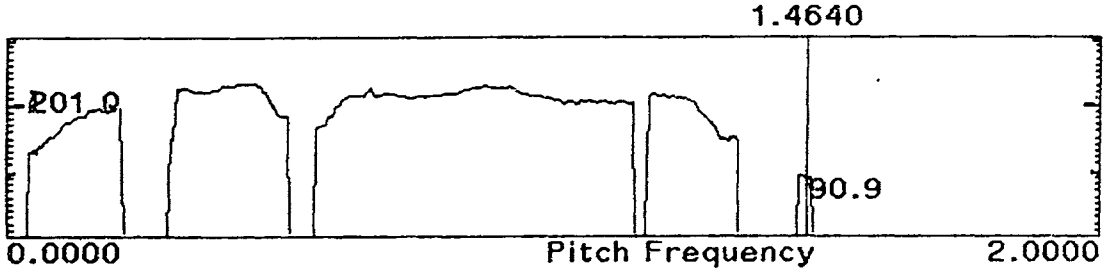
(a) hat rise set to 0



(b) hat rise set to 18 (normal)



(c) hat rise set to 50



(d) hat rise set to 100

Figure 5.10: **Effect of the Dectalk "hat rise" parameter on F0.** Pitch tracks for *"I ['']thought you ['']really ['']meant it."* with hat rise values of (a) 0 (b) 18 (normal) (c) 50 (d) 100.

way in which the transition from sound to silence is accomplished.

Duration is a function of speech rate, pause type — fluent or hesitation — and pause ranking. Pause rank designations are: FLUENT-1, FLUENT-2 OR FLUENT-3, or HESITATION-1, HESITATION-2 OR HESITATION-3. FLUENT-1 and HESITATION-1 denotes a pause of highest rank. The fluent pause duration is initially 400 milliseconds, while the hesitation pause duration is initially 250 milliseconds. These values are lowered for a faster speech rate, a pause of low rank and a small *pause discontinuity* value. They are raised for a slow speech rate, a pause of high rank or a large *pause discontinuity* value.

The type of transition into silence — *smooth, firm, firmer* and *abrupt* — is a function of the *pause discontinuity* parameter. Its value guides the selection from the phoneme alteration table of the phoneme string which follows the word but precedes the silence. Figure 5.11 shows how variations of pause location and pause onset change pause duration and the continuity of the F0 contour.

Speech rate The Dectalk's speech rate setting varies directly with the Affect Editor's *speech rate* parameter. The Dectalk measures speech rate in words per minute. The Affect Editor's *speech rate* parameter also directly affects the durations of the period pause and comma pause settings. They are lengthened as the *speech rate* value decreases. Figure 5.12 shows how increased sentence duration correlates with a decreased speech rate.

Stress frequency The Dectalk has no direct means for varying the number of stressed words per utterance. It seems to accent most content words (nouns, verbs, adjectives and adverbs) as well as some function words[2]. The result is speech which often sounds too emphatic. The Affect Editor varies the number of stressed words directly with *stress frequency*. The original sentence is marked for the highest stress frequency. Obtaining

[2]The MITalk system, in whose development Klatt participated [12], applies intonational stress to all content words, and to some function words — demonstrative pronouns, contractions, modals, quantifiers and interrogative adjectives [1]. The Dectalk appears to share this approach to intonational stress.

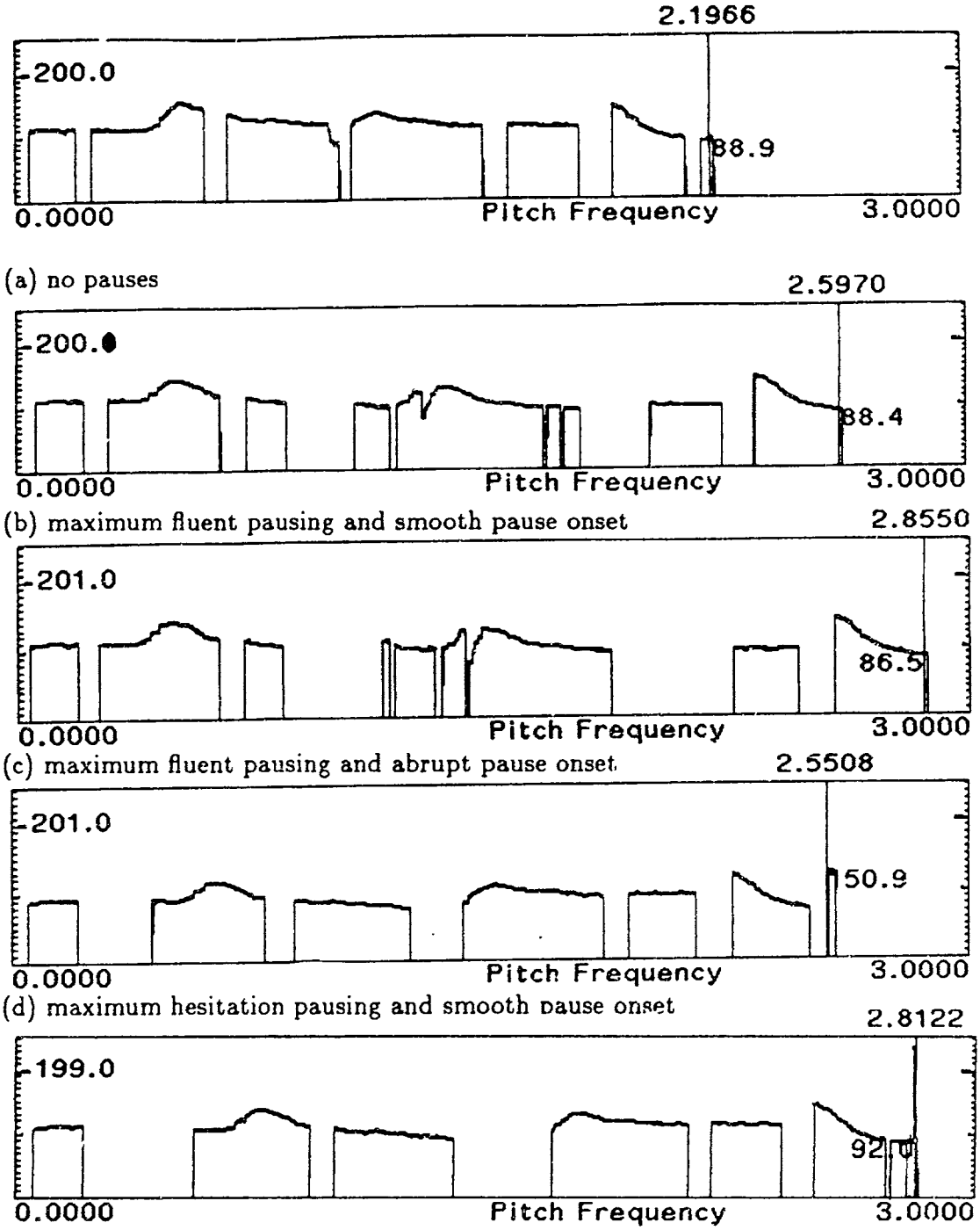
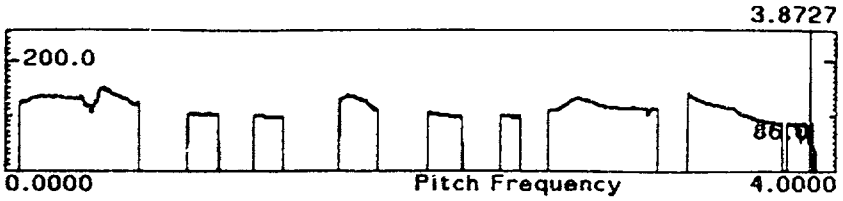
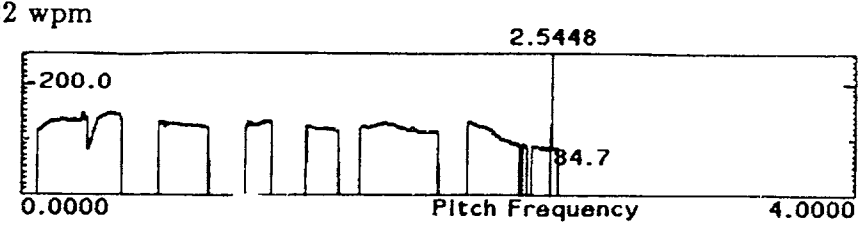


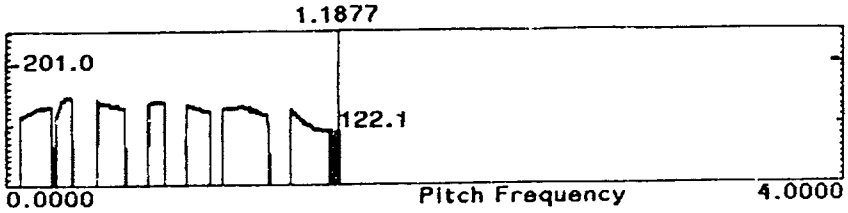
Figure 5.11: **Effect of pausing and pause discontinuity on prosody.** Pitch tracks for *"And my answer has always been the same."* with: (a) no pauses (b) maximum number of fluent pauses, and smooth pause onsets (c) maximum number of fluent pauses and abrupt pause onsets (d) maximum number of hesitation pauses and smooth pause onsets (e) maximum number of hesitation pauses and abrupt pause onsets.



(a) 122 wpm



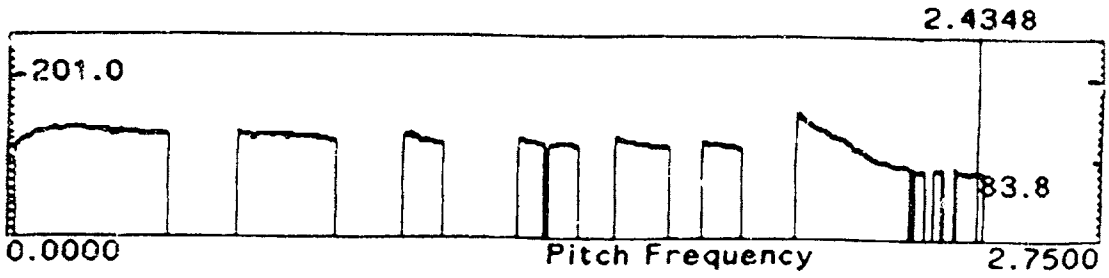
(b) 200 wpm



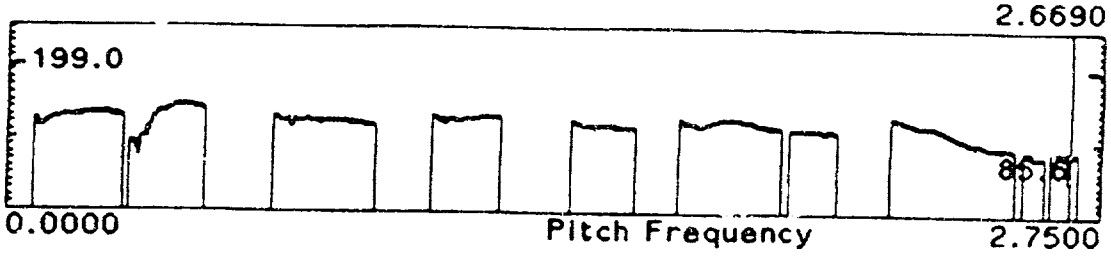
(c) 350 wpm

Figure 5.12: **Effect of the Dectalk "speech rate" setting on rhythm and sentence duration.** Pitch tracks for *"You've asked me that question a thousand times."* with speech rates of: (a) 122 wpm (b) 200 wpm (c) and 350 wpm. The duration of phonation and silence decreases as the speech rate increases.

anything less is simply a matter of marking accentable words as de-accented, in order of increasing word prominence (words with the lowest prominence are de-accented first as *stress frequency* diminishes), and then de-accenting them. Figure 5.13 shows the pitch contours for minimum and maximum *stress frequency* values



(a) one pitch accented word



(b) four pitch accented words

Figure 5.13 **Effect of increased word stress frequency on the F0 contour.** Pitch tracks for *"You've asked me that question a thousand times"*: (a) minimal pitch accenting (one content word) (b) maximal accenting (all content words). The Dectalk's word stress rules are circumvented by the use of phoneme mode. Also, the greater the *stress frequency*, the more frequent the pitch accents (realized as F0 excursions upward).

Since the Dectalk almost always accents content words, pitch accenting for these words is the default such that explicit stress markings are not necessary. However, as a precaution against unwanted de-accenting, and for the sake of consistency, the Affect Editor explicitly marks all stressed words. They are preceded with the Dectalk's primary or emphatic stress markings.

The Affect Editor achieves de-accenting of a content word by representing it in phoneme form. This hides its syntactic function and enables circumvention of the Dectalk's stress rules. De-accenting is also ensured by reducing all syllable stress markings by one level. Secondary stress markings replace those for primary stress, and original secondary stress

markings are removed entirely. This reduction of stress level lowers the F0 height on stressed syllables of de–accented words so the F0 excursion is not mistaken for a pitch accent, but perceived instead (correctly) as lexical stress.

