AN OBJECT-ORIENTED STRUCTURE FOR A SPEECH SYNTHESIS SYSTEM

Corine Bickley (1),(2) and Eric Carlson (1)
(1) Sensimetrics Corp., Cambridge, MA 02139, USA
(2) Research Lab of Electronics, MIT, Cambridge MA 02139, USA

ABSTRACT

A new approach to phonetic-string-to-speech synthesis is described. This approach uses an object-oriented framework to structure a speech synthesis system around acoustic landmarks. A landmark-based representation is used in the generation of high-level synthesis parameters. Constraints based on speech production and on the spectral and temporal properties of the speech waveform guarantee that the speech synthesized corresponds to a signal that could have been produced by a human speaker. An example of the use of the system and a discussion of its design are presented.

1. INTRODUCTION

A structure for a phonetic-string-to-speech synthesis system is presented along with examples of phonemes in various phonetic environments. An object-oriented approach was chosen because such systems have been shown to be particularly well-suited for implementing and working with systems consisting of large sets of inter-related constraints [1]. Generating high-quality synthetic speech is more a problem of constraint satisfaction than of specifying parameter tracks for a synthesizer. Synthesis-by-rule systems have traditionally used either ad hoc rule implementations [2] or have been closely modeled on standard phonological notations [3], and have typically been plagued with problems of providing a flexible environment for maintenance and enhancement of rules [4]. An appreciation for such problems has led us to alternative programming methodologies to apply to developing a speech synthesizer.

2. CONSTRAINT-BASED SYNTHESIS

Our approach to speech synthesis is based on a view of speech production as controlling the characteristics of the sound while balancing linguistic goals with speech production constraints. These constraints include, for example, limits on the rates of movement of the articulators, values of resonances of the vocal tract in the vicinity of constrictions, and conditions involving the pressure in the oral cavity and airflows through the orifices of the vocal tract as well as the spectral and temporal properties of the resulting synthesized sounds [5]. The synthesis problem is similar in some ways to the problem solved by a talker attempting to organize the sequence of articulatory movements needed to implement a series of phonemes.

The process of conversion from an input phonetic transcription (with syllabic prosodic specification) to audio waveform proceeds in several phases. The starting point for the synthesis process we describe here is a series of feature markings of the sort proposed in [6]. The transformation from phonetic and prosodic symbols to the series of feature markings is not treated in this paper. Initially this stage will be performed by hand; we plan to draw on the literature for methods of generating a series of feature markings from phonetic and prosodic information. The focus of this paper is the part of the system that generates high-level (HL) parameter tracks [7] – parameters that are based on a speech production model – from feature markings that are clustered together into landmarks. The notion of landmarks [8] is central to the implementation of the synthesis system described in this paper. Landmarks are the anchors around which the constraints are structured, and as such are the principal objects of the system. Currently, the locations in time of the landmarks are determined by hand. Another phase converts HL parameters to low-level (LL) parameters [7]. The resulting LL parameters are then used as input to a Klatt-type formant synthesizer as the final phase, which yields an audio waveform as output.

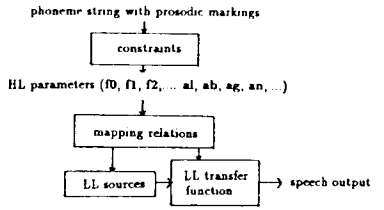

Figure 1: *Process of transformation from string of phonemes into synthesized speech.*

Figure 1 schematizes this transformation from a string of phonemes with prosodic markings into a series of landmarks, and from landmarks into sets of HL parameters. The mapping relations that transform the HL parameters into the low-level parameters that control the sources and transfer functions of a synthesizer such as KLSYN88 [9] have been described previously (for instance, [7], [10]).

A key feature of the kind of system proposed here is the characterization of the synthesis problem as one of constraint satisfaction. The synthesis program encodes constraints that must all be satisfied, and the software finds a solution automatically if possible, and if not, a reason is indicated and modifications to either the specification of the landmarks or to the rules of the system can be made by the user. The opportunity for the user to explore the relationships among the entities (objects) is one important feature of this sort of system.

3. AN EXAMPLE

A short example serves to illustrate the elegance and utility of this sort of object-oriented approach to speech synthesis. The synthesis of consonants is illustrated in this example. Synthesis of sounds involving quickly changing characteristics (such as formant frequencies and bandwidths, and amplitudes of aspiration, frication, and voicing sources) that must be carefully timed relative to one another, has often presented a difficult problem for synthesis-by-rule programs, but yet, in our opinion, is essential to the generation of highly intelligible and natural-sounding synthetic speech.

