SYNTHESIZING STYLED SPEECH USING THE KLATT SYNTHESIZER

Janet C. Rutledge[1], *Kathleen E. Cummings*[2], *Daniel A. Lambert*[2], *and Mark A. Clements*[2]

[1]School of EECS, Northwestern University, Evanston, IL 60208-3118
[2]School of ECE, Georgia Institute of Technology, Atlanta, GA 30332

ABSTRACT

This paper reports the implementation of high-quality synthesis of speech with varying speaking styles using the Klatt synthesizer. This research is based on previously-reported research that determined that the glottal waveforms of various styles of speech are significantly and identifiably different. Given the parameter tracks that control the synthesis of a *normal* version of an utterance, those parameters that control known acoustic correlates of speaking style are varied appropriately, relative to *normal*, to synthesize styled speech. In addition to varying the parameters that control the glottal waveshape, phoneme duration, phoneme intensity, and pitch contour are also varied appropriately. Listening tests that demonstrate that the synthetic speech is perceptibly and appropriately styled, and that the synthetic speech is natural-sounding, were performed, and the results are presented in this paper.

1. INTRODUCTION

Synthesizing high quality natural-sounding synthetic speech has been an important topic of research for many years. Applications such as speech aids for the handicapped, teaching aids, and automated computer-generated instructions would benefit from achieving natural-sounding speech synthesis. Since part of what is perceived as naturalness in speech is the perceived emotional content, one way to improve the perceived naturalness of synthetic speech would be to improve the ability of the synthesizer to convey various emotions and speaking styles. The objective of this research was to use an existing high-quality speech synthesizer, the Klatt synthesizer, along with the results of extensive research into the changes in the glottal excitation that occur with emotional stress, to synthesize speech with a variety of emotions and speaking styles. The styles of speech analyzed and synthesized in this research were *normal*, *angry*, *50% tasking*, *70% tasking*, *clear*, *fast*, *Lombard*, *loud*, *question*, *slow*, and *soft*.

In previously reported research [1] [2] [3], glottal waveforms were extracted from eleven commonly-encountered speech styles and analyzed. Based on six shape parameters, these eleven styles of glottal waveform were analyzed statistically, and it was shown that each of the eleven styles is both significantly different in a statistical sense and uniquely identifiable. A four parameter glottal model based on the normalized beta function, which is able to accurately model all of the salient differences in the eleven styles of glottal waveform, was developed. Finally, several applications, including speaking style modification, were developed using this new glottal model. The speaking style modification algorithms developed were simple implementations based on either Fast Fourier Transform (FFT) or Linear Predictive (LP) analysis algorithms in which the appropriately styled model glottal excitation was removed and the remaining signal or a model of the remaining signal was re-excited with a different styled model glottal excitation. Additionally, such parameters as intensity, duration, and pitch contours were varied appropriately. In this way, *normal* speech was made to sound styled and styled speech to sound *normal*.

Although listening tests demonstrated that the speech modification was indeed perceptibly and appropriately styled, the resulting modified speech suffered from artifacts that are typically associated with LP synthetic speech. In the research reported in this paper, the previous research has been extended to implement speaking style modification using an existing, high-quality speech synthesizer, KLSYN88A [7] [8] [9]. This paper will describe the implementation of speaking style modification, based on the previously-reported knowledge of variances in the glottal waveshape and other acoustic waveform parameters, using the Klatt syntheszier. Listening tests that demonstrate that *normal* speech is accurately modified to sound like each of the other ten speech styles, and that the modified speech is natural-sounding will be presented.

2. THEORY

In the standard source-filter theory of speech production [5], speech is assumed to be the result of convolving an excitation source, either quasi-periodic impulses, turbulent noise, or a combination of the two, with filters representing glottal shaping (for voiced speech), the vocal tract impulse response, and the acoustic impedance representing radiation at the lips. For voiced speech, this can be represented in terms of z-transforms as

$$S(z) = E(z)G(z)V(z)R(z), \quad (1)$$

[0]This material is based in part upon work supported under National Science Foundation Grant # CDA-9211129.

