Proceedings of The Institute of Acoustics

SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

Celia Scully

Dept. of Linguistics & Phonetics, University of Leeds, Leeds, U.K.

INTRODUCTION

Where different speakers show small differences of acoustic structure for the same broadly defined auditory goals, it may be supposed that these arise, in part, from their different individual patterns of articulator kinematics. One approach to the characterisation of rules for covarying acoustic pattern features across different speakers saying the same words of English is with analysis-by-synthesis, using a model of the physics of speech production. There is a need to go beyond linear acoustic theory, so that interactions between one acoustic source and another and between source and filter may be taken into account.

Flanagan et al. [1] employed a parametrically controlled model of speech production for the economic description of an [aɪ] diphthong said by a single speaker. Parameters in the model were adapted using criteria of minimum errors in the acoustic domain, comparing the model's output with the natural speech to be matched. The computation time required was very great, even though the analysis was limited to vowel configurations, with a rather long sample time of 12.8 ms. The capability of their speech production model to synthesise consonants was not used in this study. As the authors point out, a full understanding of articulatory contraints and acoustic behaviour of the system is lacking, especially for consonant articulations. Our aim is to gain insight into speaker-to-speaker variations in acoustic signal, rather than to investigate new methods for low bit-rate coding of speech. A composite model of speech production processes is used in which sequences containing consonants as well as vowels can be synthesised and identified by listeners. The functional models for voice and turbulence noise source generation and the kinematic descriptions of articulation are flexible, so that different speaker types may be modelled. The usual values of the parameters which determine the time paths of articulators are based on data from natural speech. The aerodynamic stage of speech production is accessible in the model, so that comparisons between synthetic and natural speech are not limited to those for acoustic outputs. A descriptive framework for the organisation of timing of articulatory events and their coordination across quasi-independent articulators has been developed [2,3]. This last simulates the planning stage of natural speech. It exhibits a rather high level of complexity. Currently about 7 or 8 expressions of coordination are needed for each phonetic unit, whether this is a vowel element or a consonant. Figure 1 shows

the events (E) and coordinations (D) proposed for voiced or voiceless fricatives. In its formal rules, this planning stage is at present underconstrained. Constraints on vocal tract configurations are introduced informally, based on data from real speech and principles such as constant tongue volume.

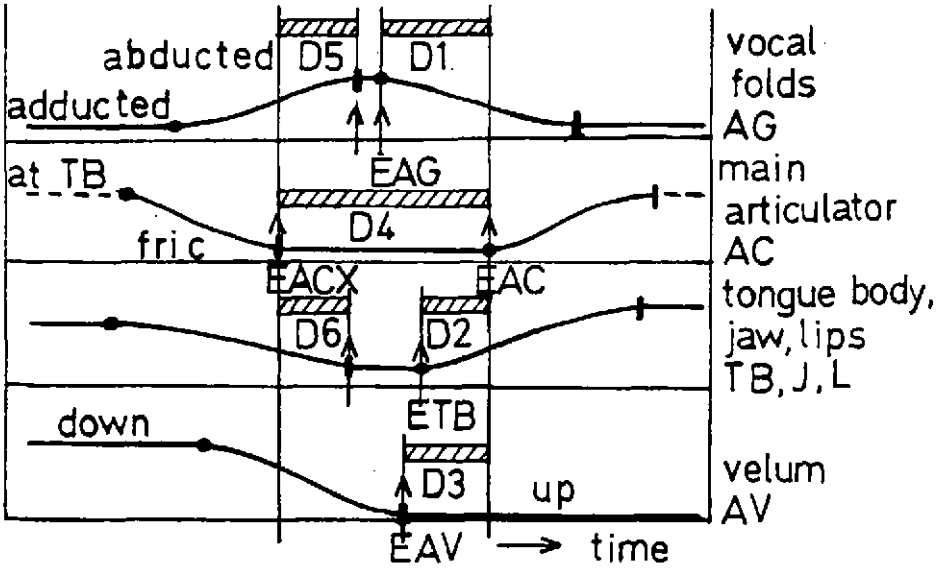

Fig.1 Articulatory transitions and coordination for a fricative consonant.

Simplifications that invoke concepts of natural phonetic classes [4, 5] need to be introduced, to characterise a number of different real speakers of English.

