ABSTRACT

This paper describes a new time-domain articulatory speech synthesis model. The system consists of a model of the speech articulators, a time-domain simulation of a lossy acoustic tube model of the vocal tract and a two-mass model of the vocal cords. The vocal tract area function is expressed as a transformation of ten articulatory variables which provide a two-dimensional geometrical description of the vocal tract in the mid-sagittal plane. The acoustic tube model is based on the Kelly-Lochbaum structure which uses the method of reflection coefficients to compute the forward and backward sound pressure waves in a lossless digital transmission line analogue of the vocal tract. The basic Kelly-Lochbaum structure has been extended to permit the modelling of viscous friction losses, yielding wall losses, radiation from the mouth, variable vocal tract length, aspiration and frication.

1. INTRODUCTION

The basic aim of most speech synthesis systems is to model the speech production process in one way or another. Although many of the most recent speech synthesis systems produce highly intelligible speech, most sound very mechanical, which makes them unacceptable for prolonged listening. One of the reasons for this lack of naturalness is that they are based on a source-system model of the speech production process, in which it is assumed that the source, i.e. the modulated airflow, is linearly separable from the system, i.e. the acoustic characteristics of the vocal tract. This is a fairly gross approximation, since it is well known that the generation and propagation of acoustic waves inside the vocal tract is governed by a single set of acoustic equations. In addition, in most current speech synthesis systems, such natural speech processes as co-articulation can only be simulated approximately in the acoustic domain. It may be expected, therefore, that a speech synthesis system in which the motion of the speech articulators, the acoustic and mechanical properties of the vocal tract and the vocal-cord/vocal-tract interaction, are taken into account, should be capable of producing much more natural-sounding synthetic speech.

This paper describes a time-domain articulatory speech synthesis model, programmed in PASCAL on a VAX minicomputer, which models directly the motion of the speech articulators and the generation and propagation of sound inside the vocal tract. A block diagram of the system is given in Fig 1. It consists essentially of three main components - an articulatory model for converting articulatory positions to a vocal tract area function, a time-domain acoustic tube model of the vocal tract and a two-mass model of the vocal cords.

2. ARTICULATORY MODEL

At present, the articulatory model is essentially that due to Mermelstein [1]. This model represents the vocal tract outline as a function of ten variables which specify the position of the jaw, tongue, lips, velum and hyoid. As shown in Fig. 2, this provides a two-dimensional outline of the vocal tract shape, viewed in the mid-sagittal plane. The position of the jaw and hyoid are expressed directly in a...
fixed co-ordinate system. Lip and tongue body positions are specified with respect to the moving jaw. The tongue body and tongue blade are considered as separate but interdependent articulators with the tongue tip being specified relative to the tongue body. The tongue body outline is modelled as a circle with a moving centre and fixed radius. The lips are allowed to open (close) or protrude (retract) relative to the jaw and maxilla.

The transformation from a vocal-tract outline to a vocal-tract area function involves a 2-stage process. Firstly, the mid-sagittal distance is extracted by superimposing a grid over the vocal-tract outline as shown in Fig. 2. The mid-sagittal distance is measured between the intersection of the outlines with the grid-lines. A line is plotted through the centre of these distances and measured to give the vocal-tract length. The mid-sagittal distance function is converted into a cross-sectional area function using a transformation which takes account of the direction of speech propagation, the distance between the centre of the mid-sagittal distances and the three-dimensional shape of the vocal-tract. Finally the vocal-tract area is spatially smoothed by low-pass filtering.

3. ACOUSTIC TUBE MODEL

The vocal tract is represented by a time domain simulation of a discrete acoustic tube model consisting of 35 uniform sections, each 0.5 cm long. The propagation of sound within the vocal tract is modelled by computing the forward and backward travelling sound pressure waves (Kelly and Lochbaum [2]). Laminar flow losses due to viscous friction between the air and the vocal tract walls is lumped at the input and output of adjacent lossless tubes as shown in Fig. 3 (Adamson and Linggard [3]). If a circular tract is assumed then the laminar or viscous resistance of the kth section, RvK, approximates to that of a circular duct (Stephens [4]). The complete vocal tract is modelled as a series of 35 lossless sections interconnected with lumped viscous loss resistances.

