An Analog Integrated-Circuit Vocal Tract

Keng Hoong Wee, Lorenzo Turicchia, and Rahul Sarpeshkar, Senior Member, IEEE

Abstract—We present the first experimental integrated-circuit vocal tract by mapping fluid volume velocity to current, fluid pressure to voltage, and linear and nonlinear mechanical impedances to linear and nonlinear electrical impedances. The 275 μW analog vocal tract chip includes a 16-stage cascade of two-port π-elements that forms a tunable transmission line, electronically variable impedances, and a current source as the glottal source. A nonlinear resistor models laminar and turbulent flow in the vocal tract. The measured SNR at the output of the analog vocal tract is 64, 66, and 63 dB for the first three formant resonances of a vocal tract with uniform cross-sectional area. The analog vocal tract can be used with auditory processors in a feedback speech locked loop—analagous to a phase locked loop—to implement speech recognition that is potentially robust in noise. Our use of a physiological model of the human vocal tract enables the analog vocal tract chip to synthesize speech signals of interest, using articulatory parameters that are intrinsically compact and linearly interpolatable.

Index Terms—Analysis-by-synthesis, electronically tunable resistor, MOS resistor, nonlinear resistor, speech codec, speech pros thesis, speech recognition, speech synthesis, vocal tract.

I. INTRODUCTION

InCREASINGLY, circuit models of biology are being used to improve performance in engineering systems. For example, silicon-cochlea-like models have led to improved speech recognition in noise and low-power cochlear-implant processors for the deaf [1]–[3]. The problem of representing speech events with robust and compact signals that describe the salient features of speech is an important area of speech communication. For example, in speech codecs and synthetic speech systems, an efficient representation of speech and naturalness of generated speech are important requirements. A promising approach to improve the naturalness of synthetic speech is to exploit bio-inspired models of speech production with physiological control parameters that are intrinsically robust, linearly interpolatable, and which enable low bit-rate transmission [4], [5]. An important technique employed in speech applications involves analysis-by-synthesis [4]: The speech is analyzed by extracting parameters from it that are used to configure a speech synthesizer to reproduce the speech.

Fig. 1 shows an analysis-by-synthesis block diagram that creates what we term a speech locked loop (SLL) in analogy with phase locked loops (PLL) used in other communication systems. The auditory processor and controller are analogous to a phase detector and loop filter in a PLL and the vocal tract is analogous to a voltage-controlled-oscillator (VCO). The speech produced by the vocal tract is analyzed and compared to that of the input, and a measure of error is computed. Different sounds are generated until one is found that produces the least error (using gradient descent techniques with pre-selected initial conditions that ensure global minimum convergence) at which time the speech locked loop locks to the input sound with an optimal vocal tract profile produced by the controller. Previous attempts that take advantage of the powerful analysis-by-synthesis method employed computationally expensive approaches to articulatory synthesis using digital computation [5]. In the past, Stevens et al. [6] built a static electrical analog of the vocal tract using discrete elements. A dynamically controllable electrical analog of the vocal tract is described in [7] using vacuum tube technology. Our strategy uses an analog vocal tract (AVT) to drastically reduce power consumption, enables real-time performance, and could be useful in portable speech processing systems of moderate complexity, e.g., cell phones, digital assistants, and bionic speech-prosthesis systems. The AVT may be used in conjunction with an analog bionic ear processor [2] to implement an ultra-low-power speech locked loop.

The organization of this paper is as follows. In Section II, we describe our circuit model of the vocal tract. In Section III, we describe an electronically tunable two-port building block of the transmission line vocal tract. In Section IV, we propose a tunable linear and nonlinear resistance that can serve to model the constrictions created by the opening and closing of the vocal folds in the glottis and thus model turbulent and laminar flow in the vocal tract. In Section V, we describe how we drive the vocal tract using articulatory controls. In Section VI, we present experimental results of speech produced by our analog vocal tract chip. In Section VII, we summarize the contributions of the paper.

II. CIRCUIT MODEL OF THE VOCAL TRACT

The vocal tract can be approximated as a non-uniform acoustic tube, with time-varying cross-sectional areas, that is terminated by the vocal cords at one end, and the lips and/or...
nose at the other. If the cross-sectional dimensions of the tube are small compared to the wavelength of sound, the waves that propagate along the tube are approximately planar. The acoustic properties of such a tube are indistinguishable from that of a tube with a circular cross section. The wave equation for planar sound propagation (1-D) in a uniform tube of circular cross section can be derived as:

$$\frac{\partial P}{\partial t} - \frac{\rho}{A} \frac{\partial U}{\partial t} + R_e U = \frac{A}{\rho c^2} \frac{\partial P}{\partial t} + G_h P$$

(1)

where $P$ is the sound pressure, $U$ is the volume velocity, $\rho$ is the density of the medium, $c$ is the velocity of sound in the medium, and $A$ is the area of cross section. The volume of air in a tube exhibits an acoustic inductance $\rho/A$ due to its mass (which opposes acceleration) and an acoustic compliance $A/\rho c^2$ due to its compressibility (which opposes changes in volume). In general, the propagation of sound in a tube is accompanied by energy losses due to viscous friction and heat conduction through the tube wall. These effects are represented by an acoustic resistance $R_e$ and an acoustic conductance $G_h$, respectively. In a lossless uniform tube, $R_e = 0$ and $G_h = 0$.

