An Analysis of Modal Coupling Effects During the Glottal Cycle: Formant Synthesizers from Time-Domain Finite-Difference Simulations

Gordon Ramsay

Institut de la Communication Parlée
UPRESA CNRS 5009 - INPG
46 avenue Félix Viallet
38031 Grenoble CEDEX 01
France.

10.21437/ICSLP.1998-362

ABSTRACT

Speech is typically modelled using time-domain or frequency-domain simulations of the acoustic field in the vocal tract. Using a biorthogonal modal decomposition, it is shown that time-domain finite-difference simulations can be transformed algebraically into equivalent formant synthesizers, the parameters of which vary in time and are calculated directly from the laws of physics. Examining the structure of the equivalent formant synthesizer, it is observed that formant excitation is largely due to internal modal coupling effects, induced by rapid perturbation of the acoustic eigenmodes caused by vibration of the glottis, and does not rely precisely on external sources provided by boundary conditions. This leads to a novel interpretation and justification of traditional models of the glottal source.

1. INTRODUCTION

Acoustic models of speech production are typically based either on frequency-domain simulations, which generate formant synthesizer parameters from a glottal waveform and quasi-static sequence of area functions, or on time-domain simulations, which calculate the detailed evolution of a spatio-temporal pressure/velocity field from a dynamic specification of the entire vocal tract shape, including the glottis.

Frequency-domain simulations are stable, easy to implement, and generate parameters that directly describe the spectral properties of synthetic speech, but cannot easily account for properties of the voice source, and are not strictly valid for time-varying area functions. Time-domain simulations generate the entire dynamic acoustic field directly from basic physical principles, without artificially separating "source" and "filter", but cannot easily be used to extract meaningful spectral parameters (e.g. formant frequencies and bandwidths) without a considerable amount of inaccurate post-processing. Neither method provides a complete account of the underlying mechanism by which glottal motion excites the acoustic eigenmodes of the vocal tract during phonation to produce voiced speech.

The purpose of this paper is to demonstrate that it is possible to establish a formal mathematical equivalence between the structure of time-domain and frequency-domain simulations of acoustic wave propagation in the vocal tract, by introducing an explicit time-varying modal representation of the sound field. Using the proposed technique, it is shown that time-domain finite-difference simulations can be used to generate equivalent formant synthesizers, the parameters of which are calculated implicitly from the underlying laws of physics, given a time-varying area function for the entire vocal tract. The sound field generated by the time-domain simulation can then be broken down into individual modal components, each of which corresponds to a single formant, with the advantage that the mechanism generating each formant oscillation is made transparent. Simulation results suggest that formant excitation is largely caused by internal modal coupling effects, induced by rapid perturbation of the acoustic eigenmodes caused by vibration of the glottis, and does not rely precisely on external sources provided by boundary conditions.

The results described in this paper are of both theoretical and practical interest, since the technique provides a means of analysing and controlling the behaviour of time-domain simulations of vocal tract acoustics directly, while contributing to an explanation of the mechanism underlying the generation of sound.

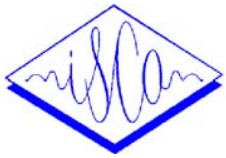
2. ACOUSTIC MODEL

Suppose that the vocal tract can be modelled as an elastic tube of length L and time-varying cross-sectional area $A : \Omega \rightarrow \mathbb{R}$ defined on a bounded rectangle $\Omega = \{(x,t) : x \in [0,L], t \in [0,T]\}$ in $\mathbb{R}^2$, where x represents the distance along the tract midline from the trachea to the lips, and t represents time.

Assume that the physical state of the air within the tube can be represented by functions $\rho, P, U, Y : \Omega \rightarrow \mathbb{R}$ representing respectively the density, pressure, particle velocity, and radial wall displacement from equilibrium measured along the tube length, averaged over the tube cross-section. If the sound field is taken to consist of isentropic compressible perturbations $\rho_1, P_1, U_1, Y_1 : \Omega \rightarrow \mathbb{R}$ of a perfect gas, superimposed on an underlying incompressible mean flow $\rho_0, P_0, U_0, Y_0 : \Omega \rightarrow \mathbb{R}$, and if the mean pressure and particle velocity are assumed to be zero, the acoustic field can be expressed in terms of dimensionless groups $\alpha, p, u, y : \omega \rightarrow \mathbb{R}$ defined on a domain $\omega = \{(\xi,\tau) : \xi \in [0,1], \tau \in [0, cT/L]\}$, where c is the local sound speed and

$$\begin{aligned} p &:= P_1/\rho_0 c^2, & \alpha &:= A/L^2, \\ u &:= U_1/c, & \xi &:= x/L, \\ y &:= Y_1/L, & \tau &:= ct/L. \end{aligned}$$

