reading

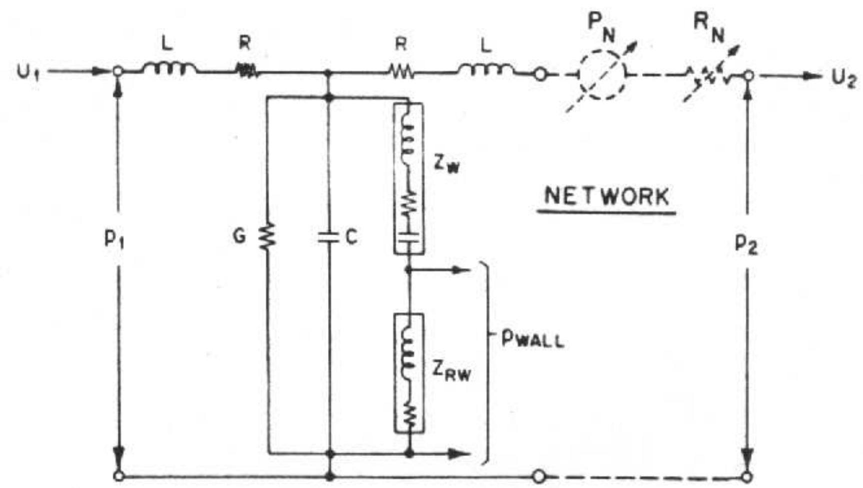
# J. Flanagan & K. Ishizaka, “Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer”

SyRG, 28 Oct 2004

Kornel Laskowski

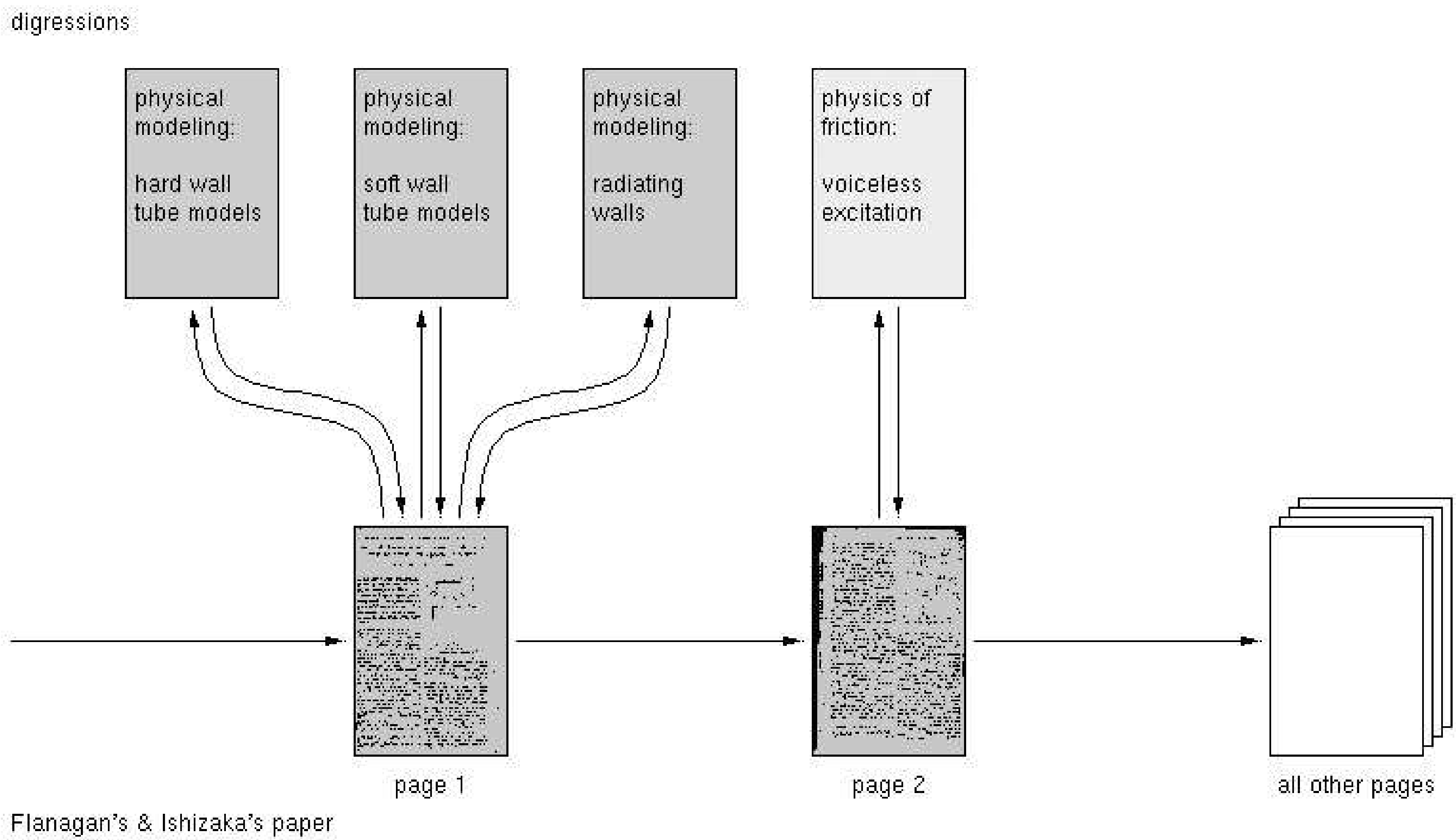
# (My) Motivations

Lots of (circa 1970) synthesis work opens with or culminates in diagrams like this one:




- Where does this come from? Why?
- How can I use it?