Acoustic wave propagation in a tube is analogous to plane-wave propagation along an electrical transmission line where voltage and current are analogous to sound pressure and volume velocity. The voltage $V$ and current $I$ for a lossy transmission line can be described by the following coupled partial differential equations:

$$-\frac{\partial V}{\partial x} = L \frac{\partial I}{\partial t} + RI$$

$$-\frac{\partial I}{\partial x} = C \frac{\partial V}{\partial t} + GV$$

(2)

where $L$ and $C$ are the inductance and capacitance per unit length. The loss elements $R$ and $G$ represent the series resistance and shunt conductance per unit length of the transmission line. The extensive modeling of mechanical systems as circuits was pioneered by Gabriel Kron, e.g., [8]. A unidirectional transmission line was first used in an integrated circuit to create a silicon cochlea [9]. An integrated-circuit analog of a bidirectional transmission line was first employed in [10] to create a two-dimensional silicon cochlea capable of propagating waves from input to output and back. Voltage-mode resistive networks have been used to model the retina [11]. Current-mode resistive networks have also been used to model the retina [12], [13]. To our best knowledge, ours is the first experimental integrated-circuit implementation of a vocal tract.

Fig. 2 shows our circuit model of the vocal tract. The AVT represents the human vocal tract as acoustic tubes (intra-oral and oral tract) using a transmission line (TL) model. The TL comprises a cascade of tunable two-port elements, corresponding to a concatenation of short cylindrical acoustic tubes (each of length $\ell$) with varying cross sections. The error introduced by spatial quantization is kept small by making $\ell$ short compared to the wavelength of sound corresponding to the maximum frequency of interest. Each two-port is an electrical equivalent of a $LC\, \pi$-circuit element where the series inductance $L$ and the shunt capacitance $C$ may be controlled by physiological parameters corresponding to articulatory movement (i.e., movement of the tongue, jaw, lips, etc). Speech is produced by controlled variations of the cross-sectional areas along the tube in conjunction with the application of one or two sources of excitation: 1) a periodic source at the glottis and/or 2) a turbulent noise source $P_{\text{turb}}$, at some point along the tube.

Vocal fold vibration produces a periodic interruption of the air flow from the lungs to supraglottal vocal tract. At most frequencies of interest, the glottal source has a high acoustic impedance compared to the driving point impedance of the vocal tract. Consequently, a current source may be used as the electrical analog that approximates the volume velocity source at the glottis. Alternatively, the constriction at the glottis may be represented...
by a variable impedance modulated by a glottal oscillator to simulate the opening and closing of the vocal folds. Both the current source and variable impedance model are illustrated in Fig. 2. The current source model is represented by an AC current source (corresponding to a volume velocity source) implemented using a wide linear range operational transconductance amplifier (WLR OTA) [14] with a cascaded output stage to obtain high output impedance. The variable impedance model is represented by a voltage source \( P_{\text{air}} \) (corresponding to the alveolar pressure source) with variable source impedance \( Z_{\text{GC}} \), that is modulated by a glottal oscillator. The variable impedance comprises a linear \((I \propto V)\) and nonlinear resistance \((I \propto \sqrt{V})\) connected in series to represent losses occurring at the glottis due to laminar and turbulent flow, respectively. Such electrical circuit models are consistent with the equations in [15].

The turbulent source \( P_{\text{turb}} \) has a source impedance comprising the constriction impedance \( Z_{\text{SCC}} \). The location of \( P_{\text{turb}} \) is downstream of the constriction location. We approximate the noise generated by \( P_{\text{turb}} \) with a signal produced by modulating the cross-sectional area of a two-port section downstream of the constriction in a noisy fashion. At the lips, the transmission line is terminated by a radiation impedance \( Z_{\text{rad}} \) and the radiated sound pressure at the mouth, \( P_{\text{rad}} \), is proportional to the derivative of the current flowing in \( Z_{\text{rad}} \).