Neglecting viscous effects, it can be shown that the conservation laws defining quasi-one-dimensional mass and momentum bal-



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

ance are then described by the following equations (cf. [1]):

$$\frac{\partial}{\partial\tau}\alpha p + \frac{\partial}{\partial\xi}\alpha u + \frac{\partial}{\partial\tau}2\sqrt{\alpha}y = 0, \quad (1)$$

$$\frac{\partial}{\partial\tau}\alpha u + \alpha\frac{\partial p}{\partial\xi} = 0, \quad (2)$$

and the walls of the vocal tract can be modelled in the usual manner as a normally-reacting elastic membrane under tension, with

$$w_m\frac{\partial^2 y}{\partial\tau^2} + w_b\frac{\partial y}{\partial\tau} + w_k y - w_t\frac{\partial^2 y}{\partial\xi^2} = 2\sqrt{\alpha}p, \quad (3)$$

where w_m, w_k, w_b, w_t are dimensionless constants related to the density, elasticity, damping, and tension of the wall tissue.

Assume that the vocal tract is initially in a state of equilibrium,

$$\tau = 0 \quad : \quad p(\xi,0) = u(\xi,0) = y(\xi,0) = \dot{y}(\xi,0) = 0; \quad (4)$$

that a constant acoustic pressure is applied at the trachea entrance,

$$\xi = 0 \quad : \quad p(0,\tau) = p_*; \quad (5)$$

and that radiation at the lips can be modelled by the equation

$$\xi = 1 \quad : \quad \frac{\partial}{\partial\tau}\alpha u - \kappa_a\frac{\partial}{\partial\tau}\alpha p - \kappa_b\alpha p = 0, \quad (6)$$

where κ_a and κ_b are dimensionless functions of the lip aperture.

The system of partial differential equations (1)-(3) together with the boundary conditions (4)-(6) define the basic physical laws governing the evolution of the sound field in the vocal tract. Solutions cannot in general be obtained analytically, but must be calculated instead by numerical simulation.

Applying the finite-difference method, a grid $\{(\xi_j,\tau_k) : j = 1,\dots M, k = 1,\dots,N\}$ is imposed on ω, and partial derivatives are replaced by discretized approximations expressed in terms of functions evaluated at the grid points. Denoting by Z_k the vector of acoustic state variables on all grid points $\{\xi_j\}$ at time τ_k, and taking $Z_0 = 0$, the resulting system of linear algebraic equations can always be written as an implicit time-varying recursion,

$$P_k Z_k = Q_k Z_{k-1} + R_k, \quad (7)$$

where P_k, Q_k are sparse banded matrices whose elements are functions of $\alpha(\xi,\tau)$, and R_k is a driving function derived from the boundary conditions. Under minor restrictions on α, a unique solution of the recursion exists, and if the chosen finite-difference scheme can be shown to be *consistent* and *stable*, the discretized solution will converge to a solution of the original partial differential equations as the discretization intervals tend to zero.

The principal interest lies in deriving a complete characterisation of the family of solutions of the finite-dimensional system of equations represented by (7), in terms of the area function α, and in using this to establish a useful physical interpretation for the structure of the corresponding family of solutions of the original infinite-dimensional system of equations (1)-(6).

Since both the original equations and their numerical approximation describe linear time-varying systems, the space of possible solutions on continuous and discretized domains can be described as a superposition of characteristic modal vibrations, corresponding to acoustic resonances of the vocal tract, and the problem essentially consists in determining the time-varying eigensystem of the recursion, onto which any solution can then be projected.

3. MODAL ANALYSIS

The use of modal analysis in characterising the behaviour of finite-difference schemes is well-known, and forms the basis for many common methods of proving stability and convergence. Titze [2] was the first to apply the technique to speech, and used it to analyse the vibrations of a linear time-invariant model of the glottis. Other authors [3] [4] [5] [6] have since applied similar methods to examine the behaviour of static acoustic models. The derivation below, detailed in a previous paper [7], applies to a general *time-varying* model, and can be used to calculate the acoustic eigenmodes of the vocal tract at each point in time.

