

SYNTHESIS OF INITIAL (/s/-) STOP-LIQUID CLUSTERS USING HLsyn

David R. Williams

Sensimetrics Corporation
26 Landsdowne Street
Cambridge, MA 02139 USA

ABSTRACT

This paper describes synthesis of English syllable-initial stop-liquid and /s/-stop-liquid clusters using the HLsyn speech synthesis program. The articulo-acoustic parameters of HLsyn permit efficient synthesis of most consonant types; the parameter specifications also capture important generalizations about how related sets of consonants are produced. Here, we discuss settings of a small number of parameters that permit synthesis of 60 different phonetic sequences.

1. THE HL SYNTHESIS APPROACH

A primary motivation for the HL synthesis approach is to combine the simplicity of control that characterizes articulatory approaches to synthesis with the accuracy and computational efficiency of traditional formant synthesis [1, 2]. This hybrid approach employs a small set of high-level (HL) parameters to construct an articulo-acoustic utterance specification which is then transformed by means of a set of physiologically- and acoustically-motivated mapping relations into a specification in terms of the larger set of lower-level (LL) acoustic parameters needed to control a KLSYN88 formant synthesizer [3]. In effect, the HLsyn synthesis system provides an articulatory interface to a formant synthesizer.

1.1. Functions of the HLsyn parameters

Ten user-settable parameters are included in the HLsyn synthesis system. The functions of these parameters can be described in terms of three broad classes:

1. Class 1 parameters control the first four natural frequencies of the vocal tract ($f1$, $f2$, $f3$, $f4$); these parameters specify acoustically the vocal tract configuration and slow movements of articulators. The $f0$ parameter specifies the fundamental frequency.
2. Class 2 parameters control cross-sectional areas of local constrictions formed by the lips (al) and the tongue tip/blade (ab). They specify the fast movements of primary articulators that rapidly decrease/increase airflow within the oral tract.

3. Class 3 parameters control cross-sectional areas of the glottal orifice (ag) and velopharyngeal port (an) and the pharyngeal volume (ue). These parameters specify opening/closing movements of the glottis and velum and active expansion or contraction of the pharynx.

1.2. HLsyn mapping relations

After the HL parameter values have been specified, the first step in determining values for the LL parameters is to calculate the pressures and flows at the supraglottal and glottal orifices using an aerodynamic model [4]. In addition to the Class 2 and 3 parameter values, inputs to the model include agx (the glottal orifice area as modified by supraglottal forces) and acx (the smallest current supraglottal constriction area). The output of the model is an estimate of the intraoral pressure (Pm) which, along with the orifice areas and a constant subglottal pressure (Ps) value, provides the basis for computing the LL source amplitudes AV , AH and AF .

Other settings and modifications of the LL parameters result from values of HL parameters specified by the user. In general, the Class 1 parameters are mapped directly to their corresponding LL parameters when the glottal area is modal and the velum is closed. An increased glottal area agx affects the LL formant bandwidths and the values of OQ and TL . The presence of voicing in the synthesis signal is conditional on agx ($AV = 0$ when $agx > 15 \text{ mm}^2$). Place-specific filtering of the friction is determined from a look-up table when $AF > 40$ based on the values of $f2$ and $f3$.

The HL parameter $f1$, the first natural resonance, plays several important roles in the synthesis specification. When a class 2 parameter specifies a local labial (al) or alveolar (ab) constriction, $f1$ is modified ($f1c$) to reflect the fact that the constriction is currently controlling the acoustic properties of the vocal tract. The value of $f1c$ is approximated as the lowest frequency of a Helmholtz resonator with constriction area acx and with a constriction length and pre-constriction volume that are determined by the place of articulation. On the other hand, $f1$ can also be used to specify a tongue dorsum (e.g., velar) constriction (acd), in which case acx directly reflects the value of $f1$.

[The parameter $f1$ also plays an important role in the synthesis of nasals. This role and other mapping relations that are operative when the velopharyngeal port is open are not discussed here.]

1.3. Mapping relations for liquids

In the current version of Hlsyn, the method for synthesizing the transfer function associated with liquids (and glides) is only an approximation. When produced with a glottal source (e.g., for liquids following a voiced stop), these sonorants are synthesized simply by specifying the time course of the formants, using default formant bandwidths. This method neglects two factors: (1) that certain formant bandwidths may be significantly widened due to increased acoustic losses in the vocal tract, and (2) that a vocal tract constriction can affect the glottal source, resulting in decreases in amplitude and increases in **OQ** and **TL**.