Voice quality

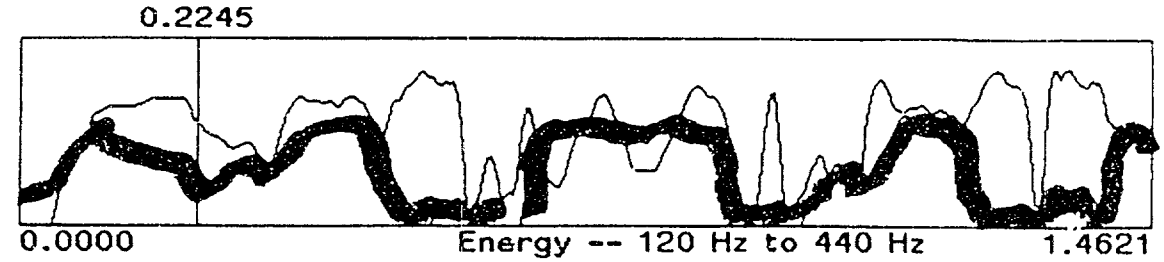
This section discusses the effect of the Dectalk on the implementation of the Affect Editor's voice quality parameters — *breathiness, brilliance, laryngealization, loudness, pause discontinuity* and *pitch discontinuity*.

Breathiness The Dectalk `breathiness` and `lax-breathiness` values vary directly with the Affect Editor **breathiness** parameter.

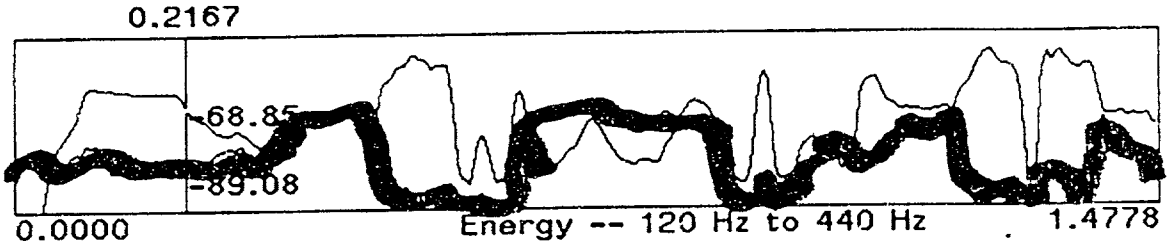
Brilliance The Dectalk `smoothness` parameter varies inversely with **brilliance**. The greater the `smoothness`, the greater the attenuation of the higher frequencies. The `richness` setting, however, varies directly with *brilliance*. The greater the richness, the stronger the lower frequencies. Figure 5.14 illustrates the progressive attenuation of higher frequencies as a result of variations in `smoothness`.

Laryngealization The Dectalk first applies laryngealization (creaky voice) to the ends of sentences, and then progressively to the rest of the sentence. Maximal laryngealization produces a voice that sounds very old and tired. Although the Dectalk permits 0 to 100% laryngealization, it is only allowed to vary from 0 to 10% with the **laryngealization** parameter in order to better retain speaker identity.

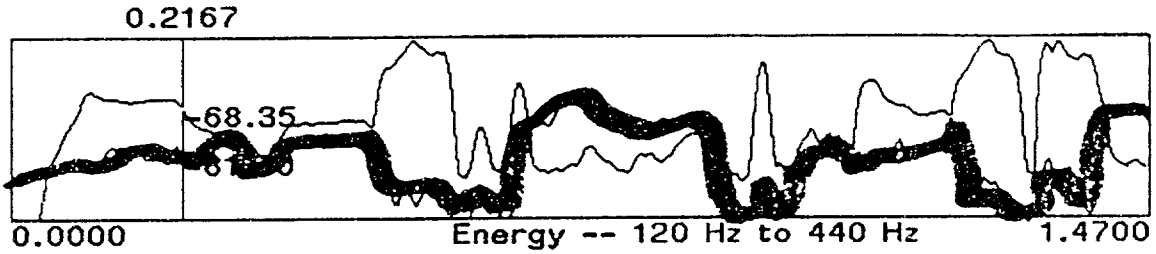
Loudness The Dectalk's `loudness` setting varies directly with the Affect Editor's *loudness* parameter. To prevent overload and squawking, the fourth and fifth formant bandwidths parameters vary inversely with *loudness*. The `gain of voicing` parameter is also partially controlled by *loudness*, varying directly.



(a) smoothness at 0



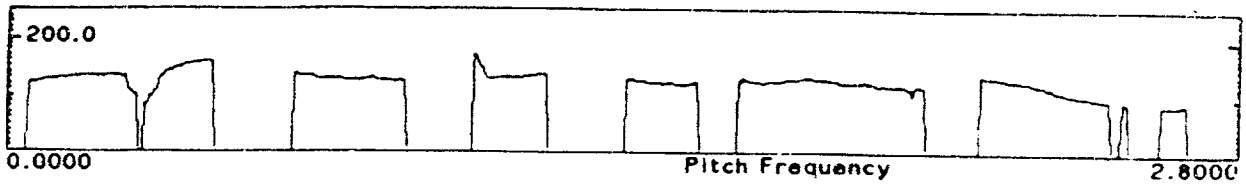
(b) smoothness at 50



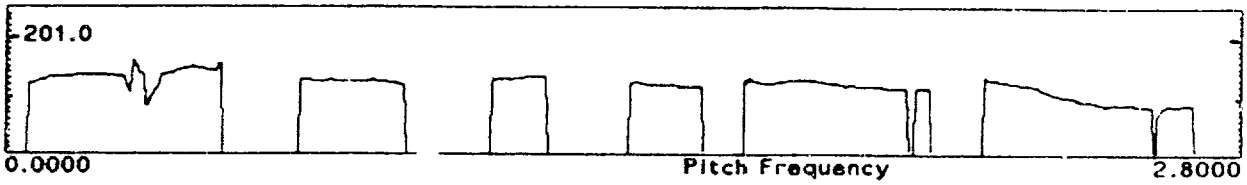
(c) smoothness at 100

Figure 5.14: **Effect of the Dectalk "smoothness" parameter on high frequency energy.** Energy plots for *"You've asked me that question a thousand times."* with smoothness values of (a) 0 (b) 50 (c) 100. The highlighted line represented energy in the 3400 to 5000 Hz band. The other line represents energy in the 120 to 440 Hz band.

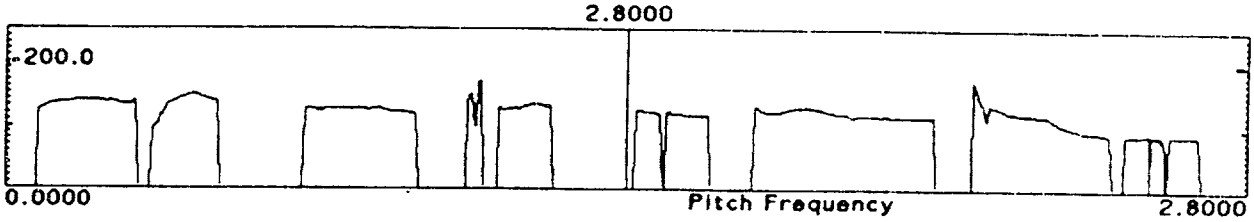
Pitch Discontinuity The Dectalk's quickness setting controls how quickly a specified F0 target is reached, and thereby affects the smoothness of F0 transitions. When quickness is low, the transitions are so smooth that the F0 target may not be reached. Quickness describes the rate of F0 change along the whole pitch contour and varies directly with *pitch discontinuity*. Figure 5.15 shows how increases in quickness produce more fluctuations and discontinuities in the F0 contour.



(a) quickness set to 0%



(b) quickness set to 50%



(c) quickness set to 100%

Figure 5.15: **Effect of the Dectalk "quickness" parameter on the F0 contour.** The greater the quickness, the more discontinuous the F0 transitions. Pitch tracks for *"You've ['']asked me that ['']question a ['']thousand ['']times."* with quickness values of: (a) 0% (b) 50% (c) 100%.

Pause Discontinuity The quality of pause onset — from smooth to abrupt — is effected by an articulatory configuration. This configuration is represented as a phoneme or set of

phonemes. Often, it is a repetition of the final phoneme of the word. The inclusion of the *pause discontinuity* parameter was originally Dectalk driven. When the silence phoneme immediately followed a word, the pause onset was abrupt, often inappropriately so. While the theoretical status of *pause discontinuity* is unknown, it is possible that the pause onset quality is a product of muscle control *and* enunciation and, therefore, that the inclusion of *pause discontinuity* is correct.

To vary the transitions from sound to silence, it was necessary to use table lookup, indexed by the final phoneme of the word preceding the pause. The pause–implementation descriptor is used to select the articulatory configuration for the phoneme, from one of four fields — pause-transition-smooth, pause-transition-firm, pause-transition-firmer and pause-transition-abrupt — of the phoneme–alteration structure. Each field contains a string composed of one or more phonemes, and occasionally, their durations.

Tremor Tremors are sometimes observed for utterances spoken in fear [24]. However, there is no means to directly implement or even approximate *tremor* with the Dectalk, so it remains unimplemented.

Articulation

Precision of articulation is the Affect Editor's only articulation parameter.

Precision The Dectalk's `gain of frication`, `gain of aspiration` and `gain of voicing` settings vary directly with **precision of articulation**. Such changes, however, convey only minimally perceptible shadings of *precision*. Further enunciation or slurring must be approximated by other means. The Affect Editor accomplished this with table lookup wherein a phoneme is either substituted for the original or inserted before or after. With some exceptions, the table entries have been guided by the following principles:

- To slur an unvoiced consonant, replace it with its voiced equivalent.

- To enunciate a voiced consonant, replace it with its unvoiced equivalent.
- To slur a vowel, reduce it:
  - Reduce a diphthong by reducing its component vowels.
  - Reduce some vowels by substituting a schwa (either "AX" or "IX").
- To emphasize a word that starts with a vowel, precede it with a glottal stop.
- To emphasize a word that begins with a consonant, repeat the consonant. However, phoneme repetition is not effective for:
  - the diphthongs "AW", "AY", "EY", "OW", "OY" and "YU"
  - the affricatives "CH" and "JH"
  - a few others — "L", "TX" and "W".

  When these phonemes are doubled, the Dectalk fully articulates them twice instead of increasing their duration. Repetitions of the consonants — "CH", "JH" and "TX" — introduce a stutter into the speech, while repeated diphthongs — "AW", "AY", "EY", "OW", "OY" and "YU" — and sonorants — "L" and "W" — sound like a poor attempt at echo simulation, so repetition for these phonemes is blocked in the table.

Precision of articulation is achieved by progressively applied alterations to a word — first to the end of the word, then to its beginning and finally to the rest of the word. This scheme is employed so that enunciation is a continuous (or at least additive) rather than binary feature. This ordered application allows changes in enunciation while preserving the word as text for as long as possible. Articulation effects at the word boundaries are achieved by following (at the word end) or preceding (at the word beginning) the word with phonemes that either emphasize or minimize (often via coarticulation effects) the word-final phoneme. Only for extremes of enunciation are phonemes *within* the word changed. In this case the entire word is sent in phoneme mode.

To effect extremely precise articulation, the silence phoneme, "_", follows the word. It is articulated as an abrupt transition into silence. Words are spoken as completely separate entities, with no coarticulation between words. In some cases, the Dectalk mispronounces schwa vowels when they are immediately by the silence phoneme. To prevent this, these words — e.g., "a", "to" — must be sent as phonemes that specify the correct pronunciation.

Figure 5.16 shows the effects of variations in articulation on phoneme clarity and duration.

5.3 Features of an Affect Synthesizer

The Dectalk was chosen for its power, especially for the richness of its parameter set. Yet, because of the specialized use to which it was put, many of its deficiencies were highlighted. This section describes synthesizer features and capabilities that will make the synthesis of affect easier and more precise. Most of the improvements concern scope of parameter control, and the methods for specifying that control.

5.3.1 Scope of control

There are two parts to scope of control. One focuses on the scope of control for *each* parameter — whether it applies to a word, phrase or utterance, and whether it can dynamically change within an utterance. Improvement in this area is achieved by expanding the power of existing parameters. The other part concerns the scope of the synthesizer parameter set. Improvement is achieved by adding more parameters to the set. Speech features over which more and precise control would be useful are parameters that affect word stress, the pitch range, pausing, precision of articulation, exaggeration and tremor.

