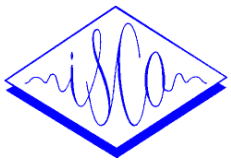

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CONTROL OF A KLATT SYNTHESIZER BY ARTICULATORY PARAMETERS

Kenneth N. Stevens[1], *Corine A. Bickley*[1], *and David R. Williams*[2]

[1]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, and Sensimetrics Corporation, Cambridge, MA 02139

[2]Sensimetrics Corporation, Cambridge, MA 02139

ABSTRACT

The movements of the articulators during speech production perform two functions: (1) to generate sources in the vicinity of constrictions within the airway, and (2) to shape the vocal tract to filter these sources. In an articulatory synthesizer, these two functions are automatic consequences of the movements and states of the articulators, whereas in a formant synthesizer such as a Klatt synthesizer control of the sources and of the filtering are performed independently. This paper describes an attempt to combine the simplicity of control based on articulatory parameters with the demonstrated capability of a formant synthesizer for generating natural-sounding and intelligible speech. A set of higher-level parameters related to articulation is selected, and these parameters are mapped into a larger set of acoustically-based parameters that control the Klatt synthesizer. Strategies for generating these parameters from a linguistic description of an utterance are described.

I. INTRODUCTION

In selecting a set of parameters for controlling an articulatory synthesizer, it is convenient to think of three classes of control parameters:

1. Parameters relating to movements of the tongue body, mandible, lip rounding, and vocal-fold stiffness;

2. Parameters specifying the formation or release of narrow constrictions in the oral cavity; and

3. Parameters describing the cross-sectional areas of velopharyngeal and glottal orifices and active expansion or contraction of the vocal-tract volume behind a constriction.

The acoustic consequences of changes in these three types of parameters interact, so that there is not always a one-to-one relation between the settings of the parameters in one domain and attributes of the sound output. Some of these interactions stem from processes involving acoustic resonators, and others are mediated by aerodynamic processes. For example, a particular adjustment of the tongue body and lip rounding (parameters in class 1 above) gives rise to a certain pattern of formant frequencies and bandwidths if there is no velopharyngeal opening and if the glottis is in a modal configuration, but this acoustic pattern is modified if there is a velopharyngeal opening or glottal spreading (class 3 parameters) or if a labial closure is formed (class 2 parameter). Or, for a particular vocal-tract shape (for example, a high tongue-body configuration) a well-defined formant pattern is observed if the glottis is in a modal configuration but a noise with one or two major spectral peaks is generated when the glottis is sufficiently spread to create turbulent airflow at the constriction formed by the high tongue-body position. These and many other examples show that, even though the inventory of patterns or movements of the parameters in each of classes 1, 2, and 3 may be quite restricted, the variety of acoustic patterns that can result from various combinations of parameters across these classes is very large.

These various interactions in the acoustic consequences of the control parameters for speech production complicate the design of speech synthesizers. A speech synthesizer based primarily on the control of acoustic parameters must have the capability of generating a wide variety of acoustic patterns, and the number of parameters that need to be controlled in such a synthesizer to produce these patterns is often very large – 40 or more in the Klatt synthesizer, for example (Klatt and Klatt, 1990). Nevertheless, it is well known that such a synthesizer is capable of producing utterances that are indistinguishable from those of a human female or male speaker if sufficient care is given to adjusting these many control parameters.

In the synthesizer described here, as in any articulatory-based synthesizer, the interactions between the parameters in the three different classes are built into the structure of the synthesizer. These interactions are accounted for through a set of mapping relations or equations that transform a compact set of higher-level (HL) articulatory-based parameters into the larger set of lower-level parameters for the Klatt synthesizer (Stevens and Bickley, 1990). These mapping relations are based on aerodynamic and acoustic principles that have been described in the literature.

We have been attempting to simplify the control of the Klatt synthesizer and, at the same time, to move toward a control strategy that is based on speech-production principles that are manifested in the three-class structure described above.

