

An Articulatory Silicon Vocal Tract for Speech and Hearing Prostheses

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

Abstract—We describe the concept of a bioinspired feedback loop that combines a cochlear processor with an integrated-circuit vocal tract to create what we call a speech-locked loop. We discuss how the speech-locked loop can be applied in hearing prostheses, such as cochlear implants, to help improve speech recognition in noise. We also investigate speech-coding strategies for brain-machine-interface-based speech prostheses and present an articulatory speech-synthesis system by using an integrated-circuit vocal tract that models the human vocal tract. Our articulatory silicon vocal tract makes the transmission of low bit-rate speech-coding parameters feasible over a bandwidth-constrained body sensor network. To the best of our knowledge, this is the first articulatory speech-prosthesis system reported to date. We also present a speech-prosthesis simulator as a means to generate realistic articulatory parameter sequences.

Index Terms—Analog bionic vocal tract, articulatory speech prosthesis, bioinspired circuits, speech coding, speech synthesis.

I. INTRODUCTION

BIOLOGICAL systems exhibit remarkable degrees of energy efficiency and robustness that are orders of magnitude better than the best engineering systems today. For example, the human cochlea performs highly complex spectral analysis with a meager 14 μW of power in a tiny volume. When applied in the proper engineering context, models of biology can provide inspiration for improving the performance of engineering systems [1]. For example, we can derive electrical circuit models of biological systems that give insightful engineering perspectives and, thus, provide an intuitive framework for analyzing and understanding biology. These electrical circuit models of biology are increasingly being used to shed light on biological systems and to improve performance in engineering systems. For example, complex biomechanical systems, such as the cochlea, vocal tract, and heart can be modeled using electrical circuits by mapping pressure to voltage, volume velocity to current, and mechanical impedances to electrical impedances. Circuit models of the heart [2] can be used to estimate cardiovascular parameters and provide insight into

Manuscript received December 02, 2010; revised April 05, 2011; accepted May 21, 2011. Date of current version July 27, 2011. This work was supported in part by Grants NIH NS-056140 and ONR N00014-09-1-1015. This paper was recommended by Associate Editor W.-F. Wong.

K. H. Wee is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: elewkh@nus.edu.sg).

L. Turicchia and R. Sarpeshkar are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: turic@mit.edu; rahuls@mit.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TBCAS.2011.2159858

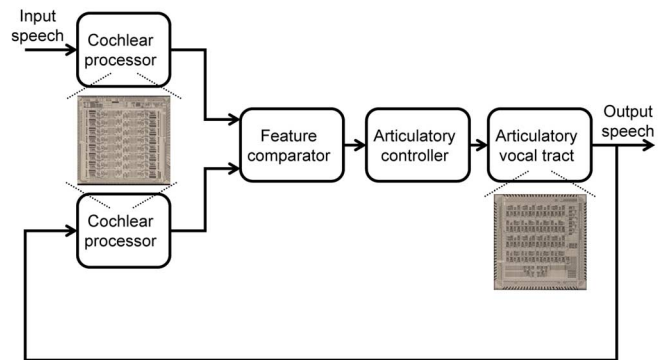


Fig. 1. Block diagram illustration of a speech-locked loop.

cardiac and circulatory functions. Cochlea-like models have led to improved processors for hearing in noisy environments [3]–[6] and for rapid radio-frequency (RF) spectrum analysis [7].

Feedback loops have a central role in ensuring that biological systems function in a robust manner. In this vein, we have combined a model of the cochlea and a vocal tract in a bioinspired feedback configuration [1], [8]–[10]. We call this feedback system, a speech-locked-loop (SLL), in analogy with the phase-locked-loops (PLL), which are widely used in other communication systems. Specifically, the cochlea and vocal tract in the SLL are analogous to a phase detector and voltage-controlled oscillator (VCO) of the phase-locked loop (PLL), respectively. Fig. 1 is a block diagram representation of the SLL that illustrates the dual relationship between analysis and synthesis: analysis is needed to fine-tune synthesis, and conversely, synthesis can be exploited for better analysis.

In our previous work [8], we introduced the concept of the speech-locked loop and described an integrated-circuit (IC) vocal tract that operates at the heart of such a system. In this invited journal paper, we present how the speech-locked loop can be applied in hearing prostheses, such as cochlear implants, to help improve speech recognition in noise. We also investigate speech-coding strategies for brain-machine-interface (BMI)-based speech prostheses and present an articulatory speech-synthesis system using an IC vocal tract that models the human vocal tract. Our system makes the transmission of low bit-rate speech-coding parameters feasible and can be used as a speech-prosthesis simulator (SPS) to generate realistic articulatory parameter sequences.