Define $M_k(\lambda) : \lambda \in \mathbb{C}$ to be the regular matrix pencil given by

$$M_k(\lambda) = P_k\lambda + Q_k. \quad (8)$$

Let U_k, V_k be the matrices of left and right latent vectors of $M_k(\lambda)$ respectively, and let Λ_k be the diagonal matrix of complex latent roots. Under the assumption that the right latent vectors are linearly independent, U_k and V_k may be chosen to satisfy

$$U_k^H P_k V_k = I, \quad (9)$$

$$U_k^H Q_k V_k = \Lambda_k, \quad (10)$$

where ^{H} denotes the conjugate transpose. Denoting by E_k the projection of Z_k onto the right eigenspace of $M_k(\lambda)$, we have

$$E_k = U_k^H P_k Z_k, \quad (11)$$

$$Z_k = V_k E_k, \quad (12)$$

and equation (7), transformed into the modal domain, becomes

$$E_k = \Lambda_k(U_k^H P_k V_{k-1})E_{k-1} + U_k^H R_k. \quad (13)$$

The diagonal elements of Λ_k are the time-varying complex poles of the acoustic model, and can be used to calculate the instantaneous formant frequencies and bandwidths of the modelled vocal tract. The columns of V_k represent the instantaneous spatial distributions of pressure, velocity, and wall displacement associated with each formant, and define the characteristic vibrations of the system. The columns of U_k define projections onto the corresponding invariant subspaces, and determine the relative proportion of energy entering each eigenmode at time τ_k from a source R_k distributed along the length of the vocal tract. Of particular interest is the term $(U_k^H P_k V_{k-1})$, which reduces to the identity matrix for a static area function, and represents a dynamic leakage of energy between the different formants when the eigenmodes of the system change in time. To make this explicit, equation (13) may be re-written for a single element e_k^i of E_k as

$$e_k^i = \lambda_k^i e_{k-1}^i + c_k^i + s_k^i, \quad (14)$$

where c_k^i and s_k^i are defined by

$$c_k^i = \textstyle\sum_{j\neq i}\lambda_k^i({u_k^i}^H P_k v_{k-1}^j)e_k^j, \quad (15)$$

$$s_k^i = {u_k^i}^H R_k. \quad (16)$$

The finite-difference recursion (7) can therefore be transformed into a bank of first-order modal oscillators with time-varying coefficients, each of which is driven by an external *source term* s_k derived from the boundary conditions and a internal *coupling term* c_k which represents a transfer of energy between different

eigenmodes. The effective excitation driving each formant is the sum of the corresponding source term and coupling term.

It is of considerable interest to compare the modal structure of the finite-difference simulation, which resembles a classical formant synthesizer and can be calculated directly from the underlying physics, with the structure of a traditional formant synthesizer, where parameters are chosen heuristically from prior knowledge, or from quasi-static frequency-domain models (e.g. [8]).

4. SIMULATION RESULTS

Simulation results are provided here to illustrate the modal analysis procedure for a simple time-varying area function representing the vowel $/u/$. Figure 1 shows the static tube shape used to model the trachea (20 grid points) and oral tract (80 grid points). Figure 2 shows the area variations imposed at the glottis (8 grid points). A sampling frequency of 80kHz was used throughout.

Illustrations of the temporal evolution of the acoustic eigenvalues and eigenvectors and corresponding formant frequencies and bandwidths for a similar time-varying area function were presented previously [7], and will be omitted here; the purpose of the present paper is to examine in detail the mechanism responsible for generating the underlying formant oscillations.

Figure 3 shows the pressure waveform/spectrum generated at the lips by the finite-difference recursion. The oral tract formants are clearly visible in both time and frequency domain representations, but cannot be separated using signal processing methods alone.

Figure 5 shows the pressure waveform at the lips obtained by projecting the entire sound field onto the time-varying eigenspace associated with the first oral formant. The result is a pure formant oscillation, with time-varying centre frequency and bandwidth, that corresponds well with the 1st spectral peak in Figure 3.

Figure 6 shows the modal amplitude e_k of the first oral formant, representing the projection of the same formant oscillation illustrated in Figure 5 onto the associated eigenspace.

Figure 7 shows the external source term s_k describing the injection of energy from the boundary conditions into the formant illustrated in Figure 5. For the oral formants, this is a slowly-varying signal that closely follows modulations in the glottal area.

Figure 8 shows the internal coupling term c_k describing the injection of energy from other formants into the formant illustrated in Figure 5. Remarkably, it consists of a sequence of rapid primary and secondary spikes located at the instants of glottal opening and closure. The relative amplitude of the spikes has been found to depend strongly on the phase and speed of glottal movement, on the formant frequency, and on the shape of the oral tract.

Figure 4 shows the total excitation driving the first oral formant, obtained by summing external and internal source terms. A sequence of half sinewaves is generally obtained, terminating in abrupt pulses at the moment of glottal closure. The excitation spectrum is dominated by the contribution of the internal source.

