ARTICULATORY AND COGNITIVE CONSTRAINTS IN SPEECH SYNTHESIS

S.J. Butler

Department of English and Linguistics
Macquarie University

ABSTRACT - The representation of constraints on speech production in speech synthesis models is considered. Constraints acting at a cognitive level and the processes which give rise to them are focused upon. The acoustic properties of vowel spaces are shown to be a product of this type of constraint.

INTRODUCTION

An important task of speech synthesis systems is to model constraints that exist in normal speech production. Only by exploiting these constraints can the requirements for speech data collection and storage be reduced to within practical limits. One identifiable source of constraints is the physiological limitations of the speech articulators. It has often been assumed that articulatory models would offer the most attractive approach to speech synthesis due to the explicit incorporation of those constraints. In reality, most successful speech synthesis systems have been based upon acoustic models of speech.

The relative difficulty in collecting articulatory speech data in comparison with acoustic data is perhaps one practical reason for this. This paper arose from a desire to more rigorously question the merits of articulatory models in speech synthesis from a theoretical point of view. The results are inconclusive in the sense that a clear preference for either articulatory or acoustic synthesis models does not emerge. However, some useful perspectives on the problem of modeling speech production constraints arise.

COGNITIVE CONSTRAINTS

When the range of articulatory gestures used in speech production is examined closely, it becomes evident that this range includes only a subset of the physiologically possible gestures. This claim is supported by the work of Lindblom (1983) who found that even a simplified articulatory model generated far more articulatory configurations than are found in speech data. Evidently, extreme articulatory settings are avoided in normal speech suggesting that articulatory constraints are often not active in shaping patterns of speech sounds. A useful way in which this fact may be restated, is that there are additional speech specific constraints which determine the articulatory gestures used in normal speech. Since speech is a learned behaviour, these constraints must arise at a cognitive level and for this reason they will be termed cognitive constraints. The consequence of these constraints is that individual articulators are no longer controlled independently but are co-ordinated together to perform specific functions (Fowler et al, 1980). It would seem advantageous to exploit these constraints in speech synthesis models. To achieve some understanding of which type of synthesis model might more easily incorporate the cognitive constraints, the process by which these constraints arise needs to be considered.

A useful insight into this process is gained if it is presupposed that the constraints are arrived at by selecting articulatory gestures according to an objective criteria. Figure 1 illustrates two examples of selection criteria which give rise to alternative subsets of the available articulatory gestures, each of which defines an alternative cognitive constraint. The two types of criteria shown represent a logical division which separates those selection criteria in which motor considerations are the dominant influence from those in which acoustic considerations are dominant. It is tacitly assumed that the terms "acoustic" and "perceptual" are interchangeable in the present context. An illustrative example of a physiological criteria is the principle of minimum articulatory effort (Lindblom, 1983) by which the articulatory gesture requiring minimum expenditure of energy is selected from available alternatives. This principle is one potential explanation for the avoidance of extreme articulatory settings in normal speech. The quantal theory of articulation (Stevens, 1972) is an example of an acoustic selection criteria. According to the quantal theory, articulatory configurations are selected according to the relative insensitivities of their acoustic properties to perturbations in articulatory settings. Stevens has, for instance, used this theory to predict observed consonant places of articulation. It seems probable that both physiological and acoustic considerations influence the selection of articulatory gestures in speech. This third alternative is represented in Figure 1 by the intersection of the previously discussed subsets, defining an appropriate cognitive constraint. To see how this might work with respect to the previous examples, consider the well known variation of velar consonant place of articulation with vowel context. The quantal theory, as mentioned earlier, is able to define an optimal place of articulation for velar consonants. A common explanation for the variation of place of articulation in different vowel contexts is that the place of maximum constriction of the vowel influences the optimal place of articulation of the consonant according to a principle of minimal articulatory distance. The simultaneous application of physiological and acoustic criteria would predict both the nominal place of articulation and the observed variation due to vowel context.

With this understanding of the relationship between cognitive constraints and the criteria which give rise to them, the manner in which the constraints might be represented in a speech synthesis model can be addressed. A relevant point to note is that the transformation between articulatory configurations and acoustic properties of a given speech gesture is a highly complex and non-linear one. The implication of this is that where a cognitive constraint has a straightforward representation in terms of articulatory configurations, there is no guarantee that its representation at the acoustic level is also straightforward. Of course, the converse of this is equally true. By virtue of this, it is reasonable to expect that physiological criteria will lead to constraints which are most easily represented in terms of detailed models of the articulators. Similarly, acoustic criteria suggest the use of acoustic speech synthesis models. Where both physiological and acoustic criteria are involved, as seems to be the case for normal speech, a preference between articulatory and acoustic synthesis models would depend on which is the dominant criteria. Unfortunately, the criteria which give rise to cognitive constraints in speech are only vaguely understood at present and the preferred type of model remains an open question.