The organization of this paper is as follows. In Section II, we study the pros and cons of various speech-coding strategies that can be used to control the speech synthesizer in a prosthesis.

We also discuss the application of the SLL to brain-machine-interface (BMI)-based speech prostheses and propose a speech-coding strategy ideal for these prostheses. In Section III, we describe our IC vocal tract that produces speech in real time using articulatory speech-coding parameters that require low transmission bandwidth and are thus well suited for use in a body sensor network (BSN). In Section IV, we describe how our SLL can improve hearing in noise and, therefore, be useful in hearing prostheses, such as cochlear implants (CIs). Since cochlear-implant and hearing-impaired patients have extreme difficulty understanding speech in noisy environments, such noise-cancellation techniques are very helpful. We also present the experimental results of speech produced by our IC vocal tract chip. In Section V, we summarize the contributions of our paper.

II. CODING STRATEGIES

The loss of the ability to communicate is considered one of the most disabling conditions a person can experience. Perhaps the most disabling condition is experienced by locked-in syndrome patients who are conscious but cannot move, due to complete paralysis of almost all voluntary muscles. This terrible condition, sometimes referred to as the “buried alive” syndrome, currently has no cure. Restoring the ability to speak would be a major improvement in the lives of these patients. In the past, brain-machine-interfaces (BMIs) have been used to help these patients communicate. BMIs based on electroencephalography (EEG) have been shown to give some sort of communication capability to these patients [11]. Unfortunately, EEG systems use electrodes placed on the scalp to measure neural activity occurring deep inside the brain; as a consequence, the recorded information is noisy. BMIs based on functional magnetic resonance imaging (fMRI) overcome this problem because they can record neural activity occurring deep inside the brain with high spatial resolution [12]. Currently, fMRI systems are large and impractical and cannot create a viable speech prosthesis. In addition, fMRI measurements are inherently slow and, consequently, it is almost impossible to achieve real-time production of speech. Perhaps in the future, some of these drawbacks can be resolved by using other promising and noninvasive techniques, such as magnetoencephalography (MEG) and functional near-infrared spectroscopy (fNIRS). On the other hand, a more direct, albeit invasive technique that employs electrodes implanted directly in the brain to measure neural activity has shown great promise [13].

Regardless of the transduction method (e.g., EEG, MEG, fMRI, fNIRS, semi-invasive subdural or cortical neural recordings), brain-activity measurements have to be decoded into features that can control the production of speech. These control signals drive a speech synthesizer, which then generates the intended speech. Together, the brain-activity transducer and decoder form a brain-machine interface (BMI). The BMI and the speech synthesizer constitute a BSN suited for people with speech disorders. Fig. 2 shows a BSN with the essential components needed to produce speech from intention.

Ideally, for speech synthesis, we want independent, precise, and reliable control signals. However, current state-of-the-art BMIs have limitations as follows.

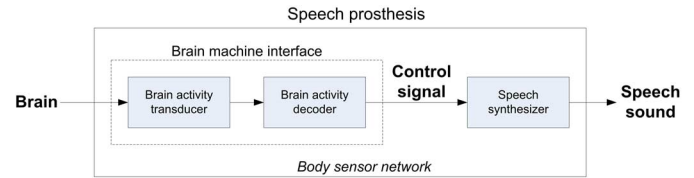


Fig. 2. Concept of a speech-prosthesis system.

- 1) Only a few control signals are independent (i.e., low-dimensional control signals).
- 2) Control signals can reliably code only a few different states (i.e., low-resolution control signals).
- 3) Reliable control signals cannot carry high-frequency information (i.e., only low-bandwidth control signals are presently possible).
- 4) Control signals are noisy.

For example, state-of-the-art BMIs are able to decode only four to seven independent control signals from a monkey’s brain for the control of a prosthetic arm [14], [15]. As a result, the speech synthesizer (in the speech-prosthesis system depicted in Fig. 2) must produce speech despite low-dimensional, low-resolution, low-bandwidth, and noisy control signals. In addition, a speech prosthesis should have the following characteristics:

- 1) The subject should be able to generate speech as naturally as possible.
- 2) The subject should be able to learn and control the prosthesis as easily as possible.
- 3) The synthesized speech should be produced in real time.
- 4) The synthesized speech should sound as natural as possible.
- 5) The speech prosthesis should be portable or implantable and, consequently, have low power consumption.

