6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MODAL SYNTHESIS AND MODELING OF VOWELS

Unto K. Laine
Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing
PO BOX 3000, FIN-02015 Espoo, Finland
Unto.Laine@hut.fi
http://www.acoustics.hut.fi/~unski/

10.21437/Eurospeech.1999-253

ABSTRACT

A new simple model for vocal tract (VT) acoustics is presented. The model is based on a systems approach applied to the modes of the VT and the subglottal cavities. Glottal interaction with the ability to produce pitch-synchronous modulation effects is included as well as the lip radiation impedance. The synthesized vowels have spectral and temporal features close to natural ones. They are almost free of the nasal quality present in many vowels synthesized by conventional methods.

Keywords: vocal tract models, speech synthesis

1. INTRODUCTION

Many studies during the last years have shown new evidence of the quality and quantity of the pitch-synchronous modulation effects in vowel sounds. Both the Q-values (i.e., time-envelopes) of the resonances and their instantaneous frequencies are quite strongly modulated in synchrony with the glottal area during vocal fold vibration. Many authors have experimented with these modulation effects and some of them agree that these temporal fine structures are of importance in terms of the quality, especially the naturalness of the produced vowels [1,2,5].

During the closed period the vocal tract resonates autonomously with the only loads being the lip radiation and the internal losses in the tract, due to, e.g., the yielding cavity walls. The losses of the first formants are relatively low during the closed period causing high Q-values for the formants. The acoustics related to the closed period is relatively easy to understand and model.

However, the situation turns much more complicated when the coupling effects of the subglottal cavities during the glottal open period are considered. The variable impedance of the glottal orifice changes the formant frequencies and brings about an additional load which dampens the formant oscillations. Simultaneously, the formants are excited by the flow through the glottis and especially by the rapid stop of the flow at the glottal closure. The closure excites the subglottal resonances, too. During the open period, when the transglottal impedance is small, the subglottal and the supraglottal modes are in interaction and they exchange energy. It is understandable that such a real-time synthesizer including all these features has been hard to implement.

The aim of this paper is to introduce a novel modal synthesis method which also models speech production in a new way. All of the essential components like sub- and supraglottal resonances and a simple glottal model with proper interactions are included. In sections 2 and 3 the VT mode approximation method is presented. In section 4 the results of the simulations are described and analyzed.

2. APPROXIMATION OF MODES

The new modal synthesizer described below is based on a radically simplified approximation of the modes of the VT and of the subglottal system as shown in *Figure 1*. In this classical view all modes of an uniform VT have their pressure maximum at the glottal end and their volume velocity (VV) maximum at the lip opening (se also *Figure 2*).

The first VT mode can be approximated by a simple second order lumped parameter system where the capacitor models the compressibility of the air and the inductor the mass effect of the air flow. The resistance models the internal losses due to viscosity and thermal conductance. It has to be noted that this type of approximation is valid only close to the mode frequency. When the component values are normalized to C=2, L=2 and R=0.1 the resonance frequency of the mode is normalized to 0.5 and its magnitude to 20 dB as shown in *Figure 2*.

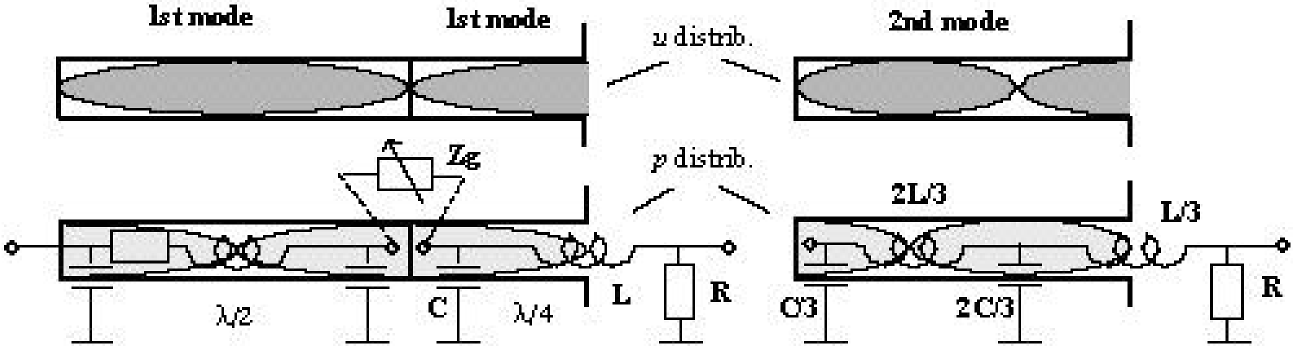

Figure 1. Left part: Simplified acoustical structures of the vocal tract and the subglottal cavities with the first modes, their VV and pressure distributions, and lumped parameter approximations. Right part: The VV and pressure distributions of the second mode of the VT with its lumped parameter approximation.