The model has a phonetic orientation, with the main turbulence noise and transient source generating constrictions specified. These are regions of small cross-section area across which a significant pressure drop develops. Aerodynamic data from real speech can give estimates for part of the articulatory descriptions required in the modelling, the time function of specified constriction areas. The orifice equation is used, with an empirical constant [6], viz.

$$A = \frac{0.00076\ U}{\Delta P^{0.5}} \qquad (1)$$

A = cross-section area of constriction in cm^2
U = volume flow rate of air through the constriction in cm^3/s
ΔP = pressure drop across the constriction in cm H_2O

Constriction areas AG, AC and AV, as shown in Figure 1, together

with cavity volumes for the lungs and the vocal tract, link the articulatory and aerodynamic blocks of the model. Estimates from real speech, input to the model, give resynthesised aerodynamic traces as outputs. These can be compared with real speech traces, generalising beyond the original contexts. The data here were obtained as in a previous study, using an earlier form of the model [7]. A trace proportional to area A was obtained by means of the Aerodynamic Speech Analyser (Electronic Instrument Design, Leeds). The traces were all LP filtered at 50 Hz to partially remove the a.c. components. Undoubtedly, the methods used are only approximate and there are many sources of error [8]. Undoubtedly, also, the model is a highly simplistic representation of the complexities of speech production. However, sharing the hope of Bridle et al. [9] that "Good, robust solutions to dramatic simplifications of a real problem can be more useful than weak solutions to a more accurate idealisation of the problem", we want to try to use in the articulatory domain something comparable to their speaker-adaptive procedures in the acoustic domain. It is hoped that, with sufficient data from natural speech, solutions to many simultaneous equations might be optimised and the results used to characterise a particular speaker, for the purpose of articulatory synthesis. Some examples of results obtained follow.

ARTICULATORY PATTERNS OF NATURAL SPEECH

Tongue tip-blade articulation for [s] and [z] in 3 vowel contexts

The phonetic class considered here is alveolar fricatives, both voiced and voiceless. Can voiced and voiceless fricatives be lumped together? Does the vowel context affect the time course of the main vocal tract constriction A_c? Are the patterns the same for different speakers? Here four adult English speakers with near-RP accents are analysed: two women A, B; two men C, D. Figure 2 shows traces of alveolar constriction A_c articulation. They exhibit the kind of speaker-specific complexity that might be anticipated on the basis of other analyses [10]. Speaker A seems to use similar articulatory paths regardless of voicing or vowel context. The other speakers could be modelled as having a long static occlusion for [s] in some contexts; for [z] also in the case of speaker D. Some of the traces for [s] have double troughs, which suggests that the actions may be more complex than simple closure-occlusion-release. There may be oscillatory paths, possibly mediated by feedback control. But alternative interpretations associated with sources of error in the experimental techniques [8] need to be explored first, before the tentative explanations offered here are considered more carefully.

Invariance across a change of speaking rate

As a first step to considering what different speakers might maintain constant across different styles of speech, traces of words in isolation said at medium and fast rates by two different speakers E and F, both women, are shown in Figure 3. There seems to be overshoot at the fast rate for both speakers. At the medium rate, speaker E makes a noticeably shorter occlusion for "eyes" than for

SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

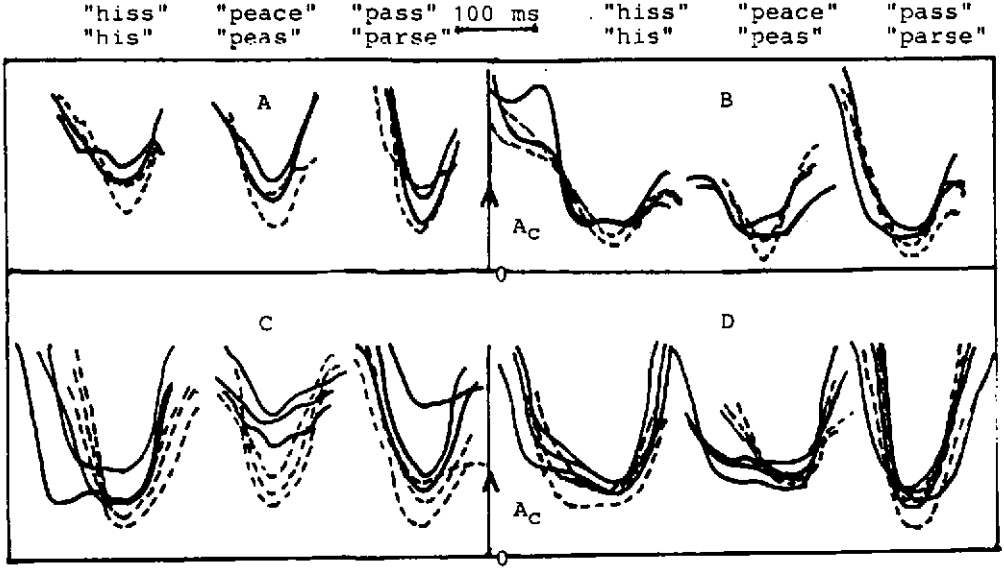

Fig.2 Tongue tip-blade articulation A_C for words in the frame "A ... it said", for 4 speakers. Solid lines [s]; dashed lines [z].

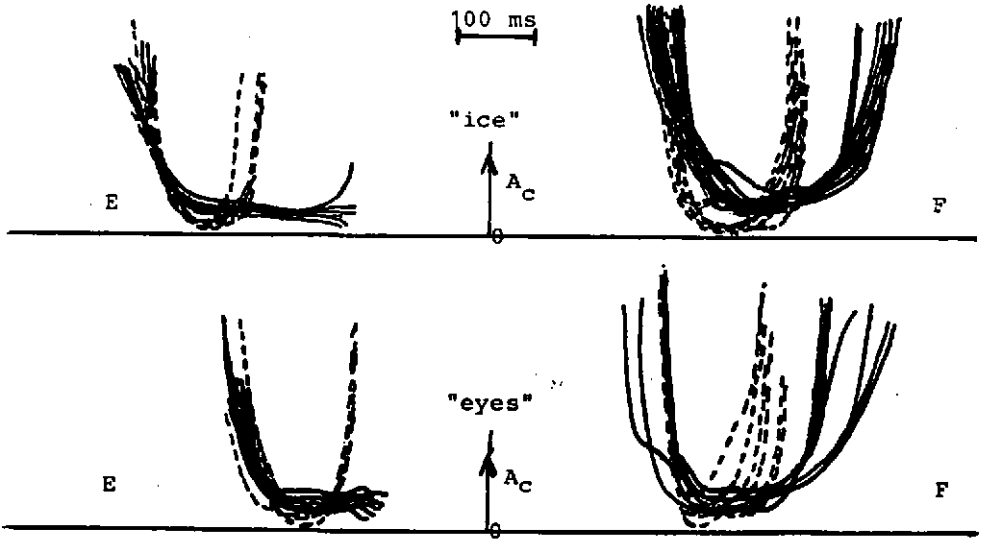

Fig.3 Tongue tip-blade articulation for words in isolation, for 2 speakers E and F. Solid lines medium rate; dashed fast.

"ice", but her patterns for the two words seem identical at the fast rate. For speaker F at the medium rate, most but not all tokens of "eyes" have shorter occlusions than those for "ice"; a small difference is maintained even at the fast rate.

Larynx articulations for vowels

It is clear that larynx actions are of central importance in speech production. Unfortunately, only invasive techniques are currently available for transducing vocal fold articulation and subglottal pressure. Pursuing the philosophy of dramatic simplification, subglottal pressure is estimated here from the more accessible variable of oral pressure. The phonetic element whose larynx articulation is to be estimated is embedded in voiceless fricatives or aspirated plosives, preferably in a high vowel context, where cavity volume changes due to jaw movements are minimised [11]. The reliability of the method has been discussed [12, 13]. Subglottal pressure in a vowel is assumed to equal oral pressure in an adjacent voiceless consonant. This method has been used to characterise different kinds of singing voices [14]. Combined with an airflow measure, it has been used to derive glottal area during vowels for dysphonic speakers before and after treatment [15]. For the 4 speakers A,B,C and D, peak oral pressure for [s] in "peace" or "pass" in the frame sentence "A ... it said" was taken to indicate subglottal pressure $\hat{P}_{sg}$ in the preceding [i] or [ɑ] vowel. The minimum value of oral airflow during the vowel (nasal airflow was zero) $U_{O\ min}$ gave an estimate for transglottal airflow; here oral pressure was close to zero, so that glottal area was estimated as:

$$\hat{A}_g = \frac{0.00076\ U_{O\ min}}{\hat{P}_{sg}^{\ 0.5}} \qquad (2)$$

Table 1 shows the minimum and maximum for 3 tokens of each item for each speaker.

