A Speech Synthesis System Based on Articulatory Modelling

Simon J. BUTLER

Speech, Hearing and Language Research Centre
Macquarie University
Sydney, New South Wales 2109, Australia

ABSTRACT : The elements of an articulatory synthesis system under development are described. Particular emphasis has been given to modelling the trans-consonantal coarticulation effect for stops in /V_1CV_2/ context that have been reported by Öhman (1966).

INTRODUCTION

One of the important roles of the rule component of a speech synthesis system is to reproduce the range of phenomena commonly referred to as coarticulation. Speech is often synthesised by generating smooth transitions between formant or articulatory targets for adjacent phonemes. The complexity of coarticulation rules is a result of the observation that neither the targets, rates of transition or timing of the transitions are completely independent of phonetic context. It has often been suggested that coarticulation rules may have a simpler representation when expressed in terms of the movements of articulators. One interesting possibility in this approach is that the transition rates between articulatory targets might be systematically related to the biomechanical properties of individual articulators. For instance, articulator motion has been conceptualised as the result of step-like changes in the motor commands to the articulators which are "smoothed" by a biomechanical lowpass filtering action (Lindblom, 1967). If this is the case, then coarticulation rules for articulatory synthesis would only need to specify a sequence of appropriately timed targets for each articulator, the transitions being generated by lowpass filters characteristic of the particular articulator.

A more general expression of these ideas is Coker's (1968) notion that vocal tract dynamics can be characterised through a separation of variables each of which is represented by an articulatory parameter. The independence of these parameters not only allows a characteristic "smoothing" filter to be associated with each parameter, but the targets when expressed as model parameters become context invariant. One apparent result of this process is that the motions of individual articulators require different parameters for vowel and consonant articulations. The lips, for instance, require one parameter for the lip rounding of vowels and another for consonantal lip closure. These parameters have characteristic filters with very different response times. The coarticulation model of Öhman (1967) achieves a similar "decoupling" of the dipthongal and consonantal components of articulation without explicitly modelling individual articulators. Coarticulation is modelled as slow dipthongal transitions in the whole vocal tract shape upon which are superimposed localised perturbations due to consonant transitions. Öhman's simple model of /VCV/ stops would seem to offer a more detailed account of coarticulation than most speech synthesis systems.

The aim has been, in the first instance, to implement a synthesis system capable of generating the coarticulation phenomena that Öhman (1966) has reported for stop consonants in asymmetrical vowel context. This sytem includes an accurate simulation of the vocal tract acoustics (the vocal tract model), an adaptation of Öhman's numerical model of coarticulation in terms of vocal tract area functions (the articulatory model) and a graphical user interface (the parameter editor) for generating parameter contours for the articulatory model.

THE VOCAL TRACT MODEL

A basic requirement for articulatory synthesis is an accurate simulation of the acoustics of the vocal tract. The most sophisticated vocal tract models are lumped element electrical network analogs of sound propogation in the vocal tract (Flanagan, 1972) which may be simulated by difference equations on a computer. As a starting point, a computer implementation of a network analog of the vocal tract

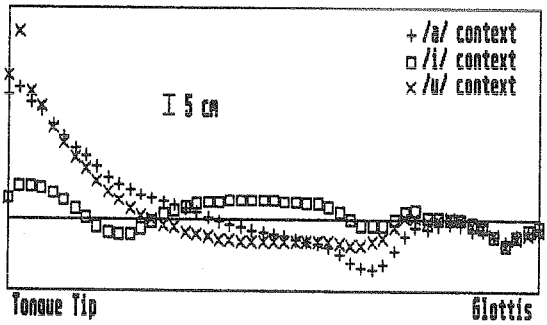

Figure 1: Öhman's perturbation function $[v(x,t) - c(x)]w_c(x)$] for alveolars in symmetrical vowel context.

due to Maeda (1982) was used. This model includes both oral and nasal tracts and allows voiced excitation of the vocal tract by external modulation of the glottal area. To provide the voiceless excitation for consonants, automatic generation of noise in the vocal tract as described by Flanagan and Ishizaka (1976) has been added to the Maeda model. As a result, if the flow velocity (or more correctly, the Reynolds number) exceeds a critical value at any point in the vocal tract, a noise source at that location is activated.