Many strategies can be used to produce speech from transduced brain activity (i.e., from control signals). In particular, it is critical to choose what information has to be learned and used by the subject to drive the speech prosthesis. The subject can control the speech prosthesis BY using various speech representations or coding strategies. For example, the subject can learn to produce different sounds by controlling the spectral poles, corresponding to speech formants, created by the speech synthesizer. Perhaps the brain can learn to use various speech representations to deliver speech; but, given the current limitation of BMIs, some representations are better than others. We discuss the pros and cons of some of the most promising speech representations that can be used.

A. Alphabet

The alphabet (26 letters + space) can be used to compose words and, hence, speech. A subject can select the letters of the alphabet in a 2-D matrix, learning to control a cursor position on a screen that displays the alphabet visually. This approach involves learning to move a cursor with two degrees of freedom. However, if only two states can be reliably selected with the BMI (due to low signal-to-noise ratios), another approach is required. This approach involves a computer sequentially asking whether the desired letter is present on the screen; the subject needs only to respond “yes” when the desired letter is shown.

Using alphabets to produce speech in this fashion is slow and does not constitute a natural way to speak; but, the production of speech can be easily learned and controlled, and the created sound is natural because a high-quality text-to-speech synthesizer can be used to convert the composed words into speech. This approach works despite the limitations of state-of-the-art BMIs and can lead to coding strategies with low-dimensional, low-resolution, low-bandwidth control signals. The 26 letters of the alphabet can be arranged in a 2-D matrix (6×5). Consequently, two independent control signals with, at most, six selectable states are required. The low number of control signals and states make this method robust to noise. However, it needs a screen in front of the subject for the selection of the letters and, consequently, the prosthesis cannot be totally implanted. The cursor can also be driven from letter to letter by coding the desired direction (e.g., left, right, up, down), making the control more robust but, at the same time, even slower.

B. Word

A preselected set of words can be used to compose speech. The subject can select the words as in the alphabet case. Here, a word-to-speech synthesizer is required and since there are more words than letters of the alphabet, the required number of states is increased.

C. Sentence

A preselected set of sentences is used to compose speech. The subject can select the sentences as explained in the alphabet case. Here, a sentence-to-speech synthesizer is required and since there are more possible sentences than words, the required number of states is even larger.

D. Phoneme

A preselected set of phonemes is used to compose speech. A subset of the international phonetic alphabet (IPA) could be used (e.g., ~ 40 phonemes plus silence). The subject can select phonemes as in the alphabet case. Here, a phoneme-based synthesizer is required. The properties of this approach are similar to those of the alphabet case. However, since phonemes are direct representations of speech sounds, this approach results in more efficient synthesis (after a learning phase). Special techniques to reduce co-articulation are needed in order to generate high-quality speech.

E. Formant

Speech signals can be represented by resonances in the frequency spectrum (i.e., formant frequencies). The Klatt formant synthesizer exploits this characteristic [15], which can also be exploited in a speech prosthesis. The most significant formants (e.g., the first two or three formants) can be used by the subject to compose speech. In addition, we need to control the voiced and unvoiced source amplitude, and, eventually, for more natural speech, we also need to encode the fundamental frequency (pitch). Consequently, this method requires at least 4–6 control variables to encode speech. Note that unlike the previous cases, the subject directly controls the production of speech in real time: a screen for selecting the parameters is not necessary and the control variables immediately modify the production

of the sound. The control signal can be fine-tuned by the user because he or she is simultaneously listening to the produced speech, thus enhancing robustness to noise. However, the control signal has to be fast. In addition, the formants need to be controlled precisely, making them difficult to learn and control. Such a speech prosthesis has been recently reported to produce four vowels (UH, IY, A, OO) [22]. Higher sound quality can be achieved by also encoding the bandwidth of each formant: at least 6–9 control variables are required. However, formant synthesizers (e.g., [15]) require abrupt changes in the control signals to produce vowel-consonant transitions, which impose impractical bandwidth requirements for current state-of-the-art BMIs. As described in the alphabet case, coding the formant trajectory instead of the absolute formant position improves robustness to noise but makes the system unacceptably slow.

F. Filter Coefficients

The subject could learn to generate speech by driving a set of control variables corresponding to filter coefficients. The resulting filter is used to filter white noise and/or a glottal signal in order to produce speech signals. This technique of encoding speech has been successfully used in LPC synthesizers, but learning to use filter coefficients is extremely hard and, at the same time, a large set of independent and very precise control variables is required. Consequently, although this technique has been used in digital speech synthesis with some success, it cannot be easily used in a speech prosthesis.