Table 1. Estimates of the articulatory (d.c.) component of glottal area during 2 vowels for 4 speakers

Vowel	Speaker	$\hat{A}_g$ range (3 tokens) in cm^2
[ɑ]	A (F)	0.024 to 0.030
	B (F)	0.038 to 0.040
	C (M)	0.088 to 0.098
	D (M)	0.069 to 0.085
[i]	A (F)	0.016 to 0.030
	B (F)	0.021 to 0.036
	C (M)	0.044 to 0.062
	D (M)	0.042 to 0.054

The difference for each speaker across vowel context may be spurious but, if enough different vowels were analysed, an overall value for each speaker might be stated. From this admittedly very small sample there is a hint that men may operate with a larger glottal area than women. Given the likely differences in vocal fold length, this suggests that glottal width is what is controlled. A possible

reason might be offered. The Bernoulli force may operate over only a very narrow range of glottal width and is crucial for the maintenance of vocal fold oscillation [16]. Estimates of subglottal pressure are useful in their own right for characterising the respiratory component of articulation for each speaker.

Larynx articulations for consonants

The errors in estimating subglottal pressure are likely to be more serious in the final articulatory component to be discussed here, that is the abduction-adduction of the vocal folds, shown in Figure 1 for fricatives. The following data are offered very tentatively, see Table 2 and Figure 4, for speaker B only. The articulatory time path of glottal area Ag(t) is estimated from

$$\hat{A}_g(t) = \frac{0.00076\ U_0(t)}{(\hat{P}_{sg}(t) - P_0(t))^{0.5}} \qquad (3)$$

Table 2 is for voiced fricatives (individual tokens) in the frame "Say [spəC'əsp] again"; an articulatory path $\hat{A}_g$ is shown in Figure 4, for one token of [v] in the frame "Just [əv'ɜst] again". An attempt is made to consider the effect upon $\hat{A}_g$ of a likely error in $\hat{P}_{sg}$.

Table 2. Estimates of maximum glottal area $\hat{A}_{g\ max}$ during voiced fricative consonants and during the following vowels

Fricative	$\hat{A}_g$ max in cm^2	$\hat{A}_g$ near mid-[ɜ] following in cm^2
[z]	0.110	0.055
[ʒ]	0.073	0.044
[z]	0.083	0.040
[v]	0.110	0.053
[ð]	0.087	0.048

These preliminary figures are compatible with fixed patterns of vocal fold articulation for voiced fricatives regardless of place of articulation.

Analysis-by-synthesis for improved matching

The next step will be to try out the articulatory patterns for each speaker in the model. Where the match between aerodynamic outputs from the model and the corresponding traces from the real speech is poor, the model itself can be used to indicate the correction needed. For example, in estimating glottal area during a voiced fricative it was assumed that essentially all the transglottal airflow appears at the mouth outlet. Clearly, this is not the case in general: airflow is absorbed by vocal tract cavity enlargement, both active and passive, and by the build-up of air pressure behind the vocal tract constriction. The aerodynamic equations of the model quantify theses processes. It is hoped that a limited number of analysis-by-synthesis cycles will yield a converging solution to the optimisation problem.

