Production of Consonants with a Quasi-Articulatory Synthesizer

Kenneth N. Stevens[†] and **Helen M. Hanson**[*]

[†]Dept. of Electrical Engineering and Computer Science, and Research Laboratory of Electronics
M.I.T., Cambridge, Massachusetts, USA 02139-4307
stevens@speech.mit.edu

[*]Research Laboratory of Electronics, M.I.T., Cambridge, Massachusetts, USA 02139-4307
hanson@speech.mit.edu

ABSTRACT

This paper describes a speech synthesizer, called HLsyn, that is controlled by articulatory parameters. A set of aerodynamic and acoustic parameters are calculated from mapping equations internal to the synthesizer, and the derived acoustic parameters are used to control the sources and filters for a Klatt formant synthesizer. An example showing the synthesis of an utterance containing a consonant cluster illustrates how the synthesis parameters for place of articulation and for glottal control are overlapped to produce a fluent output. The organization of rules controlling HLsyn from a phonological planning stage is described.

1 INTRODUCTION

In this paper we describe a speech synthesizer that is controlled by articulatory parameters. From these articulatory parameters, a set of aerodynamic and acoustic parameters are derived, and the acoustic parameters are used to control a Klatt formant synthesizer [1]. The synthesizer is called HLsyn, because it is controlled by higher-level articulatory parameters [2].

The primary aim of research with the synthesizer is to formulate principles for the synthesis of running speech based on parameters related to articulation, and therefore to gain insight into the process by which speakers coordinate the movements of the articulators. We illustrate these principles by the synthesis of a sequence containing two adjacent consonants. In human speech production, these sequences are produced with some overlap of gestures associated with the features for the individual consonants. The overlap can occur in the movements of the primary articulators that produce the consonants and in the movements of secondary articulators like the vocal folds, the soft palate, and the tongue body. This overlap can lead to some weakening of the acoustic cues for the individual consonants. However, the amount of overlap must be limited so that some cues for the features for each consonant are retained. The amount of overlap that is allowed is greater for within-syllable consonant sequences than for consonant sequences across syllable boundaries.

2 DESCRIPTION OF HLsyn

The parameters that control HLsyn, and their relation to the articulatory and laryngeal structures and to respiration are shown in Fig. 1(a). Brief descriptions of these 13 parameters are listed in Table 1. The diagram in Fig. 1(b) shows these parameters as inputs to the synthesizer. A set of mapping relations transforms the HL parameters (represented by lowercase letters) into the larger number of acoustic source and filter KL parameters (represented by capital letters) that control the Klatt formant synthesizer. For example, the KL parameter AF is the amplitude of frication noise, and several KL parameters specify the properties of the glottal source (such as the amplitude AV of the periodic component, the spectrum tilt TL, the open quotient OQ, and the aspiration noise amplitude AH). A number of KL parameters control the filtering of the sources. These include the formant frequencies and bandwidths (such as F1, B1, F2, and B2), and nasal poles, zeros, and bandwidths.

Within the 13 control parameters, there are four (f1, f2, f3, and f4) that are not strictly articulatory parameters, but are closely related to articulation. These are the natural frequencies of the vocal tract, assuming that there is no acoustic coupling to the tracheal and nasal cavities, and assuming there are no local consonantal constrictions in the vocal tract. If there is coupling to the nose or trachea, or if a local constriction exists (as for the parameters al and ab), the mapping relations make some adjustments to the natural frequencies, and, in some cases, modify the formant bandwidths relative to a set of default values. Thus the KL parameters such as F1 and F2 in Fig. 1 may be different from the HL parameters that specify the

