

MIT Open Access Articles

Biologically inspired silicon vocal tract

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Wee, Keng Hoong. "A Biologically Inspired Silicon Vocal Tract." SPIE Newsroom (2010). Copyright © 2010 SPIE

As Published: <http://dx.doi.org/10.1117/2.1201001.1807>

Publisher: SPIE

Persistent URL: <http://hdl.handle.net/1721.1/73934>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



A biologically inspired silicon vocal tract

Keng Hoong Wee, Lorenzo Turicchia, and Rahul Sarpeshkar

The first integrated-circuit vocal tract, combined with a bionic ear processor in a feedback speech-locked loop, provides a new way to process speech.

Electrical circuit models of biological systems provide an intuitive mechanism for engineers' understanding and are increasingly used to improve the performance of related technology. For example, visual processing performed by the retina can be modeled by a resistive network of interconnected photodetectors and analog processing elements. Complex bio-mechanical systems such as the heart, cochlea, and vocal tract can be modeled using electrical circuits by mapping pressure to voltage, volume velocity to current, and mechanical impedances to electrical impedances, and by representing valves with diodes. Silicon models of the retina¹ have been used in machine vision systems and circuit models of the heart have been used to shed light on cardiac and circulatory malfunction in medicine. Silicon cochlea models have led to improved speech recognition in noise² and low-power cochlear-implant processors for the deaf.³

In this vein, we have developed the first integrated-circuit vocal tract that uses a physiological model of the human vocal tract to synthesize speech. The system employs articulatory parameters that are intrinsically compact, robust, and linearly interpolatable: it could therefore be well suited to noise-robust automatic speech recognition, speech compression, audio noise cancellation, and could form the basis for future bionic speech prostheses.⁴

Figure 1 shows an analysis-by-synthesis block diagram that creates what we term a *speech locked loop* (SLL) in analogy with the phase locked loops (PLL) commonly used in communication systems. The auditory processor and controller are analogous to a phase detector and loop filter in a PLL. 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 at which time the SLL locks to the input sound with an

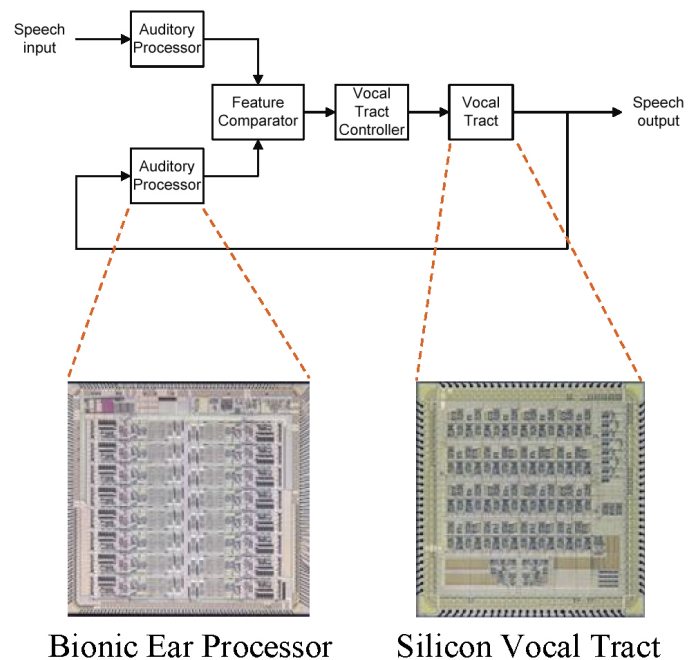


Figure 1. Concept of speech locked loop.

optimal vocal tract profile produced by the controller. Analysis-by-synthesis methods were previously implemented using computationally expensive digital techniques.⁵ Instead, our strategy employs a silicon vocal tract to drastically reduce power consumption, and thus is ideal for portable speech processing systems such as cell phones and personal digital assistants. Power consumption can be reduced further using our previously developed analog bionic ear processor³ as the auditory processors for the SLL.