# Outline


digressions
physical modeling: hard wall tube models
physical modeling: soft wall tube models
physical modeling: radiating walls
physics of friction: voiceless excitation
page 1
page 2
all other pages
Flanagan's & Ishizaka's paper

# Starting Point: Hard Wall Tube Models

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-24, NO. 2, APRIL 1976 163

## Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer

JAMES L. FLANAGAN, FELLOW, IEEE, AND KENZO ISHIZAKA

*Abstract*—A speech synthesis technique is described which incorporates acoustic models for sound propagation in a tube with yielding walls, turbulent noise generation at locations of constricted volume flow in the vocal tract, and the self-oscillatory properties of the vocal cord source. This formulation frees the experimenter from a traditional limitation, namely, the assumption of linear separability of sound source and resonant system. As a consequence, new opportunities accrue for building realistic physiological characteristics into the synthesizer. These built-in characteristics represent information that need not be overtly supplied to control the synthesizer. The system is used to synthesize test syllables from controls which are stylized models of articulation and connected speech from controls automatically derived from printed text. The synthesis technique demonstrates the feasibility of generating all speech sounds (voiced, unvoiced, nasal) from a common set of physiologically based control parameters, as the human does.

SINCE its inception, the theory of speech synthesis has assumed a linear separability between the source of sound in the vocal tract and the vocal tract system response. That is, speech synthesis traditionally has been viewed as a sou[illegible] [illegible]enerator) supplying a linearly separable (noninteracting) filter system. This formulation places certain limitations upon the physiological realism which can be incorporated into the synthesizer and, hence, in the naturalness which can be produced in the synthetic output. (This artificial constraint of source-system isolation, incidently, is largely the genesis of difficulties in automatic pitch extraction. The problem, traditionally, is ill-posed, because an acceptable specification of voice pitch is rarely formulated.)

In an effort to avoid these limitations, and, at the same time, to build more physiological realism into the synthesizer, we have approached the speech synthesis problem from the viewpoint of sound sources which interact with the resonant system [1]-[4]. We have aimed especially at accounting for phenomena associated with sound propagation and sound generation in the vocal tract. Toward this end, we model sound behavior in an incremental length of the vocal tract as shown in Fig. 1. This incremental length, $\Delta x$, is treated as a right-circular tube of cross sectional area $A$, and with input and output sound pressures and volume velocities $p$ and $U$, respectively. Further, the tube is considered to have a soft, yielding wall whose displacement is $\xi$, and where the vibrating wall radiates the sound pressure $p_{\text{wall}}$.

At frequencies of interest (namely about 4000 Hz and below), we consider one-dimensional wave propagation to suitably describe the significant acoustic effects. We, therefore, model the elemental piece of tube by the network shown in the lower part of Fig. 1. The elements $L$, $R$, $G$, and $C$ are the classical representations for one-dimensional propagation in a hard wall pipe and reflect, respectively, the inertance (mass) of the contained air, the viscous loss at the side wall, the heat conduction loss at the side wall, and the compressibility of the contained volume of air [5]. Additional to the classical representations, however, we have included several other elements. Specifically, we have measured the mechanical impedance of the yielding vocal tract wall [6], and we represent this impedance by the mechanical mass, resistance, and stiffness shown as $Z_w$. The acoustic volume velocity passing this impedance is the source of sound radiation from the vocal tract wall. We represent the radiation impedance of the vibrating wall as that appropriate for a pulsating right-circular cylinder [7], shown as the mass and resistance components in $Z_{RW}$. The sound pressure appearing across $Z_{RW}$ is the tube-element's contribution to the wall-radiated sound, $p_{\text{wall}}$. The total wall-radiated sound is obtained by integrating $p_{\text{wall}}$ over the total vocal tract length.

Further, the network of Fig. 1 includes series elements representing a sound pressure source, $P_N$, and its inherent resistance, $R_N$. This source models the generation of turbulence at constrictions in the vocal tract. It is a series source of random sound pressure with an internal resistance that depends upon

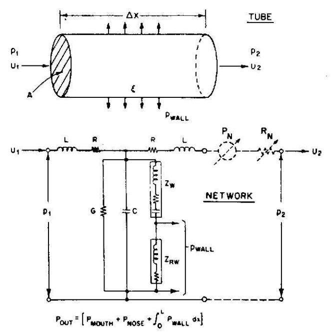



Fig. 1. Representation of one-dimensional acoustic wave propagation in a right-circular tube with yielding side wall.

Manuscript received July 15, 1975; revised October 28, 1975. This work was presented at the 89th meeting of the Acoustical Society of America, Austin, TX, April 8-11, 1975.

The authors are with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974.

- Lumped-parameter modeling of the vocal tract
- Lossless tube model: consider mass and compliance of air only
- Lossy tube model: same as lossless tube model, but also consider viscous and thermal losses

# Soft Wall Tube Models

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-24, NO. 2, APRIL 1976 163

## Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer

JAMES L. FLANAGAN, FELLOW, IEEE, AND KENZO ISHIZAKA

*Abstract*—A speech synthesis technique is described which incorporates acoustic models for sound propagation in a tube with yielding walls, turbulent noise generation at locations of constricted volume flow in the vocal tract, and the self-oscillatory properties of the vocal cord source. This formulation frees the experimenter from a traditional limitation, namely, the assumption of linear separability of sound source and resonant system. As a consequence, new opportunities accrue for building realistic physiological characteristics into the synthesizer. These built-in characteristics represent information that need not be overtly supplied to control the synthesizer. The system is used to synthesize test syllables from controls which are stylized models of articulation and connected speech from controls automatically derived from printed text. The synthesis technique demonstrates the feasibility of generating all speech sounds (voiced, unvoiced, nasal) from a common set of physiologically based control parameters, as the human does.