III. TWO-PORT \( \pi \)-SECTION

Fig. 3 shows a circuit diagram of an electronically tunable two-port \( \pi \)-section that forms the basic building block of the transmission line. Each \( LC \) \( \pi \)-section has a resistance \( R \) in series with the inductance \( L \) to model viscous losses and a shunt conductance \( G \) in parallel with the capacitance \( C \) to account for losses at the walls. A chain of basic two-port \( \pi \)-sections, concatenated end to end, yields a complete transmission line. Note that each section contains a shunt capacitance \( C \) at its left port and a shunt conductance \( G \) at its right port such that each port after concatenation with its neighboring sections has both a \( G \) and a \( C \). Fig. 3(a) is mapped to an equivalent tunable two-port \( \pi \)-section in Fig. 3(b). The latter two-port \( \pi \)-section only uses \( G_m \) and \( C \) elements and does not require resistors or inductors and is consequently suitable for a compact subthreshold low-power implementation. The transconductors \( G_1 \) and \( G_{2A} \) gyrate the parallel combination of \( R_0 \) and \( C_1 \) into a unidirectional series \( L \) and \( R \) that corresponds to the current observed at node \( P_1 \) when \( P_2 \) is grounded. The transconductors \( G_2 \) and \( G_{2B} \) gyrate the parallel combination of \( R_0 \) and \( C_1 \) into a unidirectional series \( L \) and \( R \) that corresponds to the current observed at node \( P_2 \) when \( P_1 \) is grounded. Thus, \( R, R_0, C_1, G_1, G_{2A}, \) and \( G_{2B} \) create a bidirectional series \( L \) and \( R \). The shunt conductance \( G \) in Fig. 3(a) is mapped to \( 1/R_{\text{ds}} \) in Fig. 3(b) where \( R_{\text{ds}} \) is the source-to-drain resistance of triode-operated transistor \( M \) with a gate voltage \( V_{\text{GG}} \). The transconductors \( G_{2A} \) and \( G_3 \) gyrate a grounded inductance \( Z_C = j\omega L_C \) into the shunt capacitance \( C \) in Fig. 3(a). The grounded inductance \( Z_C \) is itself created by gyrating another capacitance that is not shown in Fig. 3(b). The reason for creating a capacitance through gyration rather than via a direct passive implementation is to ensure that the product \( LC \) is independent of \( G_{2A} \) since \( L \) is proportional to \( 1/G_{2A} \) while \( C \) is proportional to \( G_{2A} \). \( G_{2A} \) may be tuned in each two-port \( \pi \)-section to alter the ratio of \( L/C \) in accordance with the cross-sectional area of the vocal tract at the given two-port \( \pi \)-section location. Fig. 3(b) lists the algebraic relationships that define the mapping of the circuit of Fig. 3(a) to the equivalent circuit of Fig. 3(b).

The values of \( G_{2A} \) and \( G_{2B} \) are designed to be equal since they are intended to represent a symmetrical \( L \) and \( R \) in series. However, unlike a real series \( L \) and \( R \), the d.c. current from \( P_1 \) to \( P_2 \) is not zero when \( P_1 \) and \( P_2 \) are equal because of circuit offsets. Such mismatch leads to a large d.c. offset in each stage. To mitigate such mismatch, copies of the output current from \( G_{2A} \) and \( G_{2B} \) are used to create a difference current that is converted to a voltage on the integrating capacitor \( C_C \). This voltage is fed back to an input of \( G_{2B} \), thus creating a negative-feedback loop that serves to equalize the output currents from \( G_{2A} \) and \( G_{2B} \) to a precision set by the degree to which the copies faithfully represent the real currents.

To increase the length of a section, we decrease \( G_1 \) and increase \( G_3 \) proportionately such that \( L \) and \( C \) both increase proportionately in the algebraic expressions of Fig. 3(b). For example, in vowels such as /u/ and /o/, where lip rounding occurs, lip protrusion increases the length of the vocal tract and we may increase the length of the two-port \( \pi \)-sections near the end of the AVT to model this effect. The overall length of the vocal tract can be varied by global adjustment of \( G_1 \) and \( G_3 \), as for example, when modeling female versus male speakers.
The radiation impedance at the mouth $Z_{\text{rad}}$ is well modeled by an inductance $L_{\text{rad}}$ implemented by gyrating a capacitor. Thus, the output of the vocal tract, $P_{\text{rad}} \propto (dU_{\text{rad}}/dt)$, is proportional to the voltage across $Z_{\text{rad}} = sL_{\text{rad}}$.

Next, we show experimental results obtained from a 16-stage cascade of two-port $\pi$-sections fabricated in a 1.5 $\mu$m AMI CMOS process. The input to the cascade is a current source denoted by $U_{\text{gl}}$ in Fig. 2 and implemented on chip using a WLR OTA [14]. The output current of the OTA is produced by setting the inverting input terminal $V^-$ to a fixed reference voltage $V_{\text{REF}}$ and applying a sinusoidal a.c. voltage centered about $V_{\text{REF}}$ to the non-inverting input terminal $V^+$. The amplitude of the sinusoidal voltage is 0.5 V. The biasing current of the OTA $I_{\text{GM}}$ is kept constant. The sinusoidal current produced by the WLR OTA serves as the input to the 16-stage cascade of two-port $\pi$-sections under test. The AVT output signal is the voltage measured across $Z_{\text{rad}} = sL_{\text{rad}}$ via an on-chip buffer circuit.

Fig. 4 shows the measured signal and noise characteristics at the output of a 16-stage cascade of two-port $\pi$-sections. In this measurement, the cascade of two-port $\pi$-sections are configured electronically to form an equivalent uniform acoustic tube corresponding to the voiced phoneme /e/. The signal frequency response is obtained by sweeping the frequency of the sinusoidal a.c. voltage source from 100 to 8 kHz. The noise characteristic is presented when the a.c source is set to zero. The measured SNR is 64, 66, and 63 dB for the first three formant resonances (F1, F2, and F3) /e/.

