

An Analog Bionic Vocal Tract

Keng Hoong Wee, Lorenzo Turicchia, Rahul Sarpeshkar

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA, USA
Email: {khwee, turic, rahuls}@mit.edu

Abstract— We present the first experimental integrated-circuit vocal tract. The 275 μ W analog vocal tract chip can be used for real-time speech production in bionic speech-prosthesis systems where low power is critical. We also describe how our vocal tract can be used with auditory processors in a feedback speech locked loop to implement speech recognition that is potentially robust in noise.

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 [1] and low-power cochlear-implant processors for the deaf [2]. Silicon models of the retina [3] have been used in machine vision systems and circuit models of the heart have been used to provide insight into cardiac and circulatory malfunction in medicine. In this work, we present the first experimental integrated-circuit analog vocal tract (AVT) by mapping fluid volume velocity to current, fluid pressure to voltage, and linear and nonlinear mechanical impedances to linear and nonlinear electrical impedances. Such silicon vocal tracts find applications in real-time low-power speech production for bionic speech-prosthesis systems, and can be used with auditory processors in a feedback loop to implement real-time, low-power robust speech recognition in noise via analysis-by-synthesis techniques. Our use of a physiological model of the human vocal tract enables the AVT to synthesize all *and only* the speech signals of interest, using articulatory parameters that are intrinsically compact, robust, and linearly interpolatable [4][5]. Below, we describe how our chip is architected and demonstrate some of its potential applications.

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 SLL 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 [4]. Our strategy uses an analog vocal tract to drastically reduce power consumption, enables real-time performance and could be useful in portable speech processing systems of moderate complexity, e.g., in cell phones, digital assistants, and laptops.

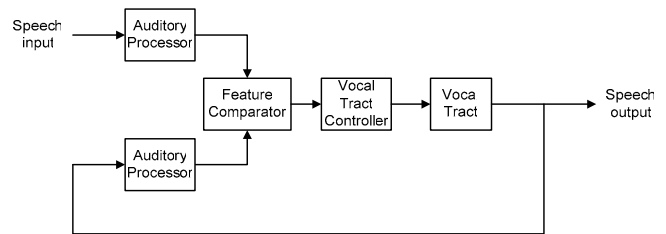


Figure 1. Concept of speech locked loop.

II. CIRCUIT MODEL OF 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 ℓ) with varying cross sections. Each two-port is an electrical equivalent of an LC π -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: (a) a periodic source at the glottis and/or (b) a turbulent noise source P_{turb} at some point along the tube. In Fig. 2, the glottal source is represented by a voltage source P_{alv} with variable source impedance Z_{GC} , that is modulated by a glottal oscillator.

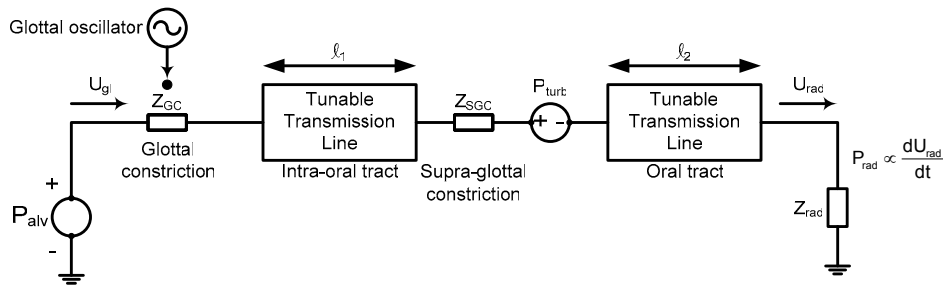


Figure 2. Schematic of transmission line vocal tract.

We use a circuit model of the glottis that 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. The turbulent source P_{turb} has a source impedance comprising the constriction impedance Z_{SGC} . The location of P_{turb} is variable, depending on the constriction location. At the lips, the transmission line is terminated by a radiation impedance Z_{rad} and the radiated sound pressure at the mouth, P_{rad} , is proportional to the derivative of the current flowing in Z_{rad} .