Sound pressure wave behaviour at the junction of any two sections can be described in terms of wave propagation and reflection. For the forward travelling wave, the reflection coefficient, r_f, is given by

\[ r_f = \frac{R_k + R_{k+1} + Z_{k+1} - Z_k}{R_k + R_{k+1} + Z_k} \]  

(1)

For the backward travelling wave, the reflection coefficient, r_b, is given by

\[ r_b = \frac{R_k + R_{k+1} + Z_{k+1} - Z_k}{R_k + R_{k+1} + Z_k} \]  

(2)

where Z_k = ρc/A_k is the characteristic acoustic impedance of the kth section, ρ is the density of air and c is the velocity of sound in air. Wave behaviour between adjacent sections can therefore be described by the discrete-time equations.

\[ P_k(n+1) = \frac{1}{1+\frac{R_k}{Z_k}} P_k(n) + \frac{\frac{1}{Z_k}}{1+\frac{R_k}{Z_k}} P_{k+1}(n) \]  

(3)

\[ P_k(n) = \frac{1}{1+\frac{R_k}{Z_k}} P_k(n) + \frac{\frac{1}{Z_k}}{1+\frac{R_k}{Z_k}} P_{k-1}(n) \]  

(4)

where \( P_k(n) \) and \( P_{k+1}(n) \) denote the forward and reverse travelling waves, respectively, at time n in the kth section. The sample period, T, is twice the time taken for sound to propagate across each section of tube, i.e. T = 20L/c.

Losses due to the vibration of the non-rigid vocal tract walls are incorporated in the model by assuming independent motion of the wall of each section of vocal tract. Assuming circular sections and denoting the pressure and the target area in the kth section by \( P_k \) and \( A_k \) respectively, the motion of its walls may be described by the differential equation

\[ 2 \pi f_k \Delta L_k = M \frac{d^2}{dt^2} (r_k - r_x) + D \frac{d}{dt} (r_k - r_x) + Q \frac{d}{dt} (r_k - r_x) \]  

(5)

where M is the mass, D is the damping and Q is the compliance of the vocal tract walls. The values of M, D and Q are assumed to be constant for each section and are estimated from the data reported by Ishizaka et al [5] who list the following per unit area parameters: M = 1.5 g; D = 1060 gm/s; Q = 33.3 x 10^3 dyn/cm. Actual values of the parameters are approximately estimated by assuming a constant surface area of 4 cm^2 in each section. The pressure change \( P_k \) due to the vibrating walls in the kth section of tract is given by

\[ P_k = P_k[1 - \frac{A_k}{(A_k + \delta A_k)}] \]  

(6)

where \( A_k \) is the target area and \( \delta A_k \) is the change in area of the kth section.

The impedance to sound radiation from the mouth is modelled as an equivalent parallel R-L circuit, where the values of R and L are those for a circular vibrating piston in an infinite plane baffle.

In normal speech production the vocal-tract length varies between about 16 cm and 19 cm. This is simulated in the model by keeping the number of sections fixed at 35 but allowing the length of each section to vary. This leads in effect to a system with a varying sample rate. Conversion to a fixed sample rate signal is achieved by signal reconstruction and re-sampling. This process is carried out by a time-varying FIR low-pass digital filter.

The nasal tract is modelled in much the same way as the oral tract. It is 12.5 cm in length and is coupled to the oral tract at a point 8 cm from the glottis. Most sections, except for a few in the vicinity of the velum, have a fixed cross-
sectional area. A single sinus cavity model is also incorporated in a similar way. This cavity model is coupled 7 cm from the velum and is 4.5 cm in length and is of fixed cross-sectional area.

4. VOCAL CORD MODEL

The vocal cord model used in the system (Fig 4) is the two-mass model originally developed by Ishizaka and Flanagan [6] and later modified by Sondhi and Schroeter [7]. The non-linear springs $s_1$ and $s_2$ model the tension in the vocal cords and the linear coupling spring $k_c$ models the effect of their flexural stiffness in the lateral direction. The glottal volume velocity $U_g(t)$ satisfies the following differential equation

$$R_{tot} U_g + L_{tot} \frac{dU_g}{dt} = P_u - P_i - P_w$$

(7)

where $P_u$ is the subglottal pressure, $P_i$ is the pressure in the first section of the vocal tract, $P_w$ is a series noise pressure source for modelling aspiration and $R_{tot}$ and $L_{tot}$ are the total resistance and inductance representing the contraction, glottis and expansion. $R_{tot}$ and $L_{tot}$ are dependent upon the glottal areas $A_1$ and $A_2$, and are calculated from analytic expressions (Ishizaka et al [5]). $A_1$ and $A_2$ are expressed in terms of the lateral displacements $x_1$ and $x_2$ as

$$A_1 = A_{01} + 2l_u x_1$$
$$A_2 = A_{02} + 2l_u x_2$$

(8)

where $A_{01}$ and $A_{02}$ are normally set to the same glottal rest area $A_0$. The lateral displacements $x_1$ and $x_2$ are determined from the equations of motion of the two masses, i.e.

$$m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} + s_1 + k_c (x_1 - x_2) = f_1$$
$$m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} + s_2 + k_c (x_2 - x_1) = f_2$$

(9)

where $r_1$ and $r_2$ are the damping resistances, $s_1$ and $s_2$ are the spring forces and $f_1$ and $f_2$ are complex functions of $P_i$ and $U_g$. Numerical values for $m_1$, $m_2$, $r_1$, $r_2$, and $k_c$, and expressions for $s_1$, $s_2$, $f_1$, and $f_2$ are those given by Ishizaka and Flanagan [6] and Sondhi and Schroeter [7].