G. Spectrum

It has been proven that a user, after a year of training, is able to generate intelligible speech using a machine with a keyboard and a foot pedal [17]. This machine is known as the Voder. The spectrum of speech is divided into ten frequency bands, each of which is associated with a key. Six additional keys and a foot pedal control other aspects of the speech production (e.g., loudness of voice, pitch). In the same manner, intelligible speech can be generated by a speech prosthesis when the keys and the foot pedal are substituted by control variables measured in the brain. In this case, 17 independent variables are needed, making the method highly impractical for current state-of-the-art BMIs.

H. Articulatory Parameters

A subject could, we suggest, learn to generate speech by driving a set of control variables corresponding to vocal-tract articulation. A detailed physiological model of the human vocal tract and a set of control variables, which correspond to the nerves that stimulate the muscles that configure the vocal tract, can be used to generate speech. In order to more efficiently represent the salient features of vocal-tract movements, we can project the high-dimensional control space of the human vocal tract (i.e., the vocal tract shape) into a smaller compressed space that is composed of the principal articulators. This is achieved by using an articulatory model. For example, the Maeda articulatory model [18], illustrated in Fig. 3, reduces the high dimensionality of the control variable space to only seven parameters that represent principal articulators (i.e., the articulatory model consolidates a large quantity of muscle actions into a few macro movements). This reduced dimensionality

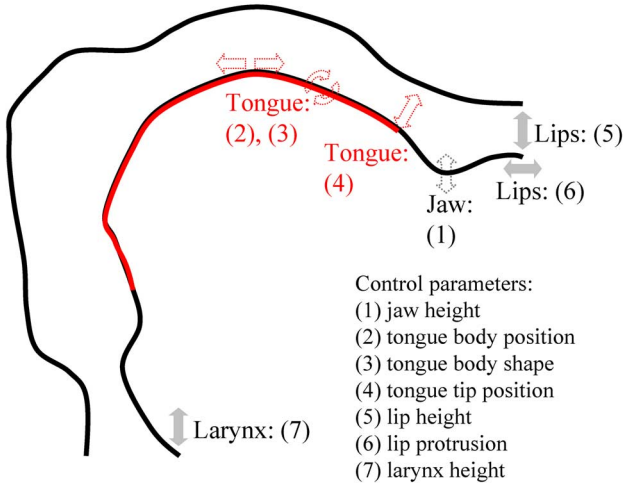


Fig. 3. We encode speech by using an articulatory model driven by seven control parameters.

of control signals is advantageous in a speech prosthesis. The seven parameters, also illustrated in Fig. 3, are: 1) jaw height, which controls the vertical position of the jaw; 2) tongue body position, which moves the tongue dorsum from the front to the back of the oral cavity; 3) tongue body shape, which indicates whether the tongue dorsum is rounded or unrounded; 4) tongue tip position, which controls the position of the tongue apex; 5) lip height, which varies the mouth opening; 6) lip protrusion, which controls the mouth protrusion; and 7) larynx height, which raises or lowers the position of the larynx.

In addition, the use of a speech-production model simplifies control of the speech because the complexity of the speech signals is subsumed under the model. Only the driving parameters of a speech-production model need to be decoded by the BMI: the spectral details are automatically generated by the speech-production model, and can be adapted to the parameters of a specific talker. The model intrinsically contains the following information that does not need to be decoded by the BMI:

- 1) vocal tract dimensions (e.g., overall length and cross-sectional areas along the entire vocal tract);
- 2) velum position;
- 3) subglottal resonances [19];
- 4) nasal resonances [19];
- 5) glottal characteristics (e.g., the parameterization of the output frequency spectrum of the glottal waveform and the pitch dynamic range);
- 6) rigidity of the vocal tract walls [19];
- 7) characteristic profiles of the vocal tract.

A vocal-tract model, controlled with articulatory parameters as depicted in Fig. 3, constrains the synthesized sounds to all and only the speech signals of a specific talker using efficient representation. This approach is similar to music-compression techniques where only musical gestures of the player are encoded and the sounds are reproduced by a model of the musical instrument. Here, a vocal-tract synthesizer, including an articulatory model, is needed to produce speech. For example, we

can drive the synthesizer using seven articulatory control variables with five states each (to control the Maeda model), one fundamental-frequency control variable with five states (to control the speech pitch contour), one noise amplitude-control variable with five states (to control the amplitude of an unvoiced noise source), and one glottal-amplitude control variable with five states (to control the amplitude of a voiced source).