III. VLSI IMPLEMENTATION

Fig. 3 is a circuit diagram showing an electronically tunable linear or nonlinear MOS resistor that can model linear or nonlinear constriction losses due to laminar or turbulent flow respectively, such as the glottal and supraglottal constriction in the vocal tract. 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. For a given I-V characteristic, the circuit senses the source-to-drain potential across an MOS device and automatically generates an appropriate bias for the gate terminal to implement the characteristic via negative feedback. The source-to-drain voltage $V_{XY} = V_X - V_Y$ of the MOS transistor M_R is sensed and converted into a current by a wide linear range operational transconductance amplifier (WLR OTA). The OTA in conjunction with a current rectifier creates a full-wave-rectified current I_m that is proportional to V_{XY} . The saturation currents $I_{X,\text{sat}}$ and $I_{Y,\text{sat}}$ of M_R are proportionally replicated in transistors M_X and M_Y by sensing voltages V_G , V_W , V_X , and V_Y on the gate, well, source, and drain terminals of M_R with source followers, and applying V_{GX} and V_{GY} across the gate-to-source terminals of M_X and M_Y . The difference between $I_{X,\text{sat}}$ and $I_{Y,\text{sat}}$ is full-wave-rectified and compared with a mirrored version of the translinear output current I_{out} . Any difference between the two currents causes the capacitor C to charge or discharge, such that the gate bias equilibrates at a point where the two are nearly equal via negative-feedback action. By using a translinear circuit that implements an appropriate function, the MOS resistor may be configured to have the desired linear (laminar) or nonlinear (turbulent) I-V characteristics (e.g., $I \propto V$ or $I \propto \sqrt{V}$). Such electrical circuit models are consistent with the equations in [6]. Further details on the MOS resistor are described in [7].

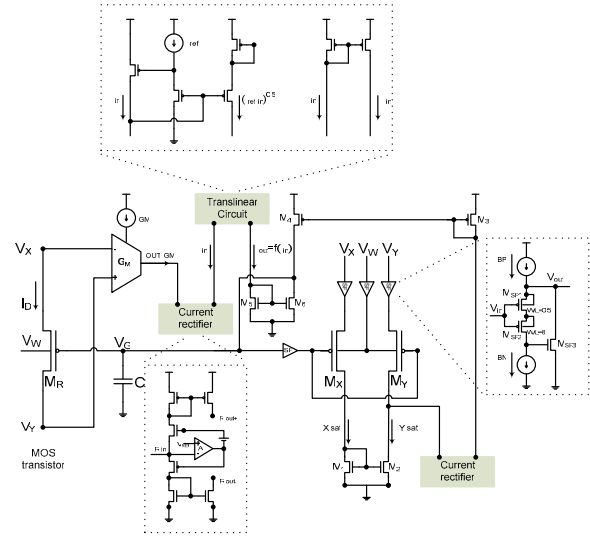


Figure 3. Circuit diagram of linear or nonlinear MOS resistor to model laminar or turbulent fluid flow respectively.

Fig. 4 is a circuit diagram showing an electronically tunable two-port π -section that forms the basic building block of the transmission line. Each LC π -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. Wide linear range operational transconductance amplifiers (WLR OTA) G_1 , G_{2A} , G_{2B} and capacitor C_1 form a tunable bidirectional gyrated inductance given by $L = C_1 / G_1 G_2$. A unidirectional gyrated inductance $Z_C = j\omega L_C$ with OTAs G_{2A} and G_3 implement a tunable shunt capacitance given by $C = G_2 G_3 L_C$. The series resistance R is given by $1 / G_1 G_2 R_0$ where R_0 is the output resistance of G_1 . The shunt conductance G is given by $G = 1 / R_{ds}$ where R_{ds} is the source-to-drain resistance of triode-operated M . The values of G_1 , G_2 and G_3 are controlled by the respective OTA bias currents and the value of G is tunable via a bias voltage V_{GG} . In order to prevent the accumulation of DC offset, mirrored copies of currents I_{1a} and I_{2a} (I_{1a}' and I_{2a}') are compared and the difference integrated to generate an offset compensation bias voltage on capacitor C_C .

