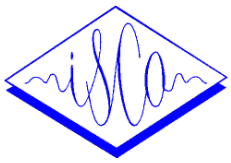

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

ON MODELLING THE PHONOLOGY PHONETICS INTERFACE FOR ARTICULATORY SYNTHESIS

Henrietta J. Cedergren*, Gilles Boulianne*, Danièle Archambault**

*Département de linguistique
Université du Québec à Montréal
C.P 8888 Montréal, H3C 3P8, Québec, Canada
**INRS-Télécommunications
3 Place du Commerce,Ile-des-Soeurs,H3E 1H6, Québec, Canada

ABSTRACT

The development of a research tool which will allow users to explicitly model the relationship between phonological representation, speech organization and physiological structure by means of an articulatory synthesis system which simulates the derivation of spoken French is described.

1. INTRODUCTION

As a result of recent technological developments, different speech synthesis systems have appeared on the market involving different techniques. The ability to produce speech using a computer is of primary interest to speech scientists such as phoneticians and linguists, who see in it an outstanding tool to aid in the detailed investigation of the complex phenomena of speech production [Browman et al., 1984; Coker 1976]. We are currently engaged in the development of a speech synthesis system which would be mainly a research tool for speech scientists.

Our primary goal is to increase our understanding of the relationship between phonological organization, speech and underlying physiological structure. The complex relationships between the different levels is investigated by means of an articulatory speech synthesizer which simulates each step involved in the production of spoken French, from the more abstract levels to that of the articulators themselves.

In this paper, we give a general description of the system. We also discuss the difficult question of the representation of the three levels of knowledge: phonological, phonetic and physiological, the "interface" between each type of representation and the general characteristics of the rule system.

2. AN INTERACTIVE SPEECH SYNTHESIS SYSTEM

The system is conceived as a scientific tool for researchers in speech and language. From this point of view, it is quite different from general public applications systems. Most text-to-speech systems strive to produce unlimited speech from any text and their applications vary from reading machines for the blind to access to large data banks via a telephone network [Allen et al.,1987]. Because they are designed for unsophisticated users, the underlying processes such as grapheme to phoneme conversion rules or the calculation of different acoustical parameters, for example, are usually not accessible to the users.

The different control parameters of our system are accessible to the users so that it should help researchers to attain a better understanding of the various multidimensional mechanisms involved in the control of speech, to test theories and to make use of the relationships between phonological representation and speech articulators. The user enters at a terminal a broad phonetic trancription of a word. From this entry, the system generates a phonological, a phonetic and an articulatory representation, before producing the actual synthesis. The user has the possibility to modify the different representations and to evaluate the results.

The system has three major components (see figure 1). First, a series of modules in which the different phonological, phonetic and articulatory knowledge structures are included. An algorithm which transforms the different articulatory positions into spoken sounds (the speech synthesizer itself). Finally, a central representation structure through which the different modules communicate with each other. The user has access to this structure and can modify it through the user interface.

At the outset, the central structure includes only the incomplete information given by the user. The different modules access this central structure, get information from it and then fill in the details until all levels of representation are complete. The organization of this structure in the form of a multidimensional delta was chosen to enable a maximum of flexibility in the representation of the different levels of information.

Each module contains a delta rule set. Rules at the phonological level account for categorical variation; at the phonetic level, rules intervene to produce assimilations, for example. We have chosen Delta as a programming language because it was specifically created to give linguists the ability to formulate and verify relatively easily different phonological and phonetic theories.

3. THE DELTA RULES

The primary characteristic of the system which has been developed is its openness. The rule sets of the system are accessible to the user for modification and/or progressive increase in numbers. Presently, the system consists of a limited rule set which allows the user to model the synthesis of simple V-C-V or V-C-C-V sequences, where the consonants are either stops or fricatives.

The rules of each module can be executed individually or as sets, thus, allowing the user to observe the evolution of the delta structure. The DELTA language version which we are currently using does not allow us to test the content of each phonological level in the delta structure. Therefore, the phonology-phonetics interface rules are currently implemented as macro definitions.

The rules of the phonological module assign the initial binary feature definitions on the phonological tiers aligned with the PHONEME tier. The assigned feature values are illustrated in Figure 2.

The phonetic rules translate the abstract phonological features into n-ary phonetic features and assign inherent duration to each segment by fixing its boundary in the time domain.

The rules of segmental duration which have been implemented,thus far, are quite simple. Each segment is assigned an initial inherent duration according to its manner of articulation, which is

modified under conditions of stress and position in a group of consonants. This is illustrated in Figure 3.

Rules of the physiological module translate the phonetic features into protoarticulator segments which occupy different tiers. Protoarticulators are underlying motor sensory goals that translate into many possible physiological states. The rules define their succession in the time domain using as points of reference the segmental boundaries set by the phonetic rules. In our preliminary rule set, the movement of almost all protoarticulators are referenced to the phonetic segmental boundaries. A small subset are referenced to protoarticulator events.

