DEVELOPMENT OF RULES FOR CONTROLLING THE HLSYN SPEECH SYNTHESIZER

Helen M. Hanson, Richard S. McGowan, Kenneth N. Stevens, and Robert E. Beaudoin*

Sensimetrics Corp.
48 Grove St.
Somerville, MA 02144
hanson@sens.com

ABSTRACT

In this paper we describe the development of rules to drive a quasi-articulatory speech synthesizer, HLsyn. HLsyn has 13 parameters, which are mapped to the parameters of a formant synthesizer. Its small number of parameters combined with the computational simplicity of a formant synthesizer make it a good basis for a text-to-speech system. An overview of the rule-driven system, called VHLsyn, is presented. The system assumes a phonetic string as input, and produces HLsyn parameter tracks as output. These parameter tracks are then used by HLsyn to produce synthesized speech. Recent work to improve the synthesis of consonants and suprasegmental effects is described, and is shown to improve the quality of the output speech. The improvements include refinement of release characteristics of stop consonants, methods for control of vocal-fold parameters for voiced and voiceless consonants, and rules for timing and intonation.

1. INTRODUCTION

This paper is an update on the development of rules to control the HLsyn speech-synthesis system. HLsyn is a system in which a small number of both articulatory and acoustic parameters are mapped to the acoustic parameters of a Klatt formant synthesizer [2, 5]. The 13 HLsyn parameters are described in Table 1. The mapping of these 13 parameters to the many Klatt parameters is based in part on a circuit model of the aerodynamics of the speech production system, shown in Fig. 1. With this model, intermediate parameters of intraoral pressure and flows through the glottal and supraglottal constrictions are calculated. These pressures and flows, along with the constriction-size parameters, are then used to calculate Klatt parameters related to sound sources and transfer functions. Adjustments are made to the first-formant frequency to reflect changes in supraglottal constrictions, and to the **f0** track to reflect intrinsic pitch and changes in subglottal pressure and vocal-fold compliance.

One benefit of using HLsyn over a traditional formant synthesizer is that, by taking advantage of the constraints among the many Klatt parameters, the number of parameters that must be controlled directly is reduced. Perhaps more important is that control of the synthesizer with the HLsyn parameters in Table 1 maps the natural control of human speech production, and the synthesizer output is constrained to have natural speechlike properties.

*K.N. Stevens is also at the Research Laboratory of Electronics and the Dept. of EECS at M.I.T., Cambridge MA.

This work was supported in part by NIH SBIR grant MH52358

Table 1: Description of HLsyn parameters

f1–f4	First four natural frequencies of vocal tract, assuming no local constrictions
f0	Fundamental frequency due to active adjustments of vocal folds
ag	Average area of glottal opening between the membranous portion of the vocal folds
ap	Area of the posterior glottal opening
ps	Subglottal pressure
al	Cross-sectional area of constriction at the lips
ab	Cross-sectional area of tongue blade constriction
an	Cross-sectional area of velopharyngeal port
ue	Rate of increase of vocal-tract volume
dc	change in vocal-fold or wall compliances

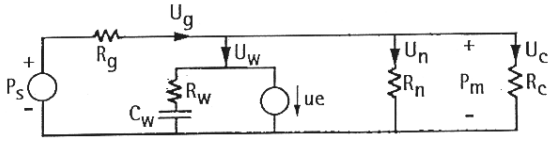

Figure 1: Low-frequency equivalent circuit used by HLsyn to calculate the intraoral pressure P_m in the vocal tract, and the flows through the glottis, nasal cavity, and supraglottal constriction (U_g, U_n, and U_c respectively), and the flow U_w due to displacement of the vocal-tract walls.

We are now writing rules that generate appropriate HLsyn parameters, given a phonetic input string. The system is referred to as VHLsyn (very high-level synthesis). A previous paper [7] described rules that control formant-track parameters, and primary and secondary articulatory parameters. In this paper, we describe new rules that improve the quality of the output speech at both the segmental and suprasegmental levels. We begin by presenting an overview of the system and reviewing the earlier rules. Next we describe the results of diagnostic listening tests, from which we determined specific areas that required improvement. New segmental rules that improve the quality of consonants, particularly the details of closure and release, are then described. Finally, we present some suprasegmental rules which further contribute to the improved quality of the output speech.

