ON THE SOURCES OF INTER- & INTRA- SPEAKER VARIABILITY IN THE ACOUSTIC DYNAMICS OF SPEECH

Xue YANG †, *J. Bruce Millar* † *and Iain Macleod*

† Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, Australia
email: xue@cslab.anu.edu.au

ABSTRACT

In this paper, we briefly report the experimental procedure and some results in testing our hypothesis that differences in size and shape of the vocal tract influence the dynamics of the formant trajectories of the speech signals.

1. INTRODUCTION

The nature of the sources of inter- and intra-speaker variability and their influence on the speech signal comprise one of the most basic issues in automatic speaker recognition research. Currently published knowledge about this issue points to two basic sources at the articulatory level: articulatory structure and articulatory dynamics. Articulatory structure is dictated by the shape and size of the vocal apparatus, while articulatory dynamics is dictated by the movement and coordination of the articulators in the vocal tract. There is an assumption that static and dynamic acoustic aspects of speaker variability derive independently from these sources respectively (Stevens 1971, Wolf 1972, Sambur 1975, Atal 1976, O'Shaughnessy 1986, Rosenberg *et al.* 1991). An analysis of the process of speech production causes us to question whether articulatory dynamics is the only source of acoustic dynamics, or whether articulatory structure also contributes.

Any acoustic parameter of the speech signal as a function of time is actually formed by two mappings: one mapping from time to the configuration of the articulatory system and one mapping from the system configuration to the acoustic parameter. The first mapping certainly relates directly to articulatory dynamics. The second mapping concerns relations between the articulatory configuration and the acoustic parameter and is not related to time. When constriction degree and lip rounding are fixed, the effects of constriction position on formants are nonlinear (Stevens *et al.*, 1955; Fant, 1960). This nonlinearity basically represents how much and in what manner formants change as the result of tongue body movement along the vocal tract length specified by the change of constriction position. These characteristics are necessarily included in the corresponding formant trajectories as a function of time as the tongue body moves. The effects of constriction degree and lip rounding on the formant trajectories as a function of constriction position show nonlinearity too.

In order to test our hypothesis that differences in size and shape of the vocal tract also influence the dynamics of the formant trajectories of the speech signals, three dynamic phonetic segments (/aI/ as in "high", /aU/ as in "how" and /wi/ as in "we") were simulated by performing the two mappings. In the limited space for this paper we report the experimental procedure and the results for the sound /aI/ only. In overall terms, these results are consistent with obtained for /aU/ and /wi/.

2. SIMULATIONS OF THE DYNAMIC SOUNDS

2.1. Mapping from Articulatory Parameters to Acoustic Parameters

The vocal tract model used to perform the mapping from articulatory parameters to acoustic parameters is based on Fant's (1960) acoustic lossless tube model. Eight parameters employed to define the model configuration are,

- L_g, representing the length in cm of the larynx portion;
- L_m, representing the length in cm of the lip section;
- L, representing the length in cm of the vocal tract;
- X_c, representing the distance in cm of the maximum tongue constriction position to the glottis;
- A_g, representing the cross-sectional area in cm^2 of the larynx portion;
- A_l (or R_l), representing the cross-sectional area in cm^2 (or the radius in cm) of the lip opening ;
- A_m (or R_m), representing the maximum cross-sectional area in cm^2 (or the radius in cm) of the vocal tract;
- A_c (or R_c), representing the cross-sectional area in cm^2 (or the radius in cm) at the point of maximum tongue constriction in the vocal tract;

The model is formed by concatenation of a finite number of short uniform sections of equal length. The cross-sectional area for each section is the average of the function over the short length of that section. The transformation from the model configuration defined by a set of cross-sectional areas to the first three formants is performed via linear prediction analysis techniques.

2.2. Mapping from Time to Articulatory Configurations — Articulatory Dynamics

Our current knowledge about articulatory dynamics for /aI/, /aU/, and /wi/ (Stevens *et al.* 1955, Fant 1960, Kent 1972, Kent *et al.* 1972) suggests that the significant model parameters are those listed in Table 1, and that their temporal variation may take the form of a sigmoidal function.

Sounds	Time-varying parameters	
/aI/	$X_c(t)$	$A_c(t)$ (or $R_c(t)$)
/aU/	$X_c(t)$	$A_l(t)$ (or $R_l(t)$)
/wi/	$X_c(t)$	$A_l(t)$ (or $R_l(t)$)

Table 1: Time-varying parameters of the model in simulation of /aI/, /aU/ and /wi/.