During the time span defined by the phonetic boundaries of each consonant, the protoarticulator segments are assumed to occupy a particular position in physiological space. Transition movements from the protoarticulator position defined for the preceding segment take place relative to the boundary of the current consonantal segment. Transition durations vary according to consonant type. Transition movements to the position of the following segment are referenced to the final boundary of the consonant. Figure 4 illustrates the typical trajectory of the AC (area of constriction) protoarticulator: transition movements toward and out of the defined segmental duration boundaries ds and fs are defined by the positive transition times t2 and t3.

At present, the rules concerning protoarticulator movements for vowels do not allow for the adjacency of several vowels. Vowel durations defined by the phonetic boundaries are not inherently static; vowel to consonant and consonant to vowel transitions take place within the phonetic temporal boundary span. Protoarticulator segments are assumed to occupy a position defined by the phonetic features.

The movements of the AG (glottal aperture) protoarticulator segment are referenced not only to the phonetic segmental boundaries, but also to defined glottal states. Figure 4 illustrates the two levels of synchronisation which define AG movement patterns. Initial AG movement is referenced as a transition t2 with respect to segment initial boundary (ds) and glottal aperture peak is situated at a time t4 before segment final boundary (fs). However glottal peak aperture duration and glottal adduction are referenced to onset of peak aperture. This manner of defining AG movement allows accounting for the observed differences of interarticulator synchronization during the production of stops and fricatives [Löfqvist & Yoshioka, 1984; Scully,1987] as a difference of temporal alignment of the onset of glottal peak aperture with segment final boundary, rather than alternative modes of glottal closure.

4. THE PHONETICS-PHYSIOLOGY INTERFACE QUESTION

The phonetics-articulatory modules interface is constrained by the qualitative differences in the characteristics of the elements of each module. Phonetic elements are discrete temporal segments, while protoarticulatory segments define continuous physiological regions. The passage from discrete to continuous space in the system under development is constrained by procedural criteria and is accomplished using classical optimization techniques: phonetic specification is translated into protoarticulatory targets (and thus physiological regions) that capture the acoustically important aspects of vocal tract shape; the position of an articulator within the physiological region is determined by an optimization algorithm constrained by principles of precision and economy of articulator movements.

4. CONCLUSIONS

As currently implemented, the knowledge structure of each module reflects fonctionally motivated theoretical decisions. The structure and content of the rule system may be modified to reflect different hypotheses on the nature of the phonology phonetics interface.

ACKNOWLEDGMENTS

This research was made possible through a grant by the Fond de développement académique du réseau de l'Université du Québec.

REFERENCES

Allen, J. et al. (1987), From text to speech: the MITalk system, Cambridge University Press, Cambridge.

Browman, C.P. et al. (1984), "Articulatory synthesis from underlying dynamics", J.Acous.Soc. Am., vol.75,S22-S23.

Coker, C.H. (1976), "A Model of Articulatory Dynamics and Control", Proceedings of the IEEE, vol.64,no.4.

Löfqvist, A. & Yoshioka, H. (1984), Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. Speech Communication 3, 279-289.

Scully, C. (1987), Linguistic units and units of speech productions. Speech Communication 6, 77-142.

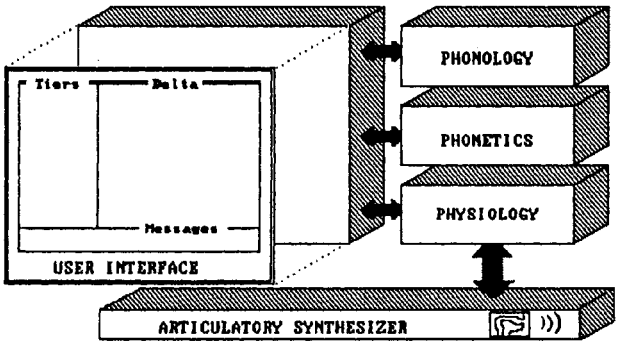

Figure 1. System components

PHONEME:		A	t		i	
COR:			+			
VOIS:			-			
ANT:			+			
CONT:			-			
SON:			-			
NASAL:		-	-		-	
LAT:			-			
ARRIERE:		+			-	
ARROND:		-			-	
HAUT:		-			+	
BAS:		+			-	

Figure 2. Example of phonological tiers

PHONEME:		A	t		i	
anter:			alv			
apical:			apic			
haut:		bas			haut	
interrupt:			occl			
nasal:		oral	oral		oral	
arrond:		arr			ecart	
ant:		post			ant	
racine:		natr			atr	
MS:		40	80		40	

Figure 3. Example of phonetic tiers

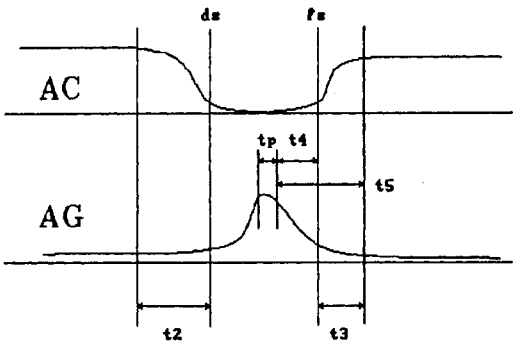

Figure 4. Protoarticulator timing