The second VT mode has two pressure and two VV maxima both of which can be approximated by corresponding lumped parameters. According to the mode pattern of *Figure 2* the capacitance and inductance of the whole VT have to be divided into two parts: C=C/3+2C/3 and L=L/3+2L/3. The resulting transfer function has two resonances. The second, the higher one, is the true mode with an exactly correct frequency and magnitude values of 1.5 and 20 dB. The normalized resonance frequency of the lower one is about 0.75. This resonance is needed to “correct” the phase and magnitude of the true mode.

In the approximation made the corrective resonance at 0.75 is left out. The true mode is modeled alone with the lumped components C/3 and L/3. The capacitor is still located at the glottis and the inductor at the lip opening. Thus we leave out the closer description of the waves along the tract. The necessary phase correction can simply be approximated by changing the sign of the VV (the phase shift equals π). The amplitude correction can be made by scaling the output VV. After this reduction the second mode is modeled with a simple resonator just like in the first mode case.

The same procedure can be extented to the higher modes, too. Finally, only the most essential features are left: All modes have a pressure maximum at the glottis (over the capacitor) and flow maximum at the lips (through the inductor). They are essentially *parallel or superimposed* in the tract and do change energy through the *finite impedances* located at both terminals, i.e., through glottal impedance and lip radiation impedance. The details along the tract are not explicitely modeled. In the case of nonuniform profiles the mode frequencies alone are varied.

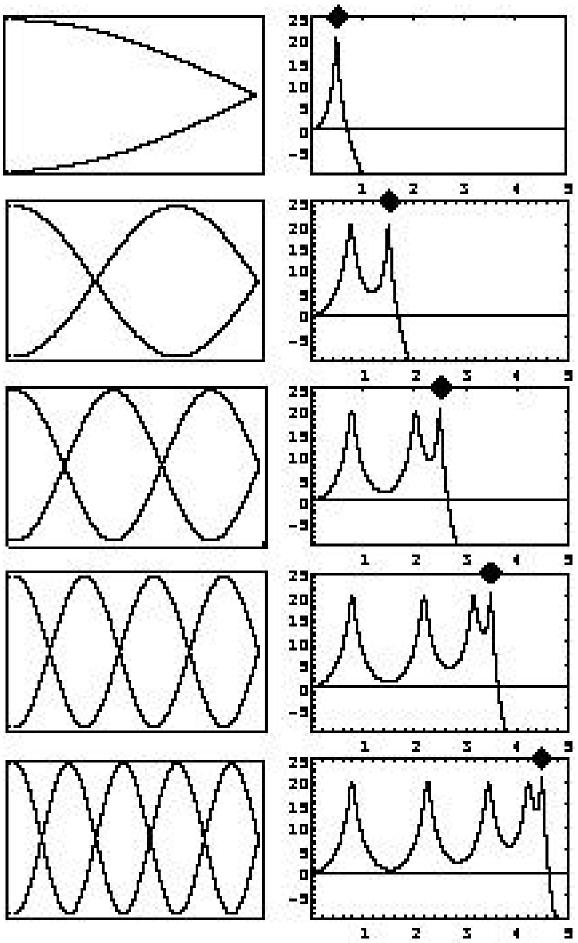

Figure 2. VT modes and transfer functions of their lumped parameter approximations. The square indicates the true mode.

The supraglottal pressure is built up by adding all the modal pressures and the lip VV by the sum of the modal VVs with the signs of the second, fourth, etc. modes inverted (the phase correction).

Correspondingly, the subglottal pressure is formed by the sum of the pressures of the subglottal modes (around 500 and 1000 Hz) when both of them are included in the model.

3. VT MODE AS A SYSTEM

Based on the simplifications made above a system model for the VT modes can be created (see *Figure 3*). Each mode is modeled with a system having two internal state variables: the pressure at the glottis and the VV at the lip opening. The system consists of two digital integrators which model a parallel type of resonator. One of the integrators simulates the behaviour of the inductor (integrates voltage to current) and the other simulates the capacitor (integrates current to voltage). Thus they produce the needed phase difference between the two state variables. This system was originally designed by Gold and Rader [3]. Here we just give a new interpretation and application for it.

