Reprinted with permission of the authors from *Proc. Fourth Intern. Congr. Acoust.*, Paper G42, 1–4 (1962)

SPEECH SYNTHESIS

10

John L. Kelly, Jr., and Carol C. Lochbaum, Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

Introduction

The goal of the investigations reported in this paper is the production of synthetic speech from an input consisting only of the names of phonemes and a minimum of pitch and timing information. The work was all done on a digital computer, the IBM 7090. Two general approaches were attempted, based respectively on a terminal analogue and an artificial vocal tract. The first approach uses for an intermediate language formant frequencies, bandwidths, and excitation parameters; while the second uses cross-sectional areas of the vocal tract, nasal coupling and excitation parameters. The two programs work on the same general principle: stored tables give values of the parameters (acoustic or articulatory, respectively) corresponding to each phoneme, and various ad-hoc rules cause smooth transitions to occur between phonemes. The ultimate output of each program is a digital tape representing the sound pressure waveform, which can be translated to analog. The terminal analog has been described elsewhere, and therefore we will concentrate mainly on the vocal tract analog.

The Artificial Vocal Tract

Several practical considerations served to determine the method of integrating the wave equation in the computer. The most important of these was running time. For reasons of economy, it was necessary that the program run in about forty times real time on an IBM 7090. Another requirement was that the output sampling rate be no greater than 20 kc. It was found that these requirements could be met be approximating the tract with 21 delay lines with one fortieth of a millisecond delay and characteristic impedance proportional to the area of the tract at a given place. Figure 1 shows a typical junction between sections and a signal flow diagram which is equivalent. The flow diagram shown in Fig. 2 is equivalent except that the ultimate output is delayed. Since the delays of Fig. 2 are the same as the sampling times, it can be simulated with difference equations in the computer.

The figures are included merely to give a rough idea of the principle involved. Many features are not shown and will be explained below:

Damping - Since there is no way to obtain frequency dependent damping with the model, a constant damping was inserted by attenuating the backward-going pressure wave at each junction. The amount (one part in 128) was chosen to give the formants the right average widths (and to avoid a multiplication).

Terminal Impedances - The reflection coefficient at the glottis is computed to correspond to source impedance of 0.2 cm^2 (a nice thing about computer simulation is that all proportionality constants may be taken to be one). This is the value used by Fant.[1] Strictly speaking the

[1] Gunnar Fant, "Acoustic Theory of Speech Production," published by Mouton and Co., 'S-Gravenhage, 1960.

radiation impedance is zero to the same order of accuracy that holds for the entire model. Nevertheless, some extra damping was included at the mouth end of the tract.

Nasal Tract - The nose was approximated by four sections three times as long as those of the oral tract and with more damping.

Voice Excitation - The excitation is obtained by operating on a pair of pulses (one for opening and one for closing of the glottis) with difference equations for spectral shaping. These correspond to several poles and zeros in the complex plane and tend to control voice quality.

Aspiration and Affrication - These are inserted at the glottis and place of greatest constriction respectively.

Control Methods

Before discussing the signal generator program, which controls the motions of the tract in response to a phonemic input, I will discuss in some detail the timing of the control system. This is not due to a desire to delve into programming techniques, but because some of the compromises made in the interest of speed might affect performance. To begin with, the tract configuration is held constant for intervals of ten samples (0.5 msec). It is changed over eight of these intervals (4 msec) by linearly interpolating the excitation parameters and reflection coefficients to new values determined by the signal generator program. Thus, it is only at the end of each 4 msec interval that new area information can be supplied.

The Signal Generator

To drive the tract, it is necessary to supply 26 parameters as functions of time. These consist of 21 cross-sectional areas, nasal coupling, pitch, buzz, aspiration, and affrication. The signal generator program accomplishes this by reading the names of phonemes off of input cards and following certain rules.

An imput card consists of 6 columns. The first contains the name of the phoneme, in a machine readable code equivalent to the IPA, and the next indicates the stress of vowels. The other 4 may be left blank or may be used to control loudness, pitch, transition time, and duration. If any of these columns are left blank, the program computes the corresponding parameters by rule.

The general principle of operation of the program consists of one simple rule and endless exceptions. The rule is that all parameters change linearly with time from their old values to a set obtained from stored tables, one for each phoneme. The transition time is a function of the old and the new phoneme. All parameters are then held constant for a duration determined by more tables. Some of the more interesting exceptions are the following:

Nasalized Vowels - Vowels adjacent to nasal consonants are partly nasalized.

Labials - B,M,P, and W have no fixed configuration. The stops are formed by closing the lips, leaving the remainder of the tract as it was. While the lips are closed the remaining sections change over to their values for the next sound. W is made similarly but with the lips only partly closed.

Neutral Tract - This is a tract configuration corresponding to a sound between æ and Λ . The tract returns to this position during silence. This serves to give a release on stops.