Our circuit model of the vocal tract, shown in Figure 2, represents the human vocal tract (composed of nasal, pharyngeal, and oral tracts) as acoustic tubes using a transmission line model. Each transmission line comprises a cascade of tunable two-port elements, corresponding to a concatenation of short cylindrical acoustic tubes (illustrated in Figure 3) of length ℓ with varying

Continued on next page

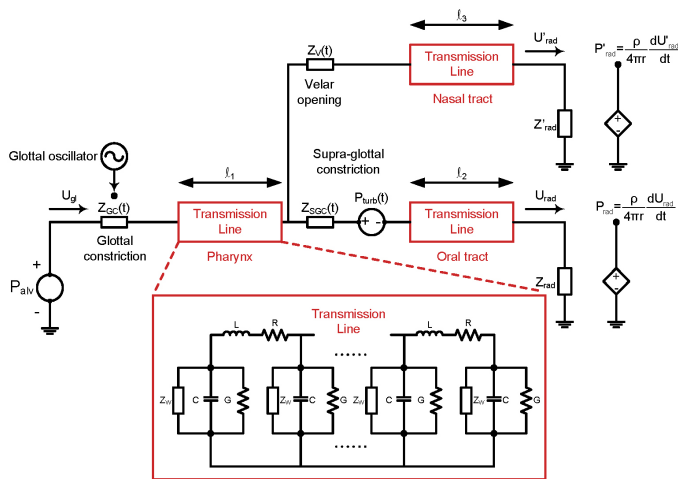


Figure 2. Schematic of transmission line vocal tract.

cross-sections. Each two-port element 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 (e.g., movement of the tongue, jaw, lips). 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, namely a periodic source at the glottis and/or a turbulent noise source P_{turb} at some point along the tube.

In Figure 2, the glottal source is represented by a voltage source P_{alv} with variable source impedance Z_{GC} , which is modulated by a glottal oscillator. We use a circuit model of the glottis that comprises linear and nonlinear resistances connected in series to represent losses occurring at the glottis due to laminar and turbulent flow, respectively. The turbulent source P_{turb} is connected in series with an impedance Z_{SGC} . The location of P_{turb} and Z_{SGC} is not fixed in the oral tract, but varies depending on the constriction location. During the production of nasal sounds, the nasal tract becomes coupled to the oral tract via the velar impedance Z_V ; otherwise, Z_V is an open circuit. At the lips and at the nose, the transmission lines are terminated by radiation impedances Z_{rad} and Z'_{rad} , and the radiated sound pressures P_{rad} and P'_{rad} are proportional to the derivative of the currents flowing in the respective radiating impedances. For simplicity, the electronic vocal tract in its first instantiation only implements the oral tract rather than the oral and nasal tracts.

Our silicon vocal tract is able to generate all speech sounds, given the vocal tract profile and the excitation sources. In order to extensively test and prove the efficacy of our SLL, we introduce a speech-coding scheme based on an anthropomorphic articulatory model⁶ that describes the vocal tract profile using

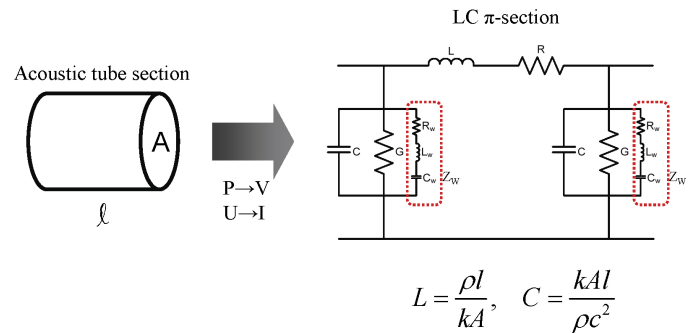


Figure 3. Electrical circuit model of short cylindrical acoustic tube of length ℓ and cross sectional area A . Acoustic inertance and compliance of the cylinder are represented by electrical inductance L and capacitance C . Viscous losses due to laminar flow (Poiseuille's Law) are modeled by electrical resistance R . Losses at the cylinder walls are modeled by electrical conductance G . Mass and compliance arising from non-rigid cylinder walls are modeled by electrical impedance Z_W . The density of air and the speed of sound in air are given by ρ and c respectively. For the purpose of practical circuit implementation, k is an arbitrary scaling constant.

seven components, each corresponding to an elementary articulator. Physiologically realistic vocal tract profiles may be represented with reasonable accuracy using these seven articulatory parameters, namely: jaw height, which controls the vertical position of the jaw; tongue body position, which moves the tongue dorsum from the front to the back of the oral cavity; tongue body shape, which indicates whether the tongue dorsum is rounded or unrounded; tongue tip position, which controls the position of the tongue apex; lip height, which varies the mouth opening; lip protrusion, which controls the mouth protrusion; and larynx height, which raises or lowers the position of the larynx. The trajectories of these seven components in time are reconstructed by our SLL, creating what we call an *articulogram*, which can be used to supplement the spectrogram to enhance the robustness of automatic speech recognition systems.