SINCE its inception, the theory of speech synthesis has assumed a linear separability between the source of sound in the vocal tract and the vocal tract system response. That is, speech synthesis traditionally has been viewed as a sou[...]enerator) supplying a linearly separable (noninteracting) filter system. This formulation places certain limitations upon the physiological realism which can be incorporated into the synthesizer and, hence, in the naturalness which can be produced in the synthetic output. (This artificial constraint of source-system isolation, incidently, is largely the genesis of difficulties in automatic pitch extraction. The problem, traditionally, is ill-posed, because an acceptable specification of voice pitch is rarely formulated.)

In an effort to avoid these limitations, and, at the same time, to build more physiological realism into the synthesizer, we have approached the speech synthesis problem from the viewpoint of sound sources which interact with the resonant system [1]-[4]. We have aimed especially at accounting for phenomena associated with sound propagation and sound generation in the vocal tract. Toward this end, we model sound behavior in an incremental length of the vocal tract as shown in Fig. 1. This incremental length, $\Delta x$, is treated as a right-circular tube of cross sectional area $A$, and with input and output sound pressures and volume velocities $p$ and $U$, respectively. Further, the tube is considered to have a soft, yielding wall whose displacement is $\xi$, and where the vibrating wall radiates the sound pressure $p_{\text{wall}}$.

At frequencies of interest (namely about 4000 Hz and below), we consider one-dimensional wave propagation to suitably describe the significant acoustic effects. We, therefore, model the elemental piece of tube by the network shown in the lower part of Fig. 1. The elements $L, R, G$, and $C$ are the classical representations for one-dimensional propagation in a hard wall pipe and reflect, respectively, the inertance (mass) of the contained air, the viscous loss at the side wall, the heat conduction loss at the side wall, and the compressibility of the contained volume of air [5]. Additional to the classical representations, however, we have included several other elements. Specifically, we have measured the mechanical impedance of the yielding vocal tract wall [6], and we represent this impedance by the mechanical mass, resistance, and stiffness shown as $Z_w$. The acoustic volume velocity passing this impedance is the source of sound radiation from the vocal tract wall. We represent the radiation impedance of the vibrating wall as that appropriate for a pulsating right-circular cylinder [7], shown as the mass and resistance components in $Z_{RW}$. The sound pressure appearing across $Z_{RW}$ is the tube-element's contribution to the wall-radiated sound, $p_{\text{wall}}$. The total wall-radiated sound is obtained by integrating $p_{\text{wall}}$ over the total vocal tract length.

Further, the network of Fig. 1 includes series elements representing a sound pressure source, $P_N$, and its inherent resistance, $R_N$. This source models the generation of turbulence at constrictions in the vocal tract. It is a series source of random sound pressure with an internal resistance that depends upon

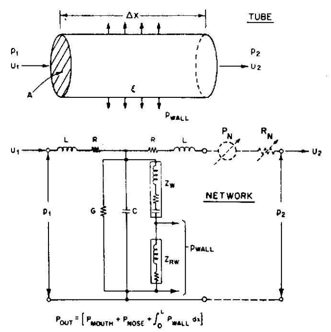


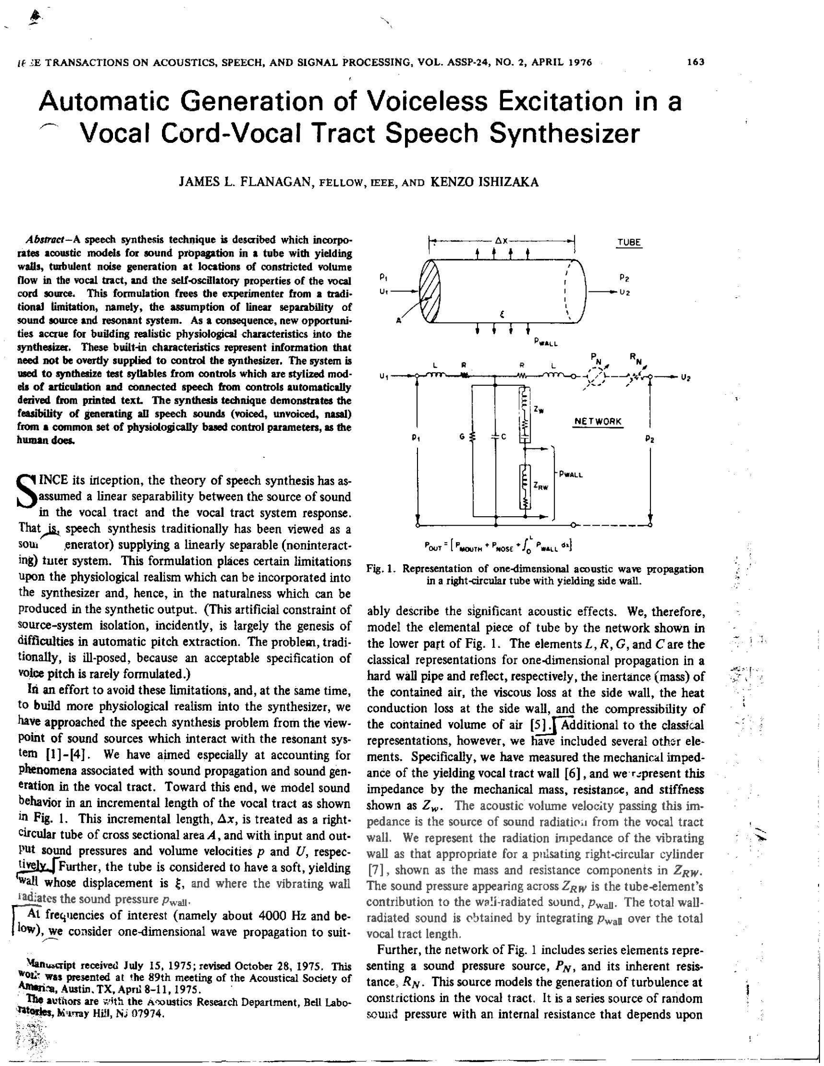

Fig. 1. Representation of one-dimensional acoustic wave propagation in a right-circular tube with yielding side wall.