When the duration is scaled to unity, $X_c(t)$ for /aI/, /aU/ and /wi/ is as follows,

$$X_c(t) = a + \frac{b}{1+e^{-c(t-d)}} \quad 0 \le t \le 1$$

where, $a = X_c(0) - \frac{b}{1+e^{c\times d}}$; $b = \frac{X_c(1)-X_c(0)}{p}$; $c = \frac{1}{d} \times \ln\frac{1+p}{1-p}$; $d = 0.5$; $p = 0.95$;

$X_c(0), X_c(1)$ are initial and final values of the constriction position during production of the sound. p controls the sigmoidal characteristics of $X_c(t)$ and, therefore, determines its slope at different time points. A value of 0.5 for d makes the function symmetrical in time. As an example, $X_c(t)$ for /aI/ with $X_c(0) = 5.5, X_c(1) = 11$ is shown in Figure 1.

For /aI/, the constriction degree as a function of time, $R_c(t)$, takes a similar form to that for $X_c(t)$. In order to simulate the value of this parameter in the straight acoustic tube model, we have
for $0 \le t \le 0.5$:

$$a = R_c(0) - \frac{b}{1+e^{c\times d}}; \; b = \frac{2(R_c(0.5)-R_c(0))}{p}.$$

and for $0.5 \le t \le 1$:

$$a = 2(R_c(0.5) - R_c(1)) + R_c(1) - \frac{b}{1+e^{c\times d}};$$
$$b = \frac{2(R_c(1)-R_c(0.5))}{p}.$$

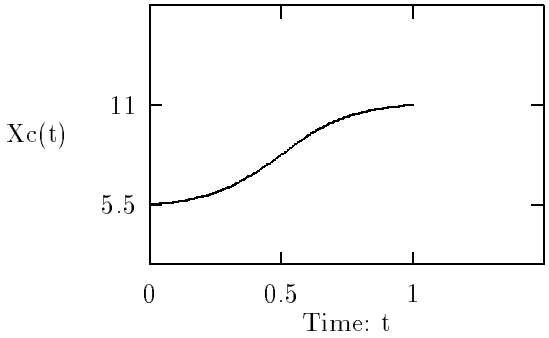

Figure 1: $X_c(t)$ for /aI/ with $X_c(0) = 5.5, X_c(1) = 11$.

$R_c(0), R_c(0.5), R_c(1)$ are the values of constriction degree at $t = 0, 0.5, 1$ respectively.

The example of $R_c(t)$ for /aI/ with $R_c(0) = R_c(1) = 0.455, R_c(0.5) = 1$ is shown in Figure 2.

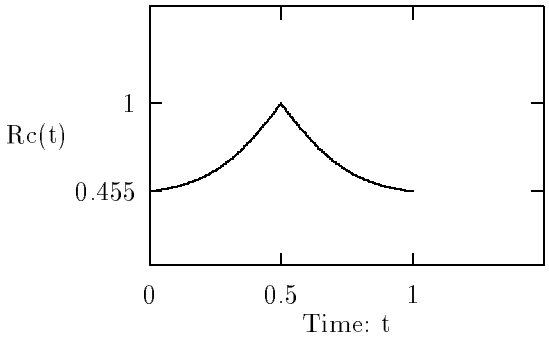

Figure 2: $R_c(t)$ for /aI/ with $R_c(0) = R_c(1) = 0.455, R_c(0.5) = 1$.

The initial and final values of $R_l(t)$ are different between /aU/ and /wi/, because the lips move from unrounding to rounding for /aU/ and vice versa for /wi/.

2.3. Boundary Conditions of Model Parameters for the Simulated Sounds

In our experiments, for each of /aI/, /aU/ and /wi/, we first perform the benchmark simulation. The values of the static model parameters are chosen either the same as or close to those used by Fant (1960). The boundary values of the time-varying parameters for each sound are chosen basically according to those for the related static vowel sounds. Table 2 lists the relevant model parameters for the benchmark simulation of /aI/ (2a), /aU/ (2b) and /wi/ (2c).