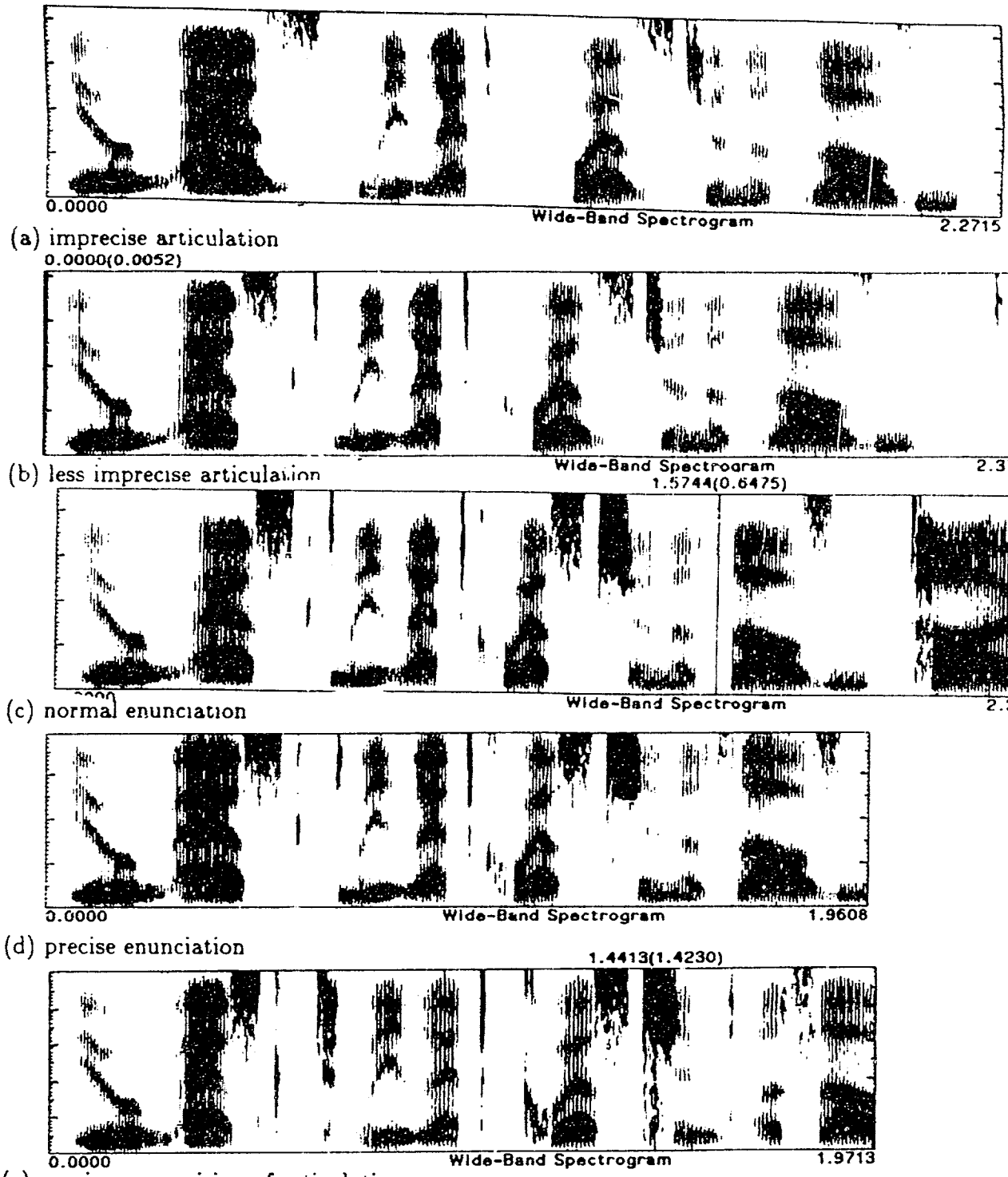


Figure 5.16: **Effects of variations in precision of articulation.** Duration and ratios of high to low frequency energy are affected. Wide-band spectrograms for *"You've asked me that question a thousand /times/."*: (a) with minimal enunciation — vowels are reduced, consonant articulation is imprecise. (b) less imprecise — word boundaries are blurred (c) with normal enunciation (d) with some precise enunciation — enunciation is applied to the last phoneme in the word. (d) with maximum enunciation — word-initial and word-final phonemes received special emphasis and vowels and consonants are fully articulated.

Parameter scope

Globally applied parameters are desirable, because they are powerful. They can be specified for a whole entity (the utterance, the speaker) rather than its parts (words, clauses, phrases)[3]. They are high level speech descriptors, whose application to parts is implemented by the synthesizer. The inclusion of a high level parameter is only possible with sufficient knowledge from which to develop the algorithms with which it is implemented. One way to develop these algorithms is to add more locally applied parameters to the synthesizer repertoire, and from repeated use, form theories which can then be implemented globally.

Besides the addition of more global parameters, another desirable feature is the addition of *dynamic parameters* that can cause changes over the course of an utterance without side effects. Dynamic parameters would improve naturalness since most speech features — brilliance, loudness, pitch range, speech rate, and voicing, for example — do not remain constant in human speech. Changes in some parameters, for example, from voicing to a whisper, could be used for dramatic effect.

Speech parameters

Side effects of Dectalk parameters seem to derive from dependencies among parameters, or from implementations that allow too little control. This section recommends improvements to the overall means of controlling intonation, and to specific parameters: *average pitch, pitch range, pausing* and *precision of articulation*.

Intonation An intonational description system should distinguish between lexical and phrase features. The meaning and scope of the descriptive tokens should be well defined in terms of intonational semantics and effects on the F0 contour. The means by which intonation is described to the Dectalk is does not meet these criteria. The meaning and scope of its parts — `average pitch`, `pitch range`, `baseline fall`, `hat rise`, `stress`

[3]The Dectalk `speech rate` and `stress rise` parameters are global in scope while the word stress markings are local in scope.

rise, **assertiveness** and its six stress markings — are indistinct. There is overlap, and especially, the effects of parameter interaction are not always predictable or desirable.

Pierrehumbert's Generative Intonation theory meets the minimum requirements for a usable intonation description language. It has tokens and a grammar and its tokens are semantically and operationally distinct. Unlike the approach taken by the Dectalk, which posits an overall contour shape (the "hat rise" contour) to which alterations are then applied, a contour is described by Generative Intonation only in terms of its parts. This approach eliminates side effects such that the overall shape of F0 contours can be accurately predicted from the abstract description.

Because the effects of pitch rise and fall can extend inappropriately beyond the word to which they apply, the Dectalk effectively allows only two pitch accents, "'" and "," . The first corresponds most closely to **H***, the second, possibly to **H*+L**. The incorporation of Generative Intonation would expand the pitch accent inventory to six. Precise control over terminal contours can then be gained by the use of phrase accents and final boundary tones and their prominences. The advantages of this theory lie in its conceptual elegance and simplicity, and in the scope and precision of the control it affords.

Pitch range The Dectalk's **pitch range** feature is a percent applied as a scalar to a default F0 contour that has already been calculated for the utterance. Instead, it should be redefined as a frequency band. Direct specification of a frequency band is cleaner, more modular and, therefore, more intuitive. Practically, it allows the pitch range to be part of the original F0 contour construction so a contour need be calculated only once.

Average pitch The chief drawback to the Dectalk's **average pitch** implementation is that, as it changes, the voice quality seems to indicate a different speaker instead of a different affect. This is in part because **pitch range** and **average pitch** are not implemented independently. Change in one affect the other. The F0 calculations are based on the assumption that the average pitch bears a fixed relation to the pitch range. It is not clear that this is the case for human speakers.

Independent control of average pitch and pitch range is a simple matter. Either pitch range — a frequency band — and speaker baseline or alternatively, speaker baseline (a fixed quantity) and topline (a changeable quantity, the highest F0 for the utterance) are all that is needed to determine average pitch. If, however, the average pitch is indeed changeable within a pitch range, three measures are needed: the baseline (an absolute F0 value) the pitch range (a relative value, in Hz) and average pitch (a percent). These together can determine the frequencies of the lowest, average and highest F0. A change in pitch range for one speaker is thus a change only to the highest F0 for the utterance, and a change in average pitch is distinct and independent of a change in pitch range.

Pausing The silence phoneme was the obvious choice for representing an unfilled pause. However, it caused phonation to stop suddenly, always resulting in an abrupt pause. Punctuation induced pausing was smoother[4]. This difference highlighted pause quality as yet another feature of pausing. However, the quality of pauses ought to be controllable and not a side effect. A "smooth pause" phoneme or a set of pause phonemes for multiple kinds of transitions into silence would add more control. Alternatively, a global parameter for controlling the quality of the transition from sound into silence can be added to the parameter set. Lacking any inherent means of control, the approximations for *smooth, firm, firmer* and *abrupt* pausing are explicitly represented in the phoneme alteration table (see Appendix B).

Precision of articulation The Dectalk's `gain of voicing`, `gain of frication` and `gain of aspiration` vary directly with the *precision of articulation* parameter. Phoneme substitution or insertions were used to approximate perceptible changes in precision of articulation. Classification of the *locations* of enunciation effects — at word offset, word onset and within the word itself — allowed for even more fine–grained variation.

Variations in precision of articulation would more easily be accomplished with a `precision`

[4]The problem with using the comma to insert a pause is that it also inserted a terminal contour, signifying the phrase end. This was fine for fluent pauses, but contrary to what is needed for hesitation pauses, which occur in the middle of an intonational phrase.

of articulation parameter that would be correctly interpreted for the phoneme, taking into account the phoneme's position — whether it is word-initial, word-final or neither — and role — whether it is in a lexically stressed syllable — within the word and the word's position and role whether it is accented, and if so, its relative prominence — within the sentence. These considerations need not be addressed all at once or to the same degree of accuracy to be useful in the finer control of articulation. Table lookup may still be necessary to obtain phoneme replacements or embellishments as long as a satisfactory articulatory model synthesizer does not exist.

Exaggeration and tremors Neither of these two Affect Editor parameters could be implemented, as discussed in sections 5.2.2 and 5.2.2. Since *exaggeration* is a word feature, it could be implemented as an annotation to a word. If it were instead implemented as a global parameter, it would apply to all stressed words.

Tremor cannot be approximated by any Dectalk feature or combination of features. It is a voice quality parameter that must simply be added to the set of global parameters.

5.3.2 Specification methods

This section discusses two aspects of parameter specification. The first concerns scale — the size of the linguistic unit over which a parameter has influence; the second, representation.

Scale of parameter effects

Specification differs for globally and locally applied parameters. Global utterance features can be set once per utterance or even once per speaker, and are applied as needed throughout the utterance.

A more local approach, applied to utterance *parts* — words, clauses, phrases — represents a compromise between ease of control and the current lack of knowledge about how the effects should be applied. In this way, more knowledge is encoded into the synthesizer, but

features must be manually assigned to the part they affect before the knowledge can be utilized. For example, if a word is to be exaggerated it is explicitly (manually) marked for exaggeration and then automatically exaggerated by the synthesizer. The synthesizer knows *how* to change features of parts, but not *when*. This approach could be combined with global parameters. For example, an articulation parameter could govern the overall quality of articulation while still allowing explicitly specified local variations.

This is a discrete approach. The units are words, intermediate or intonational phrases and discourse segments, all of which may be annotated with features and their magnitudes. The kinds of features one would assign to a *word* are: a pitch accent, pitch accent prominence, the F0 contour shape for a pitch accent, exaggeration and precision of articulation. A *phrase* or *discourse segment* would be affected by *pitch range, final lowering, average pitch* or *reference line, speech rate* and *contour slope* specifications and, possibly, *precision of articulation, word stress frequency*, frequency of *fluent pause* occurrence and frequency of *hesitation pause* occurrence.

Parameters whose effects are filtered through *stress frequency* need not be specified in absolute terms if *prominence* is allowed to apply to phrases and discourse segments. Then the set of absolute values need only be specified for the structure that dominates the segment–phrase hierarchy. As long as a constituent is marked with prominence, the absolute values of its features are derived by multiplying its prominence with those of all its ancestor nodes, until a node is found that contained not a prominence, but absolute values for phrase features — e.g., a pitch range in Hz or a speech rate in words or syllables per minute. Thus, prominence is almost always a relative value that quantifies a constituent's role relative to the other constituents in which it is contained. This hierarchical arrangement is one way to achieve the automaticity needed for globally applied parameters.

Pronunciation specification

The application of some of the Dectalk pronunciation rules is governed by a surface syntactic analysis. When the syntactic role of a phoneme representation is obscured, inferior

pronunciation — unnatural articulation, duration and pitch — is often the result. Improved processing may fix this artifact of phoneme mode. However, more information about the word is available when a word remains in text form. It seems best to keep words as annotated text, or conversely, to include a facility for annotating the phoneme form with the word's syntactic role.

Pitch contour specification

Embedding F0 and duration instructions in the phoneme string has two disadvantages. One is that phoneme mode overrides the application of the rich set of Dectalk duration and coarticulation rules. The other is that long specification strings for momentary speech events cause the Dectalk to occasionally suspend its output.

One remedy is the symbolic specification of pitch contours. This is the approach taken by Generative Intonation. As with any symbolic annotation, the means for interpretation of symbols must be built into the synthesizer so it can calculate a pitch contour from a sequence of annotated words and phrases.

A lower-level alternative is one where the pitch contour is specified not symbolically but absolutely, in Hz. The frequencies from which the contour is constructed may be sent independently of the text or as part of word annotations. If the frequencies represent only significant F0 values, that is, the F0 peaks and depths that represent pitch accents, phrase accents and boundary tones, the pitch contour specification will be sparse. To complement this, the synthesizer will need to know how to assign the significant F0 values word, and how to interpolate between significant pitches. With F0 tied to words, control over pitch accents and terminal contours is precise and also minimally specified.

5.4 Summary

This section summarizes the Dectalk's capacities and limitations, some Affect Editor remedies for Dectalk shortcomings, and suggestions for improving a synthesizer in order to facilitate affect generation in synthesized speech.

5.4.1 Synthesizer limitations

The Dectalk provides three means for varying speech output: *global parameters*, applied to the whole phrase or utterance and affecting speaker vocal tract settings and prosody; *stress markings*, applied to words, to vary the F0 contour at that point, and *phoneme mode*, in which pitch and duration of phonemes are explicitly controlled. The problems with global parameters are that there are not enough of them, they are not orthogonal and they cannot change dynamically over the course of phrase. Particularly, dynamic change cannot even be approximated by specifying new values for a subordinate phrase, because this will introduce an unwanted pause. The problems with word stress markings and phoneme mode are their side effects. The chief side effect of word stress markings is that, instead of affecting F0 for the word to which they are applied, they often affect the entire F0 contour. The chief side effect of phoneme mode is that the long ASCII strings it requires cause bottlenecks in the processing, and unwanted pauses in the speech output.

The *global* parameters are the simplest to use but the hardest to correctly implement. They provide automatic mapping of parameter values to phrasal and lexical features. However, the Dectalk mapping is not always satisfactory, nor are these parameters fully controllable or independent. For properly working global parameters, a syntactic, clausal and intonational analysis is needed, along with sufficient knowledge about how the effects of emotion show up in the word and phrase.