Manuscript received July 15, 1975; revised October 28, 1975. This work was presented at the 89th meeting of the Acoustical Society of America, Austin, TX, April 8-11, 1975.

The authors are with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974.

- What if tube walls are not hard, but yielding?
- Some of the acoustic energy previously propagated to the next acoustic subsystem section is now transduced to a mechanical subsystem
- It's either
  - Stored in the inertia/compliance of the vocal tract walls, or
  - Dissipated in some form of resistance of the vocal tract walls

# Radiation from Tube Walls



IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-24, NO. 2, APRIL 1976 163

Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer

JAMES L. FLANAGAN, FELLOW, IEEE, AND KENZO ISHIZAKA

*Abstract*—A speech synthesis technique is described which incorporates acoustic models for sound propagation in a tube with yielding walls, turbulent noise generation at locations of constricted volume flow in the vocal tract, and the self-oscillatory properties of the vocal cord source. This formulation frees the experimenter from a traditional limitation, namely, the assumption of linear separability of sound source and resonant system. As a consequence, new opportunities accrue for building realistic physiological characteristics into the synthesizer. These built-in characteristics represent information that need not be overtly supplied to control the synthesizer. The system is used to synthesize test syllables from controls which are stylized models of articulation and connected speech from controls automatically derived from printed text. The synthesis technique demonstrates the feasibility of generating all speech sounds (voiced, unvoiced, nasal) from a common set of physiologically based control parameters, as the human does.

SINCE its inception, the theory of speech synthesis has assumed a linear separability between the source of sound in the vocal tract and the vocal tract system response. That is, speech synthesis traditionally has been viewed as a sou[...] [...]enerator) supplying a linearly separable (noninteracting) filter system. This formulation places certain limitations upon the physiological realism which can be incorporated into the synthesizer and, hence, in the naturalness which can be produced in the synthetic output. (This artificial constraint of source-system isolation, incidently, is largely the genesis of difficulties in automatic pitch extraction. The problem, traditionally, is ill-posed, because an acceptable specification of voice pitch is rarely formulated.)

In an effort to avoid these limitations, and, at the same time, to build more physiological realism into the synthesizer, we have approached the speech synthesis problem from the viewpoint of sound sources which interact with the resonant system [1]-[4]. We have aimed especially at accounting for phenomena associated with sound propagation and sound generation in the vocal tract. Toward this end, we model sound behavior in an incremental length of the vocal tract as shown in Fig. 1. This incremental length, $\Delta x$, is treated as a right-circular tube of cross sectional area $A$, and with input and output sound pressures and volume velocities $p$ and $U$, respectively. Further, the tube is considered to have a soft, yielding wall whose displacement is $\xi$, and where the vibrating wall radiates the sound pressure $p_{wall}$.

At frequencies of interest (namely about 4000 Hz and below), we consider one-dimensional wave propagation to suitably describe the significant acoustic effects. We, therefore, model the elemental piece of tube by the network shown in the lower part of Fig. 1. The elements $L, R, G$, and $C$ are the classical representations for one-dimensional propagation in a hard wall pipe and reflect, respectively, the inertance (mass) of the contained air, the viscous loss at the side wall, the heat conduction loss at the side wall, and the compressibility of the contained volume of air [5]. Additional to the classical representations, however, we have included several other elements. Specifically, we have measured the mechanical impedance of the yielding vocal tract wall [6], and we represent this impedance by the mechanical mass, resistance, and stiffness shown as $Z_w$. The acoustic volume velocity passing this impedance is the source of sound radiation from the vocal tract wall. We represent the radiation impedance of the vibrating wall as that appropriate for a pulsating right-circular cylinder [7], shown as the mass and resistance components in $Z_{RW}$. The sound pressure appearing across $Z_{RW}$ is the tube-element's contribution to the wall-radiated sound, $p_{wall}$. The total wall-radiated sound is obtained by integrating $p_{wall}$ over the total vocal tract length.

Further, the network of Fig. 1 includes series elements representing a sound pressure source, $P_N$, and its inherent resistance, $R_N$. This source models the generation of turbulence at constrictions in the vocal tract. It is a series source of random sound pressure with an internal resistance that depends upon



$$P_{OUT} = [P_{MOUTH} + P_{NOSE} + \int_0^L P_{WALL}\, dx]$$

Fig. 1. Representation of one-dimensional acoustic wave propagation in a right-circular tube with yielding side wall.

Manuscript received July 15, 1975; revised October 28, 1975. This work was presented at the 89th meeting of the Acoustical Society of America, Austin, TX, April 8-11, 1975.

The authors are with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974.

