PERFORMANCE OF "GALE" USING SEMANTICALLY NEUTRAL SENTENCES

By

VUDAY NANDUR

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

UNIVERSITY OF FLORIDA

2003

This work is dedicated to my wife and my family
who have supported my every endeavor.

ACKNOWLEDGMENTS

I would like to thank Dr. Blischak for her continued encouragement and valuable guidance towards completion of this work.

I would also like to thank Dr. Brown, Dr. Rothman and Dr. Shrivastav for their timely advice.

Last but not least, I would like to thank the participants who readily took part in the study.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table page

LIST OF FIGURES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts

PERFORMANCE OF "GALE" USING SEMANTICALLY NEUTRAL SENTENCES

By

Vuday Nandur

May 2003

Chair: Doreen M. Blischak
Major Department: Communication Sciences and Disorders

"GALE" is software that interfaces with the DECtalk speech synthesizer to simulate emotions in synthetic speech. It takes a text string as input and generates synthetic speech with five emotions--neutral, happiness, sadness, fear, and anger. A preliminary study on the performance of GALE indicated that the five emotions were simulated with some success. However, the semantic content of stimulus sentences was not controlled. The purpose of the current study was to determine how well GALE simulates the five emotions with semantically neutral sentences.

The study was divided into two phases. In Phase I, a list of 10 semantically neutral sentences was compiled. In Phase II, the ten sentences were synthesized with each of GALE's five emotions. Twenty participants determined the emotion they heard in each of the 50 sentences as neutral, happy, sad, afraid, angry, or other. Sad was correctly identified the most number of times as the target emotion. Anger was identified correctly the least number of times. All emotions except sad were identified as angry the same number times as angry was identified correctly.

Truluck had similar findings, except that the angry emotion was correctly identified significantly more often. A post-hoc analysis of Truluck's stimuli revealed that this difference was likely due to use of sentences that were semantically biased towards anger.

CHAPTER 1
INTRODUCTION

Computers are probably the closest man-made replicas of human beings. Technological advances are making computers see, talk, move, and "think" more and more like humans. In the past few decades, there have been rapid advances in computer generated or synthesized speech. Synthetic speech has developed from being relatively monotonous to having rich intonation. With the advent of more powerful machines and better algorithms, it is now possible to simulate emotions in synthetic speech.

Synthetic speech systems that convey emotions have been developed in the last decade. These include HAMLET by Murray and Arnott (1995), Affect Editor by Cahn (1990), and GALE by Truluck (1998). These systems act as an interface between an input system such as a text editor and a speech synthesizer such as DECtalk (Digital Equipment Corporation, Nashua, NH) and introduce the element of emotion into synthetic speech. Each of these systems produces emotive speech with varying degrees of naturalness.

The performance of GALE was determined by Truluck (1998) using sentences synthesized with five different emotions--neutral, happy, sad, afraid, and angry. Truluck found that GALE simulated sadness, afraid, and anger with a high degree of success. However, Truluck's sentences were not controlled for semantic content, potentially confounding her results. The purpose of the present study was to determine the performance of GALE using semantically neutral sentences.

The study is divided into two phases. In Phase I, 75 sentences were created to depict five emotions--neutral (no emotion), happy, sad, fear, and anger. Participants were

given a written list of randomly distributed sentences and asked to indicate the strongest emotion conveyed by each sentence. Ten sentences that were determined to be semantically neutral at least 85% of the time were chosen for Phase II.

In Phase II, each of the 10 sentences was generated in the five emotions of GALE, for a total of 50 sentences that were synthesized. A second set of participants listened to these sentences and identified the emotion conveyed by each sentence.

Following a review of literature, procedures, results and their implications are discussed in detail in the following chapters. Future developments and applications of emotional synthetic speech are also discussed.

CHAPTER 2
LITERATURE REVIEW

Synthetic Speech

History

Speech that is generated mechanically and/or electronically, with no human vocal component, is called synthetic speech. One of the first known efforts to build a speech synthesis system was by Wolfang von Kempelen (1791), who attempted to build a mechanical model to emulate human spoken communication. Charles Wheatson (1830) reconstructed Kempelen's model by adding a flexible oral cavity and voicing control. Unfortunately, intelligibility data of these devices have not been reported in the literature (Klatt, 1987). The first electronic synthesizer to generate speech sounds was VODER (Voice Operation Demonstrator) developed by Homer Dudley in 1939. VODER was difficult to operate and its intelligibility was "marginal" (Klatt, 1987). The Pattern Playback synthesizer was developed around the same time and was based on converting broadband spectrograms into speech. The speech sounded unnatural but with "adequate" intelligibility (Klatt, 1987).

In 1960, the acoustic theory of speech was developed. The theory states that speech is the result of one or more sound sources exciting a linear filter (Fant, 1960 cited in Klatt, 1987). The most popular speech synthesis model, that of acoustic synthesis, is based on this theory. This model further improved the quality of synthetic speech. In the late 1960s, computers replaced electrical systems to produce synthetic speech. This development allowed for the use of more control parameters, which helped improve the

quality of synthetic speech (Klatt, 1987). With the advent of more powerful computers and more memory, highly intelligible synthetic speech could be generated.

Models of Synthetic Speech Production

Synthetic speech is produced according to two basic models. The articulatory synthesis model involves producing sounds by emulating the human vocal system. Speech synthesizers based on this model were popular before the acoustic theory of speech was proposed. An example of an articulatory synthesizer is IGOR, developed at Bell labs (Coker, 1967). The articulatory model has not been widely used because of the inherent difficulties in the exact and precise emulation of human vocal tract dynamics, resulting in poor quality speech (Hill et al., 1995). However, with the advent of more powerful computers, this method shows promise in producing natural sounding speech (Klatt, 1987; Venkatagiri & Ramabadran, 1995).

The second model, acoustic synthesis, has been the more popular of the two and is based on the acoustic theory of speech. In this model, speech is generated by digital to analog conversion of a synthesized acoustic waveform (Venkatagiri & Ramabadran, 1995). Mathematical algorithms are used to combine acoustic properties of an utterance with rules of pronunciation, voice inflection, and accent to generate synthetic speech (Mirenda & Beukelman, 1990). Because of its relative ease and lesser cost, most of the commercially available synthesizers are based on the acoustic synthesis model (Klatt, 1987).

Construction of a Speech Synthesizer

Construction of a speech synthesizer based on the acoustic model involves the development of a database of digitally coded speech. Briefly, the existing linguistic and paralinguistic features of the target language are extracted from an analog sample and

coded as digital data. Coding is done in two general ways--waveform coding and parametric coding.