To synthesise speech, the vocal tract model requires a specification of vocal tract area function, subglottal pressure, degree of nasal coupling and the time variation of glottal area. The subglottal pressure and nasal coupling have a one-to-one correspondence with respective parameters of the articulatory model. At present, the time varying glottal area consists of pulses of constant shape according to a suggestion of Maeda (1982). The fundamental frequency and glottal amplitude of these pulses are controlled directly by parameters of the articulatory model. A fundamental frequency of zero generates a constant glottal opening for voiceless consonants. This approach does not, however, provide an adequate range of glottal conditions for speech synthesis, particularly during transitions between voiced and voiceless excitation. For this reason, the two-mass vocal chord model of Ishizaka and Flanagan (1972) is being implemented to provide a more natural simulation of glottal conditions. The two control parameters of the two-mass model (resting glottal area and vocal chord tension) will directly replace the existing Glottal Amplitude and Fundamental Frequency parameters.

The simulation uses 37 vocal tract sections with lengths that vary in the lip and glottal regions to reflect lip rounding and larynx height adjustments. This also satisfies a practical requirement of the vocal tract model that the velum remain a constant number of sections from the ends of the tract, despite vocal tract length changes. The large number of sections is required to minimise a form of spectral distortion which results from approximating the vocal tract acoustics by a lumped equivalent circuit. For similar considerations of numerical accuracy, a 20KHz sampling rate is used.

THE ARTICULATORY MODEL

The articulatory model is an attempt to use the concepts of Öhman's numerical model of stop coarticulation in $/V_1CV_2/$ context to synthesise more complicated utterances. This includes consonant clusters involving stops and fricatives but excludes the liquids. Öhman summarised the coarticulation of alveolar stops in asymmetrical vowel context by the following formula:

$$s(x,t) = v(x,t) + k(t)[v(x,t) - c(x)]w_c(x) \quad (1)$$

The midsaggittal distance $s(x,t)$ is the result of the perturbation of an underlying vowel shape $v(x,t)$ by a characteristic consonant shape $c(x)$. The effect of the perturbation is localised to the consonant's place of constriction by the weighting function $w_c(x)$ and the time course of the perturbation is controlled by $k(t)$. Each of these latter two functions have values in the range zero to one. The underlying

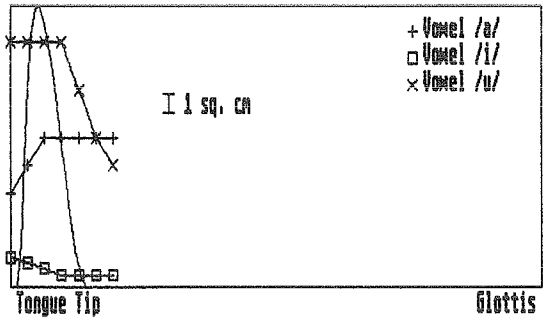

Figure 2: Simplified perturbation function $C(x)$ compared to vowel area functions.

vowel shape $v(x,t)$ is also in transition during the consonant and resembles a dipthongal articulation between the preceding and following vowels. Öhman represented the vowel shape as an interpolation between the vocal tract shapes of the extreme cardinal vowels (/i/,/a/ and /u/) according to two vowel parameters.

As an alternative to midsaggital distances, vocal tract shapes can be represented by vocal tract area functions. It is generally accepted that the two representations are directly related, albeit by a complex transformation. The principle advantage of using area functions is that they are in the required form to synthesise speech using vocal tract simulation. The approach of the articulatory model has therefore been to re-express Öhman's equations in terms of area functions. It has been shown (Butler and Wakita, 1982) that interpolation of the logarithm of cardinal vowel area functions realistically models observed vowel formant spaces. Unlike Öhman's description of midsaggital distances, the area function formulation for vowels implicitly includes the lip region of the vocal tract, reflecting the fact that the entire articulatory system is involved in vowel articulation.

One aspect of Öhman's formula (equation (1)) for consonant perturbations is that two functions are required to specify the consonant. If the notion is accepted that consonant perturbations are strictly localised, then away from the region of constriction $w_c(x)$ will be close to zero and $c(x)$ will have an arbitrary value. The perturbation function $[v(x,t) - k(t)c(x)]w_c(x)$ predicted by Öhman's model is plotted in Figure 1 for the symmetrical vowel contexts /ada/,/idi/ and /udu/. Away from the constriction, the two consonant functions do have non-trivial values which represent, at least in part, physiological constraints between the tongue tip and tongue blade. For the purposes of speech synthesis, however, there are two reasons why the perturbations away from the tongue tip can be ignored. Firstly, some passive expansion of the supraglottal cavity occurs during stops due to an increase in supraglottal pressure and is known to be essential for the maintenance of voicing during voiced stops. This might account for the consistent shape of perturbations in the glottal region of Figure 1. This is already accounted for in the vocal tract model by a wall equivalent circuit and does not require explicit representation in the articulatory model. Secondly, for tightly constricted vocal tract shapes, the sensitivity of vocal tract resonances (and hence formants) to perturbations in vocal tract area is most pronounced in the vicinity of the constriction. It is therefore unlikely that the perturbations behind the constriction significantly affect the acoustic properties of consonants.