The results shown in Figures 3-8 describe part of the input and output for a formant synthesizer that is algebraically equivalent to the original finite-difference simulation. It is remarkable that the modal excitation, which arises automatically from the internal structure of the physical simulation, has roughly the same form as a traditional model of the glottal source [9]. The crucial point to note is that formant oscillations are generated by the model in an abstract modal domain, and the equivalent excitation needed to drive each individual formant resonator does *not* then correspond to an acoustic source localized at the glottis, but instead consists of an external source term derived from an energy source distributed along the vocal tract length, and an internal coupling term arising from modal leakage between the formants caused by rapid perturbation of the acoustic eigenmodes. During the open portion of the glottal cycle, energy appears to be absorbed from the boundary conditions through the source term, and then abruptly redistributed among the different formants by the coupling term at the moment of glottal closure. It is the coupling term, not the source term, that dominates the excitation spectrum. Similar results hold for different formants and area functions.

5. CONCLUSIONS

By introducing a biorthogonal modal decomposition, it has been demonstrated that it is possible to obtain an equivalent formant synthesizer with time-varying parameters from a time-domain finite-difference simulation. Speech can indeed be modelled correctly by a source-filter structure, and the excitation does indeed have the form claimed in the literature; this can be justified directly from the underlying physics, but in order to do so, formant oscillations need to be represented in an abstract modal coordinate system, and the formant excitation needs to be carefully decomposed into external source and internal coupling components.

6. REFERENCES

1. M. R. Portnoff. *A quasi-one-dimensional digital simulation for the time-varying vocal tract*. Master's thesis, MIT, 1973.
2. I. R. Titze. The human vocal cords : a mathematical model (Part I). *Phonetica*, 28:129–170, 1973.
3. G. Fant and S. Pauli. Spatial characteristics of vocal tract resonance modes. In G. Fant, editor, *Speech Communication Seminar, Stockholm*, pages 121–132. Wiley, New York, 1975.
4. M. Mrayati and R. Carré. Relations entre la forme du conduit vocal et les caractéristiques acoustiques des voyelles francaises. *Phonetica*, 33:285–206, 1976.
5. A. Miller and V. Sorokin. Acoustic model for an articulatory-formant speech synthesizer. In *Proceedings, XIII'th International Congress of Phonetic Sciences*, volume 2, pages 466–470, 1995.
6. R. Van Praag and P. Jospa. Variational method applied to formants computation for a pharyngo-bucco-nasal tract. In *Proceedings, XIII'th International Congress of Phonetic Sciences*, volume 4, pages 456–459, 1995.
7. G. Ramsay. Modal synthesis of acoustic wave propagation in the vocal tract using a finite-difference simulation. In *Proceedings of the 1st ETRW on Speech Production Modelling*, pages 207–210, Autrans, France, 1996.
8. Q. Lin. *Speech production theory and articulatory speech synthesis*. PhD thesis, KTH, Stockholm, 1990.
9. G. Fant. Some problems in voice source analysis. *Speech Communication*, 13:7–22, 1993.

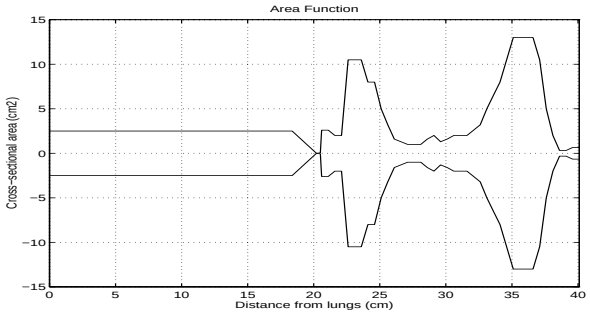


Figure 1: Static area function for vowel /u/ (F1).

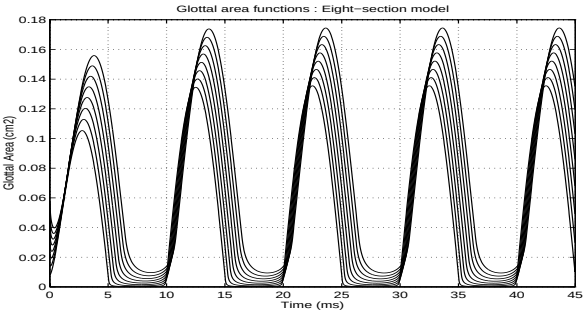


Figure 2: Dynamic glottal area waveforms for vowel /u/.