Waveform Coding

In this technique, the entire sound wave, typically encompassing words or sentences, is coded and stored. Due to a large memory requirement, this method is used only when small amounts of speech output are needed at a time. On the other hand, extensive processing or complicated electronics are not required. The coarticulatory and prosodic features of the original waveform are also preserved (Venkatagiri & Ramabadran, 1995).

Parametric Coding

In parametric coding, only the perceptually important acoustic parameters of speech are coded and stored. Therefore, unlike waveform coding, the output is perceptually similar but acoustically different from the original speech from which the digitized data were derived (Venkatagiri & Ramabadran, 1995). This technique is less taxing on memory but requires relatively more processing power. There are two types of parametric coding--formant coding and linear predictive coding (LPC).

In formant coding, various parametric values such as center frequencies, formants, bandwidths, amplitudes, and voicing, extracted from the original signal, are used to generate synthetic speech by passing them through signal processors, which simulate the glottal and vocal tract functions using resonators. Resonators are arranged in either a cascade or parallel form. Cascade form is a better model for non-nasal voiced sounds, while parallel form is better for nasal and unvoiced sounds (Holmes, 1983). Klatt (1980) combined both cascade and parallel forms in Klatt's Formant Synthesizer, which has

been incorporated into several synthesis systems, such as MITalk, DECtalk, and Votrax (Votrax, Inc., Troy, MI) (Donovan, 1996).

In linear predictive coding (LPC), the values of a speech unit are estimated based on the values of previous samples. A phoneme, for example, is divided into many frames, each containing all the spectral details necessary to reproduce the utterance (Atal & Hanauer, 1971 cited in Venkatagiri & Ramabadran, 1995). The frame sets, along with fundamental frequency and overall amplitude of the source function are used to generate speech. Speak and Spell (Texas Instruments, Austin, TX), Echo products (Street Electronics, Carpentaria, CA), and Macintalk II Pro (Apple Computer Corporation, Cupertino, CA) are some of the synthesizers based on this technique.

Text-to-Speech Synthesis

Text-to-speech is the most popular method of speech synthesis. Input provided as text in the form of letters, syllables, words, or sentences is converted to speech, based on pre-existing coded speech data in the synthesizer's memory (Venkatagiri & Ramabadran, 1995). The entire process of text-to-speech synthesis involves three steps--phonetic transcription, parameter calculation, and digital to analog conversion.

Phonetic Transcription

In this step, text input is transformed into a string of corresponding phonemes and allophones. Letter-to-sound rules numbering in the hundreds are used to assign codes to letters based on phonetic context. Words that have exceptional pronunciation rules are stored separately. Lexical stress, intonation, and timing rules are also added to ensure correct pronunciation and prosody (Allen et al., 1987). For example, DECtalk has an exception dictionary of 6000 words and rules that generate prosodic aspects, resulting in highly intelligible synthetic speech (Venkatagiri & Ramabadran, 1995).

Digital Signal Generation

Parametric values corresponding to the phonetic representations of the letters and words in the text are extracted from the coded speech database to generate a digital representation of the utterance. The parameters are calculated based on the size of the units of speech used as building blocks. DECtalk, Echo, and Votrax synthesizers use phonemes as the building blocks. Spectral smoothing across phonemic boundaries is done based on rules stored as parameters to simulate coarticulation effects on a phoneme (Venkatagiri & Ramabadran, 1995). Larger units such as diphones, which extend from the center of one phoneme to the center of the next phoneme, better preserve coarticulatory effects but require approximately 60 times more memory than do phonemes (O'Shaughnessy et al., 1988). Macintalk II Pro and Smoothtalker 3.0 (First Byte, Long Beach, CA) are two commercial speech synthesizers that use diphones as building blocks. After the extraction of the parametric values, the stage is now set for the final step of digital to analog conversion.

Digital to Analog

The final step in the speech synthesis process is the conversion of the speech data in digital form into analog form to be heard as synthetic speech. The final product then is an acoustic signal which is perceived as speech. For more information regarding details of the synthetic speech generating process, the reader is referred to Venkatagiri and Ramabadran (1995) and Kent and Read (2002).

Applications

Speech scientists may use synthetic speech to conduct basic research into speech perception (Kent & Read, 1992). The ability to manipulate various parameters of synthetic speech facilitates preparation of precise stimuli. Speech-language pathologists

work with individuals who use synthetic speech in speech generating devices (SGDs) as an augmentative and alternative communication (AAC) method. The most popular text-to-speech synthesizer used in speech generating devices is DECtalk (Mirenda & Beukelman, 1990).

Synthetic speech can also be used by visually challenged individuals, for example, to tell time, announce the entrance or exit of people from an area, the outcome of an action, and to access written text. The general population can use synthetic speech in tasks such as message reading, automated message systems, accent modification, and translation from one language to another (Klatt, 1987).

Performance of Speech Synthesizers

Several studies have been conducted to determine the intelligibility of various speech synthesizers. Intelligibility can be defined as the ability of listeners to correctly identify linguistic units such as syllables and words (Moody et al., 1987 cited in Reynolds et al., 2000). Intelligibility of synthesizers before 1983, such as Votrax and Echo, was found to be significantly less than natural speech (Greene et al., 1986). Then, with increasingly powerful processors and greater memory, intelligibility of synthetic speech has steadily improved. DECtalk and MacIntalk are examples of synthesizers that produce high quality, intelligible speech (Rupprecht et al., 1995). In fact, the intelligibility of DECtalk was found to be statistically equivalent to that of natural speech when used with meaningful sentences (Mirenda & Beukelman, 1987, 1990; Reynolds & Jefferson, 1999).

Another measure of speech synthesizer performance is comprehensibility. Comprehensibility can be defined as the ability of listeners to extract the underlying meaning from the speech signal (Duffy & Pisoni, 1992). For example, Reynolds et al. (2002) used a sentence verification task to measure comprehension of natural and

synthetic speech. DECtalk is generally significantly more comprehensible than other synthesizers but more difficult to comprehend than natural speech (Reynolds & Jefferson, 1999). This difference in comprehensibility between DECtalk and natural speech was explained as due to the relatively acoustically impoverished synthetic speech, making it necessary for increased cognitive effort (Nusbaum & Pisoni, 1985). This suggests that there is something lacking even in the most intelligible synthetic speech.

Though synthetic speech has been shown to be intelligible and reasonably comprehensible, it lacks naturalness. This may be attributed to incomplete representation of coarticulation and prosody found in natural speech. Studies have shown that simulating affect or emotiveness in synthesized speech may make it sound more natural (Cahn, 1990; Murray & Arnott, 1995; Truluck, 1998).