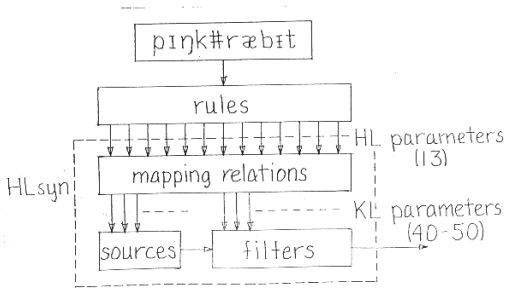

Figure 2: Schematic of the VHLsyn rule-based synthesis system, showing a phonetic input string at the top.

2. SYSTEM OVERVIEW AND PREVIOUS WORK

A schematic of VHLsyn is shown in Fig. 2. The input to the system is a phonetic string made up of phones, where a phone is defined as a speech sound that could occur in American English, including allophones. An initial step in the rules is to "parse" the phones into a sequence of landmarks. The landmarks are intended to be the times when the relevant articulators make their closest approach to certain prescribed targets. A landmark can be a release or closure of a consonant, the nucleus, and possibly offglide, of a vowel, or the nucleus of a glide. The timing of the landmarks is based on both segmental and suprasegmental factors. Table 2 shows the results of converting a phonetic string into a string of landmarks. The set of landmarks is the basis of the HLsyn parameter generation. Once the HLsyn parameters are derived, the HLsyn program maps them to Klatt (KL) parameters. The KL parameters are then input to a formant synthesizer which produces the output speech.

Table 2: Landmarks generated from the phonetic string "ax / dd ey - zz iy" ("a daisy")

Time (ms)	Type	Phonetic Segment
30	nucleus	ax
60	closure	dd
65	offglide	ax
110	release	dd
190	nucleus	ey
260	closure	zz
265	offglide	ey
310	release	zz
370	nucleus	iy
435	offglide	iy

In this paper we are focusing on the part of the system that derives the HLsyn parameters. To generate formant tracks **f1**–**f4** from landmarks, several operations are performed on the data. Approximations to HL formant parameters are first generated as if only vowels occurred in the utterance. The formant values at the vowel nuclei and offglides are set to target values from stored tables. These values are then connected using half sinewaves. The formant tracks are then adjusted to take glide landmarks into account. Finally, near the closures and releases of obstruents, the formant parameters are adjusted to rise or fall as appropriate for the place of articulation. The parameter **ag**, average area of the glottis at the membranous portion, is set at its modal value during voiced sounds, and is increased during unvoiced sounds. Similarly, parameter **an** is increased for nasal sounds. The other constriction-area parameters, **al** and **ab**, are decreased for labial and alveolar consonants, respectively

The system as described in [7] had rules for the parameters **f1**–**f4**, **ag**, **al**, **ab**, and **an**. The parameter **f0** was set by hand, and the remaining parameters were set to constant values. The landmark timing was also set by hand. As we later describe, in the current version of the system **f0** and landmark timing are derived by rule, and the parameters **ue**, **ap**, and **dc** have been incorporated.

Although the first set of rules resulted in intelligible speech, the quality of the segments required adjustment. In order to determine intelligibility of the consonants, a listening test was performed with ten subjects. One hundred monosyllables were chosen from the Harvard word lists as stimuli, and were synthesized using VHLsyn. The stimuli were randomized, and each was repeated four times. The test was open set, that is, the subjects were instructed to write down whatever word they heard.

For consonants in initial position, we found that nasals, liquids (/r, l/), and alveolar obstruents were well perceived. Labial, dental, and velar obstruents were often confused for other sounds. For stops, performance was slightly better for unvoiced segments than voiced. For consonants in final position, performance degraded somewhat for nearly all sounds. These results led us to focus our attention on improving the quality of obstruent consonants, in particular the details of closure and release.

3. FINE TUNING OF CONSONANT RULES

One source of misperception of stop consonants was the timing and amplitude of the bursts at the consonant releases. We reasoned that at the release of a stop, the active articulator (lips, tongue blade, or tongue body) makes an initial rapid release due to aerodynamic forces, followed by a brief interval (5–20 ms) in which the increase in cross-sectional area is delayed, followed finally by a continuing opening movement. In the synthesizer, the time course of the constriction areas just after release was adjusted to conform to this pattern. This led to more realistic timing of the bursts, especially for voiced stops.

