EECS225d Audio Signal Processing in Humans and Machines	Spring 1995
Lecture 7: February 1, 1995	
Lecturers: Professors Nelson Morgan and Ben Gold	*Scribe: Ted Lai*

7.1 Clarification of last lecture (Lecture 6) by Prof. Morgan

Main points about the content of last tape:

(1) It is an example of poor user interface.

(a) The instruction for user is not clear.

(b) People are not consistent in answering questions.

(2) It also shows technical limitation.

(a) Machine cannot distinguish between proper answers or mumblings.

(b) Machine's inherit error may cause a problem in recognizing a long string of number. (e.g., a machine with one-percent recognition error per digit could have an overall error in recognizing a fifteen-digit number of 1 - $(0.99)^{15}$ = 14 percent.)

(c) Voice may change over real telephone line.

7.2 Professor Morgan's lecture

A constraint representation: **production** $\longleftrightarrow$ **perception**

For the next few lectures, the focus will be on **production** and later on **perception**. For example, this Friday's lecture by Dr. John Ohala will be on articulators, and later lectures Prof. Gold on production of speech and music.

A model for production is following.

<u>sources</u>

buzz $\longrightarrow$

$\Longrightarrow$ **resonant tube** $\Longrightarrow$ **high pass filter**

hiss $\longrightarrow$

(1) The sources are following:

(a) **Buzz** is a periodic repetition of glottis closing and has a lowpass spectrum (12 dB/oct).

(b) **Hiss** is turbulant, fricative, flat, and acts like white noise.

(2) The **resonant tube** is like Prof. Gold's acoustic tube, a model of vocal tract. Another path is via nasal pathway, which acts like anti-resonance.

(3) The **high pass filter** consists of radiation characteristic of nasal pathway and lips.

7.3 Professor Gold's lecture

7.3.1 Review of equation of an acoustic tube opened at one end and excited on the other end.

$$u(x,t) = \frac{cos(\frac{\Omega(\ell - x)}{c})}{cos(\frac{\Omega \ell}{c})} U(\Omega) e^{j\Omega t}$$

where $u(x,t)$ is the velocity, $U(\Omega)$ is the magnitude of the excitation as a function of radian frequency, ℓ is the length of the tube, x is the position, c is the velocity of sound in air, and t is the time.

Also, poles are located at

$$f = c(\frac{2n+1}{4\ell})|_{n=0,1,2,3,\ldots}$$

.

If a human vocal tract length is about $\ell = 17cm$, then this equation predicts that the resonance of the human vocal tract (formant) $f = 500, 1500, 2500, 3500Hz$.

7.3.2 Viewgraphs

(1) Standing wave pattern in a string:

Two traveling waves move pass each other to produce standing wave in a string. A standing wave is characterized by its stationary nodes in a string. The viewgraph shows standing wave in one-eighth of a cycle apart. The wave's amplitude changes but its nodal points stayed in the same place.

(2) Longitudinal standing wave in an air column:

The viewgraph shows (a) at an instant of maximum displacement of the air molecules, and (b) one-half cycle later. It is different from viewgraph (1) by its longitudinal movement.

(3) Mermelstein's picture of human voicing apparatus:

Vocal tract acts like an acoustic tube. The viewgraph shows location of glottis, velum, tongue center, tongue tip, jaw, and lip.

(4) "Dog":

The view graph shows overlapping traces of tongue position in pronouncing the word "DOG".

(5) Vocal tract positions for some English vowels:

The viewgraph shows that nasal tract is closed by vellum for those English vowels. Here, the main articulator is the tongue. The vowels in *heed*, *hid*, *head*, and *had* are called *front vowels*, because the highest point of the tongue is in the front of the mouth. The vowels in *hod*, *hawed*, *hood* and *who'd* are called *back vowels*, because the highest point of the tongue is in the back of the mouth. The tongue is highest in the vowels of *heed* and *who'd*, which are called *high* or *close* vowels, and lowest in the vowels of *had* and *hod*, which are called *low* or *open* vowels. As to timbre, the vowel of *heed* seems shrill, and the timbre of the vowel in *who'd* seems low or dull.

(A question was raised about the curvature of the vocal tract and how that affects the modeling in the acoustic tubes.)

(Professor Morgan showed us a tape after Professor Gold's 5th viewgraph.)

7.3.3 A Tape of X-rays data from the University of Iowa

It shows several people's position of velum, vocal tract, tongue, and other main articulators while they are speaking words or sentences. (Due to possible side effects of X-rays, this type of tape recording has not been done since.)

7.3.4 Viewgraphs (continued)

(6) Calculations based on X-ray data of fixed sound:

The viewgraph shows Fant's method for arriving at a stepwise approximation of the vocal tract area function from the X-ray data. Cross-section figures are made for a number of planes perpendicular to the central line from the glottis to the lips.

(7) Fant's X-ray tracings of [k], [p], [t] and their corresponding area functions:

The viewgraph shows [p] is produced by closing the lip, [k] is produced by closing the back of tongue, and [t] is produced by tongue touching the back of front teeth.

(8) Fant's area function and Portnoff's frequency response for the Russian vowel /a/:

The viewgraph shows the frequency spectrum of synthetic acoustic tube based on vocal tract area function. Several formants are showned.

(9) Concatenation of lossless tubes of equal length:

Single tube gives resonance frequencies at odd harmonics. Several tubes give formants.

(10) Vowel sounds are characterized by resonances of the vocal tracts, or *formants*, which make the speech sound more intense in narrow frequency ranges:

The curves shown in this figure illustrate the effect of these resonances for the vowel *a* as in *had*. Curve A illustrates a whispered *had*. This is a noiselike sound, and a whole range of frequencies is present. *Had* spoken with a pitch frequency of 200 Hz is illustrated in part B. Only partials of frequencies 200 Hz and its harmonics are present. The dashed *envelope* curve indicates the effects of the resonances of the vocal tract on the intensities of these partials. In part C, the pitch frequency is 400 Hz. There are only half as many partials as in part B. One might guess that it is harder to find the formants from a recording of a female voice than from a recording of a male voice, and this is true. The widely spaced partials of the female voice do not indicate the formant frequencies very clearly.

(11) Kelly-Lochbaum acoustic tube section:

The viewgraph shows the digital-block diagram for acoustic tubes.

7.3.5 Listened to Tape

"To be or not to be ..." This is a late 1950's recording of speech synthesizer by an early-generation computer.

(Next lecture will be given by Dr. John Ohala.")