Stress markings, and phoneme mode provide more local control, especially of intonation. But they often have side effects. The pitch rise and fall markings, and sometimes emphatic stress, cause F0 changes over the whole contour instead of only for the word they anno-

tate. Phoneme mode tends to lower pitch, and more unfortunately, overloads the Dectalk's capacity, causing it to stop speaking while it processes another batch of ASCII instructions.

5.4.2 Short term remedies

The Affect Editor applies symbolic descriptions to *words* in order to simulate features not in the Dectalk's parameter set. The application is guided by a minimal syntactic and intonational analysis. The interpretation of the descriptors by the Affect Editor results in the *controlled* introduction of word stress markings and phoneme mode. In this way, it becomes possible to closely approximate *fluent pauses, hesitation pauses, stress frequency, pitch discontinuity, pause discontinuity* and *precision of articulation*, and some aspects of *accent shape* and *contour slope*. These are features which allowed the word to remain in text form for most values.

The other part of the remedy is to avoid using features with side effects. Thus, phrase parameters are never reset in the middle of an utterance, pitch rise and fall markings are never used to shape the pitch contour, and emphatic stress and phoneme mode are used sparingly.

5.4.3 Recommendations

To implement some of what is needed for convincing affect generation, speech synthesizers must become more robust while minimizing parameter side effects. The addition of symbolic word feature description and interpretation is a step towards developing algorithms for the automatic assignment of word features within a phrase or utterance. More powerful processing — larger buffers, better host–synthesizer synchronization — and a more comprehensive model of speech events will go a long way towards creating an affect synthesizer that will expedite the research, and eventually, commercial applications.

Chapter 6

Evaluation

6.1 An Experiment

This chapter describes the design and results of a perceptual experiment that tested the recognizability of emotions generated by the Affect Editor. Five sentences were synthesized for six emotions — anger, disgust, fear, gladness, sadness and surprise. A total of thirty sentences were presented to twenty-eight subjects, whose task was to identify the emotion with which the sentence was spoken.

The hypothesis, stimuli, subjects, experimental procedure and results are discussed in this chapter.

6.1.1 The hypothesis

The experiment tested the hypothesis that the speech correlates of an emotion could be replicated in synthesized speech such that the emotion would be recognizable by human listeners. It predicted that the difficulty of distinguishing among synthesized affects would parallel that observed for affect identification in human speech. These difficulties arise for emotions with dissimilar semantics but similar acoustical features, such as anger and joy [5],

or between emotions that have similar semantic interpretations, such as anger and disgust.

For this experiment, confusion was expected on semantic and acoustic grounds between gladness and [happy] surprise, anger and disgust, and fear and [startled] surprise, and for acoustical reasons, between anger and gladness, and sadness and disgust.

Anger, disgust, fear, happiness, sadness and surprise are among the stronger emotions. The failure of subjects to differentiate among even the extremes of affect, could be ascribed to: a false original hypothesis, an incorrect implementation in the Affect Editor of a true hypothesis or an inaccurate reproduction by the Dectalk of accurate affect–generating instructions.

6.1.2 Stimuli

The stimuli consisted of five sentences spoken with the speech correlates of six different emotions. The emotions and sentences are discussed in this section.

The emotions

The six emotions tested were basic emotions, most with a distinct configuration of physiological and acoustical correlates. Of these, anger, fear, gladness and sadness are the most semantically and acoustically unambiguous. In contrast, disgust and surprise are not as semantically extreme nor are their speech correlates as unique. For example, disgust is a kind of anger or disapproval and therefore might be mistaken for anger. Surprise is even more variable — it can be perceived or understood as joyous surprise or startled surprise, where one is a positive emotion and the other negative. Given this variability of semantics and affect, the most confusion was expected whenever these two emotions were a possible choice.

The affect configurations were generated with the Affect Editor starting with descriptions found in the literature. In cases where the descriptions conflicted, the conflicts were resolved in favor of the feature value that sounded the most authentic. The descriptions are

summarized below:

Angry speech was characterized by large F0 transitions, a generally downward inflection [8], irregular rhythm, irregular inflection [3], rising pitch contours, sharp attacks, quick rises in intensity [20], precise enunciation [3], strong high frequency energy[3], and speech that is loud, high–pitched and quick [22].

In **disgusted** speech, Fairbanks and Pronovost observed low pitch, a wide pitch range, extreme variations in inflection, and a slow speech rate (comparable to sadness) [8]. In contrast, description, Scherer noted only moderate pitch variation in disgusted speech [20]. Accordingly, the Affect Editor's simulation of disgusted speech has moderate values for *pitch range* and *stress frequency*.

Fearful speech was described as high–pitched. It had a wide pitch range, few pauses[8] and was loud, quick Rising rising pitch contours, [22], sharp attacks, quick rises in intensity and concentrations of energy in the upper ranges of the spectrum [20] were also observed.

Glad speech was characterized by regular rhythm and inflection [5]; a steady upward inflection [3], a wide pitch range and extreme variations in pitch[20]. It was loud, high-pitched, blaring, fast and with precise enunciation [3].

Sad speech was slow, with minimal variability among its features[7] and many pauses, (especially between phrases[7]). It displayed a narrow pitch range, diminished high resonance (high frequencies), a faint quiver [3], irregular pauses and a downward, irregular inflection [3]. It was soft, low-pitched and slurred [3].

The descriptions incorporated into the generation of **surprised** speech were sparse. They called for speech that is loud, high–pitched and quick [22], with a rising contour [22].

These descriptions guided the selection of Affect-Editor parameters. Their mapping to Affect Editor parameter values is displayed in Table 6.1.

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 6.1: **Affect Editor parameter values for the six emotion stimuli.**

The sentences

Each subject heard five sentences spoken with six affective colorings. Each sentence–emotion combination was presented once. 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 sentence stimuli, when read out of context, were intended to imply no particular affect. This posed two challenges. First, finding sentences that whose meaning made sense for the six emotional contexts was difficult, as a such a sentence tends to be also almost meaningless. Secondly, sentences that were too neutral sounded incongruous when uttered with a strong emotion. In consequence, the sentences chosen were declarative — conveying information but giving little clue as to how the speaker might feel about the information. However, they had some affective coloration that made them more likely to be uttered in a subset of the six emotional contexts. For example, most subjects heard *"I'm almost finished."* as a happy sentence, and *"I thought you really meant it."* as somewhat irritated[1].

6.1.3 The subjects

Twenty-eight subjects participated in the experiment. Almost all were M.I.T. students. Subject characteristics are summarized in Figure 6.1.

[1]Sentences with negative adverbs seemed to prevent people from perceiving a positive affect, e.g., glad or surprised, and so were excluded from the stimulis.

by sex	
Female	*Male*
9	19

by age			
19-22	*23-26*	*27-30*	*31-35*
8	8	6	6

by nationality	
U.S.	*foreign*
24	4

U.S. natives, by regional dialect			
New England	*Mid-Atlantic*	*Midwest*	*South/Southwest*
5	8	6	3

Figure 6.1: **The twenty–eight subjects by sex, age, nationality and, for U.S. natives, dialect.**

6.1.4 Running the experiment

This section describes the environment in which the stimuli were presented, the software and hardware used, and the experimental procedures and content.

A program controlled the presentation of the sentences and recorded the subjects' responses. This allowed the subjects to work at their own pace and without the experimenter present. Figure 6.2 shows the program interface. The program presented the stimuli in one of nine random orderings. The distributions of subjects across orderings is summarized in Table 6.2.

The experiment took place in a large office. Sentences were generated by the Dectalk3, on instruction from the testing program. The Dectalk's output was piped into four medium–quality speakers in the room. The volume, bass and treble adjustments on the amplifier were the same for all subjects.

After the instructions were given, subjects took between ten and thirty minutes to complete

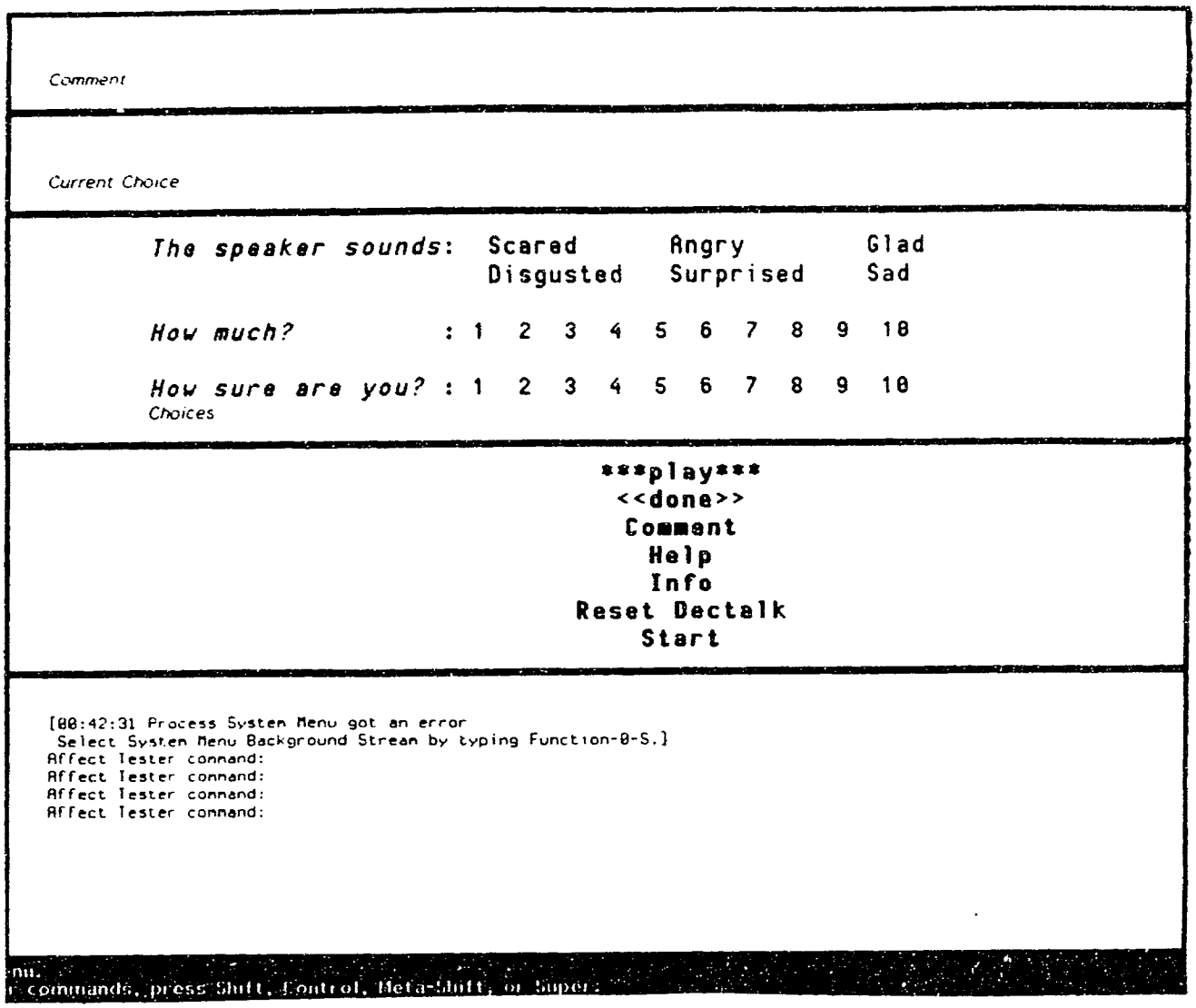


Figure 6.2: **The user interface to the program that presented the stimuli and collected the data.**

the remainder of the experiment. The average session lasted seventeen minutes, with most subjects taking about twenty minutes.

Presentation Orders									
Order	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Subjects per order	4	4	4	5	6	2	1	1	1

Table 6.2: **The number of subjects presented with sentences in each the nine random presentation orders.**

The test was set up as a forced choice. Subjects were asked to select from six emotions the one that best described what they heard. They were also asked to qualify their choice by indicating how much of the emotion they heard in the utterance, and how sure they were that their choice best described the stimulus. This served two purposes. First, because they could qualify their choices, it was easier for subjects to make a judgment in cases where either no category or more than one was appropriate. Also, it provided more data on what worked and how well.

The answers to "How much?" and "How sure are you?" took the form of ratings on a scale from one to ten, where one represented the minimal amount, ten the maximal and five the default. In addition, an optional *Comments* feature allowed subjects to type in their responses and descriptions when none of the choices seemed adequate. Most of the comments concerned two emotions that were equally possible. However, the qualifiers and comments were optional, while the choice of emotion was mandatory.

The subject could play the sentence as many times as necessary in order to make a determination. The number of replays was recorded as part of the data. Once a choice was entered, the next sentence was automatically presented.

Overall, the experiment proceeded as follows:

- The experimenter told the subject that she would be presented with synthesized ut-

terances spoken with different emotion qualities, and that her task was to choose the emotion that best described the quality she heard.

- The experimenter explained the four judgment scales.
- The experimenter explained the commands.
- The experimenter left the room.
- The subject clicked on **START** to begin the experiment.
- The Dectalk preceded the sentence presentations with a paragraph of speech,

  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.

  so that subjects could get used to synthesized speech.
- Thirty synthesized sentences, combinations of six emotions and five sentences, were presented in one of nine random orders.

When the subject finished, the experimenter explained the nature of the stimuli and the hypothesis. This was also a time when subjects commented on the aspects of the experiment that were easy, hard or confusing and on whether the stimuli sounded realistic and whether the sentence semantics interfered with their judgments.