The interaction between the vocal cord model and the vocal tract model can be expressed by

$$P_1 = 0.5(1 - r_p)(P_u - P_w) + r_p P_i$$

(10)

Equations (7), (8), (9), and (10) are all coupled and are solved simultaneously for each sample instant.

5. TURBULENT NOISE GENERATION

Turbulent noise models for the automatic production of aspiration and frication have been incorporated into the structure already described. The friction model is based on the research of Shirai [8], and allows for automatic determination of the location of the turbulence as well as the frequency characteristics of the noise source, both of which are flow-dependent.

During turbulent airflow, the microscopic resistances in the vicinity of the constriction increase due to the additional kinetic energy of the fluctuating eddies and the additional friction caused by the viscosity of the air. This additional loss resistance is added in series with the lumped viscous resistance in Fig 3. The value of the kinetic resistance $R_k$ is given by

$$R_k = \frac{K \rho U_{dc}}{2A_c}$$

(11)

where $K$ is a constant depending on the non-uniformity of the constriction, $\rho$ is the density of air, $U_{dc}$ is the d.c. airflow and $A_c$ is the area of the constriction. The value of the d.c. airflow is obtained from the d.c. equivalent circuit as shown in Fig 5.

Turbulent airflow is generated when the local Reynolds number $R_e$ at any point along the vocal tract rises above a critical value $R_e (crit)$. The local Reynolds' number, $R_e$, for a point of maximum constriction is given by

$$R_e = \frac{4 \rho U_{dc}^2}{\mu A_{crit}}$$

(12)

where $\rho$ and $\mu$ are the density and viscosity of air respectively and $A_{crit}$ is the area of the constriction. A volume-velocity noise source $U_n$ given by

$$U_n = \frac{2.178 N(R_e^2 - R_e^2(crit))}{R_t}, \quad R_e > R_e (crit)$$

(13)

$$R_n = R_e (crit)$$

where $N$ is a constant depending on the non-uniformity of the constriction, $\rho$ is the density of air, $U_{dc}$ is the d.c. airflow and $A_c$ is the area of the constriction. A volume-velocity noise source $U_n$ given by

$$U_n = \frac{2.178 N(R_e^2 - R_e^2(crit))}{R_t}, \quad R_e > R_e (crit)$$

(13)

$$R_n = R_e (crit)$$

\[\text{Fig 4 : Two-Mass Model of the Vocal Cords}\]

\[\text{Fig 5 : DC Equivalent Circuit}\]
Fig 6: Location of Noise Source

where \( g_{s+} \) is a constant amplitude control and \( N \) is a random number uniformly distributed between -0.5 and 0.5, is placed between the first and second sections after the constriction as shown in Fig 6, i.e. at the expanding portion where the turbulent airflow is at a maximum. A volume-velocity source rather than a pressure source is used because its positioning is not as critical (Shadle [9]).

The characteristics of the noise source are made to reflect the nature of the turbulence by filtering the output white noise with a second-order bandpass filter with centre frequency, \( f_0 \), given by

\[
f_0 = 0.2 |U_{xy}| / A_c^{3/2}
\]  

Therefore the frequency characteristics of the noise source is dependent both on the airflow and on the dimensions of the constriction.

For the automatic production of aspiration, a similar procedure is used except that a noise pressure source is added to the pressure difference across the glottis.

6. CONCLUSIONS

The system which has been described seems to have the potential for producing natural-sounding synthetic speech when driven by time-varying articulatory or vocal tract area functions and an appropriate pitch contour. So far we have used the system to synthesise vowels, nasalised vowels, diphthongs, plosives, fricatives and some short words. Work is in progress to optimise the performance of the model by obtaining improved estimates of physiological parameters and to incorporate a model of articulatory dynamics.

ACKNOWLEDGEMENTS

Much of the work described in this paper was carried out with the support of a CAST Award from British Telecom Research Laboratories. Also, the help and advice of Dr John Local, Dr John Kelly and Dr John Coleman at the University of York is also gratefully acknowledged.

REFERENCES


