ARTICULATORY SPEECH SYNTHESIZER

Chang-Shiann Wu
Department of Information Management
Shih Chien University
Kaohsiung, 845 Taiwan
chwu4142@ms5.hinet.net

Yu-Fu Hsieh
HarveTech DSP Corporation
LungTan
TaoYuan, 325 Taiwan
aliceufo@ms1.hinet.net

ABSTRACT

The aim of this research was to develop a flexible, high quality articulatory speech synthesis tool. One feature of this research tool is the simulated annealing optimization procedure that is used to optimize the vocal tract parameters to match a specified set of formant characteristics. Another aspect of this study is the derivation of a new form of the acoustic equations. A transmission-line circuit model of the vocal system, which includes the vocal tract, the nasal tract with sinus cavities, the glottal impedance, the subglottal tract, the excitation source, and the turbulence noise source, was constructed. The acoustic equations of the vocal system were rederived for the proposed articulatory synthesizer. A digital time-domain approach was used to simulate the dynamic properties of the vocal system as well as to improve the quality of the synthesized speech.

1. INTRODUCTION

Articulatory synthesis is the production of speech sounds using a model of the vocal tract, which directly or indirectly simulates the movements of the speech articulators. It provides a means for gaining an understanding of speech production and for studying phonetics. In such a model coarticulation effects arise naturally, and in principle it should be possible to deal correctly with glottal source properties, interaction between the vocal tract and the vocal folds, the contribution of the subglottal system, and the effects of the nasal tract and sinus cavities.

Articulatory synthesis usually consists of two separate components. In the articulatory model, the vocal tract is divided into many small sections and the corresponding cross-sectional areas are used as parameters to represent the vocal tract characteristics. In the acoustic model, each cross-sectional area is approximated by an electrical analog transmission line. To simulate the movement of the vocal tract, the area functions must change with time. Each sound is designated in terms of a target configuration and the movement of the vocal tract is specified by a separate fast or slow motion of the articulators.

A properly constructed articulatory synthesizer is capable of reproducing all the naturally relevant effects for the generation of fricatives and plosives, modeling coarticulation transitions as well as source-tract interaction in a manner that resembles the physical process that occurs in real speech production. Articulatory synthesizers will continue to be of great importance for research purposes, and to provide insights into various acoustic features of human speech.

2. REALIZATION OF THE ARTICULATORY MODEL

Geometric data concerning the vocal tract is essential to our understanding of articulation, and is a key factor in speech production. According to the acoustic theory of speech production, the human vocal tract can be modeled as an acoustic tube with nonuniform and time-varying cross-sections. It modulates the excitation source to produce various linguistic sounds. The success of articulatory modeling depends to a

large extent on the accuracy with which the vocal tract cross-sectional area function can be specified for a particular utterance. Measurement of the vocal tract geometry is difficult. Several researchers have proposed analytical methods to derive the vocal tract cross-sectional area function from acoustic data.

Articulatory models can be classified into two major types: parametric area model and midsagittal distance model. The parametric area model describes the area function as a function of distance along the tract, subject to some constraints[1][2]. The area of the vocal tract is usually represented by a continuous function such as a hyperbola, a parabola, or a sinusoid. The midsagittal distance model describes the speech organ movements in a midsagittal plane and specifies the position of articulatory parameters to represent the vocal tract shape. Coker and Fujimura (1966) introduced an articulatory model with parameters assigned to the tongue body, tongue tip, and velum. Later this model was modified to control the movements of the articulators by rules[3].

Another articulatory model was designed by Mermelstein. His model can be adjusted to match the midsagittal X-ray tracings accurately. Our articulatory model is a modified version of Mermelstein's model[4]. A set of variables is used to specify the inferior outline of the vocal tract in the midsagittal plane (Figure 1). These variables, called articulatory parameters, are the tongue body center, the tongue tip, jaw, lips, hyoid, and velum. A modification of the lower part of the pharynx and tongue-tip-to-jaw region is also provided and included in our model.