Then, we vary the values of some static parameters to per-

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	17.6	$A_c(R_c)$	.65 (.455)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1))$	.16 (.226)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9	$A_l(0)(R_l(0))$	.16(.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1)$	8.(1.6)

(c)

Table 2: Values of the model parameters for the benchmark simulation of, /aI/ in (a), /aU/ in (b) and /wi/ in (c).

turb the size and shape of the vocal tract. The varied static parameters in our study are the pharyngeal length (hence total length L), the radius of maximum cross-sectional area R_m and the radius of cross-sectional area at the maximum constriction R_c. They are increased/decreased by 33.3% of their values for the benchmark simulation to represent different size and shape of the vocal tract in its cross-sectional area. For each variation of the static parameters for the simulation of each sound, the temporal properties of the dynamic parameters are as presented in Section 2.2, thus representing identical articulatory dynamics.

2.4. Auditory Verification of Simulated Sounds

The waveforms of the simulated sounds were synthesised in order to verify their phonetic quality. Informal listening tests performed by the first author indicated that the benchmark simulation of /aI/ as in "high", /aU/ as in "how" and /wi/ as in "we" had the expected phonetic quality. The phonetic quality of each of the simulations with variation of total length L, maximum cross-sectional area A, cross-sectional area at the maximum constriction A_c was also identified as being the same as that of its corresponding benchmark simulation. Differences in auditory quality from their benchmark counterpart were recognised in some of the perturbed benchmark simulations, but for others, they were not perceived.

3. EXPERIMENTAL RESULTS: THE ACOUSTIC DYNAMICS

The formant trajectories as a function of time for each simulation of a sound were derived by performing two successive mappings. Our experimental results have shown that the first three formant patterns of simulated sounds /aI/, /aU/ and /wi/ change as certain static model parameters are varied. Here we report only the results for sound /aI/ simulated under different static conditions.

3.1. Tongue Moving from Low-back to High-front

The first three formant trajectories with variation of the acoustic tube length, of boundary values of constriction degree and of the maximum cross-sectional area of the acoustic tube model for sound /aI/ are shown respectively in Figure 3, Figure 4 and Figure 5.

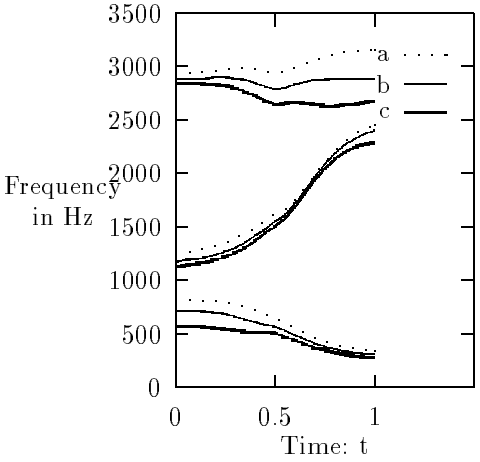

Figure 3: The first three formant trajectories for /aI/ with variation of L, a: $L = 16.5$, b: $L = 17.6$, c: $L = 18.7$.

4. DISCUSSION

We have shown how the first three formant patterns of simulated sound /aI/ change as certain static model parameters are varied. The changes in the formant pattern are seen in the position and slope of the formants. The extent of these changes are clearly non-linear with respect to the proportional perturbations made to the benchmark simulations reflecting the inherent nonlinear relationships between articulatory parameters and acoustic parameters. Trajectory features which may intuitively be associated with the dynamics of articulation are here clearly seen to be consequences of the interaction of articulatory dynamics and the static boundary conditions of the articulation.

Tosi *et al.* (1972) recognised the importance of slopes of for-

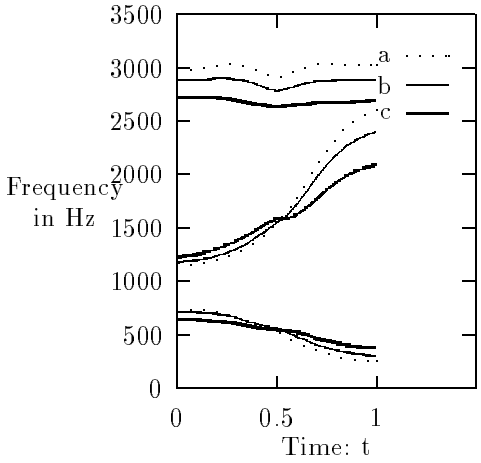

Figure 4: The first three formant trajectories for /aI/ with variation of $A_c(0)$; a: $A_c(0) = .289$, b: $A_c(0) = .65$, c: $A_c(0) = 1.16$.