In addition to the above changes having to do with articulator movements, rules have been developed for the parameters **ue** and **dc**, parameters which had not been utilized previously. When a stop consonant is produced, a complete closure is formed in the oral cavity, and pressure in the mouth increases rapidly behind the constriction. All else being the same, the result is that the transglottal pressure is quickly reduced, to a point where vocal fold vibration cannot be maintained. For voiced stops, however, some prevoicing is desired. By controlling **ue**, the rate of increase of the vocal-tract volume, from zero to a positive value, the buildup of oral pressure due to the oral-cavity closure is slowed. Therefore, the decrease in transglottal pressure is also slowed, and the vocal folds continue to vibrate for a longer time after closure. In this way, prevoicing of voiced stops is enhanced.

The parameter **dc**, change in compliance of the vocal-tract walls and vocal folds, is also used to enhance or inhibit vocal fold vibration for all types of obstruents. Increasing the compliance corresponds to slackening of the vocal folds, with a consequent decrease in the transglottal pressure that is necessary to maintain vocal fold vibration. It also means that pressure will not build up as

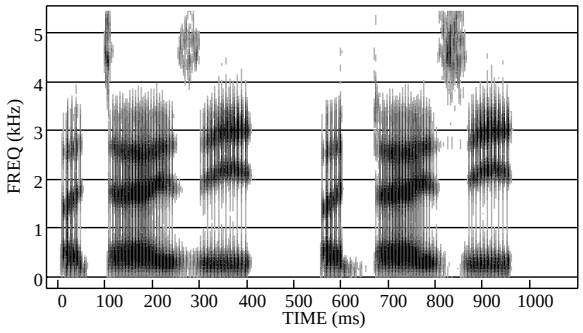

Figure 3: Spectrograms of two versions of the utterance "a daisy." To the left is the speech generated before the new rules were added to the system, and to the right is the speech generated afterwards.

quickly in the oral cavity, prolonging the vibration of the folds. On the other hand, decreasing the compliance corresponds to making the folds more taut, and therefore vibration ceases more rapidly as oral pressure builds up. Oral pressure also builds up more rapidly because the vocal-tract walls are stiffer. Our rules increase **dc** for all voiced obstruents, and decrease it for unvoiced obstruents (cf. [6]).

Note that increasing the compliance of the vocal folds reduces the fundamental frequency, while decreasing the compliance increases it. It is well known that in natural speech "pitch skips" occur during transitions between obstruents and vowels, with pitch drops occurring at the onset of a vowel preceded by an unvoiced consonant, and pitch increases occurring during the onset of a vowel preceded by a voiced obstruent. It is believed that these pitch skips are due to changes in vocal-fold compliance employed during obstruent production to either enhance or inhibit vocal fold vibration. Our rules call for the changes in **dc** to begin about 20 ms before closure, and to extend about 40 ms after the onset of voicing. (The actual adjustments to **f0** due to **dc** are implemented in HLsyn, as it maps the HL parameters to KL parameters (see Section 4.2)).

Figure 3 contrasts spectrograms for the utterance "a daisy," as produced by the system before and after the adjustment of the articulator trajectory **ab** and the addition of **ue** and **dc** rules. It can be seen that prevoicing and VOT are lengthened, and the burst is slightly weakened for /d/, due to these changes in the system.

New rules have also been developed to employ the parameter **ap** for the production of voiced fricatives. In the earlier version of the system, the KL parameter AV (amplitude of voicing) relative to AF (amplitude of frication) was too high for this class of sounds. The requirement of maintaining continued vocal-fold vibration and, at the same time, significant frication noise can be achieved through proper adjustment of **ap** (area of the cartilagenous portion of the glottis) in conjunction with **ag** (area of the membranous portion of the glottis). An increase in **ap** allows oral pressure to increase, leading to an increase in frication and a decrease in voicing, as desired. Maintaining a value of **ag** that is within the range for vocal-fold vibration, as well as an increased vocal-fold compliance, can guarantee continued glottal vibration even at a reduced transglottal pressure. Rules have been developed for control of **ag** and **ap** to produce voiced fricatives with acceptable acoustic characteristics and with calculated pressure and flow patterns that match those of natural speech. Fig. 3 illustrates