Figure 4(a) shows the spectrogram of a recording of the word 'Massachusetts' lowpass filtered at 5.5kHz. Figure 4(b) shows what we term the *vocalogram*, a 3D plot of the vocal tract profile as a function of time, extracted from analyzing the recording using the speech locked loop illustrated in Figure 1. Figure 4(c) shows the spectrogram of the same word re-synthesized by our SLL. In Figure 4(c), it is evident that high frequency speech components that were absent in Figure 4(a) have been introduced by the SLL, which inherently synthesizes only speech signals

Continued on next page

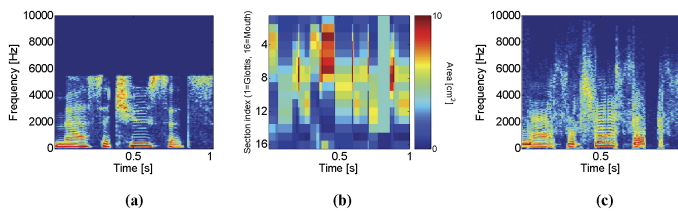


Figure 4. (a) Spectrogram of a recording of the word 'Massachusetts' lowpass filtered at 5.5kHz. (b) Vocalogram extracted from recording. (c) Spectrogram of the same word re-synthesized by our SLL.

because it is based on a physiological model. Such signal restorative properties are particularly important when dealing with noisy speech.

Conclusion

Our integrated-circuit vocal tract based on a physiological model of the human vocal tract can be used with auditory processors (e.g., analog bionic ear processors) in a feedback speech-locked loop to implement speech recognition that is robust in noise. It also has potential for future low-power, real-time speech production, speech compression, audio noise reduction, and bionic speech-prosthesis systems.

Author Information

Keng Hoong Wee

Department of Electrical and Computer Engineering
National University of Singapore
Singapore, Singapore

Keng Hoong Wee is an adjunct assistant professor. He received his PhD in electrical engineering and computer science from the Massachusetts Institute of Technology. His research interests lie in biologically inspired circuits and systems, biomedical systems, and speech processing.

Lorenzo Turicchia

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA

http://www.rle.mit.edu/avbs/whoweare_lorenzo.html

Lorenzo Turicchia is a research scientist whose main research interests are in nonlinear signal processing, especially for audio and biomedical applications, and bioelectronics. His work has included research on cochlear implants, visual prostheses, speech prostheses, speech recognition, and wearable medical devices.

Rahul Sarpeshkar

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA
<http://www.rle.mit.edu/avbs>

Rahul Sarpeshkar is currently an associate professor and heads a research group in Analog VLSI and Biological Systems. His research on bioelectronics has won several awards including the Packard award given to outstanding faculty.

References

1. C. Mead, **Analog VLSI and Neural Systems**, Addison-Wesely Publishing Co., New York, 1989.
2. B. Raj, L. Turicchia, B. Schmidt-Nielsen, and R. Sarpeshkar, *An FFT-Based Companding Front End for Noise-Robust Automatic Speech Recognition*, **EURASIP J. on Audio, Speech, and Music Proc.** 2007, p. 13, 2007. Article ID 65420, doi:10.1155/2007/65420
3. R. Sarpeshkar, C. Salthouse, J. J. Sit, M. Baker, S. Zhak, T. Lu, L. Turicchia, and S. Balster, *An Ultra-Low-Power Programmable Analog Bionic Ear Processor*, **IEEE Trans. Biomed. Eng.** 52 (4), pp. 711–727, 2005. doi:10.1109/TBME.2005.844043
4. K. H. Wee, L. Turicchia, and R. Sarpeshkar, *An Analog Integrated-Circuit Vocal Tract*, **IEEE Trans. Biomed. Circuits and Sys.** 2 (4), pp. 316–327, 2008. doi:10.1109/TBCAS.2008.2005296
5. M. M. Sondhi and J. Schroeter, *A hybrid time-frequency domain articulatory speech synthesizer*, **IEEE Trans. Acoust. Speech Signal Proc.**, pp. 955–967, 1987. ASSP-35
6. S. Maeda, *Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model*, W. J. Hardcastle and A. Marchal (eds.), in **Speech Production and Speech Modelling**, pp. 131–149, Kluwer, Dordrecht, The Netherlands, 1990.