Emotions in Synthetic Speech

Emotion has been defined in many ways for example, moving of the mind or soul; and excitement of the feelings, whether pleasing or painful. Data on the acoustical correlates of various emotions are well documented (Murray & Arnott, 1993; Scherer, 1981; William & Stevens, 1972). For example, sadness is characterized by an increase in the duration of an utterance and a decrease in the fundamental frequency, loudness, and the rate of speech. Happiness is characterized by an increase in the fundamental frequency and rate of speech. Emotions may be simulated in synthetic speech by manipulating these acoustic characteristics. Such changes in vocal parameters can be compiled and converted into values for manipulation by a synthesizer to simulate emotion.

Synthetic speech with emotion has varied applications. Emotionally rich voices may be a welcome change in the telemarketing world and in automated systems such as

credit card customer service. Emotional synthetic speech may also be used to make SGDs sound more natural. Other potential areas in our daily life where high quality synthesized speech may be welcome are e-mail readers, reminder systems in cars, cameras and watches, vocal alarm systems, and automated news readers.

Most of the emotional speech synthesis systems are based on DECtalk, as it is the most popular speech synthesizer in the market (Mirenda & Beukelman, 1990). Two such systems, the Affect Editor and HAMLET are discussed briefly. GALE, on which the current study is based, is discussed in greater detail.

Affect Editor

Janet Cahn (1990) at the MIT Media Technology Laboratory developed the Affect Editor. Emotions simulated by the Affect Editor include anger, disgust, gladness, sadness, fear, and surprise. Affect Editor was based on the phoneme and plain text modes of DECtalk 2.0. Parameters such as pitch, timing, voice quality, and articulation were varied on a scale of –10 to 10, with 0 being the neutral position (Cahn, 1990). Users can choose the emotion and type the text string to be synthesized. Appropriate annotations regarding information about the various parts of the sentence can be included in the text. The output is in the form of a text string with the appropriate acoustic and phonological cues. This output is interpreted by synthesizers based on their own speech generation rules. Cahn (1990) conducted a study with 28 participants to determine how well the Affect Editor simulates emotions. All emotions-neutral, angry, disgusted, glad, sad, scared and surprised- were identified with 50% accuracy or more. Sad was the most accurate (90%) while angry was the least accurate (43%).

HAMLET

Helpful Automatic Machine for Language and Emotional Talk (HAMLET), also based on DECtalk, was developed by Ian Murray and John Arnott (1995). Six basic emotions--anger, happiness, sadness, fear, disgust, and grief, were simulated. DECtalk's prosodic rules were ignored because they could not be set on a particular phoneme. Instead, prosodic rules were re-implemented using pitch contour rules based on the acoustic correlates of vocal emotion and actor-generated emotions (Murray & Arnott, 1993). Eleven emotion-dependent, phoneme pitch, and duration rules were defined. A subset of these rules was used for each emotion. Murray and Arnott (1995) conducted a study to determine how well emotions were simulated by HAMLET. Neutral sentences as well as sentences with semantics leaning towards an emotion were used as test stimuli. With neutral sentences, sadness was identified correctly most often and disgust was identified least often. With non-neutral sentences, anger was identified the highest number of times while disgust was identified the least number of times.

Generator of Affected Language Expression (GALE)

GALE is a DECtalk based emotional speech synthesizer that was developed by D'Arcy Truluck in 1998. Inputs to GALE include a text string, a DECtalk voice, and an emotion. Emotions generated by GALE are neutral (no change), happiness (joy, bordering on elation), sadness (very sad, close to grief), fear (very fearful), and anger (hot anger) (Truluck, 1998). Emotions are simulated by encoding the text string with the appropriate acoustic correlates. This encoded string is given as input to DECtalk, which then generates the synthesized utterance in one of its nine voices.

Truluck (1998) chose parameters to simulate emotions in GALE (Appendix A) based on previous studies to determine the vocal aspects of emotion, notably by Scherer

(1981) and Scherer et al., 1991). Scherer (1981) used the Moog synthesizer (Moog Music, Inc., Asheville, NC) to determine the vocal cues used by listeners to distinguish between emotions. Further, based on behavior observed in natural speech, Truluck (1998) made changes to the utterances to bring out better affect. For example, a period at the end of a sentence simulates a slight rise in stress for to indicate happy. Phonemes in fear and anger are made more "precise" by changing lax vowels to their tense equivalents. For sadness, phonemes are made more "slurred" by changing tense vowels to their lax equivalents.

Truluck's Evaluation of GALE

Truluck (1998) evaluated the performance of GALE. Twelve sentences were randomly selected from 14 original sentences (Appendix B), which were created to express a variety of emotions (Truluck, 1998). A set of 60 test sentences was compiled by generating the 12 sentences with each of the five emotions of GALE. The sentences were presented to 28 adult participants, who chose which emotion was expressed in each sentence. Overall results are shown in Figure 2-1.

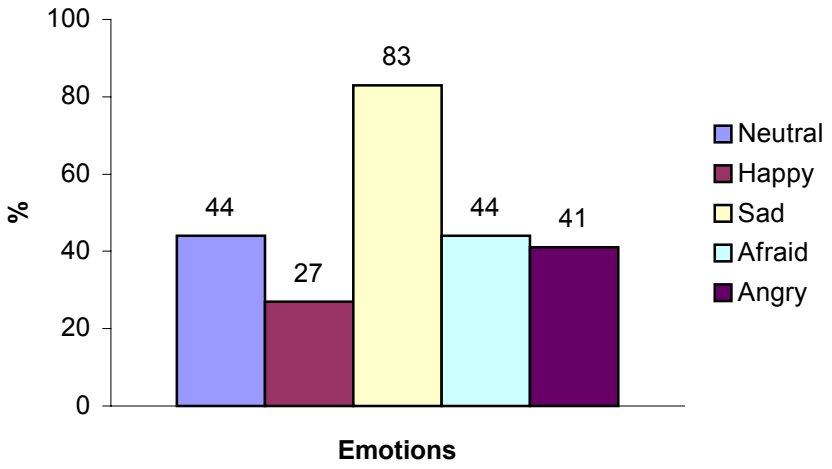

Figure 2-1: Truluck's results

Questions

Truluck, however, did not control test sentences for semantic content. This may have affected the outcome of the study. It can be hypothesized that listeners were influenced by the emotion conveyed by the semantics of the sentence when determining the emotion generated by GALE. Thus, the present study was designed to answer the following questions:

1. How does GALE perform with semantically neutral sentences?
2. How does the performance of GALE with neutral sentences compare with sentences not controlled semantically?

CHAPTER 3
METHODOLOGY

The study was conducted in two phases. In Phase I, a list of 10 semantically neutral sentences was formulated. In Phase II, these 10 sentences were synthesized with each of the five emotions in GALE to determine how well GALE generates these emotions.