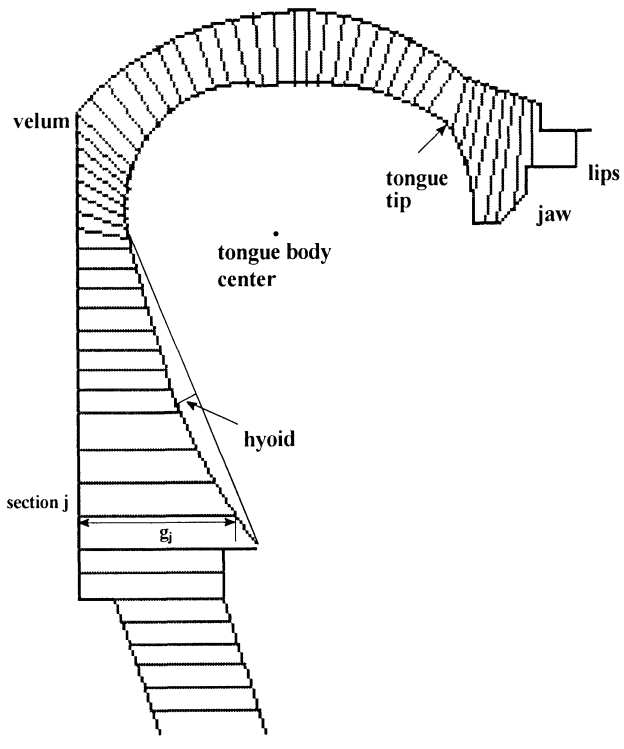


Figure 1: Articulatory model's parameters and midsagittal grids.

Once the articulatory positions have been specified, the cross-sectional areas are calculated by superimposing a grid structure on the vocal tract outline. These grid lines vary with the positions of the articulators (they are fixed in Mermelstein's model). A total of 60 sections, 59 sections for the vocal tract plus one section (fixed length and area) for the outlet of the glottis, are used in our model. The sagittal distance, g_j of section j, is defined as the grid line segment length between posterior-superior and anterior-inferior outlines. The center line of the vocal tract is formed by connecting the center points of the adjacent grid lines. The length of the center line is considered equivalent to the length of the vocal tract. The sagittal distances are eventually converted to cross-sectional areas by empiric formulas.

The calculation of formant frequencies from a given vocal tract cross-sectional area function has been well established in the acoustic theory of speech production. By computing the acoustic transfer function of a given vocal tract

configuration, we can decompose the formant frequencies from the denominator of the acoustic transfer function. One of the functions of the articulatory model is to compute the articulatory information (in particular, the vocal tract cross-sectional area) from the acoustic information (the first four formant frequencies in our study) that are obtained from the speech signal. Here we attempt a new solution using the simulated annealing algorithm, which is a "constrained multidimensional nonlinear optimization problem." The coordinates of the jaw, tongue body, tongue tip, lips, velum, and hyoid compose the multidimensional articulatory vector[5]. A comparison between the model-derived and the target-frame first four formant frequencies forms the cost function. There are two constraints: (1) the articulatory-to-acoustic transformation function, and (2) the boundary conditions for the articulatory parameters. The optimum articulatory vector is obtained by finding the minimum cost function. Once the optimum articulatory vector is determined, the articulatory model determines the vocal tract cross-sectional area function which in turn is used by the articulatory speech synthesizer[6][7][8].