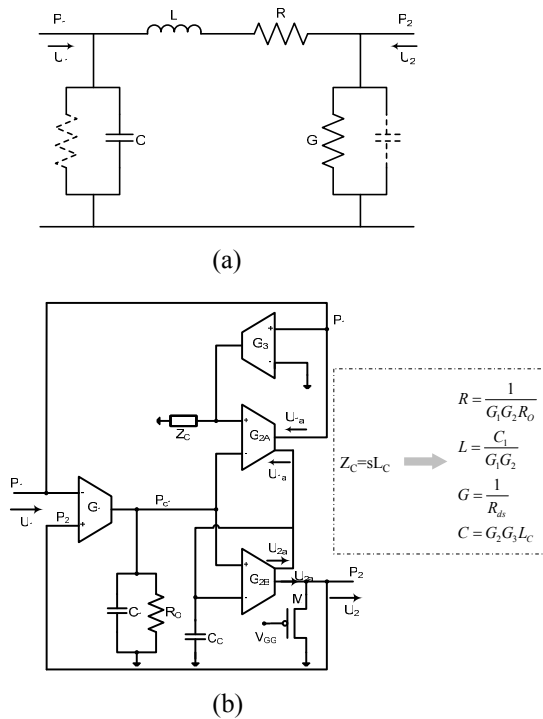


Figure 4. (a) Passive π -circuit model of a cylindrical section of acoustic tube assuming rigid walls. (b) Circuit diagram of tunable two-port π -section that is electrically equivalent to π -circuit model shown in (a).

IV. EXPERIMENTAL RESULTS

Fig. 5 shows a die photo of our AVT fabricated in a 1.5 μm AMI CMOS process. The AVT is composed of a cascade of 16 tunable two-port π -sections, each representing a uniform tube of adjustable length and cross-sectional area. The chip consumes less than 275 μW of power when operated with a 5V power supply.

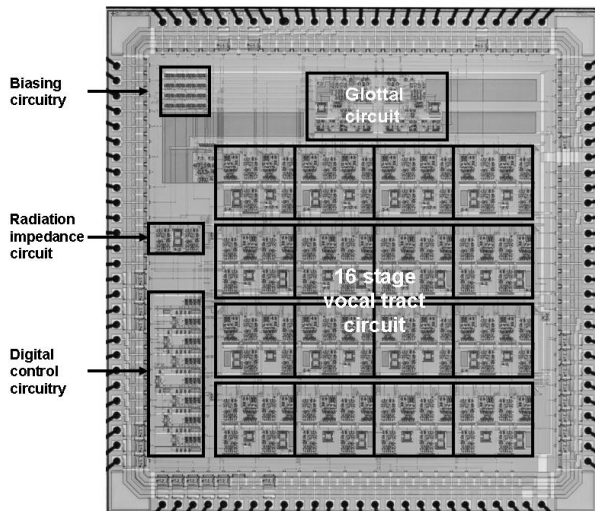
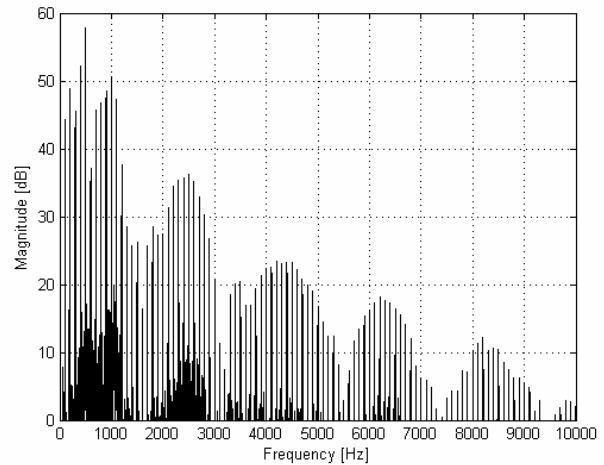
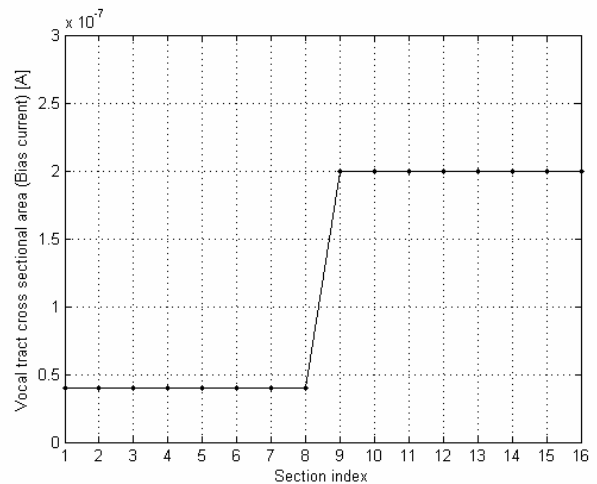