Simplications of the Öhman formula have therefore been considered which approximate the consonant perturbations using strictly localised functions. One possibility, illustrated in Figure 2, can be represented by the following formula expressing perturbations of vocal tract area $A(x,t)$:

$$\begin{aligned} A(x,t) &= V(x,t) - k(t)C(x) \quad \text{if } V(x,t) > k(t)C(x) \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned} \tag{2}$$

The saturation that occurs when the consonant perturbation $k(t)C(x)$ exceeds the underlying vowel

Vowel Quality	High/Low	a1
	Back/Front	a2
Consonant closure $k(t)$	Bilabial	a3
	Alveolar	a4
	Velar	a5
Nasalisation	Nasalisation	a6
Glottal Excitation	Subglottal Pressure	a7
	Fundamental Frequency	a8
	Glottal Amplitude	a9

Table 1: Articulatory Model Parameters

area might reflect contact pressure during stop consonant closure which is needed to counteract increased supraglottal cavity pressure. The area functions of the three extreme vowels in the region of constriction define the closure in Figure 2 which shows that consonant closure occurs for values of $k(t)$ which depend on the underlying vowel shape. Interestingly, Lindblom (1967) has put forward this kind of model as an explaination of the differences in duration of high and low vowels. Öhman's formula, on the other hand, only allows closure for $k(t) = 1$, irrespective of the underlying vowel.

The benefit of a simplified perturbation function arises in the case of velar stop consonants. For these consonants, the location of the constriction in the vocal tract is not independent of the underlying vowel (Öhman, 1966, 1967). While Öhman (1967) generalised the consonant functions $c(x)$ and $w_c(x)$ to reflect this dependence (viz. $c(x, v)$ and $w_c(x, v)$), it is no longer possible to uniquely derive these functions from articulatory data as Öhman was able to do for the alveolar case. In the case of the simplified area function formula (equation (2)), the shape of $C(x)$ is assumed to be constant for different underlying vowels but is shifted along the vocal tract to the correct constriction location for the underlyng vowel. This location is determined by interpolating between constriction locations of the extreme cardinal vowels. It may be noted that the context dependence of the perturbation process for velar consonants is "hidden" within the articulatory model, so that the separation of variables between vowel parameters and consonant parameter $k(t)$ is maintained.

The parameters of the articulatory model are given in Table 1. Two parameters control vowel quality and roughly correspond to vowel height and fronting. For each of the three stop consonant places of articulation, the value of $k(t)$ in equation (2) defines a corresponding model parameter. Finally, nasalisation and glottal excitation parameters are used directly by the vocal tract simulation. As has been noted, however, the Fundamental Frequency and Glottal Amplitude parameters are to be replaced by two parameters of the two-mass vocal chord model.

THE PARAMETER EDITOR

As an experimental tool for generating the articulatory parameters to synthesise speech, a graphically oriented user interface was developed, referred to as a parameter editor. This editor displays the time course of each model parameter in a separate window and allows parameter contours to be generated graphically using mouse input. The timing of an utterance is established first by placing a number of time markers (Figure 3(a)). Between any two markers, the parameter contour can be set to a constant value achieving step-like changes of the parameters at the marked time points. Smooth transitions between any two marked time points can then either be sketched (using the mouse), linearly interpolated or smoothed by a lowpass filter. All three methods are illustrated in Figure 3(b). Time markers can be saved in files and reloaded to allow different time plans for different parameters.

Whilst arbitrarily complex coarticulation rules can be manually simulated with the parametric editor, the motivation of this study has been to use simple lowpass filtering of model parameters. As a simple illustration, Figure 4 shows parameter contours for the Back/Front and Alveolar Closure parameters that are appropriate for synthesis of the utterance /udi/. The underlying dipthongal gesture from /u/ to /i/ has been assumed to commence simultaneously with transition to consonant closure. Consonant closure occurs before the consonant transition reaches its target, according to the simple formulation of consonant perturbation of equation (2). Critically damped second order lowpass filters with appropriate time constants are used for smoothing.