The experiment tested the *recognizability* of emotions. However, many subjects felt compelled to distinguish between recognizability and naturalness. Some utterances sounded like an emotional human speaker, while in others, the emotion was recognizable but was perceived as an emotional computer. Especially for fear, because of Dectalk introduced falsetto, the utterances sounded to some like an emotional cartoon character instead of either a human or a synthesizer.

6.1.5 Results

The results of the experiment must be regarded in light of the stimuli. Characteristics of the stimuli are described first, followed by a discussion of the results.

Stimuli characteristics

There are several reasons why the test posed in the experiment was a hard one, both for the subjects and for the Affect Editor. The reasons fall into two categories — minimal information in the stimuli and confusing information in the stimuli.

The instructions to the subjects were minimal, in order not to bias the results. For example, whether the speech sounded human and natural as well as emotional was never mentioned as a criterion for judgment. Some subjects were confused by this. However, if the emotions were detected despite the noncommittal nature of the instructions, so much the better.

The minimal information refers also to the lack of contextual or semantic information. Normally, affect is conveyed over several utterances – in the experiment it was conveyed with only one. Also, the content of a sentence for a strong emotion usually provides additional indications of the speaker's emotional state. However, the semantic content of the test utterances was almost neutral. Lexical information was of little help in forming a judgment. This is not undesirable — recognition of emotions despite the sparseness of the stimuli could only speak well for the Affect Editor and the hypothesis it tested.

Some of the information in the stimuli was confusing. One aspect, discussed in section 6.1.2 is that some of the emotions are acoustically or semantically similar. Another source of confusion is that some utterances conveyed affect that was recognizable but not well described by the six choices. This was not intended. It resulted mainly from the influence of the synthesizer. While the extremes of emotion should be easiest to recognize, they are the hardest to synthesize, because of side effects that occur consistently at extreme settings. So instead of conveying joy, for example, utterances were synthesized for gladness, a less extreme version. Thus the experiment tested whether recognizable affect could be imparted

to synthesized speech, but more precisely, whether this could happen despite incomplete or contradictory descriptions in the literature, an incorrect or incomplete set of Affect Editor parameters and parameter relations, an incorrect mapping from the Affect Editor to synthesizer capabilities or the synthesizer's inability to respond fully or consistently to the instructions.

Data

In this section, general trends in the data are described for the responses to the stimulus emotions and sentences and for individual subject biases. Overall, twenty–eight subjects were presented with about thirty sentences each in one of nine random orders. Six trials were invalid — they were never presented because of program errors. In all, there were eight hundred and thirty four valid presentations.

Emotions

For *six* emotions, a result that is significantly better than chance, wherein each emotion is recognized for *one–sixth* (17%) of the times it is presented, indicates the validity of the hypothesis. The percent of recognitions for each emotion, totaled across all subjects, is shown in Table 6.3. Sad utterances generated the greatest number of recognitions. This is generally as predicted since it is among the most unique in the set. The recognition rates for all emotions were results are well above chance.

The results must be understood in light of the confusion inherent in the stimuli. Some emotions were consistently and predictably mistaken for each other. This is shown in Figure 6.3. In particular, anger and disgust were mistaken for each other, as were gladness and surprise. Sometimes the confusion was one way. Thus, disgust was sometimes perceived as sadness, and fear as surprise but sadness was rarely perceived as disgust, nor surprise as fear. That the confusion was consistent rather than random reflects the affective or conceptual similarities between emotions. Since the responses are consistent, it may be useful to consider exact and close matches as recognitions. Responses considered close

Recognized emotions							
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	.439	.421	.482	.91	.518	.439	.535

Table 6.3: **The number of times each intended affect was recognized, for each emotion and for all emotions, totaled across all subject responses.**

matches are listed in Table 6.4. Table 6.5 shows the results after recalculation to allow close matches. The percent of recognitions improves dramatically, from fifty four to seventy eight percent.

Exact and close matches, for emotion stimuli						
stimulus	*Angry*	*Disgusted*	*Glad*	*Sad*	*Scared*	*Surprised*
Exact	Angry	Disgusted	Glad	Sad	Scared	Surprised
Close	Disgusted	Angry	Surprised	Scared	Surprised	Glad

Table 6.4: **Responses counted as recognitions because of acoustic and semantic similarities between emotions.**

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 (adjusted)	91	113	114	136	101	101	656
Percent recognized (adjusted)	.655	.807	.832	.971	.727	.727	.787

Table 6.5: **The number of adjusted and exact recognitions, for each emotion and for all emotions, totaled across all subject responses.**

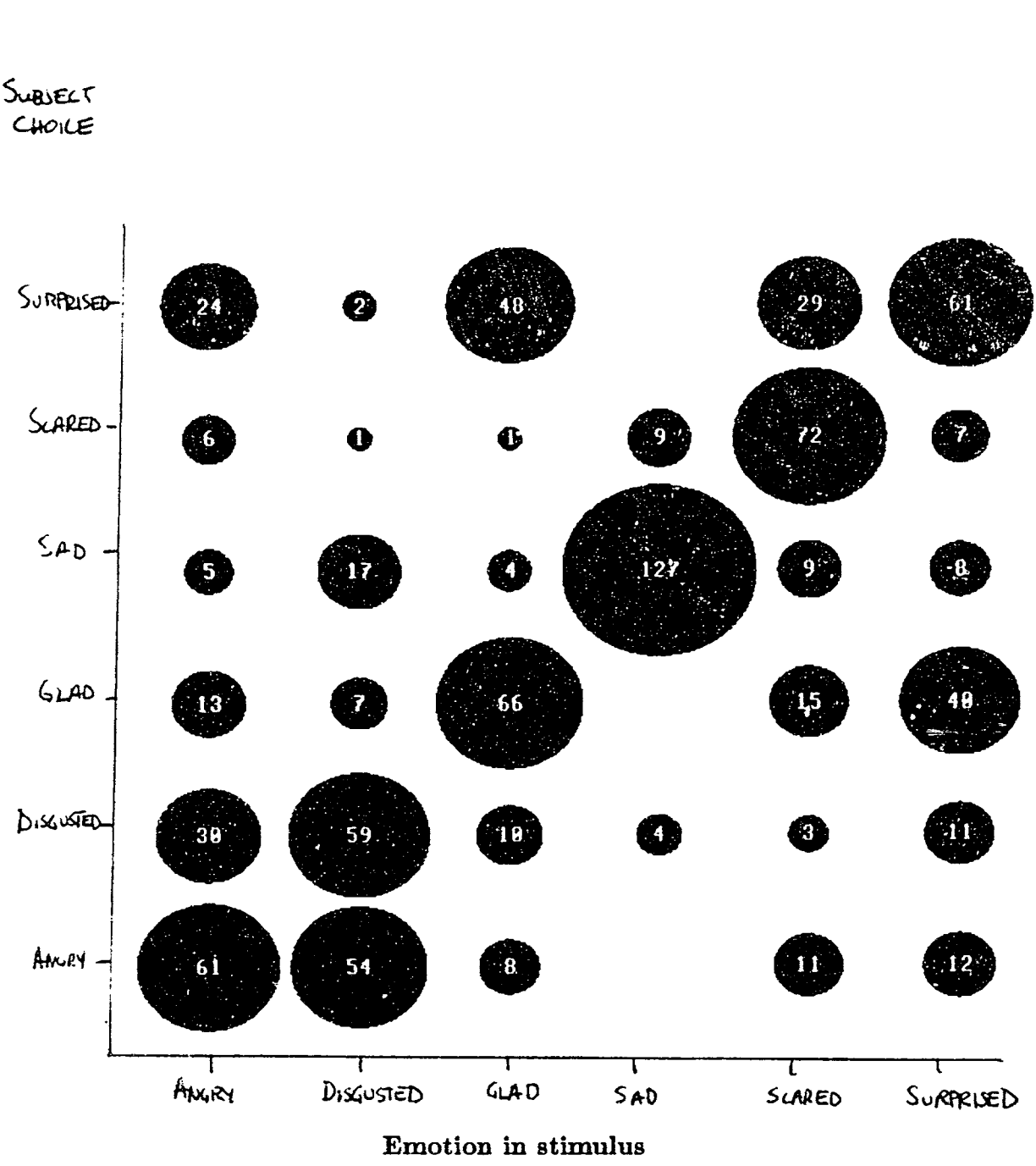


Figure 6.3: **Plot showing the distribution of emotions in the stimuli and how they were categorized, for all subjects.** The x-axis shows the emotion stimulus, while the y-axis shows the subjects' choices. Thus, an *"Angry"* utterance was perceived as angry 61 times, disgusted 30 times, glad 13 times, sad 5 times, scared 6 times and surprised 24 times.

Sentences

Despite attempts to minimize semantic influence, recognition did vary with sentence content. The number of exact and close matches for each sentences is summarized in Table 6.6. It shows that the sentence, *"I thought you really meant it."*, was responsible for most of the errors, with only a 39.3% initial success rate, and a 66.1% adjusted success rate. The semantics of *"I'm almost finished."* biased it in favor of gladness, while the semantics of *"I thought you really meant it."* and *"Look at that picture."* biased the responses away from gladness. The distribution of emotion judgments for each of the sentences is presented in Figure 6.4.

Adjusted responses per sentence						
stimulus ⇒	*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.*	Totals
Total presentations	167	165	168	167	167	834
Total exact	90	97	66	101	92	446
Total close	37	47	45	38	43	210
Total adjusted	127	144	111	139	135	656
Percent exact	.539	.588	.393	.605	.551	.535
Percent close	.222	.285	.268	.228	.257	.252
Percent adjusted	.76	.873	.661	.832	.808	.787

Table 6.6: **The number of adjusted responses, for each emotion and for all emotions, totaled across all subject responses.**

Subject Biases

The most interesting result was the subjectivity that people brought to the recognition of emotions. In the discussions following the experiment, some subjects insisted that one emotion was presented significantly more than others. However, they disagreed on which emotion dominated. The tendencies for each subject to pick a particular response are summarized in Table 6.7.

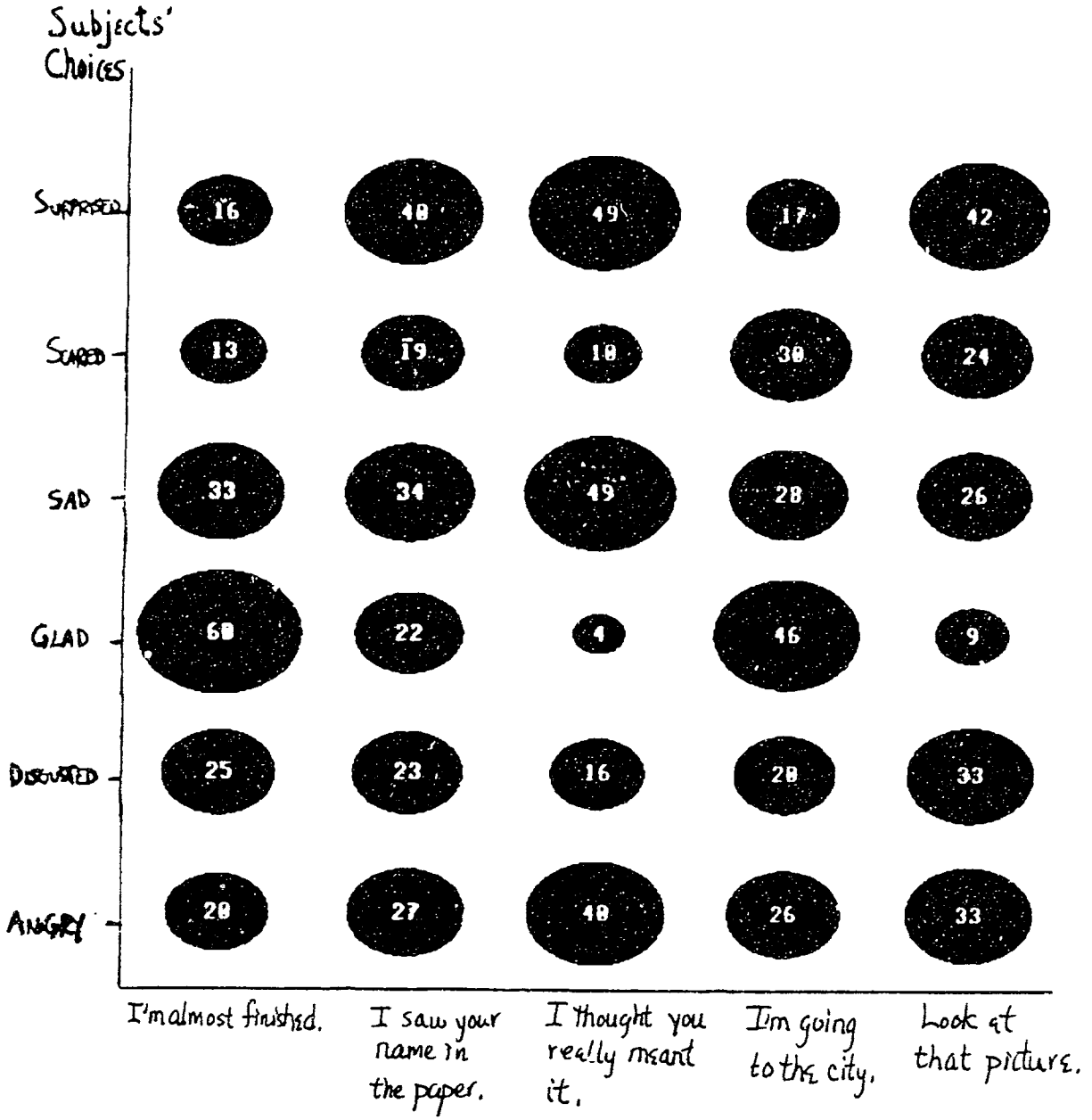


Figure 6.4: **Plot showing the distribution of sentences in the stimuli and how they were categorized, for all subjects.** The x-axis shows the sentence stimulus, while the y-axis shows the subjects' choices. Thus, ***"I thought you really meant it."*** was perceived as angry 40 times, disgusted 16 times, glad 4 times, sad 49 times, scared 10 times and surprised 49 times.