- Vibrating walls are actually causing the air outside to vibrate too (unless in a vacuum)
- Energy is once again transduced from the mechanical subsystem to the outside
- Total sound radiated along the length of the entire vocal tract is the sum of this wall radiation from each subsystem section

# Voiceless Excitation



## Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer

JAMES L. FLANAGAN, FELLOW, IEEE, AND KENZO ISHIZAKA

*Abstract*—A speech synthesis technique is described which incorporates acoustic models for sound propagation in a tube with yielding walls, turbulent noise generation at locations of constricted volume flow in the vocal tract, and the self-oscillatory properties of the vocal cord source. This formulation frees the experimenter from a traditional limitation, namely, the assumption of linear separability of sound source and resonant system. As a consequence, new opportunities accrue for building realistic physiological characteristics into the synthesizer. These built-in characteristics represent information that need not be overtly supplied to control the synthesizer. The system is used to synthesize test syllables from controls which are stylized models of articulation and connected speech from controls automatically derived from printed text. The synthesis technique demonstrates the feasibility of generating all speech sounds (voiced, unvoiced, nasal) from a common set of physiologically based control parameters, as the human does.

SINCE its inception, the theory of speech synthesis has assumed a linear separability between the source of sound in the vocal tract and the vocal tract system response. That is, speech synthesis traditionally has been viewed as a sou[rce (g]enerator) supplying a linearly separable (noninteracting) [fil]ter system. This formulation places certain limitations upon the physiological realism which can be incorporated into the synthesizer and, hence, in the naturalness which can be produced in the synthetic output. (This artificial constraint of source-system isolation, incidently, is largely the genesis of difficulties in automatic pitch extraction. The problem, traditionally, is ill-posed, because an acceptable specification of voice pitch is rarely formulated.)

In an effort to avoid these limitations, and, at the same time, to build more physiological realism into the synthesizer, we have approached the speech synthesis problem from the viewpoint of sound sources which interact with the resonant system [1]-[4]. We have aimed especially at accounting for phenomena associated with sound propagation and sound generation in the vocal tract. Toward this end, we model sound behavior in an incremental length of the vocal tract as shown in Fig. 1. This incremental length, $\Delta x$, is treated as a right-circular tube of cross sectional area $A$, and with input and output sound pressures and volume velocities $p$ and $U$, respectively. Further, the tube is considered to have a soft, yielding wall whose displacement is $\xi$, and where the vibrating wall radiates the sound pressure $p_{\text{wall}}$.

At frequencies of interest (namely about 4000 Hz and below), we consider one-dimensional wave propagation to suitably describe the significant acoustic effects. We, therefore, model the elemental piece of tube by the network shown in the lower part of Fig. 1. The elements $L, R, G$, and $C$ are the classical representations for one-dimensional propagation in a hard wall pipe and reflect, respectively, the inertance (mass) of the contained air, the viscous loss at the side wall, the heat conduction loss at the side wall, and the compressibility of the contained volume of air [5]. Additional to the classical representations, however, we have included several other elements. Specifically, we have measured the mechanical impedance of the yielding vocal tract wall [6], and we represent this impedance by the mechanical mass, resistance, and stiffness shown as $Z_w$. The acoustic volume velocity passing this impedance is the source of sound radiation from the vocal tract wall. We represent the radiation impedance of the vibrating wall as that appropriate for a pulsating right-circular cylinder [7], shown as the mass and resistance components in $Z_{RW}$. The sound pressure appearing across $Z_{RW}$ is the tube-element's contribution to the wall-radiated sound, $p_{\text{wall}}$. The total wall-radiated sound is obtained by integrating $p_{\text{wall}}$ over the total vocal tract length.

Further, the network of Fig. 1 includes series elements representing a sound pressure source, $P_N$, and its inherent resistance, $R_N$. This source models the generation of turbulence at constrictions in the vocal tract. It is a series source of random sound pressure with an internal resistance that depends upon

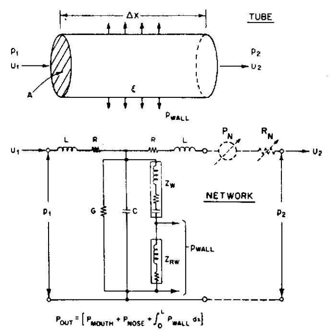



Fig. 1. Representation of one-dimensional acoustic wave propagation in a right-circular tube with yielding side wall.

Manuscript received July 15, 1975; revised October 28, 1975. This work was presented at the 89th meeting of the Acoustical Society of America, Austin, TX, April 8-11, 1975.

The authors are with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974.



the constriction size. Experiment has demonstrated [8] that the intensity of the random pressure generation in a constriction is proportional to the square of the Reynolds number, in excess of some critical or threshold value.[1] Similarly, the inherent impedance of a constriction giving rise to vorticity is primarily resistive and is flow dependent. Therefore, the variance of the random pressure source $P_N$ is proportional to $(U^2/A)$, and the loss $R_N$ is proportional to $(|U|/A^2)$, and these factors can be used for automatic generation of turbulent excitation [9]. The cross sectional area value $A$, which is supplied as an input synthesis parameter, and the resulting flow $U$ determine, uniquely, the values of $P_N$ and $R_N$. These quantities are therefore calculated on a sample-by-sample basis, along with the sound pressures and volume velocities.

These physiological factors associated with vocal tract properties are combined with a model of the vocal cords, described in detail previously [4], to provide a complete model for synthesis. The model is shown in Fig. 2. The top of Fig. 2 shows a mechanical schematic of the vocal system, in which $P_S$ is the subglottal (lung) pressure controlled in part by the muscle force of the rib cage. The vocal cords are modeled as a self-oscillating system in which each vocal cord is represented by two coupled masses, having associated nonlinear mechanical elements previously derived.