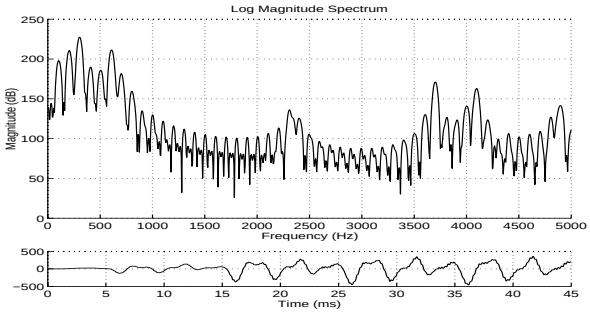


Figure 3: Pressure waveform for vowel /u/.

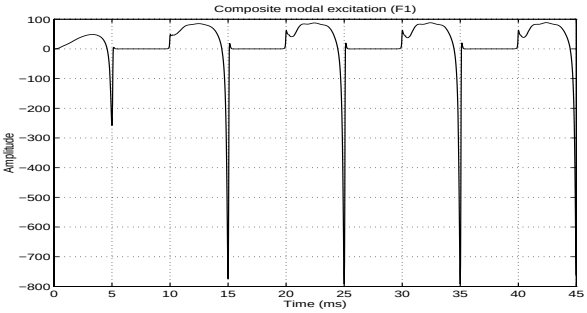


Figure 4: Modal excitation for vowel /u/ (F1).

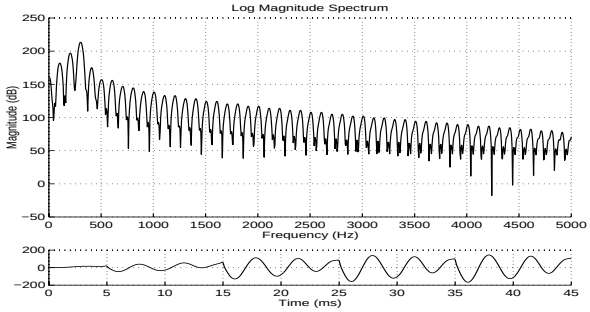


Figure 5: Modal pressure waveform for vowel /u/ (F1).

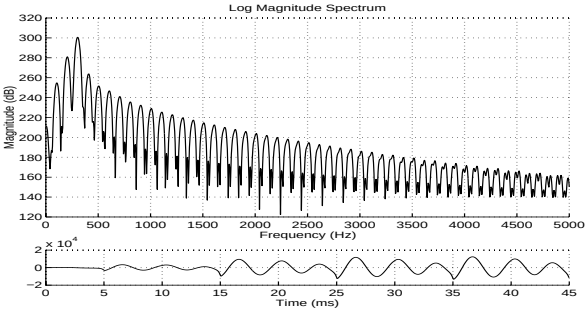


Figure 6: Modal amplitude waveform for vowel /u/ (F1).

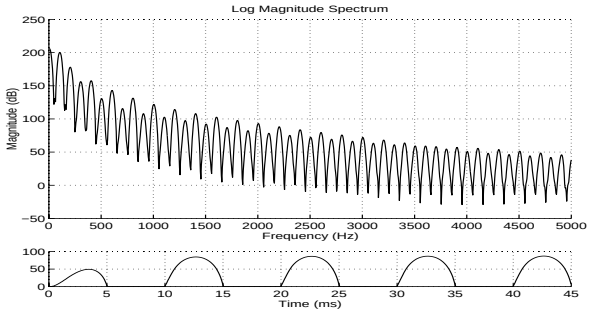


Figure 7: Modal source term for vowel /u/ (F1).

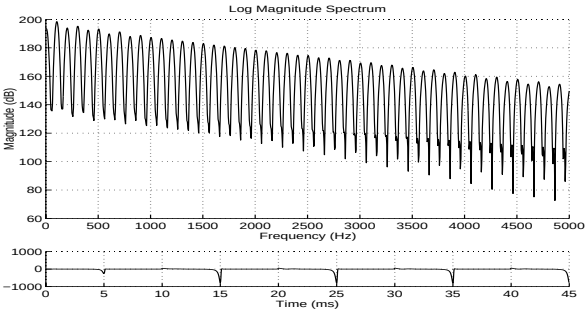


Figure 8: Modal coupling term for vowel /u/ (F1).

Acknowledgement: The simulation results described in this paper were carried out on a CRAY-T3E computer managed by the Commissariat à l'Énergie Atomique in Grenoble. Many thanks to the staff of the C.E.A. for their kind assistance.