ISBN 1-876346-48-5

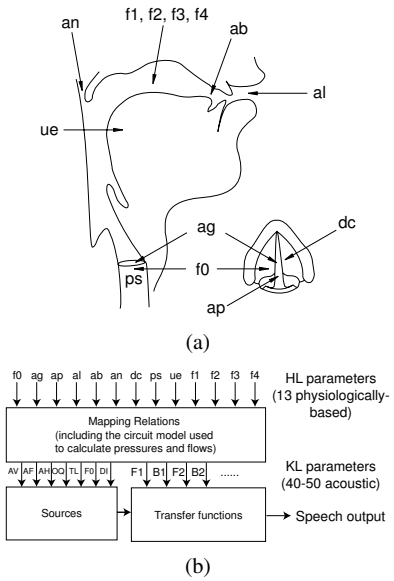

Figure 1: (a) Sketch of the vocal tract and larynx showing articulators that are controlled by HLsyn parameters. (b) Block diagrams showing HL parameters, mapping relations, and KL parameters calculated from these relations. (From [3])

f1–f4	First four natural frequencies of vocal tract, assuming no narrow local constrictions (Hz)
f0	Fundamental frequency due to active adjustments of vocal folds (Hz)
ag	Average area of glottal opening between the membranous portion of the vocal folds (mm^2)
ap	Area of the posterior glottal opening (mm^2)
ps	Subglottal pressure (cm H_2O)
al	Cross-sectional area of constriction at the lips (mm^2)
ab	Cross-sectional area of tongue blade constriction (mm^2)
an	Cross-sectional area of velopharyngeal port (mm^2)
ue	Rate of increase of vocal-tract volume (cm^3/s)
dc	Change in vocal-fold or wall compliances (%)

Table 1: Description of HLsyn parameters

"underlying" natural frequencies f1 and f2.

A central component of the mapping relations is a circuit model that calculates airflows and pressures in the vocal tract, given the subglottal pressure ps, the cross-sectional areas of constrictions at the larynx and in the supraglottal airways, the compliance of the walls of the vocal tract and glottis, and any active expansion or contraction of the vocal-tract volume [3, 4]. From these calculated flows and pressures the amplitudes and spectra of the periodic glottal source and turbulence noise sources are calculated. This aerodynamic model is particularly relevant to the synthesis of obstruent consonants.

The model calculates pressures and flows at all times in a synthesized utterance. Thus at all times the model must have knowledge of the average cross-sectional area of the laryngeal constriction, given largely by the parameters ag and ap, together with the minimum cross-sectional area in the supraglottal airway. For the supraglottal tract, the minimum of four areas is calculated: (1) the area at the lips, given by al, (2) the cross-sectional area ab formed by the tongue blade, (3) the minimum cross-sectional area as a consequence of tongue-body movement, and (4) the cross-sectional area for liquid consonants. The minimum constriction size formed by the tongue body is a simple function of f1 based on acoustic theory. For low f1, as for high vowels or velar consonants, the area is small, and is made in the oral cavity. For high f1, the area is also small, and is located in the pharyngeal region. For intermediate f1 values, there is no major tongue-body constriction. Vocal-tract configurations corresponding to liquids can be identified by particular formant patterns that are outside the range normally observed for vowels. When the formants are within this "liquid" range, a formula in the mapping relations calculates an estimate of the cross-sectional area formed by the tongue blade. Although all four types of areas are calculated at all times, the minimum vocal-tract area during vowels, glides, and liquids is usually large enough that there is no pressure buildup behind the constriction, and consequently no frication noise is generated. However, in the context of segments with an expanded glottal opening (for example adjacent to aspirated stop consonants) the tongue body or tongue blade constriction for these segments may be comparable to or less than the glottal opening, and there will be increased pressure behind the constriction and hence frication noise. Further details of the mapping relations in HLsyn are given elsewhere [3].

3 TOWARD RULES FOR SYNTHESIS WITH HLSYN

We describe here the structure of the rules that are proposed for the control of HLsyn from a discrete linguistic input. The synthesis of an utterance begins with a planning stage, which has several components. Word boundaries are marked in this planning stage, as are the beginnings and endings of phrases. The planning stage consists of sequences of phonemic segments or bundles of distinctive features. Each consonant in the sequence is labeled to indicate which vowel it is affiliated with (or it may be affiliated with two vowels, one before and one after the consonant). Vowels are

ISBN 1-876346-48-5

labeled with three degrees of prominence.