Phase I

Participants

Twenty-five female and fifteen male native English speakers 18 to 45 years of age participated in the study. They were recruited from the University of Florida campus on a voluntary basis. Participants took part individually or in groups of at most three.

Stimuli

Fifteen sentences were generating to represent each of the five emotions--neutral, happy, sad, anger, and fear, for a total of 75 sentences. These were generated by the principal investigator and the faculty advisor based on common experiences of southern United States. Five different randomly ordered sets of these 75 sentences were prepared.

Procedure

Participants were asked to provide demographic information in writing (Appendix D) and read and sign the informed consent form approved by the University's Institutional Review Board (IRB). Written instructions (Appendix E) with an illustrated example were provided and any questions were answered by the principal investigator. One of the five sentence lists was selected at random for each participant. Each sentence

had five columns, one for each emotion. Participants were asked to indicate the strongest emotion that they thought was expressed by the meaning of each sentence. They were permitted to revise a previously made choice. There was no time limit.

Results and Discussion

The 3000 obtained responses were tallied for each sentence according to the five emotion categories. Participants indicated that 55 of the 75 (73%) presented sentences depicted a particular emotion at least 85% of the time. Sentences depicting happy were identified most frequently (87%), followed by neutral (80%), afraid (73%), angry (67%), and sad (60%). Thus, nearly three-quarters of sentences created for Phase I consistently evoked the target emotion.

Phase II

Participants and Setting

Twelve female and eight male native English speakers 18 to 45 years of age participated in Phase II. They were recruited from the University of Florida campus on a voluntary basis. Data was collected in a small computer lab with one participant at a time.

Hearing Screening

Participants were asked if they recently had failed a hearing screening. Only those who answered “no” were allowed to participate in the study. Those who never had a hearing test were administered a hearing screening where pure tones were presented at 25 dB HL for the frequencies 500 Hz, 1000 Hz, and 2000 Hz. Potential participants were required to pass this hearing screening. Of the 20 participants, 13 had passed a recent hearing screening; 7 passed the administered screening.

Stimuli

Of the 12 sentences identified as "neutral" in Phase I, 10 with the highest identification rate greater than 85% were chosen for Phase II. Each sentence was synthesized using each of the five emotions of GALE in DECtalk's Paul voice. Three different randomly ordered sets of the 50 sentences, with each sentence separated by a four second gap, were recorded utilizing a digital audiotape (DAT) recorder (Sony DAT Deck DTC-60ES). Ten of the 50 sentences were randomly repeated to determine intra-subject reliability.

Procedure

Participants were asked to complete a form (Appendix D) with demographic information and, read and sign the informed consent form approved by the IRB. Written instructions (Appendix F) were provided and any questions about the procedure were answered by the principal investigator before stimulus presentation. Stimuli were presented through earphones (Pro 25 Titanium Supra aural). Two sentences not used in the actual study were used to illustrate the procedure.

Responses were recorded on a pre-designed answer sheet (Appendix G) having 6 columns, one each for neutral, happy, sad, afraid, angry, and other. Choices in the first five columns were indicated with a '√' or an 'X'. If participants wanted to indicate an alternative emotion, they could write it in the "other" column. Stimulus sentences did not appear on the answer sheet. Participants indicated the emotion they heard in the four seconds following each sentence presentation.

Chi-Squared Analysis

A chi-squared analysis was performed to determine performance above chance. The independent variable was the target emotion presented with each sentence; the dependent variable was the actual emotion chosen by participants.

CHAPTER 4
RESULTS

Overall, of the 980 responses, 419 (42.76%) were correctly identified as the target emotion. A breakdown by each emotion is provided in Table 4-1. Each emotion had a total of 200 responses (20 participants x 10 sentences) (Appendix H), except for "afraid." During analysis of the results, it was found that one sentence had been omitted for "afraid" resulting in a total of 180 samplings (20 participants rated 9 sentences).

Table 4-1: Overall responses

		Actual Response (%)						
		Neutral	Happy	Sad	Afraid	Angry	Other	TOTAL
	Neutral	**50.5**	13	5.5	10	21	0	100
Expected	Happy	39.5	**26.5**	5.5	8	19.5	1	100
Response	Sad	9.5	1	**76**	1.5	9	3	100
	Afraid	7.2	14.4	13.9	**40.6**	20.6	3.3	100
	Angry	12.5	38.5	11.5	16	**20**	1.5	100

Results of the overall chi-square analysis indicated a value of 669 ($p < 0.001$) at 20 degrees of freedom. The chi-squared values at the respective degrees of freedom and significance for each emotion are presented in Table 4-2.

Table 4-2: Chi squared results

Emotion	Chi square value	Degrees of freedom	*p* value
Neutral	94.3	45	<.001
Happy	58.72	45	<.001
Sad	337.23	45	<.001
Afraid	114.57	40	<.001
Angry	64.46	45	<.001
TOTAL	669.28	20	<.001

Chi-square values indicate that responses for each emotion were likely due to the stimuli presented and not due to chance.

Intra-Subject Agreement

Intra-subject agreement was estimated by calculating percent agreement. A point-by-point comparison of responses to ten sentences and their repetitions was conducted. Percent agreement was calculated as the ratio of the number of agreements over the sum of agreements and disagreements, times 100. Percent agreement was highest for sad (60%) and least for happy (41%) and angry (41%), as shown in Figure 4-1.

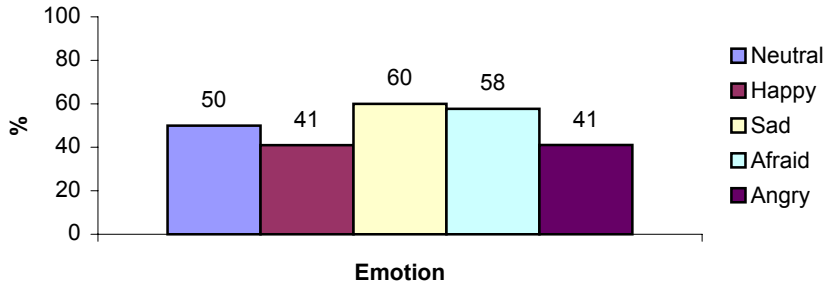

Figure 4-1: Within-subject reliability results.

A further analysis was conducted to determine the number of agreements that were also correct. This would further indicate how well the emotions were generated by GALE. Percent agreement was calculated as the ratio of agreements that were correct over the total number of agreements. Sad was found to have the highest percent agreement (87%), while angry was found to have the least agreement (12%). Results are shown in Table 4-3.