Subject response tendencies						
Subject	*Angry*	*Disgusted*	*Glad*	*Sad*	*Scared*	*Surprised*
1	.241	.069	*.000*	.172	.138	**.379**
2	**.241**	.172	**.241**	.138	*.034*	.172
3	.172	.069	.241	.207	*.034*	**.276**
4	**.429**	.107	*.000*	.107	.143	.214
5	.207	.207	.138	**.241**	*.034*	. 172
6	.167	.200	*.067*	**.333**	.167	*.067*
7	**.233**	*.100*	.200	**.233**	*.100*	.133
8	.200	.133	**.267**	.200	*.067*	.133
9	.167	*.100*	.133	.200	.167	**.233**
10	*.133*	*.133*	**.300**	.167	*.133*	*.133*
11	.200	.133	*.100*	**.233**	.133	.200
12	.233	.067	.167	**.267**	*.067*	.200
13	.133	*.100*	*.100*	**.267**	.233	.167
14	*.133*	.167	.200	**.233**	*.067*	.200
15	*.133*	.167	**.267**	*.133*	.167	.133
16	.167	.167	.167	**.233**	*.133*	*.133*
17	*.100*	.200	.200	.167	.167	.167
18	.167	.233	.200	**.267**	*.033*	.100
19	.200	.133	.167	.200	*.033*	**.267**
20	**.233**	*.067*	.133	.167	.233	.167
21	.167	**.233**	.133	**.233**	*.100*	.133
22	*.033*	.100	.167	.200	.067	**.433**
23	.200	*.033*	**.333**	.133	.100	.200
24	*.133*	*.133*	.167	**.267**	.167	*.133*
25	.133	.200	.100	.167	.133	**.267**
26	*.000*	.233	.167	.133	.167	**.300**
27	**.233**	.167	.200	.167	*.100*	.133
28	.133	*.100*	.167	.233	*.100*	**.267**

Table 6.7: **Subject tendencies to perceive emotions.** The distribution of emotions in the stimuli is even. However, the distribution of the responses per subject is not. The subject's most frequent response is presented in **boldface**, and the least frequent response is presented in *italics*.

6.2 Summary and Future Work

The experiment was designed to measure whether affect designed according to the specifications in the literature could be recognized in synthesized speech. The reasons why recognition might be difficult are summarized in section 6.1.5. An analysis of the data indicates that the intended affect was perceived in the majority of presentations when initial or adjusted responses are considered. Recognition of sadness was better than for the other emotions, whose speech correlates and subjective semantic interpretations were more similar. These emotions tended to be mistaken for each other. Overall, the initial and adjusted results are significantly above chance. This occurred despite flaws in the experiment design (confusing stimuli) or in the stimuli (affect that was imperfectly reproduced by the synthesizer). The results support the hypothesis that, through systematic manipulation of speech correlates, *recognizable* affect can be generated in synthesized speech.

The experiment answered a broad question in the affirmative. However, the Affect Editor is a tool for exploring perceptual responses to synthesized affect as well as for designing specific speech affects. Other experiments can be performed to explain in more detail what was perceived when subjects answered correctly or incorrectly. They should investigate the relationships among parameters, and the effect of each on the perception of affect. This would include the study of:

- the perceptual effects of increasing and decreasing parameter values individually and in various groupings
- the perceptual effects of increasing some parameter values while decreasing others
- the perceptual effects of changing a parameter value in proportion to changes in other parameter values

The results of these experiments should reveal thresholds at which the perceived affect changes and should better explain perceived similarity and difference among stimuli. This information can then be incorporated into the Affect Editor to improve its performance and design.

Chapter 7

Summary and Future Work

This section summarizes the contributions of the work described herein to the problem of adding and controlling affect in synthesized speech and discusses two areas in which work should proceed — improvements to the Affect Editor and the development of a generative theory of affect in speech.

7.1 Contributions

The work described by this thesis is important because it initiates the effort of systematic synthesis of affect in speech. This is its major contribution. From the development of the Affect Editor two results emerged — a model of affect in speech, upon which continued work can rest, and a more precise understanding of the features of a synthesizer that help or hinder the synthesis of affect. (They are described in Chapter 5.) The contributions of the model are discussed in this section.

Of the various representations of affect in speech, the perceptual/acoustical model is currently the most tractable. It requires the least knowledge about a speaker's internal states and processes and is supported by the current technology. It is also appropriate for gathering the perceptual information needed at this stage to improve the representation and gener-

ation of affect. Since the phenomena represented within the model are mainly perceptual, perceptual responses can be tested directly, by parameter manipulations.

The model used by the Affect Editor to represent affect in speech groups the parameters into four primarily acoustical categories. Its key features of the model are:

- **It is flat.** With the exception of *stress frequency*, its parameters vary independently — there is minimal hierarchy. This is a necessary feature of a first pass. It acknowledges that the information relating the parameters is incomplete. Part of the purpose of the Affect Editor is to serve as a tool for discovering those relationships.

- **Stressed words have special status.** This is the most significant exception to the flat parameter arrangement (the other exceptions are the *accent shape, contour slope* and *pause discontinuity* parameters). Through intonation, which clarifies the discourse and informational structure, stressed words (and phrase endings) are the prime carriers of semantic information. The Affect Editor extends this, incorporating the assumption that they also the prime carriers of *affective* information.

- **Speech correlates are represented by parameters whose presence or absence in the speech output is quantified.** Quantification of parameter influence enables the systematic changing of affect. It also simplifies mixing of affects (this is often a feature of acting instructions, e.g., *annoyed but also amused.*). The most dissimilar parameter values of the current affect closer are moved closer to those of the added affect.

- **Parameter values are balanced around a norm, (at zero) which represents the "neutral" setting.** The positive values represent the greater parameter effect; the negative, the lesser. This provides a simple means for implementing the concept of "more" and "less" of an affective presence. "More" of an affect is achieved by moving parameter values *away* from the norm value; "less", by bringing them closer. When synthesis of affect becomes as simple as imparting stage directions, fluctuations in the degree to which an emotion is expressed can (and should) still be implemented in relation to a neutral setting.

- **The inclusion of qualitative change thresholds for the extremes of emotion.** The physiological effects of extreme emotions are not always responsive to conscious control. This is especially true for emotions for which the sympathetic nervous system is highly aroused. The Affect Editor *begins* to recognize this difference with increased F0 fluctuation for the highest *accent shape* value, such that an **H*** pitch accent may be realized as a bitonal. However, the full implementation of the effects of extreme emotions is limited by synthesizer side effects which occur for extreme values.

With the exception the flatness of the model, these features should remain useful in successive implementations of the Affect Editor. Additions and changes to the Affect Editor are discussed in the next section.

7.2 Future Directions for the Affect Editor

This section discusses future directions for the Affect Editor. The Affect Editor can be improved in three ways: by improving the implementation of current parameters, by changes to the parameter set (the expansion or compression of current parameters, or addition of new ones) and by changes to the structure of the perceptual/acoustical model.

7.2.1 Improvements to current parameters

This section discusses parameters – *pause discontinuity*, *average pitch*, *reference line* and *precision of articulation* for which the implementation issues are unclear.

Pause discontinuity

Pause discontinuity was incorporated in the Affect Editor because of the inappropriately abrupt cessation of phonation induced by the Dectalk silence phoneme. Thus, its validity as a speech correlate of emotion is not assured. Investigations into: the perceptual effects

of changes in pause onset quality; the significant causes of pause quality differences (lack of muscle control? speaking style differences?); and the effects on pause quality when filled pauses, repetitions and false starts are incorporated into an affect synthesizer may resolve this.

Average pitch or reference line

The *reference line* is an intonational feature, while *average pitch* (F0) is a feature of the speech signal. The effect of intentional use upon both, and the effect of both on the perception of affect must be studied to determine whether these really are two distinct speech phenomena.

Precision of Articulation

The Affect Editor implements *precision of articulation* by combining two concepts: changes in articulation and the location of articulation change. This was motivated, originally, by the need to keep stressed words as text for as long as possible, to avoid Dectalk side effects. However, the validity of this as a speech phenomenon is not clear. The manifestation of articulation effects may be ordered instead according to phoneme class or syllable stress. To better implement *precision of articulation* the following questions should be considered:

- How is enunciation increased or decreased? Where do effects show up and in what order?
- What is the nature of the interaction between articulatory changes applied at different levels, e.g., to the phoneme or to the word?

7.2.2 Changes to the parameter set

The changes are of three types — expansion of the current parameters, compression of current parameters and the addition of new parameters.

It may be useful to expand the *precision of articulation* parameter into parameters that individually affect speech based on phoneme articulation and location features. Since articulation phenomena may vary among phoneme classes, it may be helpful to represent the influence of precision of articulation separately, for each class. It may also be helpful to represent separately parameters for the lexical and syntactic location of the articulation changes — i.e., whether (and how frequently) they occur at the beginning, middle and end of words and phrases.

Parameters that always vary together and in the same proportions may be compressed into one. New parameters should be added if the need becomes apparent from further research.

7.2.3 Changes to the model structure

The model structure may evolve such that its parameters are grouped according to other criteria and it encodes parameter relations.

Currently, the model organizes the speech correlates of emotion into those that affect features of *pitch*, *timing*, *voice quality* or *articulation*. Other means of organization may prove more useful. For example, within these categories it may prove useful to organize the features according to the amount of conscious control they allow — whether they are primarily pragmatically (linguistically) or physiologically derived. Thus, *average pitch*, which is affected by the autonomic nervous system response, is a parameter over which little conscious control may be exerted, while the *reference line*, which is often consciously employed to indicate discourse structure, allows significantly more conscious control.

The Affect Editor contains few assumptions about parameter relations. These relations — which parameters change together, proportionally or inversely — can be investigated with the current version. The encoding of parameter relations, as with improvements, is a reflection of increased knowledge about affect production and perception.

7.2.4 Towards the development of a theory of generative affect

The Affect Editor is one answer to the question of *how* to generate affect in speech. The success of its current and future versions can be measured with perceptual tests. However, it should ultimately be judged on how well it conforms to a theory which explains *why* certain parameters and levels of representation are necessary and how they are related.

The problem is that there is no generative theory for affect in speech. When it is formulated, such a theory must support theories that explain other speech phenomena — intonation, for example. For such a theory to be useful to the synthesis of affect, it must also be *generative*. It must enable a mapping between the speaker's mental state and the affect that is expressed in speech.

The physiological and acoustical correlates of emotion have been well–defined. However, to complete the description, and to add the generative component, the mental correlates of emotion must be modeled as well.

Pierrehumbert and Hirschberg have shown how intonation reveals the attentional[1] and intentional[2] *structure* of discourse [19] as described by Grosz and Sidner [10]. The integration of emotion into this analysis rests with the *intentional* component of discourse. Emotion, as a mental state (albeit a mental state with physiological components) is comprised of *beliefs*, *intentions* and *plans*. Affect is employed to convey information about these components. This information becomes part of the global context within which utterances must be interpreted. As an intentional phenomenon, it confirms the structure but especially conveys [part of] the intentional *content* of discourse. Emotion, then, can be represented as information that the speaker intends to convey to help the listener interpret her utterances. The acoustical correlates of emotion can be understood and organized in light of the information they convey about speaker plans, beliefs and intentions.

[1] The attentional state describes what is currently being discussed. It changes from utterance to utterance.

[2] The intentional component to discourse describes the reason *why* something is being discussed. It is communicated over the course of several utterances. A collection of utterances devoted to the satisfaction of one intention comprises a *discourse segment*. Discourses are divided into segments, each defined by a distinct an intention or purpose, which explains why the segment was initiated in the first place [10].

At the extremes of emotion this definition is strained because intentions are normally understood to be intentional, in other words, under the control of the person who has them. Since some emotional states evoke responses that transcend or override the person's conscious control, the justification for integration of emotion into a discourse and speech production model is weakened. However, if the degree of conscious control of intention is allowed to vary, the relation is maintained. Information about speaker plans, beliefs and intentions is still conveyed by extreme emotion. These intentions and plans are of a basically physiological nature (fight or flight) but are still plans and intentions and reflect the speaker's beliefs about a situation. Scherer's description of the role of emotion supports this view. He describes it as an adaptive phenomena which motivates coping behavior and reflects the organism's evaluation of the relevance and significance of particular stimuli as regards its needs, plans and preferences. Emotion prepares the organism for action and communicates the organism's state and intentions [20].

Exploring the relation of emotion to the *intentional* content of discourse may provide a tractable way of modeling the speaker's mental state as it affects discourse and speech generation. In this way, a theory of affect in speech can be integrated with theories about other speech and discourse phenomena, and moreover, is *generative*.

7.3 Conclusion

The goal of synthesis of recognizable affect is within our grasp. The synthesis of natural affect is, perhaps, farther off. In the meantime, however, the Affect Editor remains a tool for generating affect in speech, and, as important, for exploring the effect of acoustical features on the perception of synthesized affect. The power of the Affect Editor is that it can be used as research tool for *systematically* testing the effects of different parameters upon the speech signal and upon the perception of affect. This is an advantage over human speech, whose parameters cannot generally be systematically varied.