For each feature in a bundle there is a set of rules that specify how the various HL parameters are to be controlled [5]. These rules depend on other features in the bundle and in adjacent segments, the affiliation with a vowel (for consonants), the prominence of the vowel, the location of the word boundary relative to the segment, and the location of the segment with respect to the phrase boundaries. Three kinds of rules are involved in the synthesis of an utterance: (1) rules for prosody at the phrase level, involving the parameters ps, f0, and ag; (2) rules for controlling the parameters f1, f2, f3, and f4, which, in effect, are rules for manipulating the tongue body position and lip rounding; and (3) rules for controlling the primary and secondary articulators for consonants. For rules of type (1) and (2) the parameters must be specified at all times in an utterance. The control parameters involved in the rules of type (3) normally have some default state, and for each consonant a selection from among these parameters is made depending on the consonant features. For example, for obstruent consonants, the primary articulator must be selected, and secondary parameters such as laryngeal configuration, stiffness, and vocal-tract expansion are activated to achieve appropriate pressures and airflows, depending on the consonant voicing. For nasal consonants, the primary articulator is again selected and the parameter an is activated.

4 AN EXAMPLE: SYNTHESIS OF A CONSONANT SEQUENCE

We illustrate the control of the synthesizer for the consonant sequence /pr/ in the word pray. Figure 2 shows some of the control parameters, together with a spectrogram of the resulting synthesis and some of the derived KL parameters. For this brief utterance our main concern is not with the parameters relating to prosody. However, in Fig. 2(c) we do show the rise and fall of the ps parameter simulating the onset and termination of expiratory control of respiration. Also shown (Fig. 2(b)) is a slight spreading of the glottis (parameter ag) near the end of the word—a common gesture at the termination of expiration [6]. An appropriate f0 contour is also used in this synthesis (not shown in figure).

Our main interest here is to examine the interaction of HL parameters for the tongue body, the lips, the glottis, and the tongue blade. The rules for the synthesis of /p/ are to close the lips (parameter al), to stiffen the vocal folds and the vocal-tract walls (parameter dc with a negative value), and to spread the glottis (increased ag) during the closure and for a time interval following the labial release. The synthesis of syllable-initial /r/ involves setting f1, f2, and f3 to appropriate values, particularly with a lowered f3, and then to implement transitions of these formants to formant targets for the beginning of /e/. This diphthongized vowel has an offglide toward a lower f1 and a higher f2.

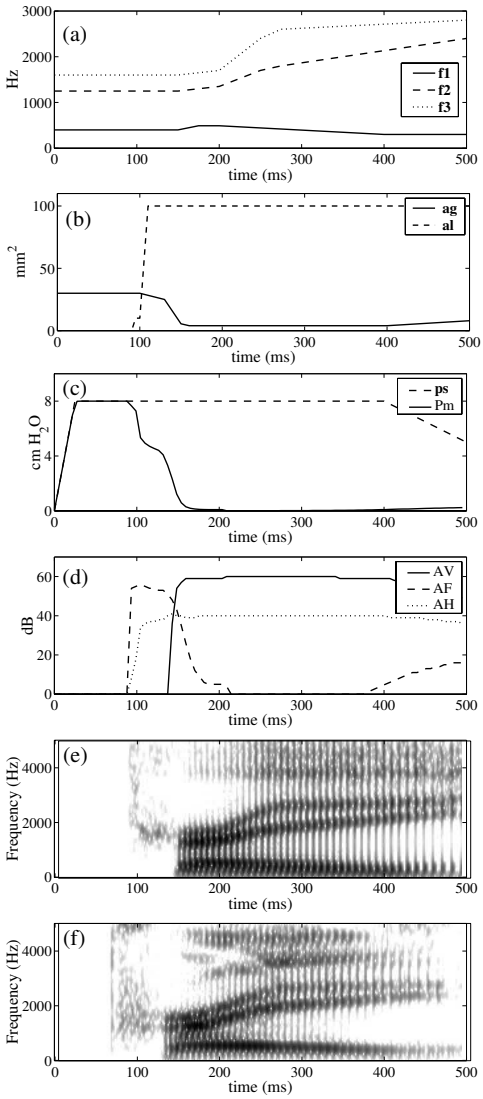

Figure 2: Synthesis of the word "pray." See text for details.