Moreover, using a vocal-tract model to produce speech, instead of a formant synthesizer, an LPC synthesizer, or a Voder, provides advantages beyond the reduction of information previously mentioned. For example, the speech formants (including the higher order ones), which contain the most important characteristics of the speech, are automatically generated by vocal-tract resonances, leading to relatively high quality and natural sounds. Another extremely important property is that the articulatory parameters are linearly interpolable [20]; as a consequence, the transition between speech sounds does not require frequent abrupt transitions in the articulatory control signals, which are necessary with other synthesizers and are not realistically achievable with current state-of-the-art BMIs. In addition, linearly interpolable control variables make the learning process easy because from phoneme to phoneme, the control-variable sequence is just a linear interpolation. This is not true in all representations. For example, with LPC synthesis, each transition has to be carefully learned because a simple interpolation can generate unrealistic sounds and even unstable filters. Learning to control the vocal tract with articulatory parameters mimics how infants learn to speak. A “babbling” phase, where the user randomly tries to generate all possible sounds, is necessary. Subsequently, all of the learned “babbles” are easily connected into words.

Compared with LPC synthesis, where the control parameters are known to be very sensitive to small perturbations, speech synthesis based on articulatory parameters is more robust to noise. In addition, fast changes in the speech spectrum can be obtained from multiple slow changes of the control variables (distributed control)—an important feature when relying on signals acquired by a BMI.

III. ANALOG BIONIC VOCAL TRACT

Fig. 4 shows our circuit model of the human vocal tract: it is represented as a spatially varying acoustic tube using a transmission-line model that comprises a cascade of tunable two-port elements, corresponding to a concatenation of short cylindrical acoustic tubes (each with a 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: 1) a periodic source at the glottis and/or 2) a turbulent noise source P_{turb} at some point along the tube. In Fig. 4, the glottal source is represented by a voltage source P_{alv} connected in series with a variable source impedance Z_{GC} that is modulated by a glottal oscillator. We use a circuit model of the glottis that comprises a linear and

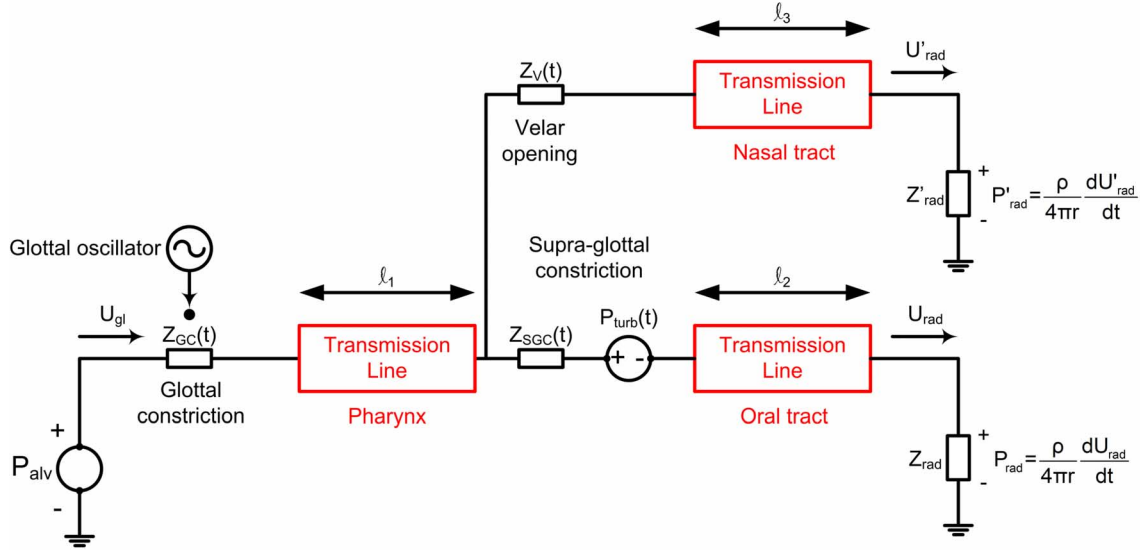


Fig. 4. Schematic of the transmission-line vocal tract.

nonlinear resistance 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. A combination of linear and nonlinear resistances is used to approximate the losses associated with laminar or turbulent flow occurring at the glottal and supraglottal constrictions in the vocal tract. An integrated-circuit implementation of these electronically tunable MOS resistors is described in [21]. These electrical circuit models are consistent with the equations in [19].

The transmission lines corresponding to the oral and nasal tracts are terminated at the lips and nose by radiation impedances Z_{rad} and Z'_{rad} . Radiated sound pressures P_{rad} and P'_{rad} are proportional to the time derivative of the currents flowing in Z_{rad} and Z'_{rad} , respectively. The IC vocal tract is described in detail in [8]. In its first instantiation, for simplicity, the electronic vocal tract only implements the pharyngeal and oral tracts. Our IC vocal tract has been fabricated in a $1.5\text{-}\mu\text{m}$ AMI complementary metal-oxide semiconductor (CMOS) process. It is composed of a cascade of 16 tunable two-port π sections, each representing a uniform tube of adjustable length. The chip consumes less than $275\ \mu\text{W}$ of power when operating with a 5-V power supply.