II. SELECTION OF CONTROL PARAMETERS

There are several approaches to the selection of a set of parameters that specify the vocal-tract shape (when it does not contain a narrow constriction), as described in class 1 above. One approach is to use parameters that describe the area function of the vocal tract (cf. Kelly and Lochbaum, 1962), and another is to devise a set of parameters describing the positions and shapes of the various structures that form the airway (Coker, 1968; Mermelstein, 1973). In the synthesis method that we are developing, we have chosen instead to specify the first four natural frequencies of the vocal tract: **f1**, **f2**, **f3**, and **f4**. There are at least two reasons for this choice. These natural frequencies are equivalent to the frequencies of the spectral prominences for nonnasal vowel configurations with modal (or pressed) voicing. When a speaker produces such a vowel, he/she probably attempts to achieve some target acoustic pattern that is defined by the formant frequencies, and there may be some variability in the articulation configuration that produces this pattern. Another advantage of selecting these natural frequencies as parameters is that they can be directly measured, at least for nonnasal vowels. As noted above, the natural frequencies **f1**, **f2**, **f3**, and **f4** translate into actual formant frequencies in a synthesizer only for nonnasal vowel-like configurations with a glottal opening that is not spread. Creation of a velopharyngeal opening, a glottal opening, or a narrow local constriction with the lips, the tongue blade, or the tongue body can cause a modification of the frequencies and bandwidths of some of these natural frequencies, according to well-defined acoustic principles. These modifications are accounted for automatically in the mapping equations in the synthesizer.

In the present implementation of the synthesizer, the control parameter representing the vocal-fold stiffness is simply the fundamental frequency **f0**. This parameter will not be discussed in this paper.

The parameters in class 2 are intended to specify the cross-sectional areas of consonantal constrictions in the vocal tract. There is a parameter **al** that gives the cross-sectional area of the lip opening, and a parameter **ab** that specifies the cross-sectional area of a constriction formed by the tongue tip or tongue blade. These two parameters are usually active only when labial or coronal consonants are being produced, and they change very rapidly as the consonantal constriction is being formed or is being released. In these local regions in the vicinity of a consonant, the parameters **al** and **ab** cause the HL first-formant parameter **f1** to to be lowered, and in some cases can also cause rapid movements of **f2** and other formant parameters. The original HL formant parameters specify the natural frequencies before these corrections for local consonantal constrictions are applied. The formation of the constriction for a velar consonant uses the tongue body as the major articulator. Such a consonantal constriction introduces a downward correction of the HL parameter **f1**, and the cross-sectional area of the constriction is estimated directly from **f1**.

The parameters in class 3 are usually (but not always) active only when a consonantal constriction is being produced. They specify the movements of "secondary" articulators: the cross-sectional area **an** of the velopharyngeal opening; the cross-sectional area **ag** of the glottis, indicating the degree of glottal adduction or abduction; and the rate of expansion (or contraction) **ue** of the vocal-tract volume behind the constriction to aid the continuation of glottal vibration during an obstruent consonant or (with a negative value) to inhibit glottal vibration. The timing of the parameters in class 3 is tied to the timing of the consonantal constrictions specified by class 2 parameters. For some sounds, such as /h/ or a nasal vowel, a class 3 parameter can be manipulated independently of a class 2 parameter.

We have presented elsewhere a number of examples in which various types of utterances can be synthesized using just the 10 HL parameters to control the synthesizer (Stevens and Bickley, 1991; Williams et al., 1992; Bickley et al., 1994). These include a variety of consonant-vowel and vowel-consonant syllables, consonant clusters, consonant sequences across syllable boundaries, and longer sequences of sentences. The proposed 10 parameters, then, appear to constitute a sufficient set for speech synthesis. To reproduce the voice of a particular speaker, it is necessary to adjust certain fixed default parameters, particularly those related to individual differences in glottal source parameters such as open quotient, spectral tilt, and glottal noise.

III. TOWARD STRATEGIES FOR SPEECH SYNTHESIS

The mapping relations from HL parameters to the lower-level acoustically-based parameters provide a set of constraints that limit the kinds of acoustic patterns that can be generated by the synthesizer. The articulatory-based HL parameters are, however, further constrained in two ways: (1) the parameters to be controlled must be constrained to vary with time in a manner that is consistent with the dynamic properties of the speech production system; and (2) the parameters must be timed and manipulated relative to each other in such a way as to generate acoustic

properties that provide evidence for the various features in the underlying linguistic representation of the utterance that is to be synthesized.

The synthesis strategy can be organized according to the three classes of control parameters listed above. Each of these classes of parameters has different timing constraints.