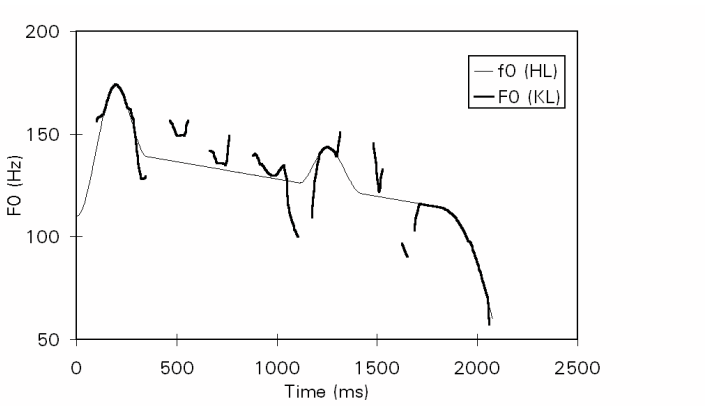

Figure 4: The intonation contour **f0** generated by VHLsyn compared to the Klatt parameter track F0 produced by HLsyn, for the sentence "Five people played basketball." The utterance has high tones on "five" and on the first syllable of "basketball." The adjustments to **f0** were made to account for subglottal pressure, variations in vocal-fold compliance, and vowel height.

the effect of the parameters **ap** and **dc** on the voiced fricative /z/. Comparing the two spectrograms, one can see that use of these parameters gives a synthesized fricative with characteristics similar to those of a naturally produced fricative. Informal listening corroborates this observation.

4. SUPRASEGMENTAL EFFECTS

Rules are also being developed to handle focus, stress, and other prosodic issues. Stress and focus are indicated in the input phonetic string by marks indicating that a syllable is either strong or weak, and that a syllable nucleus carries a high or low tone (or none at all).

4.1. Timing

Timing rules determine where along a time axis the landmarks fall. The placement of the landmarks occurs when the input phonetic string is parsed into a landmark list. During the parse, a duration for each phone (related to the time between landmarks) is extracted from a table of inherent phone durations [1, 3], and then modified by higher-level suprasegmental factors, such as syllable strength and syllable phrase position.

In the absence of suprasegmental effects, beginning and ending times of each phone are determined from the lookup table and the sequential order of the phones. Next, the landmarks are placed relative to the phone interval. For instance, the nucleus landmark of a vowel may be placed at the center of its interval, and the offglide landmark placed at a time slightly greater than its ending time.

The suprasegmental timing factors are quantified with numerical dilation factors that have been used successfully to model the rhythm of English [4]. Their effects on landmark timing are calculated using a Directed Acyclical Graph (DAG) with binary branching. A DAG is used rather than a tree because we allow for ambisyllables, that is, syllables that share phonemes. The graph is

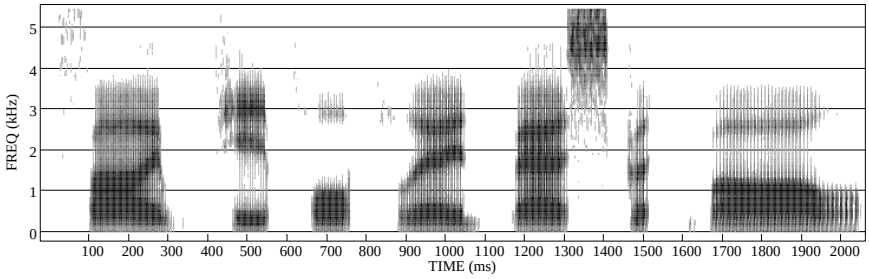

Figure 5: Spectrogram of the utterance "Five people played basketball," as generated by the rule-based system VHLsyn.

constructed during the parse of the input string, and has various levels such as intonational phrases, syllables, and phonemes. The DAG structure allows the dilation factors to be inherited by levels below the syllable level so that the durations of phones can be altered based on dilations. While syllabic dilation factors are inherited downward through the graph, they do not apply with the same strength at all terminal nodes. In particular, consonant durations are less affected by syllabic strength than are vowels.