Figure 5. Chip micrograph.

Fig. 6(a) shows the measured output spectrum when the vocal tract is configured to form an acoustic tube with an area profile depicted in Fig. 6(b). In this measurement, the input to the AVT is an LF glottal pulse train [8] with a pitch period of 10ms. The harmonics of the periodic glottal pulse train are clearly illustrated in Fig. 6(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 used to produce it. Fig. 7 shows that the measured SNR at the output of the AVT is 64dB, 66dB, and 63dB for the first three formant resonances of the voiced phoneme /e/. Fig. 8(a) shows the spectrogram of a recording of the word “Massachusetts” lowpass filtered at 5.5kHz. White noise was added to the signal to intentionally obtain a degraded SNR of 25dB. Fig. 8(b) shows the spectrogram of the same word re-synthesized by our AVT using the scheme illustrated in Fig. 1.



(a)



(b)

Figure 6. (a) Measured spectrum at output of AVT when configured to form an acoustic tube with vocal tract area profile shown in (b).

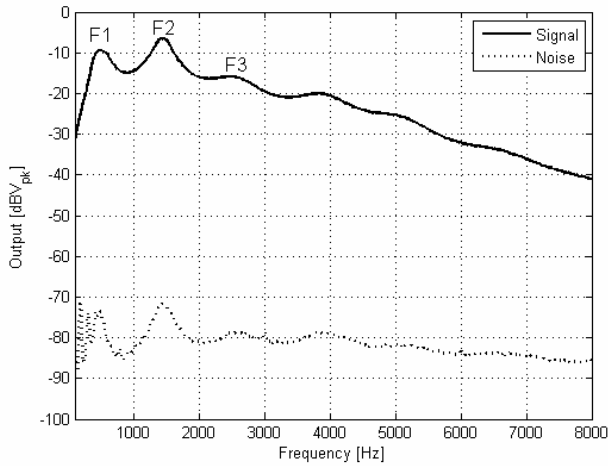


Figure 7. Measured signal and noise characteristics as a function of sinusoidal input frequency.