Table 4-3: Consistent and correct responses

	Neutral	**Happy**	**Sad**	**Afraid**	**Angry**
# agreements	17	25	24	15	16
agree + correct	9	6	21	8	2
% agreement	52.94	24.00	87.50	53.33	12.50

Procedural Integrity

A checklist of the sequential steps involved in Phase II was developed for ensuring procedural integrity. A research assistant checked off the steps as the principal

investigator performed them. This was done with eight of the twenty subjects (40%). Procedures were correctly performed 100% of the time.

Comparison to Truluck

A comparison of Truluck's results (Figure 4-2) with the results of the current study (Figure 4-3) showed that the average difference between the two studies was 7%, with angry having the most variation (21%) and happy having the least (0.5%).

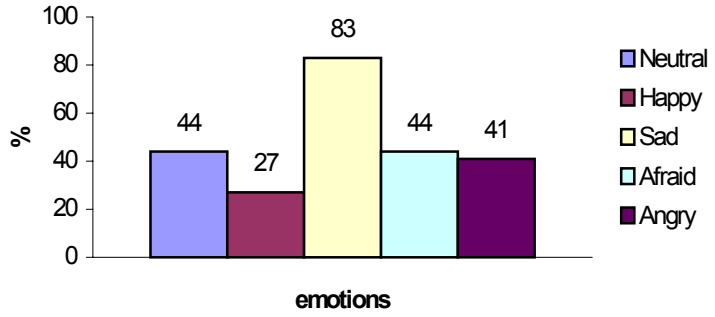

Figure 4-2: Truluck's results.

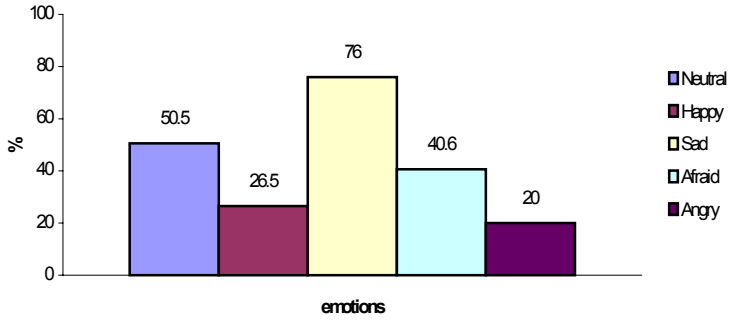

Figure 4-3: Results of current study

This observation raised the following questions:

3. Did Truluck inadvertently use semantically neutral sentences that would explain similarity across students for four of the five emotions?

4. Did the sentences used by Truluck have an inherent angry bias, resulting in the relatively large difference for angry between the two studies?

A post-hoc analysis was conducted to answer these questions. Twenty adult native English speakers recruited from the University of Florida campus were asked to determine the emotions of Truluck's 14 stimulus sentences (Appendix B). None of the participants from Phase I or Phase II were selected. The procedure was identical to that used in Phase I of the present study. Results are shown in Figure 4-4.

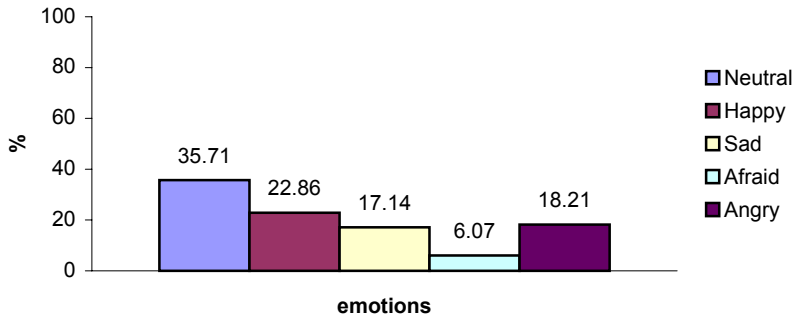

Figure 4-4: Semantic distribution of Truluck's sentences.

Only 36% of the sentences used by Truluck were semantically neutral. The other sentences were biased towards one of the other five emotions to some degree. Since Truluck did not use neutral sentences, but her results were similar to that of the current study for neutral, happy, sad, and afraid, it indicates that the performance of GALE is similar with both semantically neutral as well as non-neutral sentences for these emotions. To explain the relatively larger difference in the results for angry, semantically angry sentences were isolated and studied in more detail. Results for the three most semantically angry sentences generated with the angry emotion of GALE are shown in Table 4-4.

Table 4-4: Distribution of angry sentences with angry emotion

Rank of angry sentence	Sentence number (Appendix B)	% times sentence determined angry	% times determined angry with GALE
1	13	80	75
2	7	45	70
3	12	40	40

Sentences that were determined semantically angry the most number of times were also the ones where the angry emotion of GALE was correctly identified the most number of times. Similar analysis of other emotions did not show such a pattern (APPENDIX I). This suggests that the elevated results for angry in Truluck's study are likely due to the semantically angry sentences.

Discussion

GALE's ability to synthesize emotions appears to be mixed. Sad and afraid were identified correctly a relatively high number of times, suggesting that GALE synthesizes these emotions with some degree of success. Happiness was often identified as neutral. This may be expected, as happy was synthesized with the least number of transformations to the neutral voice (Truluck, 1998). Thus, happiness appears to need further improvement to distinguish it from the neutral emotion.

Of all the emotions, angry was identified correctly the least number of times. In fact, angry sentences often identified as happy. This is in accordance with what has been reported in literature, and may be attributed to similar values of acoustical correlates between happy and angry (Davitz, 1964). Furthermore, the emotions neutral, happy, and afraid were identified as angry the same number of times that angry was identified

correctly. This suggests that angry emotion synthesized by GALE needs the most improvement.

The pattern of results for neutral, happy, sad, and afraid was similar with semantically neutral and non-neutral sentences. However, the results for angry were considerably different, suggesting that the angry emotion was the most affected by the semantic content of sentences.

CHAPTER 5
DISCUSSION

The emotion best synthesized by GALE using neutral sentences is sad. Angry was identified correctly the least number of times, suggesting that GALE's angry emotion needs the most improvement. A comparison between semantically neutral sentences (current study) and non-neutral sentences (previous study by Truluck, 1998) showed that the performance of GALE was similar for all emotions except for angry. This difference may be as a result of bias caused by anger evoked by the meaning of some sentences used by Truluck (1998). Overall, the results suggest that there is room for improvement in the synthesis of all emotions by GALE.