Consider the phonetic string /wʌnsəpɔnə/ which has sequences containing [+consonantal] segments (/ʌnsə/ and /əpɔn/). One kind of constraint applies to [+consonantal] segments (the /ns/ and the /p/ in /ʌnsəpɔ/, for instance). In each case, a closure/opening gesture for a major articulator (or articulators) must be generated. The rates of closure and subsequent opening as well as the time required for reversal of direction depend on the particular articulator(s) and are based on limits of movement of the physical structure(s). For each [+consonantal] segment, constraints for secondary articulators (such as the glottis and the soft palate) apply. These constraints place limits on the rates of glottal abduction or adduction, rates of change of vocal-fold stiffness, rates of change of the velopharyngeal opening, and rate of change of vocal-tract volume expansion during obstruents.

Figure 2 shows parameter tracks for the individual lip and tongue blade articulators **al** and **ab** (in the top panel). The opening/closing gestures for the lips and tongue blade are similar. The parameter tracks for the secondary articulators of the velum and the glottis are shown in Fig. 3. The complete parameter tracks for the area of each orifice (shown in the bottom panel) are constructed so that the following constraints are all satisfied: closure and opening of the tongue-blade constriction aligned with the landmarks for /n/ and /s/; lip closure followed by opening (approximately 85 ms later) for the /p/; rate of closure/opening is 50 cm^2/sec for both the lips and the tongue blade; for a [+consonantal,−continuant] segment, the minimum area is zero; for [+consonantal,+continuant] segments, the minimum area is such that the amplitudes of the sources of the synthesized sound are appropriate in relation to the adjacent vowel; the area parameter tracks are constructed from individual opening/closing gestures that are derived from the landmarks – the complete track is constructed using the minimum value at each point in time; the values of **f2** and **f3** near each [+consonantal] land-

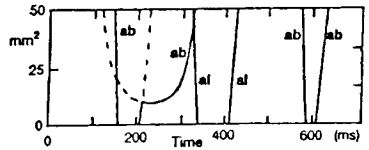

Figure 2: *Parameter tracks for lip* **al** *and tongue blade* **ab** *opening/closings.*

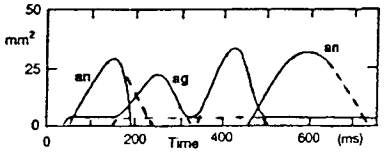

Figure 3: *Parameter tracks for the velum* **an** *and the glottis* **ag**.

mark are constrained to be appropriate for the corresponding oral-cavity constriction in relation to the features for the adjacent vowel; the parameter **ag** must reach a large enough value so that the airflow at the oral cavity constriction causes an appropriate amplitude of aspiraton to be calculated; for any [+consonantal,+nasal] segment, **an** reaches a peak at the first landmark of the pair of [+consonantal] landmarks, and the rate of opening and closing is about 10 cm²/sec. The vocal-tract resonances, target areas, and rates of change of area are adjusted automatically so that appropriate airflows and pressures are maintained to meet the constraints listed above. For any situation in which is it not possible to satisfy all of the specified constraints, the system flags the condition and an adjustment to either the landmark timing or the features markings can be made by the user.

4. SYSTEM DESIGN

Many complex systems may be modeled as interactions among a collection of less complex objects. Object-oriented systems are usually decomposed into a static component - the classes, types, or structures of objects - and a dynamic component - the methods or functions and the evaluation of interactions. The design of an object-oriented system consists of (1) choosing appropriate objects and (2) specifying how the objects interact.

In converting from the landmark representation to HL parameters, four major classes of objects are involved: landmarks, features, articulators, and parameters. Landmarks may be one of vocalic, glide, consonantal closure, or consonantal release, where the consonantal landmarks occur in pairs. The landmark objects combine a time with a list of features, which are binary (+/−) values and an associated label. The articulator-bound features add an associated articulator, which may be either the primary or secondary articulator. The output parameter objects are lists of control points, where control points are time/value pairs. There are ten HL parameters: **f0, f1, f2, f3, f4, ag, an, ab, al,** and **ue** [7]. For each of the three different types of landmarks (vocalic, glide, and consonantal) the object-oriented procedure uses one underlying, basic landmark object. This basic object is a *generic landmark*, one that captures the behaviors and attributes intrinsic to any landmark. The generic landmark consists of attributes such as phonetic-features, preceding-phonetic-context, and following-phonetic-context, as well as information representing the prosodic context.