IV. APPLICATIONS IN SPEECH AND HEARING PROSTHESIS

The SLL analyzes recordings of speech and extracts what we call an articuogram, which is a 3-D plot of the articulatory parameter trajectories as a function of time [8]. Since the SLL is based on a physiological model of the human vocal tract, it inherently synthesizes all and only speech signals. Consequently, it has the ability to restore speech that has been corrupted in noise. These signal-restorative properties are particularly important for the hearing impaired who often have difficulty understanding speech in noisy environments. We have previously proposed a companding algorithm [3] that, functioning as a pre-processor, improves the recognition of noisy vowels, conso-

nants, and sentences with cochlear-implant (CI) subjects [23], [24]. The companding algorithm emphasizes the formants after they are partially corrupted by noise using a two-tone-suppression strategy. Here, we further improve the noise reduction by completely resynthesizing the formants by using a model of the cochlea and the vocal tract in a feedback configuration instead of simply emphasizing the formants.

Fig. 5(a) shows the spectrogram of a recording of the word “hid” produced by a male speaker. Fig. 5(b) shows the same recording degraded in -2-dB SNR white noise. Fig. 5(c) shows the spectrogram of the word “hid” re-synthesized by our SLL from the noise-degraded version of Fig. 5(b). In Fig. 5(c), the property of the SLL to resynthesize even the formants immersed in noise is clearly seen.

Fig. 6 depicts the motor-domain “articuogram”—analogous to the spectrogram in the auditory domain—of a recorded vowel sequence $/a/-e/-i/$ as a vector time series of articulatory parameters. The articuogram is produced by our SLL. With a target sound, the SLL generates a sequence of articulatory parameters similar to what could be extracted by the BMI from the brain activity of the subject. In this manner, the SLL acts as a speech-prosthesis simulator (SPS), which is a useful tool for preliminary speech-prosthesis experiments when reliable BMI data are not yet available. The extracted articuogram is used to derive the vocal-tract profile that drives the ICt vocal tract to produce speech. Articulatory parameters 1–7 in this figure are the same used in Fig. 3 and defined in the text.

Fig. 7 shows the “vocalogram,” a 3-D plot of the vocal tract profile as a function of time, of the same vowel sequence $/a/-e/-i/$. Large cross-sectional areas are shown in red and small cross-sectional areas are shown in blue. The spectrogram of the synthesized speech obtained by using the articuogram of Fig. 6 is shown in Fig. 8. The regions in red indicate the presence of high-intensity frequency components, whereas the regions in blue indicate low intensity. The articuogram and vocalogram show very clearly the tongue and mouth movements during the production of the vowel sequence: at