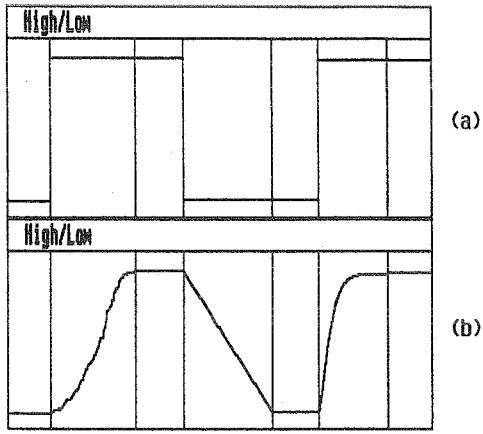

Figure 3: Methods for Generating Transitions using Parameter Editor.

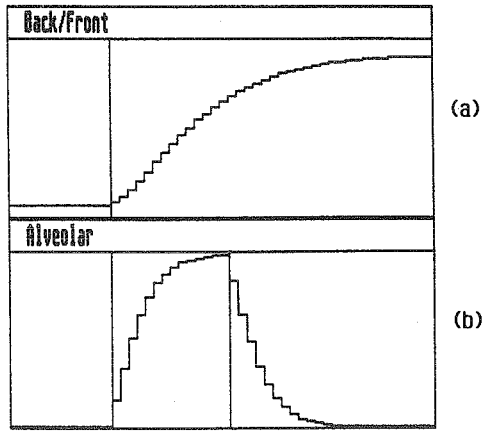

Figure 4: Parameter Contours for Utterance /ida/ (Total duration 250 ms).

DISCUSSION

The appearence of independent parameters for vowels and consonants in the articulatory model implies that they may be coproduced. Similarly, consonants with different places of articulation have separate independent parameters and also may be coproduced. Experimental data on anticipatory coarticulation (Kozhevnikov and Chistovich, 1965) indicates that some articulatory features such as lip rounding for rounded vowels may begin as early as the first consonant in the preceding consonant cluster. Öhman's data and model (which exclude the lips) indicate that an underlying dipthongal gesture toward the post-consonantal vowel takes place during stops which involves the vocal tract shape as a whole. In the articulatory model presented here, lip rounding does not have an explicit representation and anticipation occurs because the lips are constrained to take part in the underlying dipthongal gesture. There is, therefore, phenomenological agreement with the Kozhevnikov and Chistovich findings.

At the present time, no specific account is given in the articulatory model of fricative consonants. Öhman (1967) has suggested that fricative constrictions might be represented by a target value of $k(t)$ that is less than unity. Since fricative consonants are characterised by accurate control of constriction size, this would imply a target value that is dependent on vowel context, at odds with the notion of separation of variables. A current proposal is to use an additional fricative perturbation $k_2(t)F(x)$ (which for alveolar fricatives would represent tongue grooving) controlled by a distinct articulatory parameter $k_2(t)$ as follows:

$$\begin{aligned} A(x,t) &= V(x,t) - k_1(t)C(x) + k_2(t)F(x) \quad \text{if} \quad V(x,t) > k_1(t)C(x) \\ &= k_2F(x) \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{3}$$

This solution would require two additional fricative parameters in the articulatory model for the alveolar and bilabial closures. A bilabial fricative is a suitable approximation to the labio-dental fricatives for the purposes of speech synthesis.

REFERENCES

Butler, S.J. and Wakita, H. (1982) "Articulatory constraints on vocal tract area functions and their acoustic implications", J. Acoust. Soc. Am., 72, Suppl. S79(A).

Coker, C.H. (1968) "Speech synthesis with a parametric articulatory model", Kyoto Speech Symposium, Paper A-4.

Flanagan, J.L. (1972) *Speech Analysis, Synthesis and Perception*, 2nd Ed., Springer-Verlag, New York.

Flanagan, J.L. and Ishizaka, K. (1976) "Automatic generation of voiceless excitation in a vocal cord-vocal tract speech synthesizer", IEEE Trans. Acoust. Speech Sig. Proc., 24, 163-170.

Ishizaka, K. and Flanagan, J.L. (1972) "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell Syst. Tech. J., 51,1233-1268.

Kozhevnikov, V.A. and Chistovich, L.A. (1965) *Speech: Articulation and Perception* (English translation from Russian, U.S. Dept. of Commerce, Washington).

Lindblom, B. (1967) "Vowel duration and a model of lip mandible duration", Speech Transmission Lab. STL-QPSR 4/1967, 1-28.

Maeda, S. (1982) "A digital simulation method of the vocal tract system", 1, 199-229.

Öhman, S.E. (1966) "Coarticulation in VCV utterances: Spectrographic measurements", J. Acoust. Soc. Am., 39, 151-168.

Öhman, S.E. (1967) "Numerical model of coarticulation", J. Acoust. Soc. Am., 41, 310-320.