Class 1 parameters, the vocal-tract natural frequencies **f1**, **f2**, **f3**, and **f4**, are directly related to gross tongue-body movements and lip rounding for vowels and glides. These parameters are also constrained to various degrees during the production of consonants. The gross articulatory movements are relatively smooth, and are limited to a bandwidth of 5-10 Hz. The formant parameters are, for the most part, similarly constrained since rapid formant movements that may occur immediately adjacent to a consonantal constriction are derived from the specification of the consonant that produces the constriction. The HL formant parameters have, in fact, been defined to avoid rapid transitions, so that their trajectories are slowly-moving. The rules for specifying these parameters for a given utterance, then, require only the generation of smooth trajectories through values at sparse locations in time. Detailed rules for generating these parameters have yet to be worked out, and will draw on past work in this area (cf. Allen, Hunnicutt, and Klatt, 1987).

Examples of these HL formant parameters for synthesis of the word anika are given in Fig. 1a. The smooth, slowly-varying attributes of these parameters is evident as the parameters move between targets defined for the vowels and consonants. These parameters undergo further modification when other HL parameters are introduced and when the place of articulation for the consonant is specified. Two such modifications indicating how the lowest vocal-tract natural frequency is influenced by the consonant are shown by dashed lines in the figure. (The actual first-formant frequency is further adjusted by the velopharyngeal or glottal opening, but these influences are not shown in the figure.)

The parameters in class 2 specify the time variation of the cross-sectional areas of narrow constrictions that are formed with the lips or the tongue blade. (The cross-sectional area **ad** of the narrow tongue-body constriction for the velar consonant is derived from the modified **f1** parameter.) These parameters come into play only when the constriction is narrow enough to form a consonant. They change very rapidly when the constriction is formed or released, and the rate of change is determined by the articulator and by the manner of articulation (fricative or stop). When a constriction for a consonant is to be produced, then, it is only necessary to specify the place and manner of articulation for the consonant and the times at which the constriction is to be formed and released. In the example of the word anika, the time course of the tongue-

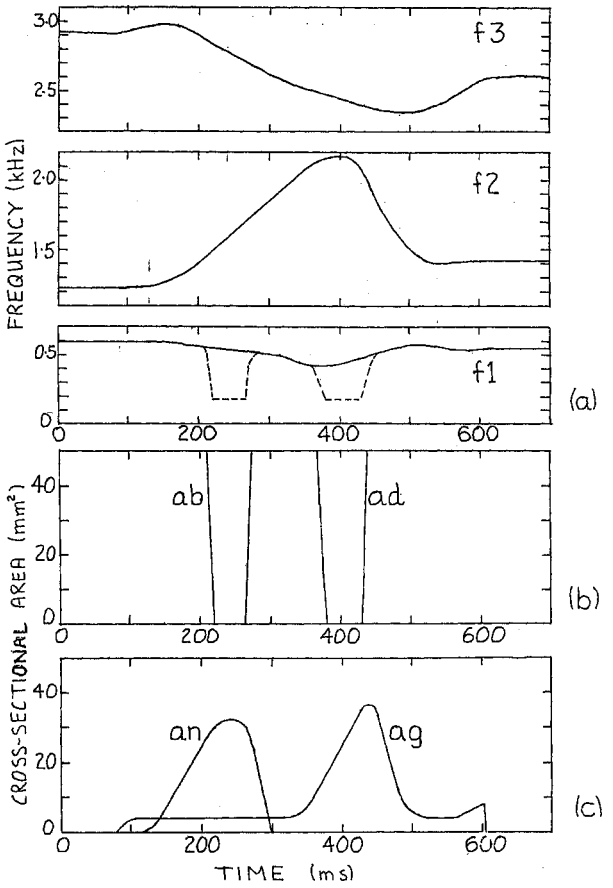

Figure 1: *Control parameters to synthesize the word anika. The three classes of parameters are given in (a), (b), and (c), as described in the text. The dashed lines for* **f1** *show how this parameter is modified when a consonantal constriction is formed.*

blade constriction for [n] is shown in Fig. 1b. This figure also gives the area of the velar constriction as calculated from **f1**.