Because the production of lateral and retroflex consonants can result in a relatively constricted airway, turbulence can be generated at the constriction when the glottis is open and airflow is sufficiently high. The relationship between the distinctive pattern of formants for these consonants (for males: $350 < f1 < 500$, $f2 < 1400$ and $f3 < 1800$ for /r/, or $f2 < 1300$ and $f3 > 2700$ for /l/) and constriction size (acl) is modeled as a Helmholtz resonator. When acl is sufficiently small and its value is also the smallest current oral constriction (acx), the noise source is shaped by **A3F**, reflecting the fact that the natural frequency of the cavity in front of the constriction is always the third formant.

2. /s/-STOP SYNTHESIS

In the following two sections, the parameter settings needed to synthesize all /s/-stop-liquid, voiceless stop-liquid and voiced stop-liquid clusters before the three vowels /i, a, ε/ are discussed. First, we examine the oral and glottal constriction and formant parameter settings for the stop and preceding (optional) fricative.

2.1. Oral constriction parameter settings

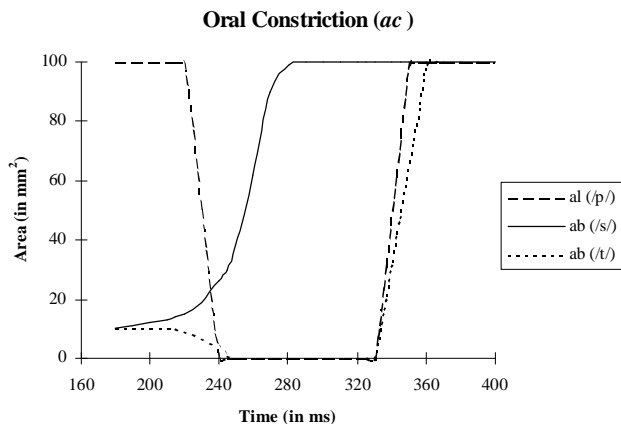


Figure 1: Oral constriction parameters for /s/-release, release of /s/ into /t/ with following /t/-release, and closure/release of /p/.

Rapid articulator movements associated with alveolar and labial stop closure and alveolar fricative constriction are specified with the Class 2 parameters al and ab . The trajectory values shown in Fig. 1 have been derived from analysis, simulation and perceptual studies [5,6]. Somewhat faster rates are used for labial vs. alveolar stops (50 vs. 33 cm^2/s). Only the fricative release trajectory is shown for /s/; the closing trajectory is symmetrical and is followed by a 50 ms held constriction that has an area of 10 mm^2 . If the fricative and stop have the same place of articulation, their HL specifications are sequential and combine (e.g., /st/); if not, the specifications overlap (e.g., /sp/).

2.2. Glottal constriction parameter settings

Settings of the parameter ag determine the spectral shape of the glottal source as well as the voicing classification of obstruents. In (/s/-) stop-liquid clusters, all three variants of stop occur. Voiced unaspirated and voiceless aspirated stops may occur in English stop-liquid clusters (excluding /l/ following alveolars). When preceded by /s/, the stops are voiceless, but unaspirated.

For voiceless stops, ag increases from a default value of 4 mm^2 to just above 15 mm^2 at stop closure (at time: 240 ms) and then to about 28 mm^2 at stop release (at time: 330 ms); the voice-onset-time, is the time between stop release and the point at which ag falls below 15 mm^2 . A similar glottal trajectory is used for the voiceless fricative, although its peak tends to be in the middle of the constriction interval rather than at the release.

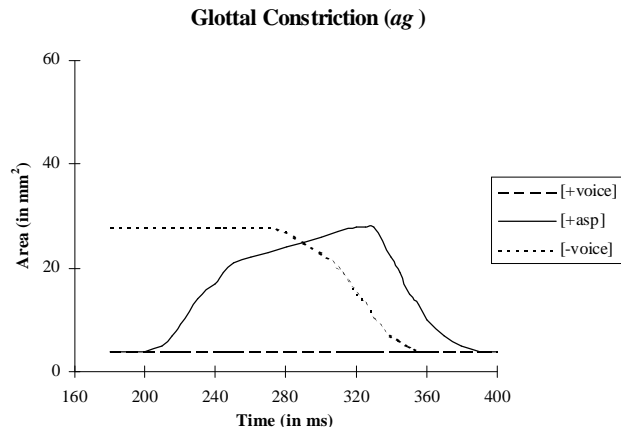


Figure 2: Glottal constriction (ag) parameter values for voiced, voiceless aspirated and unaspirated (following /s/) stops.

2.3. Formant parameter settings

Because of the change in place of articulation, formant transitions occur as the /s/ constriction is released into labial and velar stops (see Fig. 3). Synthesizing this boundary successfully with Hlsyn requires that values of $f2$ and $f3$ (which determine the noise shaping) move in coordination with the Class 2 parameter that is effecting the stop constriction. To simulate a “labial tail”, $f2$ and $f3$ are shifted from the **A5F** region (for /s/) through the **A4F**

region just as *al* gets smaller than *ab*. To simulate a “velar tail”, the formant movements are coordinated with the *f1* trajectory so that they occur as the tongue dorsum (*acd*) constriction becomes smaller than the tongue blade (*ab*) constriction. [Note that an *f1* value of 180 Hz corresponds to an *acd* of 0 mm².]