Synthesized speech would be greatly improved with better synthesizers and more knowledge about the perception of affect in speech. The role of affect in speech and discourse

would become clearer with a generative theory of affect, which would relate the information conveyed by affect to its realization in acoustic features of speech. This thesis is a step in those directions.

Appendix A

Structures

The key data structures are either Common-Lisp structures, or Zeta-lisp flavors.

A.1 Parameters

A.1.1 The param flavor

```
(defflavor param
    ((name          "")
    (value       0)
    (max            10)
    (min           -10)
    (id            (gensym))
    (documentation  "")
    )
    ()
    :readable-instance-variables
    :writable-instance-variables
    :initable-instance-variables)
```

The param flavor is the parent to the synth-param and grouped-param flavor. Param methods apply to synth-params and grouped-params. The parameter's **value** can be no less than **min** and no greater than **max**. The param's instance variables and their contents are:

- **name**: the human readable name as it appears on the Affect Editor display.
- **value**: a value between *max* and *min*.
- **max**: the maximum possible parameter value. It is static. The default is 10, customized for emotion descriptors.

- **min**: the minimum possible parameter value. It is static. The default is -10, customized for emotion descriptors.
- **id**: an internal symbol, used throughout the program to identify this parameter.
- **documentation**: the parameter definition, in English.

A.1.2 The grouped-param flavor

```
(defflavor grouped-param
    ((group (gensym))
    )
    (param)
    :initable-instance-variables
    :readable-instance-variables)
```

An emotion's descriptor set is composed of grouped params. They are grouped by features speech — pitch, timing, voice quality and articulation. The **group** instance variable is the only addition to the instance variables inherited from the param flavor.

A.1.3 The synth-param flavor

```
(defflavor synth-param
    ((synth-value 50)           ;Interpretation from abstract VALUE into
                                ;real SYNTH-VALUE
     (norm        .5)           ;The NORM is not necessarily at .5 of the total
                                ;range.  A SYNTH-VALUE  = VALUE scaled around
                                ;the NORM.
    )
    (param)                     ;inherit from param
    :readable-instance-variables
    :writable-instance-variables
    :initable-instance-variables)
```

The synth-param flavor maps directly to synthesizer parameters and is synthesizer specific. **synth-value** is value that is sent to the synthesizer. The **norm** is a percent. The default value for this parameter — its value for "neutral' affect — is obtained by multiplying the range —

$$(\mathbf{max} - \mathbf{min})$$

— by the **norm**.

The **id** indexes a table of mappings (see Table 4.1) from the listener model parameters to synthesizer settings, to determine:

- which descriptors affect the synth-param value
- by how much (a percent)
- and in what direction (some synthesizer parameters vary inversely with the descriptors that affect them).

The final **value** is the result of successive applications of the speech correlate parameter values, and lies always within the range bounded by **max** and **min**. After this calculation, the **value** is adjusted around the **norm** for the synthesizer parameter, to obtain its **synth-value**. This is the value sent to the synthesizer. The steps are as follows:

- If the ratio of the **value** to its range is greater than or equal to .5, subtract .5 from the ratio to obtain the percent above the **norm**, and calculate as follows:

$$DefaultValue + (2 \times PercentAboveNorm \times (\mathbf{max} - DefaultValue))$$

- If the ratio of the value to its range is less than .5, calculate as follows:

$$2 \times ValueRangeRatio \times DefaultValue$$

These equations express one rate of change for synthesizer values below the **norm** and another for those above. In some cases — for example, *laryngealization* and *breathiness* — the **norm** will be set to 0. This all **values** falling below the halfway point become the minimal synthesizer value for the parameter.

A.2 The Emotion

A.2.1 The emotion structure

```
(defstruct  (emotion)
  (name            "")
  (base-emotion    "")  ;Editing history
  changed-params        ;Editing history
  correlates               ;The descriptor set that best conveys this emotion
  )
```

The emotion structure has four fields. The **name** field is a string that is the emotion name. The **base-emotion** and **changed-params** field trace the editing history. The **correlates** field is the set of nineteen parameters that correspond to features of the perceived affect.

A.2.2 The correlates structure

The correlates structure contains the nineteen parameters of the perceptual model. Each parameter is a grouped-param.

A.3 The sentence

A.3.1 The constituent structure

```
(defstruct (constituent)
    type
    object
    parts
    parent
    evoked
    relations
    pitch-range
    final-lowering
    intermediate-phrase
    intonational-phrase
    rate
  )
```

A sentence is a constituent of type :S. A constituent is a discourse constituent. It is a structure imported from the Discourse package [1]. The information that the Affect Editor inscribes in its fields is not always of the same type as that intended by the Discourse package. It has these fields:

- **type**: the syntactic or semantic category. There is some confusion as to the ultimate best use of this category. Except for the sentence - a syntactic classification — most constituents fed to the Affect Editor are semantically–typed.
- **object**: not used by the Affect Editor.
- **parts**: a list of words and constituents forming leaves and lower branches of the tree or subtree headed by the constituent.
- **parent**: If the constituent is part of a larger tree structure, this field points to its immediate parent.
- **evoked**: not used by the Affect Editor.
- **relations**: not used by the Affect Editor.
- **pitch–range**: a decimal value. It reflects the constituent's discourse role, and scales the pitch range derived from affect.
- **final–lowering**: a decimal value. It reflects the constituent's discourse role, and scales the final-lowering value derived from affect.
- **intermediate–phrase**: If this constituent is an intermediate phrase, its phrase accent, either **H** or **L** will be placed here.

[1] written by Jim Davis, MIT Media Lab

- **intonational-phrase**: If this constituent is an intermediate phrase, its phrase accent, either **H%** or **L%** will be placed here.
- **rate**: A number from 1 to 4. 1 represents the quickest speech, 4 the slowest. The rate reflects the constituent's discourse role. The numbers are mapped to decimal values (1.25 for 1, .75 for 4) which are then used to scale the synthesizer speech rate as derived from affect.

A.3.2 The pause structure

```
(defstruct (pause)
  type
  implementation
  )
```

The **type** field of a pause structure contains a pause type symbol, which indicates the pause type — hesitation or fluent — and its ranking. The **implementation** field stores the symbol that describes the transition from sound to silence as the preceding word ends and the pause begins.

Legal pause types are the symbols :FLUENT-1, :FLUENT-2, :FLUENT-3, :HESITATION-1, :HESITATION-2 and :HESITATION-3. Legal pause implementation descriptors are the symbols :SMOOTH, :FIRM, :FIRMER and :ABRUPT.

A.3.3 The word structure

```
(defstruct (word)
  string
  accent
  prominence
  pronunciation
  category
  )
```

The word structure is also imported from the Discourse package. The information in its fields is not always of the same type as that inscribed by the Discourse package. It has these fields:

- **string**: the word as plain text.
- **accent**: the word's *pitch accent* if it has one. Where the original word has a pitch accent, the copy used to implement the current affect, may instead have the symbol :DE-ACCENT. This indicates that the word is a content word that does not receive a pitch accent under the influence of the current affect.

- **prominence**: a decimal. It indicates the relative semantic importance of the word to the sentence. This is reflected in the final peak F0 for the word, also derived from the prominence value.
- **pronunciation**: In the Discourse package, this field stores the word's phonemes grouped into syllables. The Affect Editor augments this by also storing here symbolic descriptions of the word's pronunciation, intonation or timing when they differ from the normal implementation. The descriptors that can legally reside in the **pronunciation** field are, in order of ascending magnitude:
  - **enunciation descriptors**: the symbols :SOFTEN-CONSONANTS-AND-VOWELS++, :SOFTEN-CONSONANTS-AND-VOWELS+, :SOFTEN-CONSONANTS, :SOFTEN-EDGES, :ENUNCIATE+, :ENUNCIATE++, :ENUNCIATE-AND-SEPARATE-WORDS+, and :ENUNCIATE-AND-SEPARATE-WORDS++.
  - **exaggeration descriptors**: the symbol :EXAGGERATE.
  - **prominence descriptors:** the symbol :PROM-MAX interpreted as the application of emphatic stress.
  - the symbol :SLOPE-PEAK, applied to the beginning of the utterance for downward-sloping contour, and to the end for an upward slope.
- **category**: the word's grammatical category — noun, verb, etc.

A.3.4 The phoneme structure

```
(defstruct (phoneme)
  orig
  accent
  replacement
  duration
  )
```

The phoneme structure fields and their contents are:

- **orig**: the original phoneme, from citation form.
- **accent**: whether it is preceded by an accent. The accent designates either primary or secondary stress for the syllable and precedes vowels and syllabics only.
- **replacement**: Some operations on word pronunciation, such as slurring or enunciation, replace the original phoneme with an allophone. When selecting phonemes for word construction, a non-nil value in this field has precedence over the **orig** value.
- **duration**: the phoneme duration in milliseconds.

A.3.5 The phoneme-alteration structure

```
(defstruct (phoneme-alteration)
  soft
  hard
  soft-leading
  hard-leading
  hard-trailing
  (clipping-duration 1)
  ;;PAUSE-TRANSITION - what to append to the phoneme to
  ;;effect a transition into silence
  pause-transition-smooth
  pause-transition-firm
  pause-transition-firmer
  pause-transition-abrupt
  )
```

The phoneme-alteration structure catalogues for one phoneme the alterations to the its pronunciation induced by context or the articulation style (slurred, enunciated) of the word to which it belongs. Its fields and their contents are:

- **soft**: The reduced vowel or less–articulated consonant that replaces the phoneme for decrease enunciation. Unvoiced consonants are often replaced with their voiced equivalents.
- **hard**: The fully enunciated vowel or consonant that replaced the phoneme for increased enunciation. Voiced consonants are often replaced with their unvoiced equivalents.
- **soft-leading**: When the phoneme is the first in the word, the contents of this field denote an articulatory configuration that precedes the word to makes its onset sound less distinct. If this field is empty, the **soft** phoneme is usually suitable.
- **hard-leading**: When the phoneme is the first in the word, the contents of this field denote an articulatory configuration that precedes the word in order to add emphasis to the word onset. If this field is empty, the **hard** phoneme is usually suitable.
- **soft-trailing**: (see *pause-transition-smooth*).
- **hard-trailing**: When the phoneme is the last in the word, the contents of this field denote an articulatory configuration that follows the word in order to add emphasis to word offset. This field is used to effect a firm or very firm word ending.
- **clipping-duration**: When the word offset is abrupt, the word is followed by the Dectalk silence phoneme. Normally, the silence phoneme abruptly stops phonation and then introduces silence. To stop phonation without also pausing, the silence phoneme is minimally sustained for **clipping-duration** milliseconds (usually 1 millisecond).
- **pause-transition-smooth**: When the phoneme is word–final, inserting the contents of this field between the phoneme and whatever follows effects a smooth transition

to silence or to the next word. The functionality of a separate, unique **soft-trailing** configuration is subsumed by this field.

- **pause-transition-firm**: When the phoneme is word–final and the word is followed by a pause, inserting the contents of this field between the phoneme and the ensuing silence effects a firm transition to silence. If this field is empty the contents of the **pause-transition-firmer** are appropriate.
- **pause-transition-firmer**: When the phoneme is word–final, and the word is followed by a pause, inserting the contents of this field between the phoneme and the ensuing silence effects a firmer transition to silence. If this field is empty the contents of **pause-transition-firm** are appropriate.
- **pause-transition-abrupt**: When the phoneme is word–final, and the word is followed by a pause, inserting the contents of this field between the phoneme and the pause effects an abrupt transition to silence. This field is filled only when the silence phoneme alone is insufficient. If the silence phoneme *is* used, its duration is the **clipping-duration**.

A.3.6 Interpreting for the synthesizer

A.3.7 The dectalk-word structure

```
(defstruct (dectalk-word )
  string                ;The text
  syllables             ;List of phonemes grouped into syllables
  accent                ;Either "'" or "\""
  (separator " ")       ;Dash or a space
  leading-phoneme       ;For enunciating or slurring
  trailing-phoneme      ;For enunciation or slurring
  pause-transition      ;Leading phoneme
  pause                 ;Pause or punctuation (comma)
  pause-duration        ;Pause duration
  phonemes?             ;Use phoneme pronunciation?
  )
```

The word accent and pronunciation descriptors are interpreted for the Dectalk. The results are stored in the dectalk-word structure.

- **string**: the word in plain text.
- **syllables**: the word phonemes, grouped into syllables.
- **accent**: a string representing primary or emphatic stress.
- **separator**: the character(s) that separate this word from the following word.
- **leading-phoneme**: effects enunciation or slurring at the beginning of the word.

- **trailing-phoneme**: effects enunciation or slurring at the end of the word.
- **pause-transition**: the phonemes representing the articulatory configuration as phonation stops and a pause begins.
- **pause**: the silence phoneme or a comma.
- **pause-duration**: duration of the silence phoneme in milliseconds.
- **phonemes?**: a boolean that signifies whether the word is sent as plain text or as phonemes.

Each word of the string eventually sent to the Dectalk is assembled in the following order:

- the leading phoneme, if it exists.
- then the accent, if it exists
- the phonemes, from **syllables** if **phonemes?** is true, or the string,
- the trailing phoneme if it exists, or the **pause-transition** phoneme(s) if they exist,
- the pause, if it exists,
- the pause duration, if it exists,
- and finally, the word separator unless the word is the last in the utterance.