[0]The authors would like to thank Kelli Long, at NU, for her assistance in running the listening tests.

where,

$S(z)$	–	the z-transform of the speech segment
$E(z)$	–	the z-transform of the impulse train
$G(z)$	–	the z-transform of the glottal waveform
$V(z)$	–	the z-transform of the vocal tract impulse response
$R(z)$	–	the z-transform representation of the acoustic impedance at the lips.

Formant synthesis is based directly on the source-filter theory of speech production and the short-time stationary properties of speech. Assuming that the excitation source is separable from the vocal tract resonant system, the short-time spectrum of speech can be modelled with a small number of poles and zeros representing the resonances, or formants, of the vocal tract. In formant synthesis, an excitation source is convolved with the impulse response of the vocal tract transfer function and with the impulse response of the radiation characteristics of the lips and/or nose to generate speech.

The Klatt synthesizer is a high-quality cascade/parallel formant synthesizer [7]. Speech is synthesized using a set of 60 parameters, 12 of which are fixed for the synthesis of the phoneme and 48 of which can be varied during the synthesis of the phoneme. The synthesizer offers several options, including three different voicing sources and the choice between a cascade and a parallel branch. The details of using the Klatt synthesizer in this research to generate styled speech will be described more fully in the next section of this paper.

3. IMPLEMENTATION OF SPEAKING STYLE MODIFICATION

In previously reported research [3] [4], the application of speech modification using simple linear predictive (LP) or FFT-based speech synthesis was described. All of these algorithms were based on decoupling the vocal tract and glottal excitation signals, and re-exciting the vocal tract signal or a model of the vocal tract signal with a different style of excitation. The decoupling and re-exciting of the vocal tract signal was accomplished using a new glottal model, based on the beta function, that is able to represent all eleven styles of glottal excitation accurately with four parameters.

In addition to varying the glottal excitation, phoneme duration, phoneme intensity, and pitch contour were also varied. The application of this speech modification was to change styled speech such that it was perceived as being *normal*, or neutral, speech. Subjective listening tests were performed to determine whether the modified speech was more neutral than the original utterance. For the styles of speech that were perceptually different from *normal* (*question*, *loud*, *soft*, *slow*, *fast*, *angry* and *Lombard*), listeners selected the modified utterance as the more neutral at least 90% of the time.

While these results were very promising, the LP- and FFT-based algorithms are not capable of synthesizing high-quality speech. The objective of the research that is reported in the remainder of this paper was to use the same basic concept, i.e., parametric modification of speaking style, with a synthesis method that is capable of producing high-quality speech. This is not a modification method in which the synthetic speech parameters are directly matched to parameters for a given style of speech. Rather, this method changes the parameters from the average value for the original speech to the average value for the new style of speech. This method is thus much more useful. For this research, formant synthesis was selected as the base synthesis method. Again, the overall speech modification scheme is straightforward. Given a set of control values that synthesize a *normal* version of an utterance, vary some subset of these control values in a manner appropriate for a new speaking style. Among the parameters that are varied, relative to the *normal* values, are those that control the glottal waveshape, phoneme intensity, phoneme duration, and pitch contour.

Specifically, KLSYN88A, a version of the Klatt synthesizer was chosen. The Klatt synthesizer is capable of generating very high quality synthetic speech and is completely controlled by a set of 60 parameters which are usually updated every 5 msec, making parametric speech modification reasonably straightforward. Another advantage of the Klatt synthesizer is that it allows the user to select one of three sources: a low-pass filtered impulse, a modified version of the Rosenberg parametric glottal volume velocity pulse model, and a modified version of the Liljencrants-Fant (L-F) glottal volume velocity derivative model. The last of these, the L-F glottal model, has enough parameters to represent the variances that occur in the glottal excitation of styled speech (although not as accurately as the beta function model that was specifically developed to represent styled glottal excitation).