Figure 2(a) shows the movements of the formant parameters f1, f2, and f3. The low f3 is implemented from the beginning of the utterance, indicating that the tongue is in the /r/ position during the /p/ closure. The labial consonant places little constraint on the tongue movement for the following segment. The time course of the labial parameter al is given in Fig. 2(b). It shows a rather rapid rise. (The mapping relations automatically compute a lowering of the first formant frequency relative to the rule-generated value of f1 due to the local labial constriction.) Figure 2(b) also displays the time course of the glottal opening ag, which remains wide for about 40 ms before returning to a modal value for the vowel. At the release of the pa-

ISBN 1-876346-48-5

rameter al, a brief initial peak in airflow through the lip opening is calculated, together with a frication noise source. Since it occurs at the lips, this source is not filtered by a formant resonator.

In the time up to about 200 ms, the formants are in the range for liquids, and the minimum cross-sectional area in the vocal tract is assigned to the tongue-blade region. This area turns out to be comparable to ag, and consequently there is pressure buildup behind the constriction for /r/, and a frication noise source is generated. This source excites a resonator with frequency set to f3, as specified by the mapping relations.

The calculated KL source parameters AV, AF, and AH (amplitudes of voicing, frication, and aspiration) are displayed in Fig. 2(d). It is noted that during the 40-ms interval following the release of al, the dominant noise source is frication rather than aspiration. A brief interval of weak frication noise appears at the end of the synthesized word. This is a consequence of the increased ag at utterance termination and the constriction formed by the high tongue body as f1 decreases at the end of the vowel /e/. A spectrogram of the synthesized utterance is shown in Fig. 2(e). A spoken version of the same utterance is displayed in Fig. 2(f) for comparison.

5 DISCUSSION

Our experience with HLsyn highlights some advantages of articulatory synthesis over more conventional source-filter synthesis or concatenative synthesis. As a research tool, it certainly helps the user to gain deeper insights into the human speech production process—what kinds of gestural overlap are allowed and what are not; the role of subglottal pressure changes in shaping acoustic events at phrase boundaries and at prominences and lenitions; the potential role of manipulation of vocal-tract wall stiffness in implementing the voicing distinction for obstruent consonants; etc. One can experiment with these and other aspects of speech production following an analysis-by-synthesis approach.

It is also evident that articulatory synthesis can be achieved with control of a relatively small number of parameters. For example, synthesis of the word pray simply involves manipulation of the lips, the glottis, and adjustment of the formants for /r/, and yet a rather complex sequence of acoustic events emerges from the synthesizer. A change from pray to bray, for example, could be made simply by adjustment of the parameter agand possibly also dc (not shown in Fig. 2). Synthesis with acoustically based parameters would involve a more complex adjustment.

Attempts to synthesize with articulatory controls comes with a cost, however. Speech articulation has not been studied as extensively as speech acoustics, and there are a number of aspects of articulation that are not as thoroughly measured as the acoustic pattern. For example, extensive data on the changes in glottal configuration for voiced and voiceless consonants or for vowels in different prosodic environments are not available, and there are essentially no data on changes in stiffness of the vocal-tract surface. Some of these parameters in the synthesizer need to be adjusted by trial and error so that they yield appropriate pressures and flows, and appropriate speech patterns.

This highlights an inherent limitation of present-day synthesizers. Unlike an adult speaker, or, more precisely, unlike a developing child, a synthesizer is not equipped to evaluate its own speech and to automatically make corrections to its production based on what it has heard. Feedback, followed by correction of the parameters, must be done by the experimenter. The ultimate synthesizer should be able to hear itself and to be self-correcting.

ACKNOWLEDGMENTS

Development of HLsyn was supported by NIH grants No. NS-27407-01 and No. MH52358. Preparation of this paper was supported by NIH grants No. DC00075 and No. DC04331.

REFERENCES

[1] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, pp. 820–857, 1990.

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

[3] Helen M. Hanson and Kenneth N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," *J. Acoust. Soc. Am.*, vol. 112, pp. 11580–1182, 2002.

[4] Kenneth N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1998.

[5] Corine A. Bickley, Kenneth N. Stevens, and David R. Williams, "A framework for synthesis of segments based on pseudoarticulatory parameters," in *Progress in Speech Synthesis*, Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, Eds., pp. 211–220. Springer-Verlag, New York, 1997.

[6] Janet Slifka, *Respiratory Constraints at Prosodic Boundaries in Speech*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.