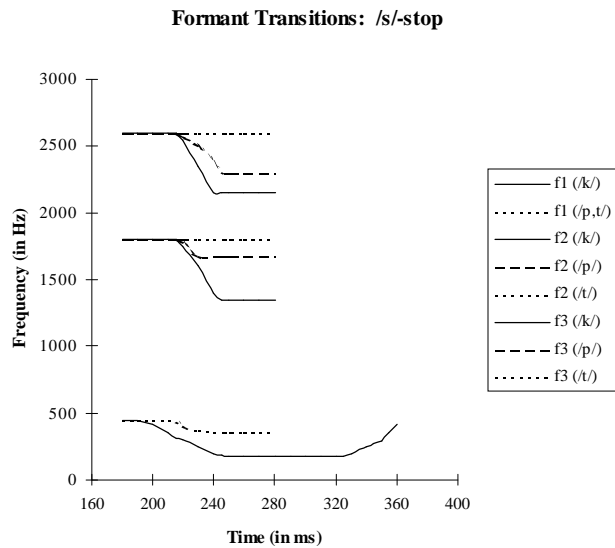


Figure 3: Formant transitions for /s/ released into labial, alveolar and velar stops. For velar stops, *f1* specifies oral constriction size due to the tongue dorsum (*acd*).

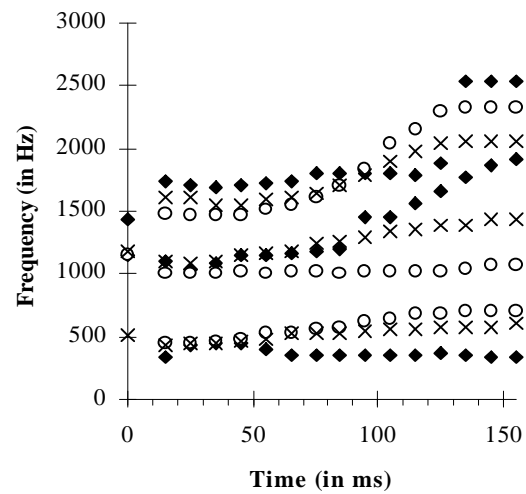
3. STOP-LIQUID OBSERVATIONS

As stated in section 1.3, the synthesis of sonorants produced with a glottal source is currently achieved in HLsyn by specifying the time course of the formants. The five graphs in this section display values of the first three formants for single tokens of labial-liquid (Fig. 4), velar-liquid (Fig. 5), and alveolar-liquid (Fig. 6) syllable-initial clusters measured at each pitch pulse from stop release (at 0 ms) to vowel nucleus. In these graphs, values marked at the origin signify measurable peaks in the release burst; the gap between burst frequencies and subsequent formant measurements is the voicing onset delay.

An overall consistency observed in these data was that movement away from the liquid configuration toward the vowel began about 80-100 ms after stop release independent of the stop place of articulation. With the exception of velars, voice-onset times were similar within stop place, being shorter for labials (15 ms) than for velars (25-35 ms before /r/, 40-65 ms before /l/) and alveolars (30-35 ms).

Also apparent in these data are the spectral hallmarks of liquids: low third formant (below 2000 Hz) for /r/, low second and high third formants for /l/. Particularly striking are the abrupt upward shifts in the second and third formants as the liquids are released into the vowel /i/, reflecting the changing cavity affiliations.

Clusters with /br/



Clusters with /bl/

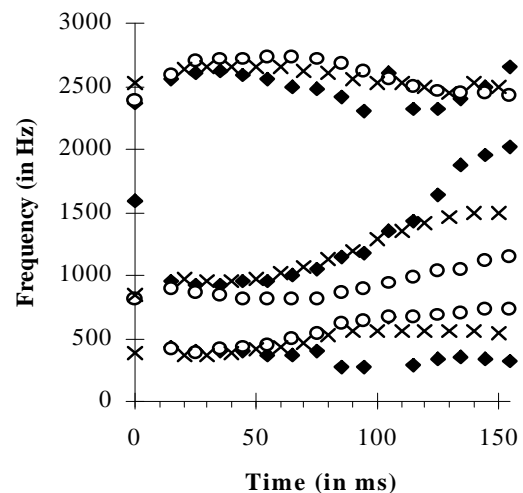


Figure 4: Formant frequency values following labial stop release into /r/ and /l/. For each graph, the following vowels are /i/ (filled diamonds), /a/ (open circles), and /ε/ (crosses).