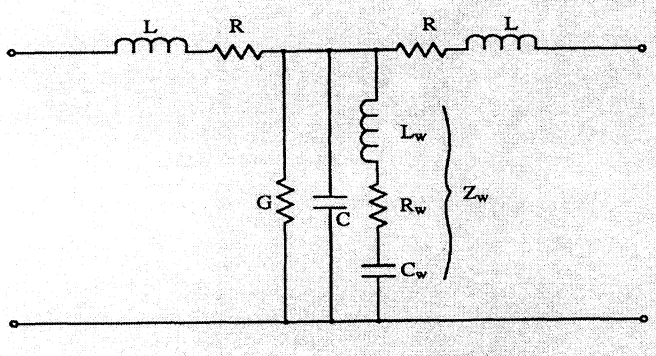
3. REALIZATION OF THE ACOUSTIC MODEL

Basically, the acoustic model of the human vocal system embodies several submodels. Both the vocal tract and nasal tract models simulate the sound propagation in these tracts. The excitation source model represents and generates the voiced excitation waveforms for the vocal tract. The turbulent air flow at a constriction for fricatives and plosives is generated by the noise source model. The radiation model simulates the acoustic energy radiating from the lips and the nostrils. A transmission-line circuit model of the vocal system, which includes the vocal tract, the nasal tract with sinus cavities, the glottal impedance, the subglottal tract, the excitation source, and the turbulence noise source, was constructed. The acoustic model of each subsystem of the vocal system was analyzed.

Acoustical parameter	Electrical parameter
p – Pressure	v – Voltage
u – Volume velocity	i – Current
ϱ/A – Air mass inertia (acoustic inductance)	L – Inductance
$A/(\varrho c^2)$ – Air compressibility (acoustic capacitance)	C – Capacitance
Viscous loss	R – Series resistance
Heat conduction loss	G – Shunt resistance
Yielding wall	Z_w – Shunt impedance

Table 1 Acoustical / electrical analogues.

Figure 2: the equivalent circuit representation.



Transmission-line analogs of the vocal tract (or equivalent electrical circuit model) is based on the similarity between the acoustic wave propagation in a cylindrical tube and the propagation of an electrical wave along a transmission line. The derivation from the basic equations of acoustic wave propagation to an equivalent electrical quadripole representation is well known [9][10]. The analogs are summarized in Table 1. Figure 2 is an equivalent circuit representation of a soft-wall, lossy cylindrical tube. The series resistor R is used to represent the acoustic loss due to viscous drag in which the energy loss is proportional to the square of the volume velocity. The shunt conductance G represents the loss due to heat conduction, which is proportional to pressure squared. The shunt impedance is the acoustic equivalent mechanical impedance of the yielding wall. This wall impedance, which represents a mass-compliance-viscosity loss of the soft tissue, has three components. Note

that both R and G are a function of frequency.

The vocal tract was approximated by a non-uniform, lossy, soft wall, straight tube with 60 concatenated elemental sections (circular or elliptic). The transmission-line analogy approach was used to model the vocal tract as an equivalent circuit network. A series resistor represents the viscous loss and a shunt conductance represents the thermal loss. The yielding wall vibration loss was modeled by a shunt impedance. The effect of the sinus cavities on the nasal consonants and nasalized vowels was discussed. The sinus cavity was regarded as a Helmholtz resonator and was modeled as a shunt impedance. Flanagan's model (1972) was considered the most appropriate radiation model for the time-domain articulatory synthesis.

For the non-interactive excitation source, we simplified the unified glottal excitation model[11] that includes the jitter model and shimmer model into the LF model. For the interactive excitation source, we proposed a new model, which consists of the unified glottal excitation model, the subglottal model, and the glottal area model. The subglottal system was modeled by three cascaded RLC Foster circuits (Ananthapadmanabha and Fant, 1982). The triangular, sine, and raised-cosine functions were used as options to model the time-varying glottal area function[12].

For the turbulence noise source model, the distributed and series pressure noise source model[13] and the downstream parallel flow source model[14][15] were discussed. The parallel flow source model was adopted for this study. The turbulence noise source can be located 1) at the center of, 2) immediately downstream from, 3) upstream from, and 4) spatially distributed along the constriction region.

A practical articulatory synthesizer was proposed that included the vocal tract, the nasal tract with sinus cavities, the glottal impedance, the subglottal system, the excitation source, and the turbulence noise source. The acoustic equations of the vocal system were derived for the proposed articulatory synthesizer. The time-domain approach was used to simulate the dynamic properties of the vocal system as well as to improve the quality of the synthesized speech. The vocal tract cross-sectional area or the articulatory parameters were interpolated between two consecutive target frames using a linear or arctan function.