Improvements

A preliminary study must be conducted to determine if the emotion conveyed by the unchanged, neutral Paul voice is indeed neutral. In the current study, the neutral emotion was correctly identified 50% of the time. Though this value is above chance, it warrants examination because semantically neutral sentences, when presented with the neutral emotion of GALE, were also perceived as other emotions. For example, "Where is 2nd Street?" was very often perceived as conveying afraid (75%). "He is in a meeting." was perceived as happy a number of times (65%). An improvement to Phase I may be to identify those sentences that sound neutral when synthesized with GALE's neutral emotion, to be used as semantically neutral sentences in Phase II. Such sentences will likely better determine how well GALE simulates the emotions.

Results of both studies provide a baseline upon which better emotional speech synthesis systems may be developed. Improvements can be made on various fronts. Finding more accurate values for the parameters used may result in better simulation of emotions. Encoding other parameters may result in the simulation of nuances that may potentially improve the perception of an emotion. However, these may be limited by the speech synthesizer used.

Truluck (1998) suggested acclimating participants to synthetic speech before evaluation, to improve consistency of responses. But it may be argued that acclimation may confound results, as responses may not be based on comparisons to emotions of natural speech, but to those of synthetic speech. Comparing responses to samples of emotions in natural and synthesized speech would make for an interesting study.

Another issue to consider is the effect of adding emotion on the intelligibility and comprehensibility of synthetic speech. If this results in the deterioration of intelligibility and/or comprehensibility, further changes in methods of synthesizing emotions will need to be made.

Future Trends

Work on emotions in synthesized speech has been going on for over a decade. After achieving highly intelligible synthetic speech with intonation, stagnation seems to have set in. There is very little extension of this research to improve the naturalness of synthetic speech, most of which has been done in computer science and engineering labs.

Synthesizing emotionally rich speech is a challenge that needs to be addressed seriously. There is a need for a multi-disciplinary approach that includes consumers and professionals such as speech scientists, speech-language pathologists, linguists, computer scientists, and cognitive scientists. Psychologists may also be a part of the group,

contributing with insight into the relationship between emotions and the state of mind, and how it translates to vocal cues. Effectively synthesizing basic emotions such as happy, anger, sad and afraid will lay the foundation for the synthesis of more complex, but necessary aspects of natural speech, such as sarcasm. With the current state of technology, it does not seem to be a question of if, but when.

APPENDIX A
ACOUSTICAL CORRELATE VALUES OF EMOTIONS

Parameter (value change in)	Happy	Sad	Afraid	Anger
Average Pitch (%)	7	-9	40	10
Pitch range (%)	50	-20	100	50
Stress Rise (%)	50	-31	***	119
Quickness (%)	40	-50	100	100
Assertiveness (%)	***	-70	0	100
Baseline fall (Hz)	0	***	***	12
Hat Rise (%)	-10	-15	***	***
Loudness (dB)	2	-1	-1	14
Gain of Aspiration (dB)	***	***	8	***
Gain of Frication (dB)	***	-8	8	10
Nasalization (dB)	***	-8	***	***
Breathiness (dB)	15	***	***	10
Lax Breathiness (dB)	15	***	***	10
Smoothness (%)	25	-70	***	***
5th Formant Frequency (Hz)	Very high	***	200	***

*** No change from neutral value.

Source: Truluck (1998)

APPENDIX B
TRULUCK'S SENTENCES

1. I didn't know that he is in town.
2. They cut down the tree next door.
3. I saw him today after twenty years.
4. When I took the book from the bookcase, I realized the cover was green.
5. I made a 75 on my math test.
6. Aunt Mabel said I should stay home tonight.
7. Her car doesn't go any faster.
8. I can't believe you have 30 credit cards. Where are we going to go?
9. When I asked what we are going to do, Sam said it was up to me.
10. I know that I will see them again. It can't be any other way.
11. It is over forty miles to the nearest town.
12. I got another assignment from my boss.
13. I thought you said there weren't any left.
14. It won't be long before he comes home.

Source: Truluck (1998)

APPENDIX C
PHASE I SENTENCE DATA

Target emotion=> NEUTRAL					
	NEUTRAL	HAPPY	SAD	AFRAID	ANGRY
This is a pen.	37	0	0	0	3
Where is 2nd Street?	40	0	0	0	0
The mirror is on the floor.	33	1	1	2	3
The television is in the living room.	40	0	0	0	0
He is going to Detroit tomorrow.	37	1	1	1	0
The door is open.	37	1	0	2	0
It is a star filled night.	8	31	0	0	1
The balloons are filled with helium.	34	6	0	0	0
The mail came at 3:00 PM.	35	5	0	0	0
There is a basketball game today.	26	14	0	0	0
Can you type?	38	2	0	0	0
I'm going to the library.	38	2	0	0	0
He is in a meeting.	39	0	1	0	0
He went to school after lunch.	38	1	1	0	0
Where is the church?	37	2	0	1	0

TARGET EMOTION => HAPPY					
	NEUTRAL	HAPPY	SAD	AFRAID	ANGRY
I got a job.	1	0	37	0	2
I won the lotto.	2	38	0	0	0
It's my 21st birthday today.	2	37	1	0	0
I've lost 10 pounds in 30 days.	4	34	1	0	0
I'll be visiting my family over the Christmas break.	5	35	0	0	0
I have three cartons of my favorite ice cream.	1	38	1	0	0
It's our 25th anniversary.	2	38	0	0	0
It's a boy!	7	33	0	0	0
This is a perfect dress.	2	38	0	0	0
I got straight A's this year.	0	40	0	0	0
I have two tickets to the Super Bowl.	6	34	0	0	0
We had a blast at the party.	0	40	0	0	0
I made the basketball team.	1	39	0	0	0
They make a nice couple.	6	34	0	0	0
The cake is delicious.	3	37	0	0	0

TARGET EMOTION => SAD					
	NEUTRAL	HAPPY	SAD	AFRAID	ANGRY
I lost my wallet.	0	0	16	2	22
No one survived the tornado.	1	0	38	0	1
She never saw her family again	2	0	37	1	0
It's been raining all day.	17	4	18	0	0
He died in a plane crash.	2	0	37	0	1
He has no money to buy food.	3	0	36	0	1
He cannot walk any more.	0	0	39	1	0
They did not get any christmas presents.	2	0	36	1	1
I lost my wedding ring.	0	0	28	3	9
I was stuck in the traffic for over 2 hours.	2	0	0	1	37
His daughter did not visit him in 20 years.	5	0	31	0	4
Her thanksgiving dinner was a can of cold soup.	2	0	35	0	3
I cannot be with my mom on her birthday.	2	0	38	0	0
There is no Spring break this year.	5	1	14	0	20
I didn't get the scholarship.	1	0	33	1	5