Appendix B

Tables

B.1 Dectalk Exceptions

This table is consulted whenever an alteration to the text might encourage incorrect accenting or pronunciation so that corrective actions could be applied. The list of words needing special treatment is incomplete, as they are drawn only from the prepared sentences.

Prominence exceptions from word category are words for which "*/'/text*" has greater prominence than "*/phoneme-representation/*". The greater prominence is implemented as higher than normal pitch accent height, sometimes to the point of sounding like emphatic stress rather than default stress. For these words, the primary stress accent is applied to its phoneme representation.

Prominence exceptions from context are those function words which, when immediately preceded (with no space separating) by anything in phoneme mode, are erroneously accented. Sending these words in phoneme mode as well keeps them free of pitch accents.

Vowel alteration exceptions are words whose vowels are altered when the next item is in phoneme mode. The final vowel is given its full weight when it should be pronounced as a schwa. The remedy is to send the word in phoneme mode.

Doubling exceptions are those phonemes that when repeated produce speech that sounds either like stuttering, for consonants, like some strange echo, for vowels and syllabics. Normally, doubling a phoneme adds duration for a vowel or sonorant or emphasis for a consonant. Consonants that are unrepeatable should neither be doubled nor preceded by their voiced or unvoiced equivalent. The phoneme categories represented in this list are:

- all diphthongs
- all syllabics

- some sonorants
- all affricatives
- and the flap t allophone, "TX"

Type	*Affected words or phonemes*
Prominence from word category	The words: "always", "not".
Prominence from context	The words: "not", "in".
Vowel alteration	The words: "a", "to".
Phoneme doubling	The phonemes: "AR", "AW", "AY", "CH", "ER", "EY", "IR", "JH", "L", "OR", "OW", "OY", "TX", "UR", "W", "YU".

Table B.1: **Dectalk pronunciation exceptions.**

B.2 Descriptors

The *word pronunciation descriptors* are ordered and correspond to decrease and increase, or absence or presence of the feature. The *contour slope descriptor* describes a phrase feature, but is implemented by marking *one* word as the site of the contour maximum F0 value. *Pause descriptors* are ordered by increasing abruptness. The descriptors are listed in Table B.2.

B.3 Phoneme Alteration Table

This table contains substitutions or additions employed to effect phoneme slurring and enunciation, and various types of transitions from the phoneme into silence. Entries marked "NONE" disallow the alteration for the phoneme and override any other rule or table driven prescriptions. The original table is quite large. Its contents are distributed across four tables — the first set shows alterations that accomplish precision of articulation, the second the alterations that accomplish smooth, firm, firmer and abrupt pause onsets.

The **hard-trailing** field states explicitly the outcomes of the following rules:

- To enunciate a word ending in a consonant, follow the word with the same consonant if doubling is allowed (check the Dectalk exception table).

Word	
Feature	*Ordered Descriptors (minimally to maximally present)*
Articulation	:SOFTEN-CONSONANTS-AND-VOWELS++ :SOFTEN-CONSONANTS-AND-VOWELS+ :SOFTEN-CONSONANTS :SOFTEN-EDGES :NORMAL :ENUNCIATE+ :ENUNCIATE++ :ENUNCIATE-AND-SEPARATE-WORDS+ :ENUNCIATE-AND-SEPARATE-WORDS++
Exaggeration	:NORMAL :EXAGGERATE
Prominence	:NORMAL :PROM-MAX
Contour	
Feature	*Ordered Descriptors*
Slope	:SLOPE-DOWNWARD :NORMAL :SLOPE-UPWARD
Pause	
Feature	*Ordered Descriptors*
Transition to Silence	:SMOOTH :FIRM :FIRMER :ABRUPT

Table B.2: **Feature descriptors sets for words, contours and pauses.**

- To enunciate a word ending in a retroflex vowel – "AR", "ER", "IR", "OR" or "UR" — follow it with "R_".

- To enunciate a word ending in neither of the above, follow the word with the silence phoneme. The duration of the silence phoneme is the clipping duration for the word-final phoneme. It is set to 50 milliseconds for the phoneme, "IY". For all other phonemes it is set to 1.

Other heuristics are explictly stated rules and so are not encoded in the table. These rules state that:

- to enunciate a word beginning with a vowel, syllabic or glottal stop precede it with "HX".

- to enunciate a word beginning with a consonant, precede it with its **hard** phoneme equivalent if the word is in phoneme mode, or precede the word with the word-initial phoneme if doubling is allowed.

- to reduce the precision at word end, follow the word with the **pause-transition-smooth** phoneme(s) for the word-final phoneme.

- if the **pause-transition-firm** field is empty, use the contents of the **pause-transition-firmer**, and vice versa.

Phoneme	*soft*	*hard*	*soft leading*	*hard leading*	*hard trailing*
AA				Q	-
AE	AX			(HX 10)	-
AH				Q	-
AO				Q	-
AR				Q	R-
AW					-
AX				Q	-
AY	AAIX			Q	-
B		PB			B-
CH	JH				-
D	DX				D-
DH		TH			DH-
DX		D			DX-
EH				Q	-
EL				Q	-
EN				Q	-
ER				Q	R-
EY		(EH 45)		Q	-
F	V				F-
G		K			G-
IH	AX			Q	-
IR				Q	R-
IX	AX			Q	-
IY	yy			Q	-

Table B.3: **Phoneme alterations for accomplishing precision of articulation variations, for most phonemes from "AA" to "IY".**

Phoneme	*soft*	*hard*	*soft leading*	*hard leading*	*hard trailing*
JH		CH			-
K					K-
L					-
M					M-
N					N-
NX					NX-
OR				Q	R-
OW				Q	-
OY				Q	-
P	BB		NONE		P-
R					R-
RR				Q	-
S	ZS				S-
SH					SH-
T	TX				T-
TH	DH				TH-
V		VF			V-
UH				Q	-
UR	UH			Q	R-
UW	UH			Q	-
YU	YAX		HX		-
Z					Z-
ZH					ZH-

Table B.4: **Phoneme alterations for accomplishing precision of articulation variations, for most phonemes from "JH" to "ZH".**

Phoneme	*pause transition smooth*	*pause transition firm*	*pause transition firmer*	*pause transition abrupt*
AE				
AH				
AO	AO		AX	
AR	RX	(HX 5)	Q	
AW			(UW 50)	(UW 2)
AX	(AX 5)	(HX 5)		
CH		(HX 5)		
D	DX	D		
DH	(DH 5)	DH		
DX	DX	D		
EN	N			
ER	R			
EY	(IY 1)			
F	(F 5)	F		
G		G		
IH				
IR	RX	(HX 5)	Q	
IX				

Table B.5: **Phoneme alteration information for pause transitions for most phonemes from "AA" to "IX".**

Phoneme	*smooth pause transition*	*firm pause transition*	*firmer pause transition*	*abrupt pause transition*
JH		(HX 5)		
K	(HX 4)	K		
L	L			
M	(MHX 5)	M	(M 5)	
N	(N 45)	(N 5)	NHX	
NX	NXNX	NX	NXHX	
OR	RX	(HX 5)	Q	
P	(HX 4)	P		
RR	(R 45)	(HX 5)	Q	
S	(S 5)	S		
SH	(SH 5)	SH		
T	(THX 4)	T		
TH	(TH 5)	TH		
V	V	V		
UR	RX	(HX 5)	Q	
UW	UW			
Z	(Z 35)	Z		
ZH	(ZH 5)	ZH		

Table B.6: **Phoneme alteration information for pause transitions for most phonemes from "JH" to "ZH".**

Appendix C

Acknowledgments

The problem with having taken so long to finish is that the list of people who have proffered support continually grows. Even though I have finished, I hope they will not feel obligated to cease performing acts of kindness and amusement. They are always appreciated. While the long list is surely incomplete, I gratefully acknowledge:

My advisor, Chris Schmandt, for believing in the work and the person doing it, for simplifying the complex task I set for myself, for cutting me generous portions of slack.

Jim Davis, for the use of his Discourse software, numerous improvements to my own software and for blinding clarity of thought that illuminated the way out of the thesis darkness.

Dennis Klatt, for painstakingly reviewing my proposal and forcing me to clarify my ideas, for entrusting me with his software and for his help in configuring the Dectalk3.

Ken Stevens, for a critical and useful review of my proposal..

Ted Adelson for clarifying my experiment design, and Karen Wynn for general enlightenment on testing and analysis procedures.

Ed Bruckert, at DEC, for sending the Dectalk 3 card schematics, and Wad and Kael, for interpreting them.

The developers of SPIRE, with which the synthesized speech was analyzed. The pitch tracks, spectrograms, spectral slices and energy plots were made with this tool.

My sponsors, N.T.T. and DARPA.

The 11 subjects who particpated in my initial experiments, and the 28 subjects from the final version. I promised to be eternally grateful and I am. Thank you!!!!!!

and then, ...

Tom, for his reassurance, and gentle and total attention, sorely missed; and Kael, for encouragement, and for clearing a path through the Latex jungle; Bill and Plin, living proof that old age need not lead to serious decline in mental faculties; Janette, Elaine, Carol and Clea, for wonderfully correcting a grievous Y-chromosome imbalance; Ben L., for preventing cultural atrophy with whale–sized feedings of poetry, music and encouragement; Pat and Plin, winners of the World's Best Officemate competition (Pat for 2 years in a row!), for their good humor and boundless tolerance; the West Coast contingent (Amy, Anna, Cathleen, Cathy, Curt, Kaiti, Lisa, Lydia, Pat, Stephanie, Vicki, Toni, etc, etc!!), whose telephone, paper and e-mail voices kept me sane; the many at HP Labs who facilated my extended leave of absence, chief among them — Ted Wilson, Arden Taylor, and Dan Flickinger; the staff of the Media Lab Hotel, for booking me a great room with a stunning view, and in general, for running a fine operation and providing all the amenities except showers; Sally Alley, for showing up always and only when needed, without prompting; my family, immediate and extended, for continual support and occasional food. And of course, Lasky for laughter (mine) and adulation (his); Henry for, well, being Henry; Odin, for conversation and legal aid, Lacsap, for motivation-by-annoyance (it worked!); Dan A., for roses! and finally... the various species of mammal that populate the Terminal Garden, for keeping company.

Bibliography

[1] Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt, Robert C. Armstrong, and David B. Pisoni. *From text to speech: The MITalk system.* Cambridge University Press, 1987.

[2] Mark D. Anderson, Janet Pierrehumbert, and Mark Y. Liberman. Synthesis by rule of english intonation patterns. *ICASSP*, 2.8.1–2.8.4, 1984.

[3] Gary Collier. *Emotional Expression.* Lawrence Erlbaum Associates, 1985.

[4] Digital Equipment Corporation. *Dectalk DTC03 Text-to-Speech System Owner's Manual.* Digital Equipment Corporation, 1985.

[5] Joel Davitz. *The Communication of Emotional Meaning.* McGraw-Hill, 1964.

[6] Allen T. Dittmann. The body movement–speech rhythm relationship as a cue to speech encoding. In Weitz, editor, *Non Verbal Communication*, pages 168–177, Oxford University, 1974.

[7] G. Fairbanks and L. W. Hoaglin. An experimental study of the duration characteristics of the voice during the expression of emotion. *Speech Monographs*, 8:85–90, 1941.

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

[9] D. B. Fry. Experiments in the perception of stress. *Language and Speech*, 1(2):126–152, 1958.

[10] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[11] Julia Hirschberg and Janet Pierrehumbert. The intonational structure of discourse. In *Proceedings of the Association for Computational Linguistics*, pages 136–144, July 1986.

[12] Dennis H. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustic Society of America*, 82(3):737–793, Sept 1987.

[13] John Laver and Robert Hanson. Describing the normal voice. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 51–78, Grune and Stratton, Inc., 1981.

[14] Mark Liberman and Janet Pierrehumbert. Intonational invariance under changes in pitch range and length. In *Language Sound Structure*, chapter 10, MIT Press, 1984.

[15] John J. Ohala. Nonlinguistic components of speech. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 39–50, Grune and Stratton, Inc., 1981.

[16] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, 1957.

[17] Janet Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, Dept of Linguistics, 1980.

[18] Janet Pierrehumbert. Synthesizing intonation. *JASA*, 985–995, Oct 1981.

[19] Janet Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In Cohen, Pollack, and Morgan, editors, *Intentions and Plans in Communication and Discourse*, 1989.

[20] Klaus Scherer. Speech and emotional states. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 189–220, Grune and Stratton, Inc., 1981.

[21] Klaus Scherer and Ursula Scherer. Speech behavior and personality. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 115–135, Grune and Stratton, Inc., 1981.

[22] Klaus R. Scherer. Acoustic concomitants of emotional dimensions: judging affects from synthesized tone sequences. In Weitz, editor, *NonVerbal Communication*, pages 105–111, Oxford University, 1974.

[23] J. M. Sorensen and W. E. Cooper. Syntactic coding of fundamental frequency in speech production. In R. A. Cole, editor, *Perception and Production of Fluent Speech*, pages 399–440, Lawrence Erlbaum, 1980.

[24] Carl E. Williams and Kenneth N. Stevens. Emotions and speech: some acoustical correlates. *JASA*, 52(4 (Part 2)):1238–1250, 1972.

[25] Carl E. Williams and Kenneth N. Stevens. On determining the emotional state of pilots during flight: an exploratory study. *Aerospace Medicine*, 40(12):1369–1372, Dec 1969.

[26] Carl E. Williams and Kenneth N. Stevens. Vocal correlates of emotional states. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 189–220, Grune and Stratton, Inc., 1981.