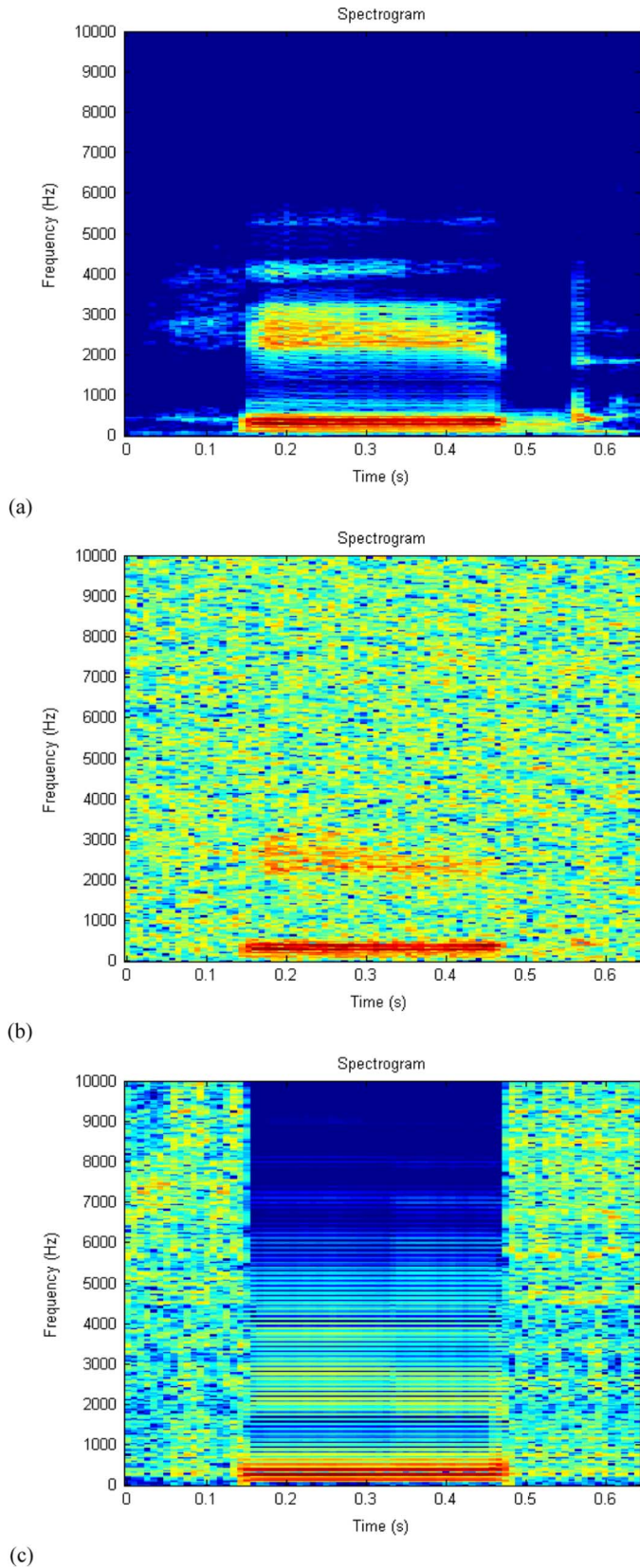


Fig. 5. (a) Spectrogram of a recording of the word “hid.” (b) Spectrogram of the same recording degraded in -2 -dB SNR white noise. (c) Spectrogram of the word “hid” re-synthesized by our SLL from the noise-degraded version.

the beginning of the sequence, the jaw is in a low position and it is raised during the sequence; simultaneously, the tongue

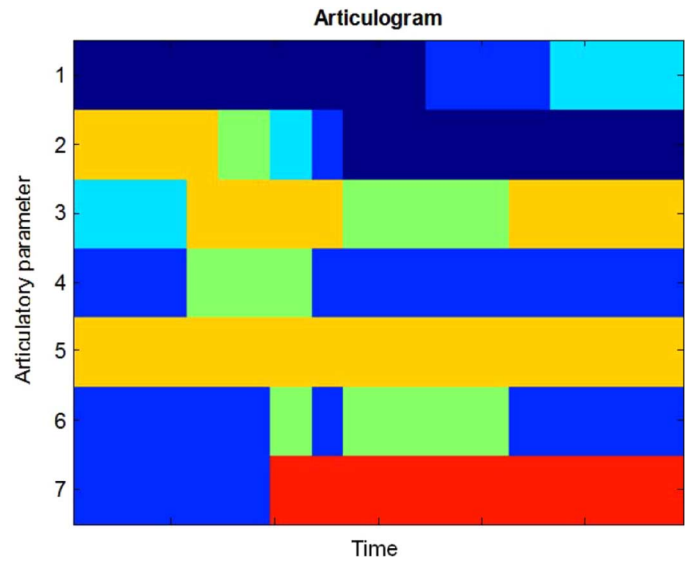


Fig. 6. Articulogram derived by the SLL for the vowel sequence /a/-/e/-/i/.

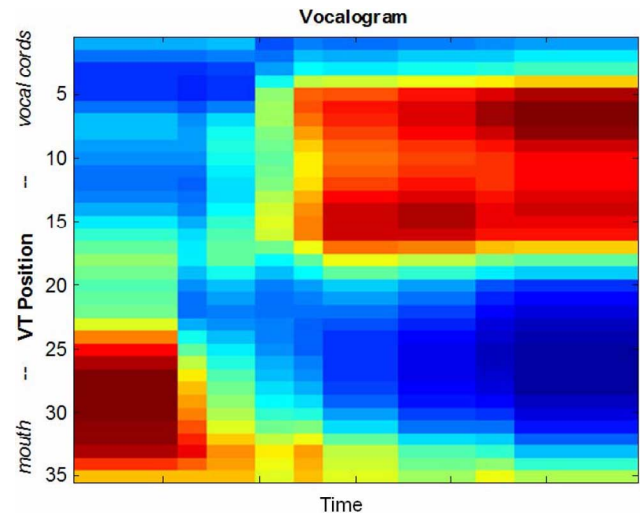


Fig. 7. Vocalogram derived from the articulogram for the vowel sequence /a/-/e/-/i/.

dorsum is moved forward. Hence, the SLL is able to derive the vocal tract movements precisely. The excellent results are also consistent with the synthesized sound, which is practically identical to the original recorded vowel sequence as shown in Figs. 8 and 9.