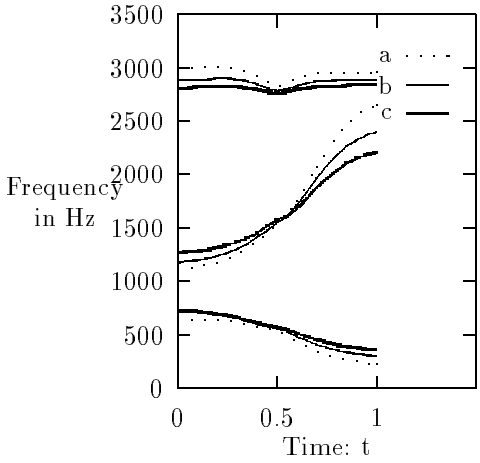

Figure 5: The first three formant trajectories for /aI/ with variation of A_m; a: $A_m = 14.2$, b: $A_m = 8$, c: $A_m = 3.57$.

mants during liquids, glides and diphthongs in speaker identification from visual examination of spectrograms. Stevens (1971) speculated that the sources of variability in the slope of the formants for those sounds "*may be the result of learned articulatory habits rather than differences in anatomical or physiological characteristics of the articulatory structure*" (note that he linked articulatory dynamics to *learned articulatory habits*). In studying the acoustic features for speaker identification, Sambur (1975) claimed "*the F_2 slope in /aI/ was also quite variable among speakers and demonstrated excellent identification potential*", via "*measurements related to the dynamic properties of the talker's voice patterns that reflect his learned behavior of speaking*". Certainly, different articulatory dynamics causes differences in the corresponding acoustic dynamics. But our experimental results in simulating acoustically dynamic sounds have shown that the size and shape of the vocal tract and the static boundary conditions can also lead to changes in the slope of the first three formant trajectories. In particular, when the cross-sectional area contrast (i.e. $\frac{A_m}{A_c}$) of the vocal tract is increased in producing /aI/, the slope of $F_2(t)$ increases.

We conclude that variations in the dynamic aspects of real speech signals can be caused by the differences in articulatory static aspects (e.g. differences in size and shape of the vocal tract), as well as by the differences in articulatory dynamic aspects (e.g. differences in speed of tongue or lip movements).

5. REFERENCES

1. Atal, B. S. 1976,"Automatic recognition of speakers from their voice", Proc. of the IEEE, Vol. 64, No. 4., pp. 460-475.
2. Fant, G. 1960,"Acoustic Theory of Speech Production", Mouton: The Hague.
3. Kent, R. D. (1972) "Some Considerations in the Cinefluorographic Analysis of Tongue Movements during Speech", Phonetica Vol. 26, pp. 16-32.
4. Kent, R. D. and Moll, K. L. (1972) "Tongue Body Articulation during Vowel and Diphthong Gestures", Folia Phoniatrica Vol. 24, pp. 278 - 300.
5. O'Shaughnessy, D. 1986, "Speaker Recognition", IEEE ASSP Magazine (October), pp. 4-17.
6. Rosenberg, A. E. and Soong, F. K. 1991, "Recent research in automatic speaker recognition", in Advances in Speech Signal Processing, ed. by Furui, S. and Sondhi, M. M..
7. Sambur, M. R. 1975, "Selection of acoustic features for speaker identification" IEEE Trans. on ASSP, Vol. 23, pp. 176-182.
8. Stevens, K. N. and House, A. S., 1955, "Development of a Quantitative Description of Vowel Articulation", J. Acoustic Soc. Am., Vol. 27, No. 3, pp.484-493.
9. Stevens, K. N. and House, A. S., 1963,"Perturbation of Vowel Articulations by Consonantal Context: an Acoustical Study", J. of Speech and Hearing Research, Vol. 6, No. 2, pp. 111 - 128.
10. Stevens, K. 1971, "Sources of Inter- and Intra- speaker variability in the acoustic properties of speech sounds", Proc. of 7th International Congress of Phonetic Sciences, pp.206-232.
11. Tosi, T., Oyer, H. J., Lashbrook, W. B., Pedrey, C. and Nichol, J. 1972, "Experiment on voice identification", J. Acoust. Soc. Amer., Vol. 51, pp. 2030-2043.
12. Wolf, J. J. 1972, "Efficient acoustic parameters for speaker recognition", J. Acoust. Soc. Amer., Vol. 51, pp. 2044-2055.