Twelve of the 60 parameters are constant during the generation of a given phoneme, while 48 are variable. The twelve constant parameters include such values as overall phoneme duration and gain factors, parameter update rate, number of formants, type of source, etc. The remaining 48 parameters control such values as fundamental frequency, glottal waveshape, frequency and bandwidth of the poles and zeros of the vocal tract function, etc. For a more complete description of the Klatt synthesizer, the reader is referred to [7] [8] [9].

A natural utterance of the word "hot" in the *normal* style was analyzed and then synthesized using the Klatt synthesizer. Given the synthesis parameter tracks for the *normal* version of "hot", those parameters that correspond to known acoustic correlates of style were varied appropriately to synthesize the ten other styles of "hot." All of the changes made to the parameter tracks were done using the results of the statistical analyses of the glottal waveform and speech waveform variances previously reported [4] [6].

The modified L-F (Liljencrants-Fant) model of the derivative of the glottal volume velocity pulse was used as the excitation source in this research. In the KLSYN88A, the shape of the L-F model is controlled with five parameters: F0, the fundamental frequency of voicing, AV, the peak amplitude of the glottal pulse, OQ, the ratio of the open-glottis time to the total period duration, SQ, the ratio of the durations of the rising to the falling portions of the glottal open phase, and TL, an additional spectral change

	fund freq (F0) (Hz)	amp voicing (AV)	speed quo (OQ)	open quo (SQ)
normal	140.4	9600	154	69
angry	266.7	12700	110	70
clear	150.9	9800	170	70
50% tasking	140.4	8860	145	70
70% tasking	145.5	9000	147	70
fast	150.9	9400	145	71
Lombard	163.3	9700	164	69
loud	250.0	12000	244	80
question	205.1	10200	159	64
slow	142.9	9700	152	70
soft	135.6	9300	127	70

Table 1. L-F glottal model parameters for the eleven styles of speech

	Clsing Slope	Opning Slope	Clsd Dur	Clsing Dur	Opning Dur	Top Dur
angry	-9910	9198	9.1	6.3	6.9	2.0
50%	-4522	2321	17.3	11.1	16.0	9.8
clear	-5011	2686	15.8	9.5	16.0	6.9
70%	-4100	2138	16.7	10.7	15.7	9.9
fast	-3972	2376	15.5	11.0	16.0	8.4
loud	-9298	3532	6.3	6.9	17.0	2.9
Lom	-5430	2871	15.2	9.3	15.2	7.6
norm	-4798	2643	17.7	10.2	15.6	9.9
ques	-4831	3034	14.0	9.4	14.9	7.0
slow	-4786	2692	16.9	10.2	15.5	8.7
soft	-2632	1921	17.7	14.7	18.6	9.9

Table 2. Mean values of the six glottal waveshape parameters for each style (durations are in samples; slopes are in (normalized amplitude)/(samples)

associated with "corner-rounding."

The L-F model parameters for each of the eleven styles are shown in Table1. The values for these parameters were determined using the Gaussian means of the extracted glottal waveshape parameters for Speaker One, shown in Table 2 [2] [4].

In addition to exciting the vocal tract resonant structure with different excitation sources, four other parameter tracks were varied according to known variances for styled speech as opposed to *normal* speech. These four parameters are vowel duration, consonants duration, word duration, and consonant intensity.

The variances of each parameter track for each style are shown, relative to *normal*, in Tables 3 and 4. Table 3 shows the variances in the L-F model excitation parameters, while Table 4 shows the variances in the overall acoustic waveform parameters. Additionally, pitch contours that were representative of the expected shape of the pitch contour for each style were used.