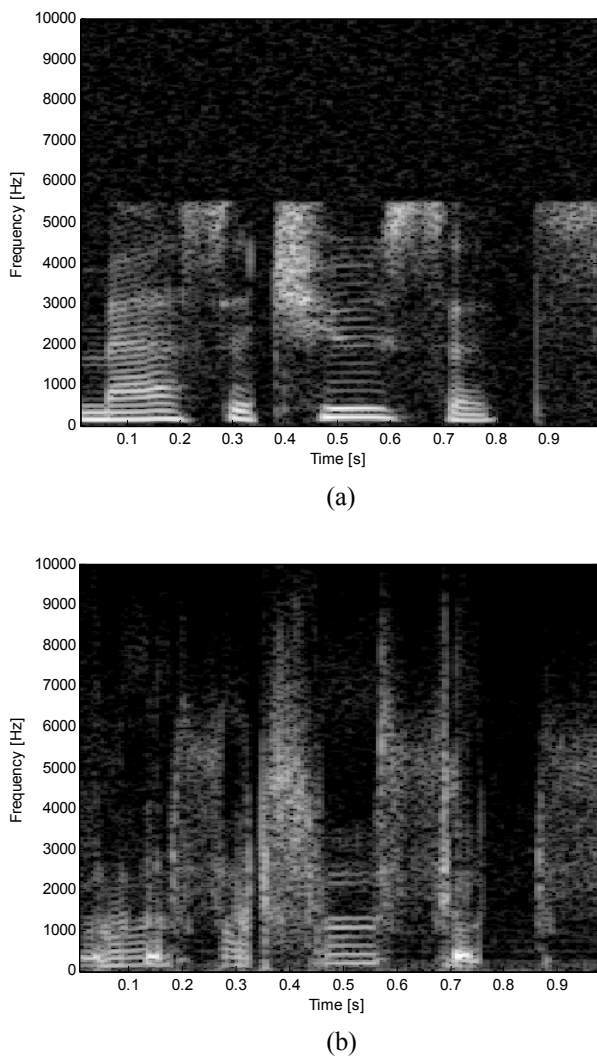


Figure 8. Spectrogram of (a) "Massachusetts" recording and (b) "Massachusetts" re-synthesized by the AVT.

In Fig. 8(b), it is evident that high frequency speech components that were absent in Fig. 8(a) have been introduced by the AVT. It is also noteworthy that the noise added to the recording of Fig. 8(a) is reduced in Fig. 8(b). The apparent noise reduction may be attributed to the inherent property of the AVT to synthesize only speech signals and not random noise as in regression-based systems that attenuate noisy data.

V. CONCLUSIONS

We presented the first experimental integrated-circuit vocal tract. The $275\mu\text{W}$ 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 presented a bidirectional electronically tunable linear and nonlinear MOS resistor implemented in CMOS technology that can be used to model the constriction at the glottis or at the supraglottal constriction. Our MOS resistor exploits the symmetry of an MOS device and has inherently zero DC offset. Our negative-feedback biasing architecture enables the resistor to have arbitrary linear and nonlinear I-V characteristics.

REFERENCES

- [1] B. Raj, L. Turicchia, B. Schmidt-Nielsen, and R. Sarpeshkar, "An FFT-based Companding Front End for Noise-Robust Automatic Speech Recognition," *EURASIP J. Audio, Speech, and Music Process.*, Aug 2007.
- [2] R. Sarpeshkar, M. Baker, C. Salthouse, J.J. Sit, L. Turicchia, and S. Zhak, "An Analog Bionic Ear Processor with Zero-Crossing Detection," *Proc. IEEE ISSCC*, San Francisco, CA, February 6-10, 2005.
- [3] C. Mead. *Analog VLSI and Neural Systems*. Addison-Wesely Publishing Co. New York, 1989.
- [4] M.M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-35, 955-967, 1987.
- [5] C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *J. Acoust. Soc. Am.*, pp. 1725-1736, 1961.
- [6] K. N. Stevens. *Acoustic Phonetics*. The MIT Press, Cambridge, MA, 1998.
- [7] K. H. Wee and R. Sarpeshkar, "An electronically tunable linear or nonlinear MOS resistor," *IEEE Trans. Circuits and Systems I*, in press, 2008.
- [8] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, no.4, 1985.