4. RESULTS AND CONCLUSION

The vocal tract tube can be described by two coupled partial differential acoustic equations. These two acoustic equations are functions of both time and space. Approximating the vocal tract as a sequence of elemental sections corresponds to digitizing the vocal tract in space, i.e., spatial sampling. For each elemental section, the transmission-line analog approach is applied to form the equivalent circuit model, as seen in Figure 2. Connecting the equivalent circuit of each section together in combination with the equivalent circuit models of the other parts of the vocal system (subglottal system, glottis, and nasal sinus cavities), a lumped circuit network representation of the vocal system can be formed, as shown in Figure 3. For the time-domain approach, the Kirchoff's and Ohm's laws are applied to the circuit network to obtain sets of differential equations. These differential equations, which correspond to the equivalent acoustic equations that govern the generation and the propagation of acoustic waves inside the vocal system, are transformed into discrete-time representations. This appendix provides a detailed derivation of the discrete-time acoustic equations, i.e., the difference matrix equations. The discretization scheme is similar to the work of Maeda (1982a)[16]. Our model, however, provides more features, such as the subglottal system, nasal sinus cavities, and turbulence noise source.

Figure 4 presents the articulatory characteristics for /I/ and /i/ vowels. The midsagittal vocal tract outline and the corresponding synthetic speech waveform are obtained from sustained vowel phonations by using the simulated annealing algorithm. The articulatory synthesis windows (see Figure 5) is divided into twelve subareas. During the synthesis of speech these subareas are used to display the following messages and waveforms: (1)the target- and excitation-frame messages, (2)the vocal tract cross-sectional area function, the acoustic transfer function, and the midsagittal vocal tract outline of the current target frame, (3)the

excitation waveform and its power spectrum, (4)the pressure and the volume velocity waveforms at different places in the vocal tract, (5)the articulatory trajectories, and (6)the synthetic speech waveform.