TARGET EMOTION => AFRAID					
	NEUTRAL	HAPPY	SAD	AFRAID	ANGRY
I heard the door creak in the middle of the night.	2	0	0	37	1
It was pitch dark when I got there.	24	0	0	14	2
I saw a shadow in the bushes.	3	0	1	35	1
I was alone in the jungle.	5	1	3	30	1
This plane is going to crash.	1	0	3	36	0
We were going at 70 mph when the brakes failed.	3	1	5	29	2
I woke up with a snake by my side.	2	0	0	36	2
They are shooting everyone.	0	0	4	28	8
Please don't hurt me.	0	0	2	38	0
Our ship is sinking.	1	0	5	34	0
My car stopped on the tracks as the train was coming.	2	0	0	38	0
A shark appeared suddenly when I was swimming.	3	0	0	37	0
The bomb is going to go off any moment.	1	0	1	37	1
A shriek came out of the empty house.	2	1	0	36	1
Suddenly a lion appeared out of nowhere.	2	1	0	36	1

TARGET EMOTION => ANGRY					
	NEUTRAL	HAPPY	SAD	AFRAID	ANGRY
She is never satisfied with my work.	2	0	8	1	29
He made me do all the chores.	2	0	1	0	37
You made me late again.	1	0	0	2	37
This is the 4th time the bus is late.	3	0	0	0	37
He hurt my little brother.	0	0	5	1	34
My neighbor empties his trash in front of my house.	1	1	1	0	37
His dog dirtied my carpet.	2	1	1	0	36
My cousin called me names.	3	0	15	2	20
This is the last time I want to see you.	2	0	10	0	28
You're fired!!	4	0	9	1	26
I'll skin you alive.	2	0	0	10	28
Get out of my house!!	2	1	0	2	35
You drive me crazy!	4	6	0	1	29
My blood boils at the thought.	1	0	0	4	35
This program is driving me crazy!	3	0	0	1	36

APPENDIX D
INFORMATION SHEET

_____ #

Please do not write above this line.

Please complete the following items. All information will be held in confidence.

Name: _____ Age: _____ Sex: M F

Address: _____

Tel#: _____

Email: _____

Native Language: _____ Other languages:_____

Major: _____ Grad Undergrad

Knowledge of speech synthesis: Expert good fair none at all.

Did you fail a hearing test recently? Yes No

Please do not write below this line.

Notes:

_____

_____

_____

_____

APPENDIX E
PHASE I INSTRUCTIONS

"Read each sentence from the list below. A choice of emotions (neutral, happy, sad, fear and, anger) is given for each sentence. Indicate with 'X' or '√' the strongest emotion that you think is conveyed by each sentence. You may go back and revise a previously made choice. There is no time limit. If you have any questions, you may ask them before you begin with the sentences. A sample is given below."

	Neutral	Happy	Sad	fear	anger
I am having great fun.		X			
He lost his father				X	

APPENDIX F
PHASE II INSTRUCTIONS

PLEASE READ THE INSTRUCTIONS CAREFULLY.

---

1) The goal of the study is to determine the emotion you hear in each of the 60 sentences that will be presented.
2) The sentences are in English and are of varied length. They are emotionally neutral. That is, they do not have emotions in their meaning.
3) There is a gap of 4 seconds between each sentence. In this time, you must determine the emotion in the sentence and record it on the given recording sheet.
4) On the recording sheet, you can indicate ('X' or '✓') one of 5 emotions ('Neutral', 'Happy', 'Sad', 'Afraid' and 'Angry') for each sentence. If you think that the emotion in the sentence is not one of the 5, then you may record the emotion you hear, in the 'Other' column.
5) The duration of stimuli presentation is approximately 6 minutes.
6) Before we start, you will listen to 2 sample sentences along with the correct response for each sentence.

PLEASE ASK ANY QUESTIONS YOU MAY HAVE

ABOUT THE STUDY AT THIS TIME.

APPENDIX G
PHASE II ANSWER SHEET (PARTIAL)

Sno	NEUTRAL	HAPPY	SAD	AFRAID	ANGRY	OTHER
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						

APPENDIX H
PHASE II DATA

Target emotion -> NEUTRAL							
sentence	**Neutral**	**Happy**	**Sad**	**Afraid**	**Angry**	**Other**	**total**
He went to school after lunch.	8	4	1	3	4	0	20
He is going to Detroit tomorrow.	4	2	1	3	10	0	20
The television is in the living room.	12	4	1	1	2	0	20
Where is the church?	15	0	1	2	2	0	20
Where is 2nd Street?	7	3	1	3	6	0	20
The door is open.	15	0	3	1	1	0	20
He is in a meeting.	10	3	2	0	5	0	20
Im going to the library.	8	5	1	2	4	0	20
This is my pen.	14	3	0	2	1	0	20
Can you type?	8	2	0	3	7	0	20
Total	101	26	11	20	42	0	200
Average	**50.5**	13	5.5	10	21	0	100

Target Emotion -> HAPPY							
sentence	**Neutral**	**Happy**	**Sad**	**Afraid**	**Angry**	**Other**	**total**
He went to school after lunch.	7	8	0	1	4	0	20
He is going to Detroit tomorrow.	5	10	2	0	3	0	20
The television is in the living room.	6	6	0	2	6	0	20
Where is the church?	10	2	1	4	2	1	20
Where is 2nd Street?	3	3	0	6	7	1	20
The door is open.	14	3	0	0	3	0	20
He is in a meeting.	14	4	0	0	2	0	20
Im going to the library.	4	10	0	1	5	0	20
This is my pen.	9	7	1	0	3	0	20
Can you type?	7	0	7	2	4	0	20
Total	79	53	11	16	39	2	200
Average	39.5	**26.5**	5.5	8	19.5	1	100

Target Emotion -> SAD							
sentence	**Neutral**	**Happy**	**Sad**	**Afraid**	**Angry**	**Other**	**total**
He went to school after lunch.	2	0	14	1	2	1	20
He is going to Detroit tomorrow.	2	0	15	0	2	1	20
The television is in the living room.	0	0	17	0	1	2	20
Where is the church?	1	0	13	0	6	0	20
Where is 2nd Street?	3	0	15	1	0	1	20
The door is open.	1	1	17	0	1	0	20
He is in a meeting.	1	0	17	0	2	0	20
Im going to the library.	1	0	19	0	0	0	20
This is my pen.	2	0	17	0	0	1	20
Can you type?	6	1	8	1	4	0	20
Total	19	2	152	3	18	6	200
Average	9.5	1	**76**	1.5	9	3	100