VOWEL CONSTRAINTS

An important acoustic property of vowels is that the first three formants for an individual speaker are clustered around a two-dimensional surface. A piecewise-planar approximation to this surface was determined by Broad and Wakita (1977) which is illustrated by third formant contours in Figure 2. This surface represents a cognitive constraint on vowel production. This must be the case, since there are numerous physiologically possible articulatory adjustments which would allow greater utilisation of the formant space. Particular examples are lip rounding, larynx height and retroflexion. From earlier discussion, it is expected that a selection criteria can be associated with this cognitive constraint. A plausible acoustic criteria is suggested by observing that the projection of the vowel surface onto the first and second formant space leads to an unambiguous relationship between vowel quality and formants. This is consistent with the common notion that it is the first two formants which are important for vowel perception. Evidently, vowel articulations are selected to achieve maximum utilisation of the two-dimensional space without ambiguity. This only defines a partial criteria, for it is also necessary to explain the behaviour of the third formant in terms of the selection criteria. A more complete explanation can be obtained by considering a simple articulatory model of vowels developed by Butler and Wakita (1982). This model presupposes that all vowel articulations are intermediate between the extreme cardinal vowels /i/, /a/ and /u/. Each vowel articulation is characterised by the logarithim of its area function which is calculated by linear interpolation between extreme vowel articulations according to the equation:

$$\log A(x) = a(1)*\log A(x,/i/) + a(2)*\log A(x,/a/) + a(3)*\log A(x,/u/) \quad (1)$$

where

$$a(1) + a(2) + a(3) = 1 \quad (2)$$

Equation (2) implies that the third model parameter is dependent on the first two so that there are two degrees of freedom in the model. Vocal tract length is interpolated in a similar manner. That such a simple model is capable of correctly reproducing features of the observed vowel formant space, can be observed by comparing model generated formants (Figure 3) with the Broad and Wakita data (Figure 2). Apparently, the cognitive constraints applying to vowels are captured in the model. Linear interpolation of vowel articulations can be interpreted in a useful way if it is noted that a linear path in articulatory space between vowel configurations is, in fact, the shortest path. If minimisation of energy expenditure in moving between articulatory configurations is important, then the most direct path in articulatory space would be an obvious candidate. It seems likely then, that a physiological criteria such as the principle of minimum articulatory effort plays an important role in determining the vowel constraints. In combination with the acoustic criteria of maximum utilisation of the two-dimensional formant space, a complete picture of the origin of vowel spaces can be drawn.

REFERENCES

BROAD, D.J., WAKITA, H. (1977) "Piecewise-planar Representation of Vowel Formant Frequencies", J. Acoust. Soc. Am. 62, 1467-1473.

BUTLER, S.J., WAKITA. H (1982) "Articulatory Constraints on Vocal Tract Area Functions and Their Acoustic Implications", J. Acoust. Soc. Am. 72, Suppl. S79(A).

FOWLER, C.A., RUBIN, P., REMEZ, R.E., TURVEY, M.T. (1980) "Implications for Speech Production of a General Theory of Action" in Butterworth, B. (Ed), Language Production, Academic Press.

LINDBLOM, B. (1983) "Economy of Speech Gestures" in Macneilage, P.F. (Ed), The Production of Speech, Springer-Verlag.

STEVENS, K.N. (1972) 'The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data" in Denes, P.B., David, E.E. (Eds.), Human Communication: A Unified View, McGraw-Hill.

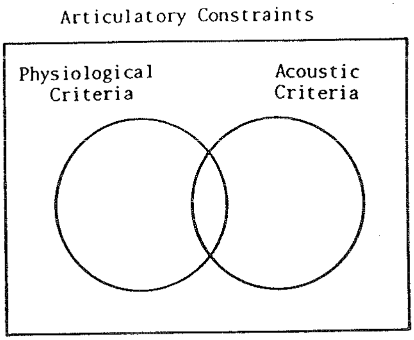

Figure 1. Articulatory gesture selection

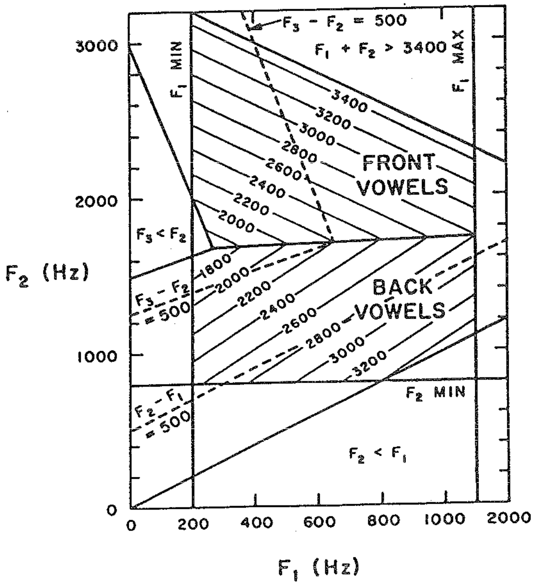

Figure 2. Piecewise planar vowel surface
(Broad and Wakita, 1977)

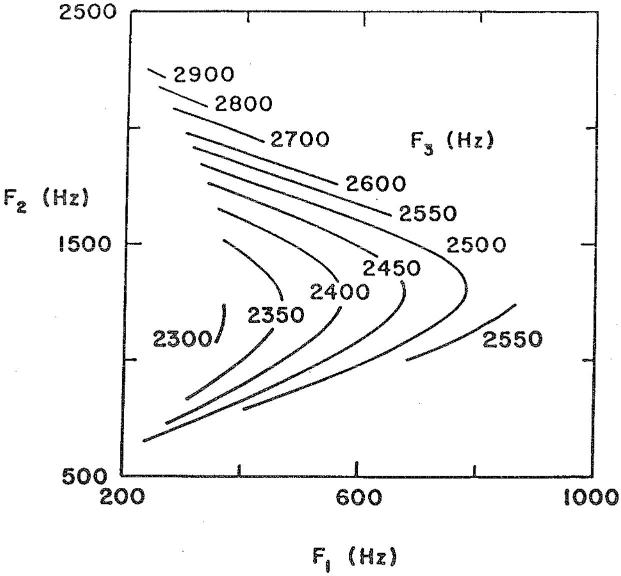

Figure 3. Acoustic properties of vowel model
(Butler and Wakita, 1982)