The acoustic volume velocity that passes the vocal cord opening is $U_G$, and this flow, when periodically interrupted, is the excitation source for voiced sounds. The vocal tract proper can be coupled to the nasal tract by the opening area at the velum. The acoustic volume velocities radiated from the mouth and from the nostril are $U_M$ and $U_N$, respectively.

The lumped-element network representation for this system is shown in the lower part of Fig. 2. The lung volume is represented by a lossy variable capacity, charged to the pressure $P_S$. A classical $T$-section represents the bronchi–trachea tube leading between the lungs and the vocal cord opening. Each $T$-section of the vocal tract network is precisely the circuit given in Fig. 1, and is specified by its cross sectional area value $A$.[2] The radiation loads at the mouth and nostril are $Z_M$ and $Z_N$, respectively, and both are shown in series with the atmospheric pressure, $P_A$. (The model, therefore, can also simulate respiration!) The sound pressure developed across the radiation loads, when combined with the wall-radiated pressure, gives the total synthetic output.

These details aside, the important aspect is that the input control data to the synthesizer are the physiologically based parameters representing, respectively, the subglottal lung pressure $P_S$, the vocal cord tension $Q$, the vocal cord neutral or rest area $A_{g0}$, the area of nasal coupling $N$ (to the fixed-shape nasal tract), and the cross sectional area of the vocal tract along its length $A(x)$. The model in this form is represented for computer simulation by a set of difference equations which involve all sound pressures and volume velocities as variables. By simultaneous solution of these equations, as described previously [3], [4], the programmed model calculates Nyguist samples of all pressures and velocities, including the output synthetic sound pressure.

We have used the synthesizer in earlier experiments to generate vowel-consonant-vowel syllables /vcv/ [13]. In this synthesis we stylize the physiological control functions which are given as input to the synthesizer. An example is shown in Fig. 3. At the top of the figure are shown the time variations of subglottal pressure $P_S$ and vocal cord neutral area $A_{g0}$. The subglottal pressure is caused to increase in articulating the second vowel, so that it becomes stressed. The vocal cord neutral area is moved to a large open value during the intervocalic voiceless stop consonant /p/. The area shape of the vocal tract is made to vary linearly from a configuration for the initial vowel /ə/ to that for the final vowel /a/. The only area function displayed in Fig. 3 is the area of the mouth opening $A_M$, which reflects the labial closure. The synthesizer calculates the remaining functions shown in the figure. The spectrogram of the synthetic output sound reflects the intervocalic voiceless stop, the aspiration following the labial release, and the initiation of voicing into the final stressed vowel. (The spectrogram is produced with an expanded time scale to correspond to the computer plots.) The calculated vocal cord opening $A_G$ and glottal volume velocity $U_G$ show the cessation of voicing during the voiceless stop, as does the output sound pressure $p$. The bottom trace is the Reynolds number at the mouth constriction computed, as discussed, from the mouth volume velocity and area. The relative timing

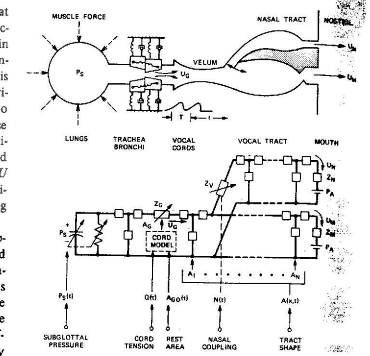
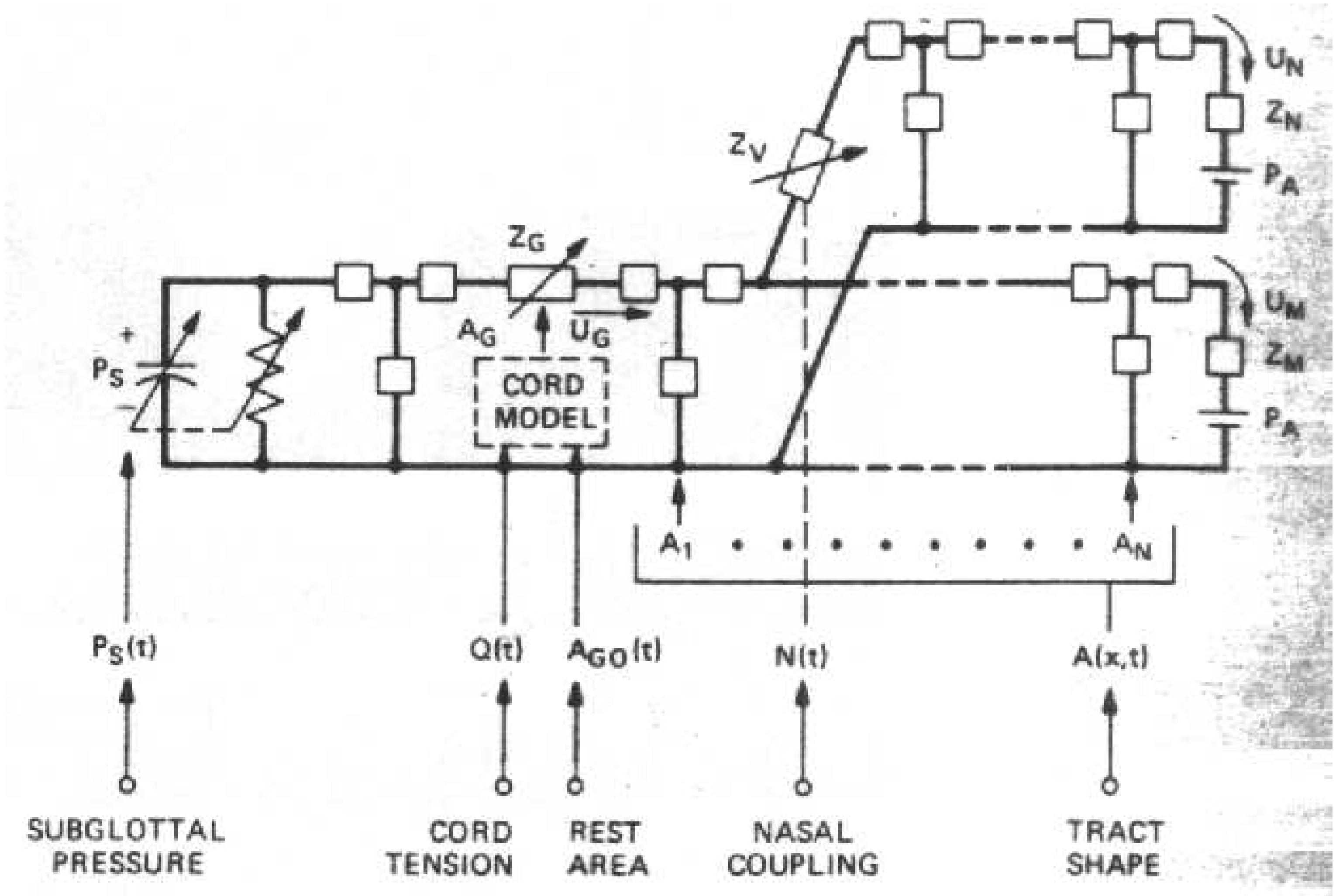