4. RESULTS

The word "hot" was first synthesized in the *normal* style, using analysis-by-synthesis and a natural utterance. "Hot" was then synthesized in each of the remaining ten styles by varying the appropriate parameter tracks according to the average deviation from *normal* for each style, as shown in Tables 3 and 4. The natural utterance and all of the synthetic utterances were sampled at 8 kHz.

Qualitatively, the results of synthesizing the 10 styled versions of the word "hot" by varying the *normal* parameter tracks were very good. All eleven styles were successfully synthesized. All of the styles except *angry* were perceptually natural-sounding. It is believed that *angry* has significant energy at frequencies above 4 kHz.

Listening tests were performed in which twenty untrained listeners were presented with eleven utterances, one in each of the eleven styles, of the word "hot" in random order. Listeners sat at a SUN workstation wearing headphones. They were given a list and a description of the eleven styles and were asked to appropriately identify each of the eleven unknown utterances which they heard. They were instructed to use each style only once. The listener was allowed to listen to the various utterances as often as they chose. On average, it took between five and seven minutes to complete the identification of each of the eleven styles. Each listener performed this test twice: once with synthetic speech and once with natural, human speech. This is a difficult task for the listener. He/she is not ask to select one of a pair of words that is "more" styled. He is asked to identify the style of a single, unknown word. Furthermore, it is difficult to accomplish this identification with a single word rather than a multi-word utterance.

The results from these listening tests are shown in Table 5 for both synthetic and natural speech. The results showed that listeners were able to correctly identify very different styles. Listeners generally made the same sort of errors for the synthetic speech that were made for the original speech. *Clear*, *50% tasking*, and *70% tasking* were confusable for both natural and synthetic speech. In fact, most of the errors made for both natural and synthetic speech involved *50% tasking* and *70% tasking*. *Loud* and *Lombard* were confused for each other for both natural and synthetic speech. Recall that *Lombard* speech tends to be a combination of *loud*, *clear* speech. There were three notable differences made in identifying natural and synthetic speech. *Fast* and *slow* were much more identifiable for synthetic speech, while *soft* speech was much more identifiable for natural speech. Several listeners volunteered the information that they could not tell the difference between the human and the synthetic speech. Other comments included several observations that *angry* sounded more like "shocked" or "unpleasantly surprised."

5. CONCLUSION

This paper has presented a method for synthesizing high-quality, styled speech. The method is based on previous knowledge of the manner in which the glottal waveform and other acoustic speech waveform parameters vary across eleven speech styles. Using the Klatt synthesizer, the parameter tracks are determined for the synthesis of a *normal*

	F0	AV	OQ	SQ	TL
norm	1.00	1.00	1.00	1.00	1.00
angry	1.90	1.32	1.01	0.71	0.85
50%	1.00	0.92	1.01	0.94	1.05
70%	1.04	0.94	1.01	0.95	1.05
clear	1.07	1.02	1.01	1.10	1.05
fast	1.07	0.98	1.03	0.94	0.95
loud	1.78	1.25	1.16	1.58	0.75
Lom	1.16	1.01	1.00	1.06	1.05
que	1.46	1.06	0.93	1.03	1.00
slow	1.02	1.01	1.01	0.99	1.00
soft	0.97	0.97	1.01	0.82	1.18

Table 3. Excitation parameter track values, relative to *normal*, that are varied to synthesize styled speech.

	Vowel Dur	Word Dur	Cons Dur	Cons Intensity
norm	1.00	1.00	1.00	1.00
angry	1.69	1.38	0.87	1.12
50%	0.92	1.01	1.11	1.14
70%	0.92	1.05	1.21	1.25
clear	1.26	1.39	1.80	0.92
fast	0.72	0.74	0.73	1.02
loud	1.58	1.36	1.03	0.84
Lom	1.24	1.20	1.03	1.01
que	1.13	1.10	1.04	1.15
slow	1.83	1.73	1.50	1.06
soft	0.92	1.06	1.23	1.34

Table 4. Acoustic waveform parameter track values, relative to *normal*, that are varied to synthesize styled speech.