Target Emotion -> AFRAID							
sentence	**Neutral**	**Happy**	**Sad**	**Afraid**	**Angry**	**Other**	**total**
He went to school after lunch.	0	1	5	11	3	0	20
He is going to Detroit tomorrow.	3	1	7	5	2	2	20
Where is the church?	2	2	5	8	3	0	20
Where is 2nd Street?	1	0	2	17	0	0	20
The door is open.	0	5	1	7	7	0	20
He is in a meeting.	4	5	0	5	4	2	20
Im going to the library.	1	3	2	3	9	2	20
This is my pen.	0	7	2	6	5	0	20
Can you type?	2	2	1	11	4	0	20
Total	13	26	25	73	37	6	180
Average	7.222	14.4	14	**40.6**	20.6	3.33	100

Target Emotion -> ANGER							
sentence	**Neutral**	**Happy**	**Sad**	**Afraid**	**Anger**	**Other**	**total**
He went to school after lunch.	2	5	3	2	8	0	20
He is going to Detroit tomorrow.	3	9	3	2	3	0	20
The television is in the living room.	2	6	2	1	6	3	20
Where is the church?	8	2	4	5	1	0	20
Where is the 2nd street?	0	1	3	15	1	0	20
The door is open.	1	12	0	1	6	0	20
He is in a meeting.	2	13	2	0	3	0	20
Im going to the library.	1	11	1	4	3	0	20
This is my pen.	3	8	2	0	7	0	20
Can you type?	3	10	3	2	2	0	20
Total	25	77	23	32	40	3	200
Average	12.5	38.5	12	16	**20**	1.5	100

APPENDIX I
TRULUCK'S SENTENCE ANALYSIS

Rank of **NEUTRAL** sentence	Sentence number (Appendix B)	% times sentence determined neural	% times determined neutral with GALE
1	4	95	35
2	9	80	50
3	1	55	55
Rank of **HAPPY** sentence	Sentence number (Appendix B)	% times sentence determined happy	% times determined happy with GALE
1	3	95	50
2	14	80	25
3	8	40	45
Rank of **SAD** sentence	Sentence number (Appendix B)	% times sentence determined sad	% times determined sad with GALE
1	5	75	85
2	2	50	80
3	12	30	80
Rank of **AFRAID** sentence	Sentence number (Appendix B)	% times sentence determined afraid	% times determined afraid with GALE
1	10	30	65
2	11	15	25
3	8	15	45

LIST OF REFERENCES

Allen, J., Hunnicutt, M. S., & Klatt, D. H. (1987). From text to speech: The MITalk system. New York: Cambridge University Press.

Cahn, J. E. (1990). The generation of affect in synthesized speech. Journal of the American Voice I/O Society, Volume 8, 1-19.

Coker, C. H. (1967). Synthesis by rule from articulatory parameters. Speech Communication Process, 52-53.

Davitz, J. (1964). The communication of emotional meaning. New York: McGraw Hill.

Donovan, R. (1996). Trainable speech synthesis. Ph.D. Thesis. Cambridge University Engineering Department, England. ftp://svr-ftp.eng.cam.ac.uk/pub/reports/donovan_thesis.ps.Z. 11/2002.

Duffy, S. A., & Pisoni, D.B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. Language and Speech, Volume 35, 351-389.

Greene, B.G., Logan, J.S., & Pisoni, D.B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. Behavior Research Methods, Instruments and Computers, 18, 100-107.

Hill D. R., Leonard M., & Schock, C., (1995). Real-time articulatory speech synthesis-by-rules. Proceedings of AVIOS '95, the 14th Annual International Voice Technologies Applications Conference of the American Voice I/O Society, Volume 11, Number 14, 27-44.

Holmes, J. N. (1983). Formant synthesizers: cascade or parallel. Speech Communication, Volume 2, 251-274.

Kent, R. D. & Read, C. (1992). The acoustic analysis of speech. San Diego: Singular Publishing Group, Inc.

Kent, R. D. & Read, C. (2002). The acoustic analysis of speech, San Diego: Singular Publishing Group, Inc.

Klatt, D. (1980). Software for the cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, Volume 67, 971-975.

Klatt, D. (1987). Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, Volume 82, 737-793.

Mirenda, P., & Beukelman, D. R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. Augmentative Alternative Communication, Volume 5, 84-88.

Mirenda, P., & Beukelman, D. R. (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. Augmentative Alternative Communication, Volume 6, 61-68.

Murray, R. L., & Arnott, J. (1993). Towards the simulation of emotions in synthetic speech. A review of literature on human vocal emotion. Journal of the Acoustical Society of America, Volume 93, 1097-1108.

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

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. Behavior Research Methods, Instruments and Computers, Volume 17, 235-242.

O'Shaughnessy, D., Barbeau, L., Bernardi, D., & Archambault, D. (1988). Diphone speech synthesis. Speech Communication, Volume 7, 55-65.

Reynolds, M.E., & Jefferson, L. (1999). Natural and synthetic speech comprehension: A comparison of children with normal and impaired language skills. Augmentative Alternative Communication, Volume 15, 174-182.

Reynolds, M. E., Isaacs-Duvall, C., Sheward, B., & Rotter, M. (2000). Examination of the effects of listening practice on synthesized speech comprehension. Augmentative Alternative Communication, Volume 15, 174-182.

Reyolds, M. E., Isaacs-Duvall, C., & Haddox, M., L., (2002). A comparison of learning curves in natural and synthesized speech comprehension. Journal of Speech, Language, and Hearing Research, Volume 45, Number 4, 802-810.

Rupprecht, S. L., Beukelman D. R., & Vrtiska, H. (1995). Comparative intelligibility of five synthesized voices. Augmentative Alternative Communication, Volume 11, 244-247.

Scherer, K. R. (1981). Speech and Emotional States. Speech evaluation in psychiatry education, John K. Darby (Ed.), New York: Grune & Stratton, Inc., 189-220.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. Motivation and Emotion, Volume 15, Number 2, 123-148.

Truluck, D'Arcy K. (1998). GALE: Emotion in synthetic speech. Master's thesis, University of Florida, Gainesville, Florida.

Venkatagiri, H. S., & Ramabadran, T. V. (1995). Digital speech synthesis: Tutorial. Augmentative Alternative Communication, Volume 11, 14-25.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: some acoustical correlates. Journal of the Acoustical Society of America, Volume 52, 1238-1250.

BIOGRAPHICAL SKETCH

Vuday Nandur was born on 12 July 1976 in Amalapuram, a small costal town in Andhra Pradesh, India. Most of his schooling was in Bharatiya Viday Bhavan's public school, Hyderabad, India. Vuday procured a bachelor's degree in speech and hearing from the Ali Yavar Jung National Institute for the Hearing Handicapped, Southern Regional Center, Hyderabad, India, in 1998.

In May 2003, Vuday will be receiving a master's degree in computer science along with one in speech language pathology and plans to pursue a Ph.D. in simulating affect in synthetic speech using the articulatory model.