Fig. 2. Network representation of the vocal cord–vocal tract synthesizer.

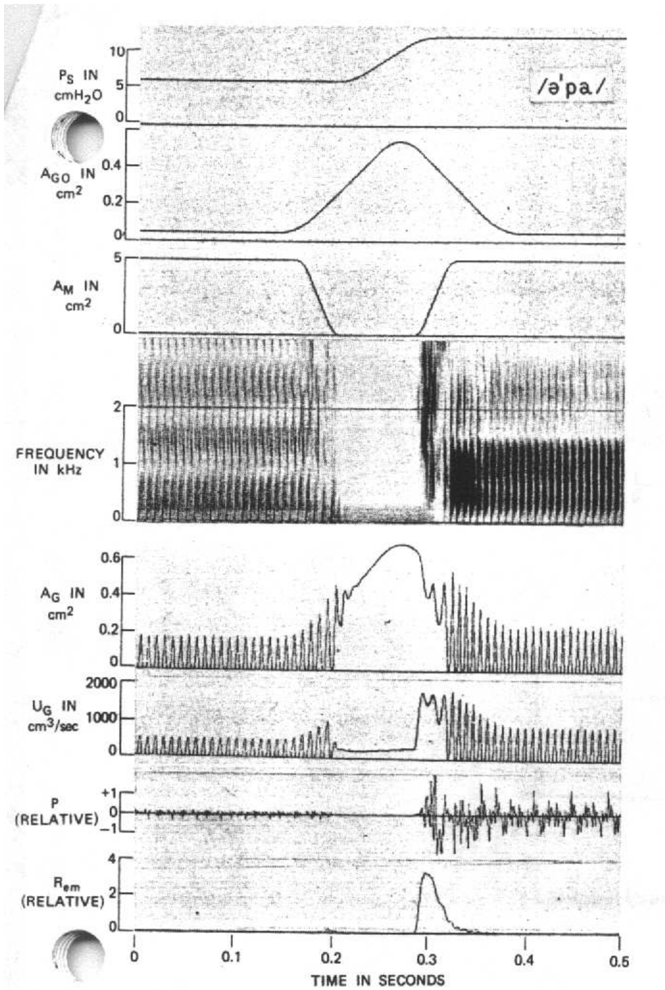
[1] Reynolds number is a dimensionless quantity equal to $(\rho w u/\mu)$ where $\rho$ is air density, $u$ is particle velocity (equal to $U/A$), $\mu$ is the coefficient of viscosity, and $w$ is the width of the constricting passage.

[2] In the present work 20 vocal tract sections are used corresponding to $\Delta X \cong 0.85$ cm.