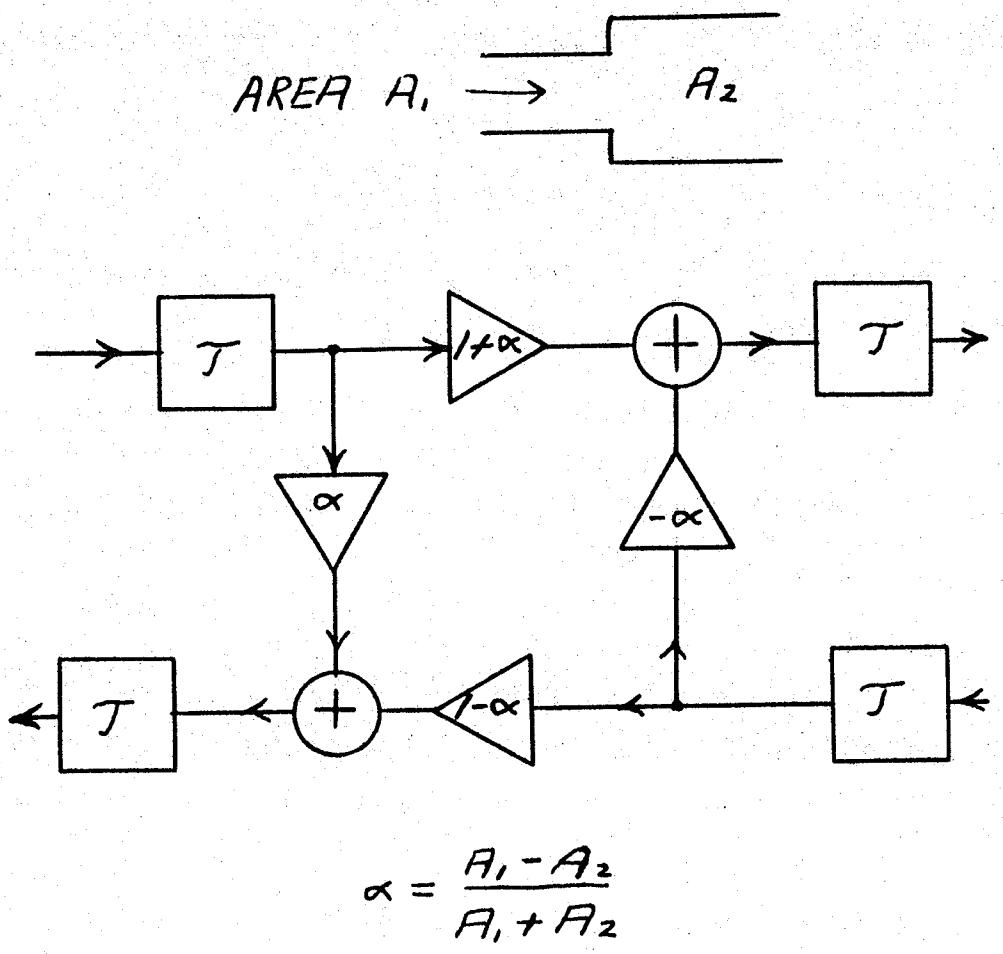
Unstressed Vowels - These are formed by averaging the configuration corresponding to the stressed vowel with that of the neutral tract. The duration is zero, i.e., these sounds are all transition.

Context - Some sounds (other than the labials mentioned above) are made by contextual rules. K,G, and NG use different configurations according as the adjacent vowels are front, back or middle. If they are different, the tract is changed during closure.

Conclusions and Results

At the time of writing of this paper the speech quality is far from satisfactory. The authors believe, however, that the trouble is mainly in the stored configurations and not in the principles involved. Accurate area information for all the sounds involved is difficult to acquire. The X-ray data in Fant's excellent book[1] served as a starting point, but experimentation is laborious with a digital computer. Also there is no simple way to correct configurations by the examination of spectrograms. The terminal analog is much better off in this respect. In fact the authors are convinced that a superior talking program could be written using a terminal analog driven by a program with the contextual features of the one described here.

Nevertheless, the vocal tract analog has many clear advantages. The complex sequence of spectral events occurring in the production of a consonant comes about quite naturally. For example, if excitation containing a DC component is left on during the closure time of a stop consonant, a burst will occur automatically upon release. It is quite possible that experimentation with this model will prove valuable in the design of a more satisfactory terminal analog.



$$\alpha = \frac{A_1 - A_2}{A_1 + A_2}$$

Figure 1 - Junction of two sections and equivalent flow diagram.

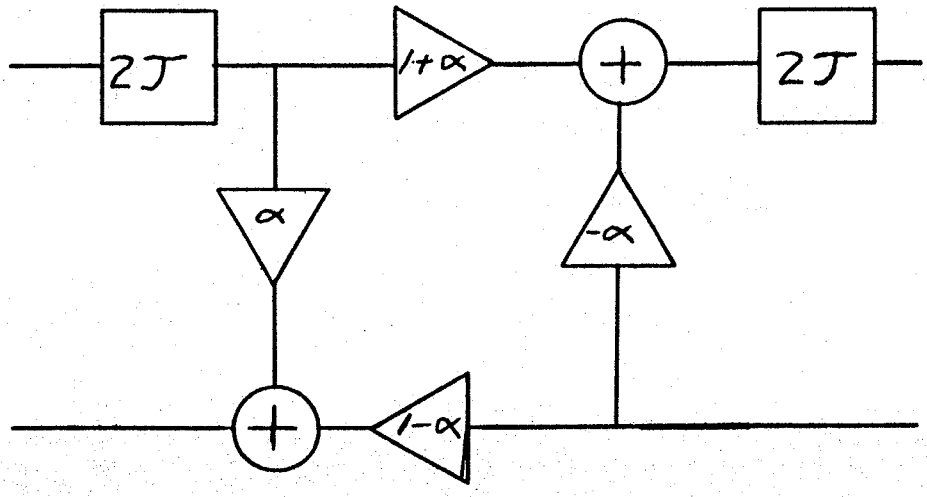


Figure 2 - Simplified flow diagram.