Fig. 5(a) shows the measured output spectrum when the two-port $\pi$-sections are configured to form an acoustic tube with vocal tract area profile shown in (b). The vocal tract area profile is obtained by raising the glottal pulse train of 10 ms. The harmonics of the periodic glottal pulse train are clearly illustrated in Fig. 5(a) by vertical lines. The spectral envelope is the product of the vocal tract transfer function (characterized by the formant resonances) and the source transfer function (determined by the glottal spectrum). The formant frequencies of the synthesized sound are consistent with the articulatory profile and correspond to the formants in a typical vowel.

Fig. 6(a) shows an equivalent $\pi$-circuit model of a cylindrical section of acoustic tube with non-rigid walls. It has a shunt impedance element $Z_{\text{W}}$ comprising a series combination of $R_{\text{W}}$, $L_{\text{W}}$, and $C_{\text{W}}$ that is connected in parallel with capacitance $C$ and conductance $G$ to model the mass and compressibility of the non-rigid walls [15]. Fig. 6(b) is a circuit diagram of the tunable two-port $\pi$-section of Fig. 3 modified to incorporate the effect of $Z_{\text{W}}$. To this end, an additional OTA $G_{\text{A}}$ is employed to gyrate an impedance $Z_{\text{RLC}}$ comprising a parallel combination of $R_{\text{W}}$, $L_{\text{W}}$, and $C_{\text{W}}$. The impedance $Z_{\text{RLC}}$ could be implemented as a bandpass $G_m - C$ impedance.
IV. LINEAR AND NONLINEAR RESISTORS USING MOS TRANSISTORS

In this section, we discuss how to implement the glottal constriction resistance $Z_{GC}$ in Fig. 2 as a series combination of linear and nonlinear resistors that model laminar and turbulent flow respectively. Electronically tunable bidirectional resistors can be implemented with MOS transistors whose source and drain terminals are symmetric and whose gate or bulk voltages may be varied to provide electronic control of the resistance. Fig. 7 explains our idea for using an MOS transistor as a resistor with an arbitrary $I$–$V$ characteristic. The $I_D$–$V_{DS}$ curves of a typical nMOS transistor for various gate voltages are shown in Fig. 7(a). To obtain any desired $I$–$V$ characteristic, the gate potential of the MOS device must be biased to the appropriate value given by the intersection of the MOS device curves and the desired $I$–$V$ curve. As an example, Fig. 7(a) illustrates the case for a linear $I$–$V$ characteristic as the desired $I$–$V$ curve. The concept of the proposed biasing scheme is illustrated in Fig. 7(b).

The current $I_D$ through an MOS device may be modeled using the following well-known bulk-referenced expressions:

Weak inversion:

$$I_D = I_O \exp \left( \frac{\kappa_0 (V_G - V_{TO})}{\phi_t} \right) \times \left( \exp \left( \frac{-V_X}{\phi_t} \right) - \exp \left( \frac{-V_Y}{\phi_t} \right) \right)$$

Strong inversion:

$$I_D = \frac{\kappa_0 \mu C_{ox}}{2} \frac{W}{L} \left[ (V_G - V_{TO} - \frac{V_X}{\kappa_0})^2 - (V_G - V_{TO} - \frac{V_Y}{\kappa_0})^2 \right]$$

where $I_O$ and $\phi_t$ are the size-dependent pre-factor and the thermal voltage ($kT/q$), respectively, and $V_{TO}$ and $\kappa_0$ are the threshold voltage and the subthreshold exponential parameter when $V_{DS} = 0$, respectively. Specifically, $\kappa_0$ is given by

$$\kappa_0 = \frac{1}{1 + \frac{\gamma}{2\sqrt{\phi_t}}}$$

where $\gamma$ is the body effect factor and $\phi_t$ corresponds to the surface potential at $V_{GB} = V_{TO}$. Equation (3) is in a form that reflects the symmetry of the source and drain terminals and may be viewed as the sum of a forward current and a reverse current as follows [17]:

$$I_D = I_{X,\text{sat}} - I_{Y,\text{sat}}$$

where $I_{X,\text{sat}}$ and $I_{Y,\text{sat}}$ are forward and reverse saturation currents determined by $V_{GX}$ and $V_{GY}$, the gate-to-source and gate-to-drain potentials respectively. For the MOS device to have an arbitrary $I$–$V$ characteristic given by:

$$I_D = g(V_{XY})$$

an appropriate $V_G$ must be applied to the gate terminal such that:

$$I_D = I_{X,\text{sat}} - I_{Y,\text{sat}} = g(V_{XY})$$

In (6), $g()$ denotes an arbitrary function and the argument $V_{XY}$ denotes the potential difference across the source-drain terminals ($V_X - V_Y$). We propose a biasing scheme that senses $V_{XY}$ across the device terminals and automatically generates the required gate bias $V_G$ by employing a negative-feedback loop that enforces the equality of (7).