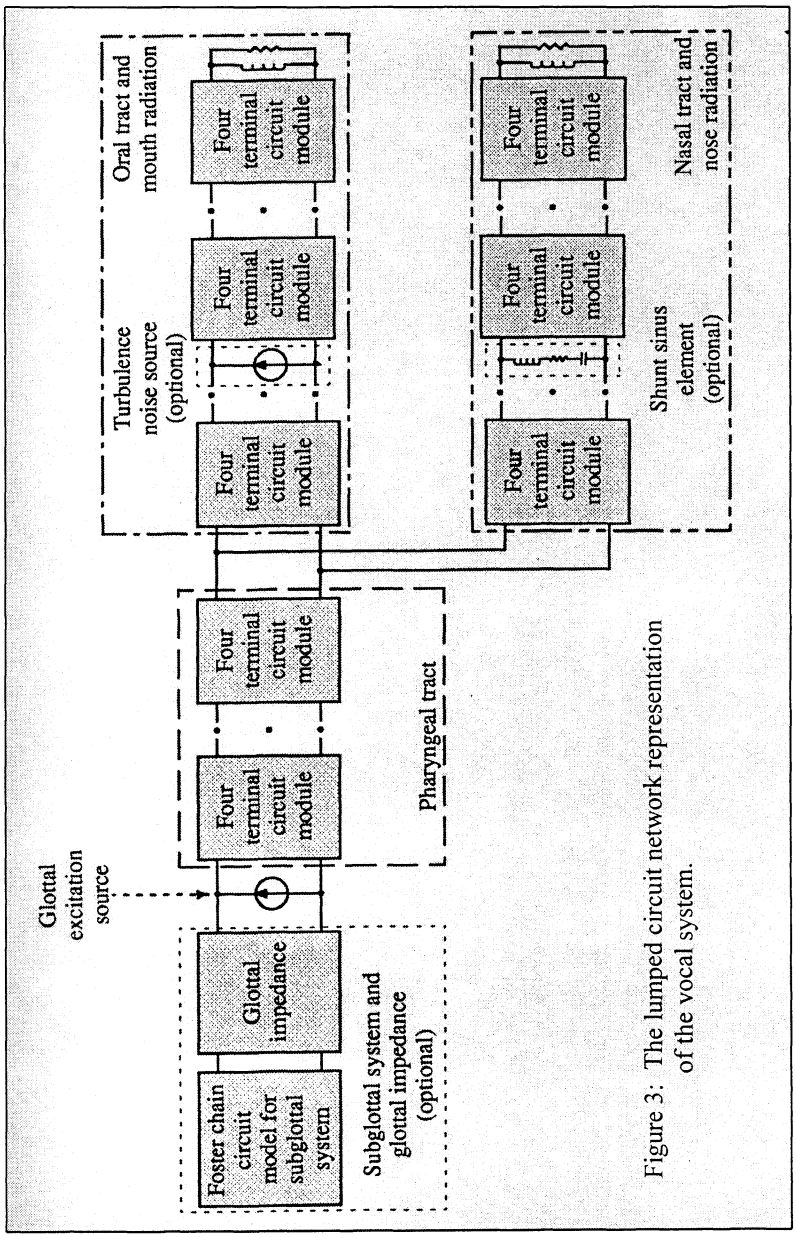


Figure 3: The lumped circuit network representation of the vocal system.

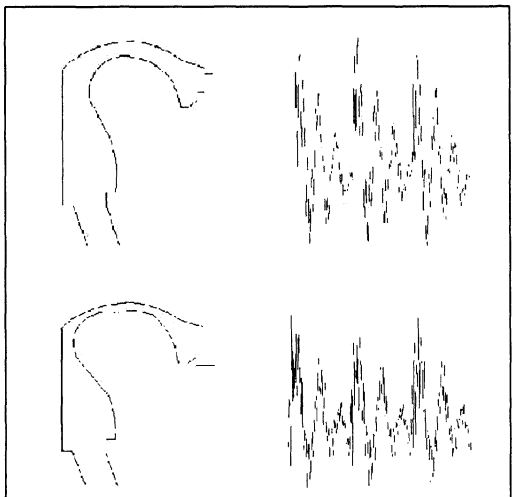


Figure 4:The articulatory characteristics for /I/ and /i/ vowels.