When rules are stated using traditional phonological notations, such as "A − > B / X _ Y", evaluation takes the form of searching a database of rules for any which are applicable to the current context. In essence, we begin by turning this system inside-out. Each landmark object notifies its predecessor and successor of its features, and is likewise notified by them of their features. Feature messages bypass landmarks unconcerned with that particular feature, providing a natural means of expressing underspecification. Within a landmark object, the cumulative effect of the incoming messages is to apply any applicable rules, sometimes causing additional features messages to be sent to the surrounding landmarks. The rules may therefore be stated subjectively from the viewpoint of each type of landmark, allowing an object-oriented system to simplify much of the rule organization and selection.

The list in Section 3 describes a group of interrelated constraints. It is not an ordered list of calculations. Instead, the system satisfies all of the constraints in a demand-driven order, that is, each calculation is performed only when the value it defines is needed (demanded) by some other calculation or

user action. An important aspect of demand-driven evaluation is that there is no need for the user to specify the particular order in which the computations are to be performed, as this sequencing is done automatically. For instance, if the user requests to display the LL parameter track AF (amplitude of frication), then the sytem demands the values for the oral-cavity pressure P_m and for the minimum cross-sectional area of an oral cavity constriction (acx) as seen in the equation for AF [5]: $AF = 20\log[K_f P_m^{1.5} acx^{0.5}]$, where K_f is a scaling factor. The pressure P_m is calculated using a low-frequency equivalent circuit model of the vocal tract [7]. The minimum area acx is determined by comparing the areas of the glottis, the tongue-body constriction, the tongue-blade constriction, and the lip opening.

Because the input to the synthesis process is in terms of feature markings and not areas and pressures, a range of values for the areas is (usually) possible, and the system calculates the values of P_m and acx that result for each choice of area allowed within the range (in appropriate increments, such as 1 mm²). A rule for "best" configuration (such as "the amplitude of the noise at the constriction is closest to the amplitude of voicing") is used to select automatically the values of P_m and acx, and therefore AF. In this way, the demand-driven feature of the sytem mirrors in some ways the behavior of a talker who adjusts the vocal-tract configuration in order to produce the "best" production of a series of phonemes.

5. SUMMARY

The development of this system has only recently begun, so there remain many unasked as well as unanswered questions, and our understanding will no doubt evolve along with the system. Nonetheless, we believe that this approach represents a significant step toward taming the complexity of existing models of speech production, and it may provide fresh insights into that complexity. For example, we can at this time only mention the apparent connection between feature geometries and object-oriented systems. This approach to speech synthesis also demonstrates the efficacy of both the HL and acoustic landmark representations of speech.

ACKNOWLEDGEMENTS

We gratefully acknowledge K.N. Stevens for helpful discussion during the preparation of this paper. This work was supported in part by NIH Grant R43 MH52358-01.

REFERENCES

[1] Wagner, M. (1988), *Understanding the ICAD System.* Cambridge, MA: Concentra Corp.

[2] Klatt, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Am. **82**, 737-793.

[3] Hertz, S.R. (1990), "The Delta programming language: an integrated approach to nonlinear phonology, phonetics, and speech synthesis", in *Between the Grammar and Physics of Speech (Papers in Laboratory Phonology I).* Eds. J. Kingston and M.E. Beckman. Cambridge: Cambridge Univ. Press.

[4] Local, J. (personal communication)

[5] Stevens, K.N. (forthcoming) *Acoustic Phonetics.*

[6] Stevens, K.N. (1993) "Lexical access from features", in *Speech Technology for Man-Machine Interaction.* Eds. P.V.S. Rao and B.B. Kalia. New Dehli: Tata McGraw-Hill, 21-46.

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

[8] Stevens, K.N., S.Y. Manuel, S. Shattuck-Hufnagel, and S. Liu (1992) "Implementation of a model for lexical access based on features", in *Proceedings of ICSLP*, October 12-16, Banff, Canada.

[9] 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**:2, 820-857.

[10] Bickley, C.A., K.N. Stevens, and D.R. Williams (1994), "A framework for synthesis of segments based on articulatory parameters", in *Conference Proceedings of ESCA/IEEE Workshop on Speech Synthesis*, Sept. 12-15, 1994, New Paltz, New York.