utterance. The parameter tracks that control the glottal waveshape, the pitch contour, the vowel duration, the word duration, the consonant duration, and the consonant intensity are then changed, relative to *normal*, such that the mean value is appropriate for the new style. This method does not depend on matching the parameters track directly to a natural styled utterance; rather, given one style of speech, any of the ten other styles can be automatically synthesized. A natural utterance of the word "hot," spoken in the *normal* style, was analyzed and synthesized using the Klatt synthesizer. By varying the appropriate parameter tracks for the *normal* synthetic version of "hot," ten other styles were synthesized. Listening tests demonstrated that listeners were able to correctly identify the synthetic utterances as accurately as the natural utterances. Those styles of speech that are very different from the other styles of speech were easily identified for both natural and synthetic utterances. Certain styles of speech that are perceptually very similar were confusable for both natural and synthetic speech.

	Synthetic Speech									
	a	50	70	c	f	l	Lm	q	sl	sf
ang	9	2	2	0	0	2	3	0	0	1
50%	1	3	4	4	0	3	1	0	3	0
70%	0	6	5	1	0	1	2	0	0	4
cle	0	2	2	9	0	0	3	0	0	3
fas	0	0	0	0	19	0	0	0	0	0
lou	4	0	0	0	0	10	5	0	0	0
Lom	3	3	2	2	0	3	5	0	1	0
que	0	0	0	0	0	0	0	19	0	0
slo	0	0	0	1	0	0	0	0	18	0
sof	2	3	4	2	0	0	0	0	0	8

	Natural Speech									
	a	50	70	c	f	l	Lm	q	sl	sf
ang	8	0	0	0	1	6	4	0	0	0
50%	0	6	5	4	2	0	1	0	0	1
70%	1	11	5	1	0	0	0	0	1	0
cle	0	1	4	12	0	0	1	0	1	0
fas	0	0	3	0	15	0	0	0	1	0
lou	2	0	0	0	0	9	8	0	0	0
Lom	8	0	0	1	0	4	4	0	2	0
que	0	0	0	0	0	0	0	19	0	0
slo	0	1	1	1	1	0	1	0	14	0
sof	0	0	1	0	0	0	0	0	0	18

Table 5. Listening test results. Vertical is actual style; horizontal is listener selection.

REFERENCES

[1] Cummings, K. E. and Clements, M. A. "Analysis of Glottal Waveforms Across Stress Styles," *Proceedings, IEEE ICASSP*, 1990, pp 369-372.

[2] Cummings, K. E. and Clements, M. A. "Improvements to and Applications of Analysis of Stressed Speech Using Glottal Waveforms," *Proceedings, IEEE ICASSP*, 1992, vol 2, pp 25-28.

[3] Cummings, K. E. and Clements, M. A., "Application of the Analysis of the Glottal Excitation of Stressed Speech to Speaking Style Modification," *Proceedings, IEEE ICASSP*, 1993, vol 2, pp 207-210.

[4] Cummings, K. E., *Analysis, Synthesis, and Recognition of Stressed Speech* , Ph. D. Dissertation, Georgia Institute of Technology, September, 1992.

[5] Fant, G., *Acoustic Theory of Speech Production*, Mouton: The Hague, 1970.

[6] Hansen, J. H. L., *Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition*, Ph. D. Dissertation, Georgia Institute of Technology, July, 1988.

[7] Klatt, D. H., "Software for a Cascade/Parallel Formant Synthesizer," *Journal of the Acoustical Society of America*, vol 67 no 3, pp 971-994, March, 1980.

[8] Klatt, D. H., "The Klattalk Text-to-Speech Conversion System," *Proceedings, IEEE ICASSP*, 1982, pp 1589-1592, May, 1982.

[9] Klatt, D. H. and Klatt, L. C., "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers," *Journal of the Acoustical Society of America*, vol 87 no 2 pp 820-857. February, 1990.