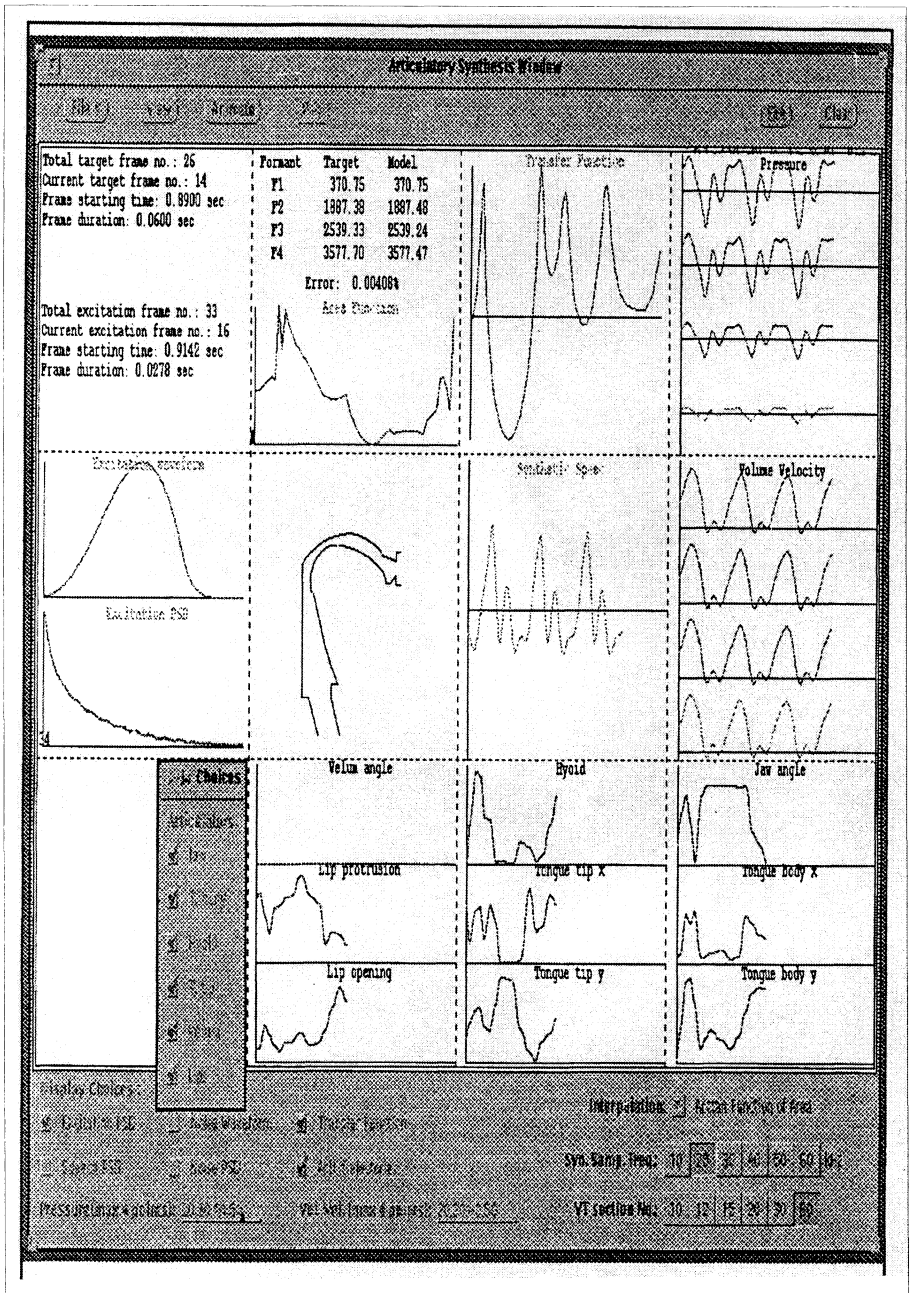


Figure 5: The synthesis windows.

5. REFERENCES

[1] Fant, G. (1960). Acoustic Theory of Speech Production, Mouton and Co., Gravenhage, The Netherlands.

[2] Lin, Q. G. (1990). "Speech production theory and articulatory speech synthesis," Ph.D. dissertation, Royal Institute of Technology, Stockholm, Sweden.

[3] Coker, C. H. (1976). "A model of articulatory dynamics and control," Proc. IEEE, 64(4), 452-460.

[4] Mermelstein, P. (1973). "Articulatory model for the study of speech production," J. Acoust. Soc. Am., 53(4), 1070-1082.

[5] Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am., 63(5), 1535-1555.

[6] Wakita, H., and Fant, G. (1978). "Toward a better vocal tract model," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 1, 9-29.

[7] Badin, P., and Fant, G. (1984). "Notes on vocal tract computation," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 2-3, 53-108.

[8] Fant, G. (1985). "The vocal tract in your pocket calculator," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 2-3, 1-19.

[9] Flanagan, J. L. (1972). Speech Analysis, Synthesis, and Perception, Springer-Verlag, Berlin, Germany.

[10] Linggard, R. (1985). Electronic Synthesis of Speech, Cambridge University Press, Cambridge, England.

[11] Lalwani, A. L., and Childers, D. G. (1991). "Modeling vocal disorders via formant synthesis, " Proc. ICASSP, 1, 505-508.

[12] Ananthapadmanabha, T. V., and Fant, G. (1982). "Calculation of true glottal flow and its components, " Speech Communication, 1, 167-184.

[13] Flanagan, J. L., and Cherry, L. (1968). "Excitation of vocal-tract synthesizers, " J. Acoust. Soc. Am., 45(3), 764-769.

[14] Sondhi, M. M., and Schroeter, J. (1986). "A nonlinear articulatory speech synthesizer using both time- and frequency-domain elements, " Proc. ICASSP, 1999-2002.

[15] Sondhi, M. M., and Schroeter, J. (1987). "A hybrid time-frequency domain articulatory speech synthesizer, " IEEE Trans. Acoust., Speech, and Signal Processing, 35(7), 955-967.

[16] Maeda, S. (1982a). "A digital simulation method of the vocal-tract system, " Speech Communication, 1(3-4), 199-229.