Fig. 8 shows a general circuit implementation of the proposed MOS resistor. In this and subsequent circuit diagrams, the bulk connections of nMOS and pMOS devices are connected to $V_{SS}$ (ground) and $V_{DD}$ respectively, except where indicated. In Fig. 8, the MOS resistor is implemented using a MOS transistor denoted as $M_R$. The source-to-drain potential difference $V_X - V_Y$ across the main MOS device $M_R$ is sensed and converted into a current $I_{OUT,GM}$ by a WLR OTA [14] denoted by $G_M$. The output current $I_{OUT,GM}$ is linearly related to the sensed input voltages as follows:

$$I_{OUT,GM} = G_M (V_X - V_Y) = G_M V_{XY}.$$
and $I_{V_{\text{sat}}}$ of $M_R$ are proportionally replicated by sensing $V_{G_X}$, $V_{W}$, $V_X$, and $V_Y$ on the gate, well, source, and drain terminals of $M_R$ with source follower circuits (denoted as SF in Fig. 8) and applying $V_{G_X}$ and $V_{G_Y}$ across the gate-source terminals of transistors $M_X$ and $M_Y$. The source follower circuits serve as buffers to prevent loading on $M_R$. The replica current reproduced by $M_X$ and $M_Y$ is full-wave rectified and compared with the output of the translinear circuit. Any difference between these two currents will cause the capacitor $C$ to charge or discharge such that the gate bias voltage $V_G$ equilibrates at a point where the two are nearly equal via negative-feedback action.

Fig. 10 shows a die micrograph of a testchip fabricated in AMI 1.5 µm CMOS technology. The testchip contains a linear and nonlinear MOS resistor. Fig. 11 shows the circuit diagram of a Wilson-mirror version of the WLR OTAs first described in [14] and used to implement the $G_M$ transconductor of Fig. 8. The WLR OTA is biased by a current $I_{G_{\text{M}}}$ through the transistor gated by $V_B$. The wide input linear range is achieved by: (a) using the wells of the input pair $M_1$, $M_2$ as inputs, (b) source degeneration through $M_3$ and $M_4$, (c) gate degeneration through $M_5$ and $M_6$ and (d) $p$-type linearization through $B_1$ and $B_2$. The measured value of $V_Z$ is 1.7 V. The current sources $I_{G\text{C}}$ and $I_{G\text{C}}$ serve to compensate for current offsets that may arise due to device mismatch.

Fig. 12 shows the measured $I$–$V$ characteristic of our linear MOS resistor electronically configured to have a resistance of $100 \, \Omega$. The tiny currents flowing through the MOS resistor are accurately sensed and measured using an on-chip current integration technique [19]. The potential difference $V_{XY}$ across its source–drain terminals is varied in 10 mV increments. Fig. 13 shows the measured $I$–$V$ characteristics for various values of WLR OTA biasing current $I_{G_{\text{M}}}$ which generate resistances varying from 24 MΩ to 120 MΩ. The slope of the $I$–$V$ characteristic, i.e., conductance, is determined by $I_{G_{\text{M}}}$. Fig. 14 shows a plot of conductance $G$ with $I_{G_{\text{M}}}$ revealing that $G$ may be varied from $10^{-11} \, \Omega^{-1}$ to $4 \times 10^{-7} \, \Omega^{-1}$, which corresponds to the resistance varying from 100 GΩ to 2.5 MΩ. The conductance $G$ varies linearly with $I_{G_{\text{M}}}$ when the WLR OTA operates in subthreshold because $G$ is determined by the transconductance $G_M$ of the WLR OTA, which is proportional to $I_{G_{\text{M}}}$ in the subthreshold regime. Further details on the resistor are described in [20].

Fig. 15(a) shows the measured $I$–$V$ data for the compressive resistor having an $I$–$V$ relation given by

$$I_D = \pm \sqrt{I_{\text{red}} G_{M} |V_{XY}|} \quad V_{XY} \geq 0$$

$$= - \sqrt{I_{\text{red}} G_{M} |V_{XY}|} \quad V_{XY} < 0$$

The nonlinear resistor is implemented using the general circuit architecture depicted in Fig. 8. The compressive resistor having

![Image](image_url)

Fig. 7. (a) Idea behind MOS resistor and (b) its biasing concept.

The proportionality constant $G_M$, the transconductance of the WLR OTA, is given by

$$G_M = \frac{I_{G_{\text{M}}}}{V_L}$$

where $I_{G_{\text{M}}}$ and $V_L$ are the biasing current and input linear range of the WLR OTA respectively. Hence, $G_M$ is electronically tunable via $I_{G_{\text{M}}}$. The translinear circuit [18], [24] takes a full-wave rectified copy of the output current of $G_M$, i.e., $I_{R_{\text{cut}}}$ in Fig. 8, and produces an output current $I_{\text{out}}$ that is a function of $I_{\text{in}}$. By using a translinear circuit that implements an appropriate function, the MOS resistor may be configured to have linear or nonlinear $I$–$V$ characteristics. Translinear circuits which eventually result in compressive, linear and expansive $I$–$V$ characteristics for the resistor are shown in Fig. 9.