The transglottal impedance is reduced to a variable resistance, the instantaneous value of which depends on the cross-sectional area of the glottal orifice. Classically, a cascaded variable inductance is included. However, the acoustical measurements have revealed that the glottal inductance almost vanishes when air flows through the orifice [4]. Thus the glottal inductance was not included in this preliminary model.

The instantaneous mode related the VV component through the glottis and can be solved from the transglottal pressure component and the transglottal resistance. The glottal VV component must be added to the VV variable of the mode. After that the integrators are updated and the new values for the mode pressure and VV are given at their outputs.

The mode VVs are fed into the lip radiation impedance in order to solve for the pressure p_L over it. The pressure over mode inductors (the input of the upper integrator in *Figure 3*) must be lowered by p_L in order to take into account the effect of the lip radiation load.

Since all of the supraglottal mode pressures affect the glottal flow they do change the subglottal pressure as well. This change is reflected back to the glottal flow. Thus the modes are interconnected via the finite glottal impedance. Note that this mutual coupling is strong only during the open glottal period.

Analogously, all of the mode loop VVs are fed to the lip radiation load where they build up a pressure p_L. This pressure will lower the pressure difference over the mode inductances thus lowering the mode VVs and causing mode interaction.

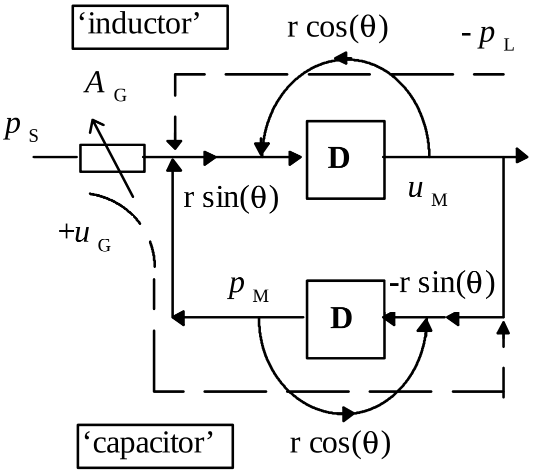

Figure 3. System model for a VT mode. State variables: p_M mode pressure, u_M mode volume velocity. Other variables: p_S subglottal pressure, A_G glottal area, , p_L pressure over the lip radiation impedance, and q normalized mode frequency. Parameter r defines the internal losses of the mode.

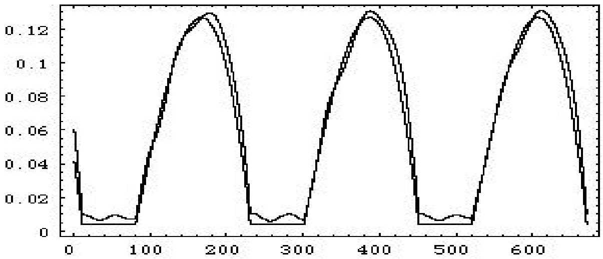

Figure 4. Glottal area function and VV (simulation of vowel /oe/).

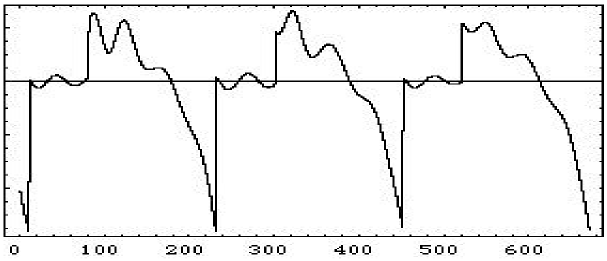

Figure 5. Differentiated glottal VV component of the first mode (simulation of vowel /oe/). x:time in samples, y: amplitude.

4. SIMULATIONS

A third order polynomial was used to model the glottal cross-sectional area. It was assumed that the glottal orifice has a tiny opening even during the closed phase as seen in *Figure 4*. The VV pulse is somewhat delayed due to the inductive load of the VT. The subglottal resonance at 500 Hz is seen during the closed (leaking) period and the VV component of the first VT mode during the open period.

A differentiated glottal VV generated in the similar simulation is depicted in *Figure 5*.