# Putting It All Together

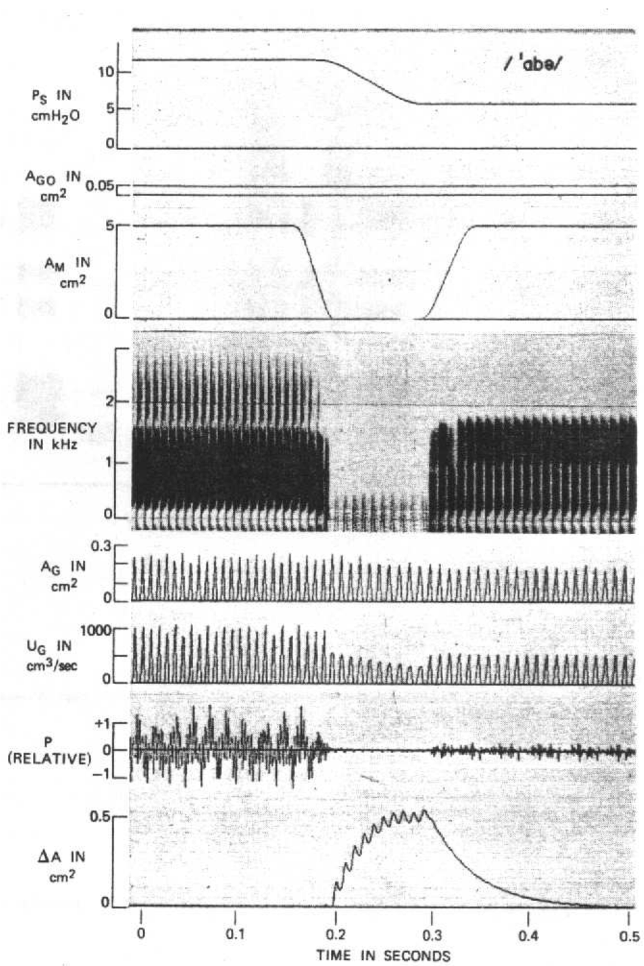


# Example: Aspiration




- Demonstrates ability of model to produce, based on “articulatory controls”, naturally occurring aspiration following unvoiced stops

# Example: Wall Vibration




- Demonstrates “voice bar”
- Even though mouth is closed, low frequencies associated with voiced stops are present

# Coupling the Model with a Coker and Umeda machine

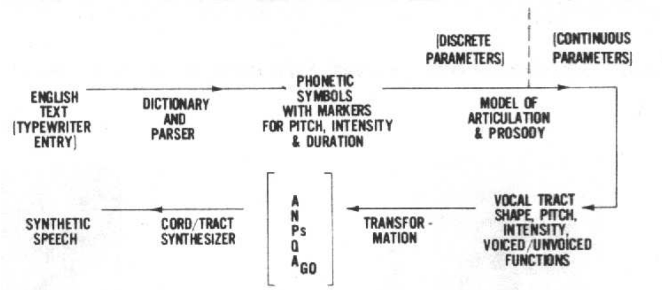



Fig. 6. System for automatic conversion of typed English text into control variables for cord-tract synthesizer.

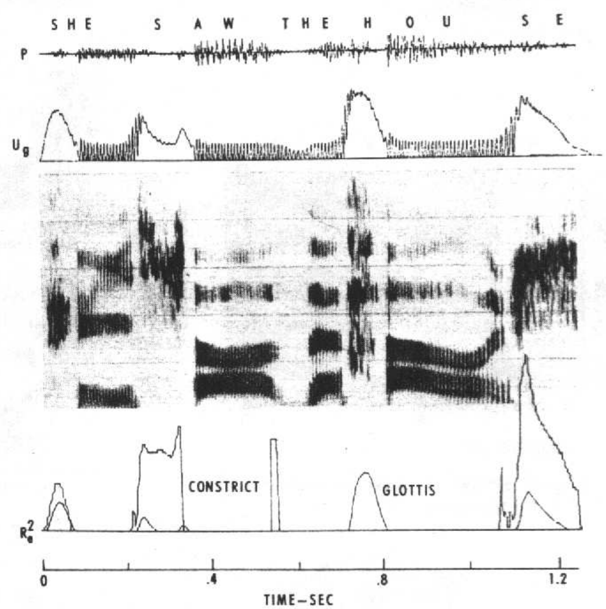



Fig. 7. Automatic synthesis from typed text using the cord-tract synthesizer.

- Appears that could produce speech provided articulatory controls available
- Said machine performs an orthographic to articulatory mapping for speech synthesis
- Have never heard of this machine. Alan?

# Frication in the Glottis

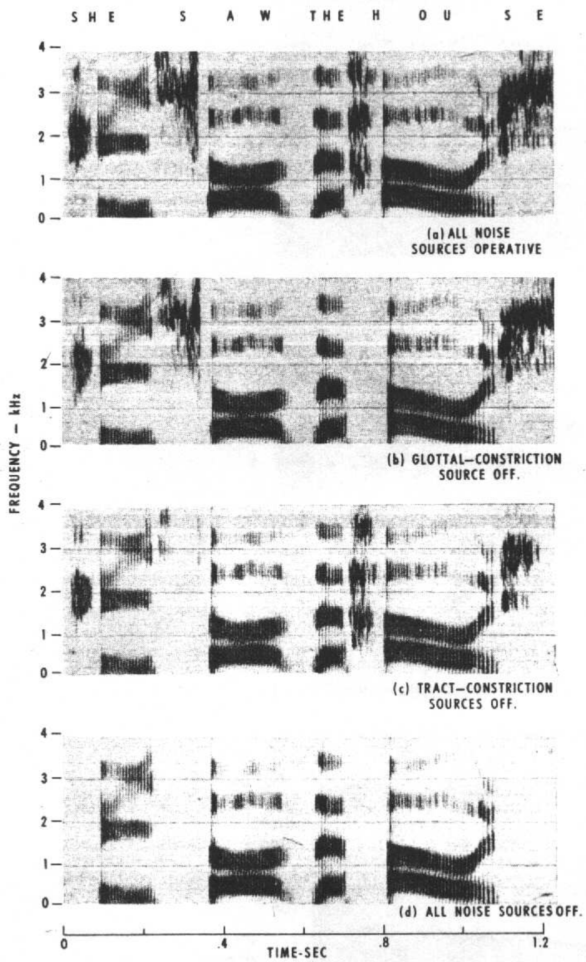



Fig. 8. Four syntheses demonstrating the relative contribution to fricative excitation from the vocal cord constriction and the vocal tract constrictions.

- /h/ is produced by duplicating fricative noise source also in the glottal model
- Experiment: turn off the ability to generate turbulent excitation somewhere in the tract model, then look at the spectrogram

# The End

Thanks!