![Image](image_url)

In general:

$$I_D = \frac{W}{L} \left[ F(\psi_X) - F(\psi_Y) \right]$$

$$= I_{X_{\text{sat}}} - I_{Y_{\text{sat}}}$$

where $I_{G_{\text{M}}}$ and $V_L$ are the biasing current and input linear range of the WLR OTA respectively. Hence, $G_M$ is electronically tunable via $I_{G_{\text{M}}}$. The translinear circuit [18], [24] takes a full-wave rectified copy of the output current of $G_M$, i.e., $I_{R_{\text{cut}}}$ in Fig. 8, and produces an output current $I_{\text{out}}$ that is a function of $I_{\text{in}}$. By using a translinear circuit that implements an appropriate function, the MOS resistor may be configured to have linear or nonlinear $I$–$V$ characteristics. Translinear circuits which eventually result in compressive, linear and expansive $I$–$V$ characteristics for the resistor are shown in Fig. 9.
a square-root $I-V$ characteristic employs the translinear circuit of Fig. 9(a). The output current of the WLR OTA given by $I_{LM}V_{XY}$ is compressed by the translinear circuit in a square-root manner to produce the desired $I-V$ relation: The negative-feedback loop serves to produce an output current $I_{LM}$ that is compressed by the translinear circuit in a square-root manner to produce the desired $I-V$ relation: The negative-feedback loop serves to produce an output current $I_{LM}$ that is compressed in a square-root manner to produce the desired $I-V$ relation. The measurement was repeated with $V_X$ and $V_Y$ interchanged in Fig. 15(b). The results show that there is good circuit symmetry. The plots also show that the $I-V$ relation may be scaled electronically by varying the biasing current $I_{BM}$ of the OTA. The same effect may also be achieved by varying $I_{REF}$ in the translinear circuit.

The variable impedance model of the glottis depicted in the top left corner of Fig. 2 is implemented using a series combination of the linear and nonlinear resistors described in this section. The values of the resistances are modulated through the biasing current $I_{BM}$ of the WLR OTA shown in Fig. 8 by a glottal oscillator. The pressure (voltage) drop across the glottal constriction is given by the difference between $P_{AdV}$ and the pressure at the entrance of the intra-oral tract in Fig. 2. The results show that there is good circuit symmetry. The plots also show that the $I-V$ relation may be scaled electronically by varying the biasing current $I_{BM}$ of the OTA. The same effect may also be achieved by varying $I_{REF}$ in the translinear circuit.

The variable impedance model of the glottis depicted in the top left corner of Fig. 2 is implemented using a series combination of the linear and nonlinear resistors described in this section. The values of the resistances are modulated through the biasing current $I_{BM}$ of the WLR OTA shown in Fig. 8 by a glottal oscillator. The pressure (voltage) drop across the glottal constriction is given by the difference between $P_{AdV}$ and the pressure at the entrance of the intra-oral tract in Fig. 2. The results show that there is good circuit symmetry. The plots also show that the $I-V$ relation may be scaled electronically by varying the biasing current $I_{BM}$ of the OTA. The same effect may also be achieved by varying $I_{REF}$ in the translinear circuit.

The variable impedance model of the glottis depicted in the top left corner of Fig. 2 is implemented using a series combination of the linear and nonlinear resistors described in this section. The values of the resistances are modulated through the biasing current $I_{BM}$ of the WLR OTA shown in Fig. 8 by a glottal oscillator. The pressure (voltage) drop across the glottal constriction is given by the difference between $P_{AdV}$ and the pressure at the entrance of the intra-oral tract in Fig. 2. The results show that there is good circuit symmetry. The plots also show that the $I-V$ relation may be scaled electronically by varying the biasing current $I_{BM}$ of the OTA. The same effect may also be achieved by varying $I_{REF}$ in the translinear circuit.
V. Driving the Vocal Tract

Our analog vocal tract, presented in Fig. 2 and comprising the circuits explained in Sections III and IV, is able to generate (decode) all the speech sounds of interest given the area function describing the vocal tract profile, the glottal excitation source, and the turbulent noise source. In order to extensively test and prove the efficacy of our AVT beyond the simple stationary sounds presented in the Section III, we introduce a speech-coding scheme.

The production of a given speech signal is determined primarily by the area function. Unfortunately, the area function space that describes a discrete acoustic tube, consisting of quantized sections that are 1 cm in length, has a large number of degrees of freedom. In order to reduce the dimensionality of the articulatory description and restrict the area function to physiologically realistic vocal tract shapes, we employ the Maeda articulatory model [21] to specify the vocal tract area function. The Maeda articulatory model is a statistical anthropomorphic model, developed from cineradiographic and labiofilm data of the human vocal tract, which describes the vocal tract profile using seven components, each corresponding to an elementary articulator. The seven articulatory parameters are: jaw height, which moves the jaw vertically up or down; tongue body position, which moves the tongue dorsum roughly horizontally from front to back of the oral cavity; tongue body shape, which indicates whether the tongue dorsum is rounded (arched) or unrounded (flat); tongue tip, which deforms the apex part of the tongue by moving it up or down; lip height, which affects the area of the mouth opening by moving the lips together or apart; lip protrusion, which extends the vocal tract slightly during the production of rounded vowels; and larynx height, which raises or lowers the position of the larynx. Physiologically realistic vocal tract profiles may be represented with reasonable accuracy using these seven articulatory parameters [21].