The most common function of the class 3 parameters is to serve as secondary articulators for the consonants. Thus the parameter **ag** is manipulated to inhibit or to facilitate glottal vibration during the consonant interval and, if required, to cause the introduction of aspiration noise. The parameter **ue** is introduced to enhance glottal vibration in the consonant interval. And **an** is increased if a nasal consonant is required. The time interval over which one of these parameters is activated is generally in the range 150–200 ms, presumably limited by physiological constraints. For each of these parameters, the timing must be adjusted relative to the consonantal events.

Two examples of these class 3 parameters are shown in Fig. 1c. For [n], the parameter **an** begins to open in the first vowel, remains open during the closure for the nasal consonant, and then closes after the tongue blade is released. The parameter **ag** begins to increase near the time of the [k] closure, increases to a maximum at the time of release, and then decreases to

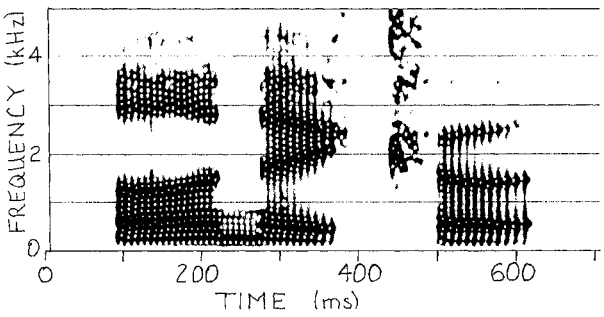

Figure 2: *Spectrogram of utterance synthesized with the HL control parameters shown in Fig. 1.*

about 4 mm^2 (the modal value) about 80 ms after the release.

Given the set of HL parameters in Fig. 1, the mapping equations calculate the pressures and flows, from which the source parameters for the Klatt synthesizer are then calculated. The various Klatt filtering parameters are then computed, and the utterance is generated with the synthesizer. A spectrogram of the final utterance is given in Fig. 2.

The above discussion suggests that the time variations of the class 3 parameters for consonants are largely determined by both acoustic and articulatory dynamic requirements once the times of formation and release of the consonantal constriction are established. If this is the case, then, the formulation of rules for synthesis of an utterance from HL parameters when the linguistic description of the utterance is given reduces to rules for just a few parameters. The formant parameters must be derived, and the times at which consonantal constrictions are formed or released must be determined.

(This research was supported in part by the National Institutes of Health.)

REFERENCES

1. Allen, J., S. Hunnicutt, and D.H. Klatt (1987) **From text to speech: The MITalk system.** Cambridge: Cambridge University Press.

2. Bickley, C.A., K.N. Stevens, and D.R. Williams (1994) *Synthesis of consonant sequences using a Klatt synthesizer with higher-level control.* **J. Acoust. Soc. Am. 95**, part 2, 2815.

3. Coker, C.H. (1968) *Speech synthesis with a parametric articulatory model.* **Speech Symposium**, Kyoto, paper A-4. Reprinted in J.L. Flanagan and L.R. Rabiner (eds.) (1973), **Speech Synthesis**, Stroudsburg PA: Dowden, Huchinson and Ross, 135-139.

4. Kelly, J.L. Jr., and C.C. Lochbaum (1962) *Speech synthesis.* **Proc. Fourth International Congress on Acoustics,** Copenhagen, Paper G42, 1-4. Reprinted in J.L. Flanagan and L.R. Rabiner (eds.) (1973), **Speech Synthesis**, Stroudsburg PA: Dowden, Huchinson and Ross, 127-130.

5. Klatt, D.H. and L.C. Klatt (1990) *Analysis, synthesis, and perception of voice quality variations among female and male talkers.* **J. Acoust. Soc. Am. 87**, 820-857.

6. Mermelstein, P. (1973) *Articulatory model for the study of speech production.* **J. Acoust. Soc. Am. 53**, 1070-1082.

7. Stevens, K.N. and C.A. Bickley (1991) *Constraints among parameters simplify control of Klatt formant synthesizer.* **J. Phonetics 19**, 161-174.

8. Williams, D.R., C.A. Bickley, and K.N. Stevens (1992) *Inventory of phonetic contrasts generated by high-level control of a formant synthesizer.* **Proc. International Conference on Spoken Language Processing**, Banff, Alberta, Canada, October 12-16, 1992, 571-574.