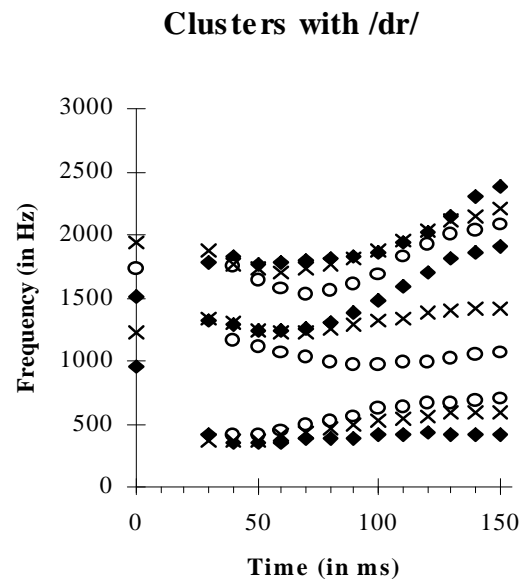
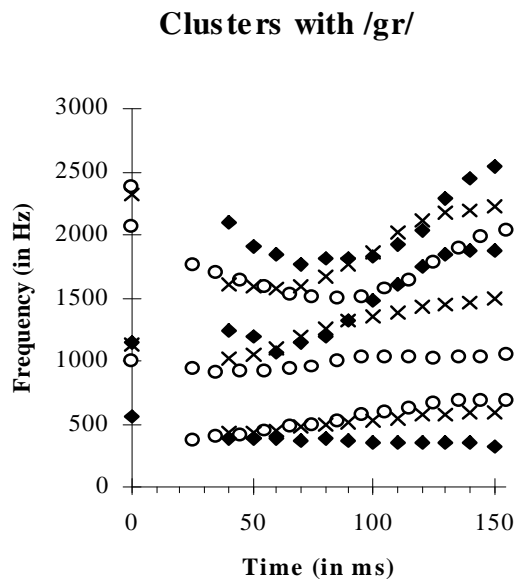


Figure 6: Formant frequencies following alveolar stop release into /ri/ (filled diamonds), /ra/ (open circles), and /re/ (crosses).

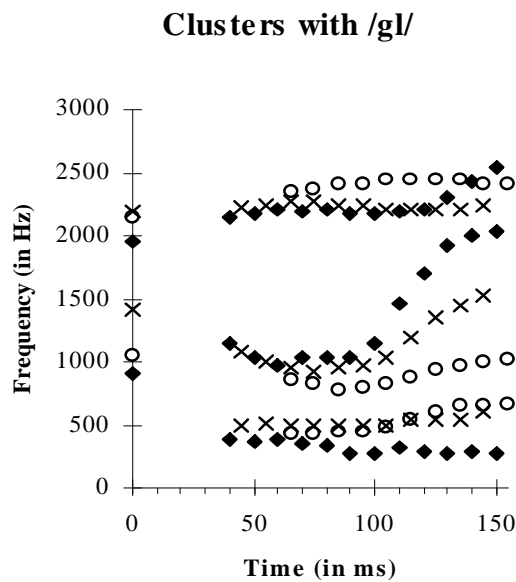


Figure 5: Formant frequency values following velar stop release into /r/ and /l/. For each graph, the following vowels are /i/ (filled diamonds), /a/ (open circles), and /e/ (crosses).

4. SOUND FILES

Attached to this paper are three synthesis examples: “a screed” [SOUND A785S01.WAV], “a spleen” [SOUND A785S02.WAV], and “a strode” [SOUND A785S03.WAV]. Comments (directed to williams@sens.com) are welcome.

5. ACKNOWLEDGEMENTS

Research supported by NIH. For more information about Hlsyn, email sensimetrics@sens.com, telephone (617) 225-2442, or visit the World Wide Web page at <http://www.sens.com>.

6. REFERENCES

1. Stevens, K. N., and C. A. Bickley, (1991) “Constraints among parameters simplify control of Klatt formant synthesizer.” *J. Phonetics* 19: 161-174.
2. Williams, D. R., K. N. Stevens, and C. A. Bickley, (1992) “Inventory of phonetic contrasts generated by high-level control of a formant synthesizer.” *Proceedings 2nd Int'l. Conf. Spoken Language Processes*, Banff, Alberta, Canada, 571-574.
3. Klatt, D. H, and L. C. Klatt (1990) “Analysis, synthesis, and perception of voice quality variations among female and male talkers.” *JASA* 53: 1070-1082.
4. Stevens, K. N. (1993) “Models for the production and acoustic of stop consonants.” *Spch. Comm.* 13: 367-375.
5. Williams, D. R. (1994, June) “Modelling changes in magnitude and timing of glottal and oral movements for synthesis of obstruent consonants.” *JASA* 95: 2815(A).
6. Williams, D. R., (1994, November) “Perception of fricatives synthesized by higher-level control of a Klatt synthesizer.” *JASA* 96: 3227(A).