We derive an articulatory codebook containing mappings from the articulatory to acoustic domains using such an articulatory model. A set of vocal tract profiles is generated by systematically stepping through the articulatory parameters. A stationary speech sound, known as the babble, is synthesized using each profile. In other words, each vocal tract profile in the set is associated with a babble. The synthesized babbles are analyzed to produce acoustic features and compiled into a look-up table to produce a codebook. We call the process of building up such an articulatory codebook “babbling.” Our codebook contains approximately 16 000 physiologically realistic entries, specified by 12 mel-frequency cepstral coefficients [22] representing the babble in the acoustic domain and the corresponding seven articulatory parameters that produce the appropriate vocal tract profile (area function).

Fig. 16 is consistent with the empirical relations derived in [15] for constrictions: The I–V relation is approximately linear when the voltage difference across the constriction is small, and gradually becomes compressive (square root) as the pressure difference across the constriction becomes large.

Authorized licensed use limited to: IEEE Xplore. Downloaded on November 23, 2008 at 16:25 from IEEE Xplore. Restrictions apply.
Fig. 17 shows the synthesis process using a circuit model of the vocal tract driven with an articulatory representation supplemented with energy and pitch contour information. The input is a target sound to be reconstructed. The mel-frequency spaced filter bank (center frequencies ranging from 130 to 6500 Hz) decomposes the target sound into its constituent frequency components, which are represented by a set of 30 mel-frequency spaced filter coefficients. A discrete cosine transform (DCT) is applied on the mel-frequency spectrum to generate the set of 12 mel-frequency cepstral coefficients representing the sound. The mel-frequency cepstrum gives a description of the shape of the mel-frequency spectrum. The set of cepstral coefficients of the target sound are compared against a codebook that contains the cepstral coefficients of babbles synthesized by the vocal tract. The set of cepstral coefficients that produce the best acoustic match is found and the corresponding articulatory parameters are forwarded to an articulatory model to produce a vocal tract area profile that serves to drive the vocal tract such that the target sound is reconstructed. The pitch contour and the energy envelope of the target sound provide information about how the pitch and loudness of the sound varies with time and are extracted to create the appropriate glottal excitation waveform.

An additional codebook is derived using the same babbling procedure for the purpose of producing consonant sounds using a turbulent noise source located downstream of the vocal tract constriction. In order to produce consonant sounds, bandlimited noise is injected at the appropriate position of the vocal tract relative to the constriction location derived from the articulatory parameters.

A sequence of vocal tract profiles derived from only the best acoustic matches is not always perceptually optimal for non-stationary sounds because two dissimilar vocal tract profiles may produce similar acoustics resulting in abrupt variations in the articulatory control as they transition between speech frames. Imposing an articulatory constraint on the vocal tract movement allows us to avoid unrealistic vocal tract transitions and alleviates the problem of abrupt changes occurring in the articulatory control that are not physiological. Specifically, at every time step, a set of vocal tract candidates is shortlisted based on acoustic match and a cost is associated with each articulatory movement between speech frames. A dynamic programming search is used to select a sequence of articulatory parameters from the articulatory codebooks such that the articulatory trajectory is optimized [23].

VI. RESULTS

Fig. 18 shows a die photo of our AVT fabricated in a 1.5 μm AMI CMOS process. The AVT comprises a cascade of 16 tunable two-port π-sections, each representing a uniform tube of adjustable length. When configured to have a length of 1.0 ~ 1.1 cm per section, the AVT has a length of 16 ~ 17.6 cm, which is approximately the vocal tract length of a male speaker. For many speech synthesis applications, a maximum frequency of 5 ~ 7 kHz is sufficient. Thus, with 16 sections, the minimum wavelength of interest is large compared to the length of each
section. In the current design, the vocal tract may be configured to comprise 8, 12, or 16 sections via digital control. It may also be connected internally to either a current source or a variable impedance circuit as the glottal driver. The digital circuitry also serves to select various points along the vocal tract cascade to be monitored. The chip consumes less than 275 µW of power when operated with a 5 V power supply. Fig. 19 depicts the spectrogram of a recording of the word “Massachusetts” lowpass filtered at 5.5 kHz. The recording has a female voice. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity. During the training phase, the AVT undergoes the babbling process to produce an articulatory codebook. Note that the training phase also serves to calibrate the AVT. Using the recording as a target sound, an optimal articulatory trajectory, given by the vocalogram of Fig. 20, is derived through dynamic programming. The motor-domain vocalogram is a vector time series of areal cross sections of the vocal tract and is analogous to the spectrogram in the auditory domain. The vocalogram is used to drive the AVT to produce speech. Synthesized speech waveforms correspond to voltage waveforms obtained at the output $P_{out}$ of Fig. 2. The spectrogram of the synthesized speech obtained using the vocalogram of Fig. 20 is shown in Fig. 21. The length of each section of the AVT was adjusted such that the total length corresponds to a female vocal tract. Comparing the spectrograms of the original recording (Fig. 19) and the synthesized sound (Fig. 21), it is evident that the principal formants and the trajectories are well matched. It is also evident that high-frequency speech components that were missing in Fig. 19 have been re-introduced by the AVT in Fig. 21. This effect is attributed to the inherent property of the AVT to synthesize all and only speech signals and thus provides a measure of signal restoration. Such signal restorative properties are particularly important when dealing with noisy speech and robust speech recognition in noise.
Fig. 20. Vocalogram of the word “Massachusetts.” The white plus sign markers on the vocalogram indicate the position of the constriction when the speech segment is a consonant.