SPEAKER-SPECIFIC PATTERNS FOR ARTICULATORY SYNTHESIS

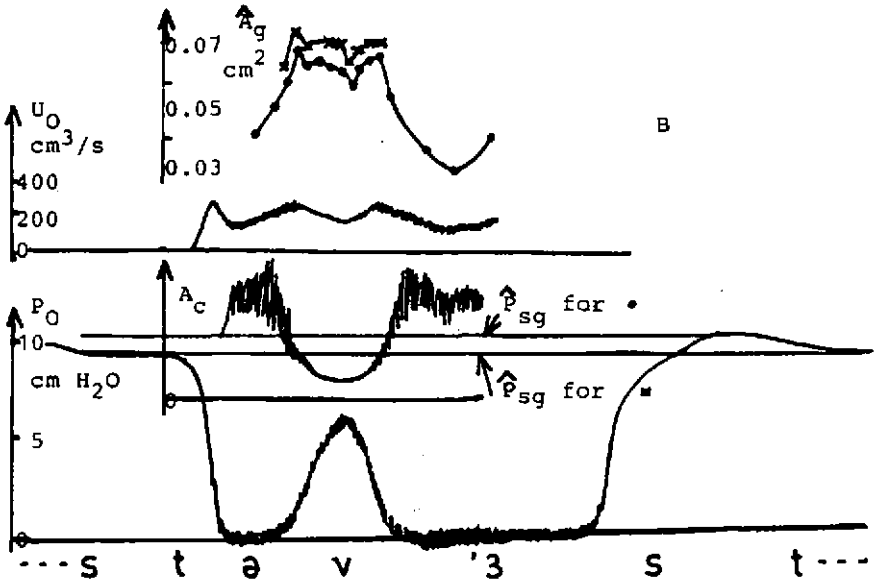

Fig.4 Estimated glottal articulation $\hat{A}_g$ and its coordination with the main vocal tract articulator A_c during [v], for speaker B.

REFERENCES

[1] J.L. Flanagan, K. Ishisaka and K.L. Shipley, 'Signal models for low bit-rate coding of speech', J.A.S.A., Vol. 68, 780-791,(1980)

[2] E. Allwood and C. Scully, 'A composite model of speech production', Conf. Rec. ICASSP 82, Paris, Vol.2, 932-935, (1982).

[3] C. Scully and E. Allwood, 'The representation of stored plans for articulatory coordination and constraints in a composite model of speech production', Speech Comm., Vol. 2, 107-110,(1983)

[4] IPA (Revised to 1979), 'The Principles of the International Phonetic Association', see J.I.P.A., Vol. 8, 1-2, (1978).

[5] K.N. Stevens, 'Bases for phonetic universals in the properties of the speech production and perception systems', Proc. 9th. Intl. Congr. of Phon. Sciences, Vol. 2, E. Fischer-Jørgensen et al., eds., Univ. of Copenhagen, 53-59, (1979).

[6] D.W. Warren and A.B. DuBois, 'A pressure-flow technique for measuring velopharyngeal orifice area during speech', Cleft Palate J., Vol. 1, 52-71, (1964).

[7] C. scully, 'Model prediction and real speech: fricative dynamics', in 'Frontiers of Speech Communication Research', B. Lindblom and S. Ohman, eds., Academic Press, London, 35-48, (1979).

[8] C. Scully, 'Problems in the interpretation of pressure and airflow data in speech', Phonetics Dept. Rep. Univ. of Leeds, No. 2, 53-92, (1969).
[9] J.S. Bridle and M.P. Ralls, 'An approach to speech recognition using synthesis-by-rule', in 'Computer Speech Processing', F. Fallside, ed., to be published 1984.
[10] J. Vaissiere, 'Prediction of articulatory movement of the velum from phonetic input', Bell Laboratories, (1983).
[11] J.R. Smitheran and T.J. Hixon, 'A clinical method for estimating laryngeal airway resistance during vowel production', J. Sp. and Hear. Dis., Vol. 46, 138-146, (1981).
[12] M. Rothenberg, 'Interpolating subglottal pressure from oral pressure', and T.J. Hixon and J.R. Smitheran, 'A reply to Rothenberg', J. Sp. and Hear. Dis., Vol. 47, 219-223, (1982).
[13] A. Löfqvist, B. Carlborg and P. Kitzing, 'Initial validation of an indirect measure of subglottal pressure during vowels, J.A.S.A., Vol. 72, 633-635, (1982).
[14] T. Cleveland and J. Sundberg, 'Acoustic analysis of three male voices of different quality', STL-QPSR 4/1983, RIT Stockholm, 27-38, (1984).
[15] B. Fritzell, J. Gauffin, B. Hammarberg, I. Karlsson and J. Sundberg, 'Measuring insufficient vocal fold closure during phonation', STL-QPSR 4/1983, 50-59, (1984).
[16] I.R. Titze, 'Comments on the myoelastic-aerodynamic theory of phonation, J. Sp. and Hear. Res., Vol. 23, 495-510, (1980).

Acknowledgements
The assistance of Mr. E. Brearley in the experiments on real speech is gratefully acknowledged.

The computer modelling of speech production is supported by the SERC.