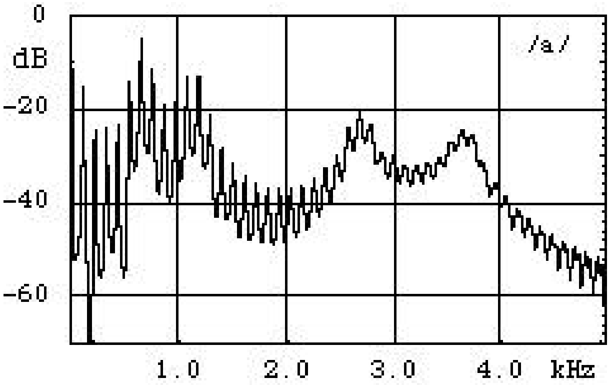

Figure 6. Spectrum of simulated Finnish /a/ vowel.

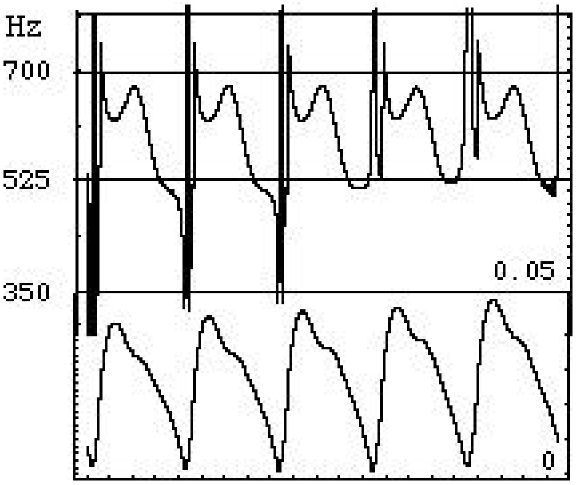

Figure 7. Instantaneous frequency (upper frame) and instantaneous amplitude of the first mode of simulated Finnish /a/ vowel.

Here the first VT mode is present alone. The simplified glottal area function leads to a somewhat unnaturally shaped waveform.

A spectrum of simulated Finnish /a/ vowel is depicted in *Figure 6*. All modes have a very small loss factor (r close to one). However, due to the interactions between the modes and the loads at both terminals the average bandwidths of the formants are broadened having quite realistic values.

The instantaneous formant frequency and the Hilbert envelope (instantaneous amplitude) of the first mode of /a/ are shown by *Figure 7*. The average closed phase frequency is below 700 Hz and the frequency is lowered to about 525 Hz during the open phase. The instantaneous amplitude is strongly attenuated during the open phase. These simulation result are close to those monitored in real speech samples [5]. Thus this simple model is able to create all the most essential features found in vowels even including fast pitch-synchronous modulation effects.

The synthetized vowels were compared to those produced with a conventional parallel type of synthesizer and by informal listening a clear quality improvement in the vowel sound was perceived especially in the low pitch male voices. The nasality of the synthetic vowels without these pitch-synchronous modulations do almost disappear when the modulations are included.

5. DISCUSSION

The starting point of this model was the uniform vocal tract. In nonuniform cases the mode pressures at the glottis and the mode VV at the lip opening may not follow this simple scheme. In those cases it may be needed to attenuate the mode pressure in order to indicate large C value or to attenuate the mode VV to indicate the large L value. This means that the LC product determing the mode frequency is not enough and the L/C must also be taken into account.

There are many other details which must be studied furher and in more details. More sophisticated glottal model is needed in order to study the perceptual meaning of its parameters. However, the synthesis model showed nice behaviour with different glottal area functions. Those variations were like variations between different persons or like changes in the phonation type.

6. ACKNOWLEGEMENT

This study has partially been financed by the Academy of Finland.

7. REFERENCES

[1] Hanson H. M., Maragos P. and, Potamianos A. (1993), Finding speech formants and modulations via energy separation: with application to a vocoder, *Proc. of IEEE ICASSP-93*, pp. II-716-719, Minneapolis, Minnesota,.

[2] Potamianos A., Maragos P. (1997), Speech analysis and synthesis using an AM-FM modulation model, *Proc. of Eurospeech'97*, pp. 1355-1358, Rhodes, Greece.

[3] Gold B. and Rader C. (1969), *Digital Processing of Signals*, McGraw-Hill, NY.

[4] Laine, U. K., Karjalainen M. (1986), Measurements on the effects of glottal opening and flow on the glottal impedance, *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP'86), vol. 3, pp. 1621-1624, Tokyo, Japan, April 7-11, 1986.

[5] Laine U. K., Laukkanen A.-M., and Leino T. (1999), Analysis of vowel production by monitoring pitch-synchronous modulations in formants, (to be appear in) *Proc. of Int. Workshop on Models and Analysis of vocal Emissions for Biomedical Applications*, Firenze.