The timing rules lead to realistic coarticulatory effects if the relative positions in time of the landmarks change. For example, when a reduced vowel occurs between two consonants, the closure of the second consonant occurs soon after the release of the first, so that the formant transitions at release may overlap with the formant transitions at closure. In such a case, a weighted average of the two trajectories is used to calculate the output formant trajectory, which can result in formant undershoot. Thus, as commonly seen in natural speech, the formant trajectories do not attain the vowel nucleus target.

4.2. Intonation contour

The high and low tones assigned to a syllable nucleus are used to derive the **f0** contour. Presently the rules are rather primitive. The default **f0** track for a declarative statement is a straight line that decreases slowly throughout a phrase until the last syllable nucleus, after which it drops off quickly. When a syllable nucleus is marked *high* or *low*, the default line is perturbed by peaks or valleys as appropriate. For now, the peak or valley is centered on the syllable nucleus. Following a high tone, the default **f0** track is adjusted upward, whereas after a low tone, it is adjusted downward. This adjusted default **f0** track is shown by the light line in Fig. 4. Any tones that follow are imposed on this adjusted default track. Although the current implementation is not sophisticated, it can increase the naturalness of the output speech.

It is important to note that the **f0** track generated by the rules is not identical to the F0 track used by the formant synthesizer to produce the utterance. That is because the **f0** track serves as input to HLsyn, which modifies it to reflect segmental effects of vowel and consonant quality. When **f1** is low, HLsyn assumes that the tongue body is high. Because high vowels are known to be produced with higher fundamental frequencies than nonhigh vowels, HLsyn adjusts the input **f0** track to reflect the high tongue body height. In addition, as discussed in the previous section, increasing or decreasing the compliance of the vocal folds will decrease or increase, respectively, the fundamental frequency. Thus, when HLsyn encounters nonzero values of **dc**, it adjusts the **f0** track accordingly. Finally, increases and decreases in subglottal pressure **ps** also result in adjustments to the **f0** track. The heavy contour in Fig. 4 shows the fundamental frequency after those adjustments are made. This adjusted track is the F0 contour that is input to the Klatt synthesizer.

5. SUMMARY

The ongoing development of a rule-based speech synthesis system, based on a formant synthesizer, has been described. The value of such a system lies in the low storage needs, flexibility, and computational simplicity of formant synthesis. Fine tuning of rules related to consonant production, and addition of suprasegmental effects have increased the intelligibility and naturalness of the output speech. An example of a complete sentence produced by the system is given in Fig. 5. Future work will focus on continued development of rules for timing, prosody, and vowel quality.

6. REFERENCES

[1] Crystal, T. H. and A. S. House (1990). "Articulation rate and the duration of syllables and stress groups," *J. Acoust. Soc. Am.* 88, pp. 101–112.

[2] Hanson, H. M., K. N. Stevens, and R. E. Beaudoin (1997). "New parameters and mapping relations for the HLsyn speech synthesizer," *J. Acoust. Soc. Am.* 102, p. 3163.

[3] Klatt, D. H. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Am.* 59, pp. 1208-1221.

[4] Local, J. and R. Ogden (1996). "A model for timing for nonsegmental phonological structure," in *Progress in Speech Synthesis*, J. P. H. van Santen et al., Eds. New York: Springer-Verlag.

[5] Stevens, K. N. and C. A. Bickley (1991). "Constraints among parameters simplify control of Klatt formant synthesizer," *J. Phon.* 19, pp. 161–174.

[6] Svirsky, M. A., K. N. Stevens, M. L. Matthies, J. Manzella, J. S. Perkell, and R. Wilhelms-Tricarico (1997). "Tongue surface displacement during bilabial stops," *J. Acoust. Soc. Am.* 102, pp. 562–571.

[7] Williams, D. R., K. N. Stevens, E. Carlson, and C. A. Bickley (1996). "Multilevel approach to rule-based speech synthesis using quasiarticulatory parameters," *J. Acoust. Soc. Am.* 100, pp. 2760–2761.