Fig. 21. Spectrogram of the word “Massachusetts” synthesized by the AVT. A female vocal tract and the original extracted pitch contour are used.

Fig. 22. Spectrogram of the word “Massachusetts” synthesized by the AVT. A male vocal tract is used and the extracted pitch contour is scaled down by a factor of 1.5.

Fig. 23. Spectrogram of the word “Massachusetts” synthesized by the AVT. A female vocal tract is used and the extracted pitch contour is scaled down by a factor of 1.5.

Compared to concatenation and formant synthesis, the AVT allows us to easily change the synthesized speaker voice by varying the vocal tract length. This property is very useful for speaker identification. Using a single current to control the length of each vocal tract section allows us to very easily change the overall length of the vocal tract. It is less straightforward to achieve similar results without a model of the vocal tract. Fig. 22 shows the synthesis results when a male vocal tract is used and when the extracted pitch is scaled down by a factor of 1.5 to produce a realistic male pitch. Compared to Fig. 21, the resulting speech has a lower third formant in the voiced segments, which is consistent with a longer vocal tract. Note that with only pitch scaling an unrealistic voice is obtained.

Fig. 23 shows the synthesis results when the female vocal tract is used and when the pitch contour is scaled down by a factor of 1.5 in an attempt to produce a male voice. The synthesized speech resembles a female speaker with a low-pitched voice. In particular, the spectrogram shows a third formant that is located at approximately the same frequency as the one in Fig. 21. The result is not surprising as the speech sounds in Fig. 21 and Fig. 23 are produced by female vocal tracts of the same length albeit with different pitch periods.

VII. CONCLUSION

We presented the first experimental integrated-circuit vocal tract. The 275 µW analog vocal tract chip can be used with auditory processors in a feedback speech locked loop to generate speech. We showed examples of words synthesized by our AVT. We presented electronically tunable two-port equivalents of LC π-sections that are used as building blocks for our transmission line vocal tract. Our two-port topology produces the correct change in the L/C ratio while keeping the LC product constant by varying a single circuit parameter that is used to control cross-sectional area variations along the transmission line vocal tract. We also showed how to incorporate the effect of non-rigid
back biasing architecture enables the resistor to have arbitrary device and has inherently zero d.c. offset. Our negative-feedforward model the constriction at the glottis or at the supraglottal constriction implemented in CMOS technology that can be used to correct an electronically tunable linear and nonlinear MOS resistor, with reciprocal synapses.

Our analog vocal tract could be potentially applied in speech synthesis, speech recognition, speech compression, and speaker identification systems.

REFERENCES


Lorenzo Turicchia received the Laurea degree in electrical engineering from the University of Padova, Italy, and the Ph.D. degree in computer science from the Department of Mathematics and Computer Science, University of Udine, Italy. In 2002 he joined the Analog VLSI and Biological Systems group at the Massachusetts Institute of Technology (MIT), Cambridge, where his doctoral research was completed, and where he is currently a Postdoctoral Fellow. His main research interest is in nonlinear signal processing, such as audio signal processing and biomedical signal processing. He is currently working on robust speech recognition in noise and algorithms for spike processing in brain-machine interfaces.

Rahul Sarpeshkar (M’01) received the B.S. degrees in electrical engineering and physics from the Massachusetts Institute of Technology (MIT), Cambridge, and the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena.

After receiving the Ph.D. degree, he joined Bell Laboratories as a member of the technical staff. Since 1999, he has been on the faculty of MIT’s Electrical Engineering and Computer Science Department where he heads a research group on analog VLSI and biological systems, and is currently an Associate Professor. He holds over 25 patents and has authored several publications including one that was featured on the cover of Nature. His research interests include biologically inspired circuits and systems, biomedical systems, and mixed-signal VLSI, ultra-low-power circuits and systems, neuroscience, molecular biology, and control theory.

Dr. Sarpeshkar has received several awards including the Packard Fellow award given to outstanding young faculty, the ONR Young Investigator Award, the Junior Bose Award for Excellence in Teaching at MIT, and the NSF Career Award.

Authorized licensed use limited to: IEEE Xplore. Downloaded on November 23, 2008 at 16:25 from IEEE Xplore. Restrictions apply.