Fig. 10 shows the articulogram of the word “Massachusetts.” The length of each section of the IC vocal tract was adjusted so that the total length corresponds to a female vocal tract. The synthesized speech waveforms correspond to voltage waveforms obtained at the output P_{rad} of Fig. 4. The spectrogram of the synthesized speech obtained by using the articulogram of Fig. 10 is shown in Fig. 11. Fig. 12 depicts the spectrogram of the recording of a female voice saying the word “Massachusetts” used to extract the articulogram of Fig. 10 by the SPS. The original recording was low-pass filtered at 5.5 kHz. Comparing the spectrograms shown in Figs. 11 and 12, it is evident that the principal formants and the trajectories are well matched. It is also evident that high-frequency speech components that were

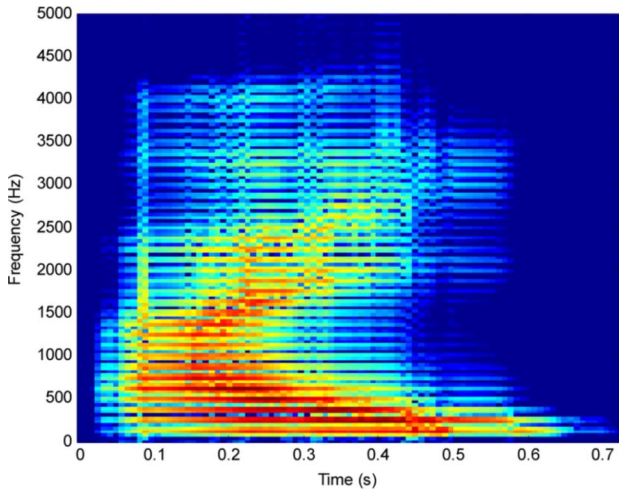


Fig. 8. Spectrogram of the vowel sequence /a/-/e/-/i/ derived from the synthesized speech.

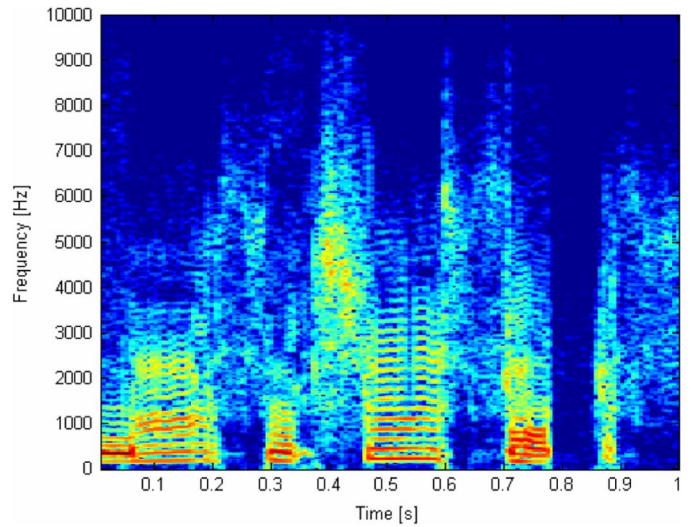


Fig. 11. Spectrogram of the word “Massachusetts” synthesized by the IC vocal tract.

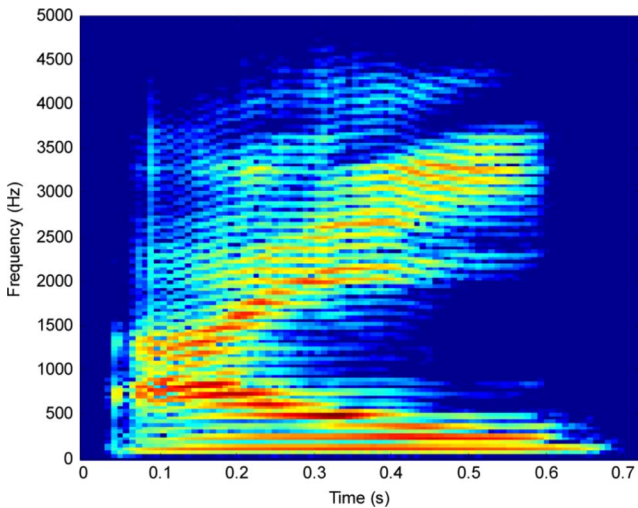


Fig. 9. Spectrogram of the vowel sequence /a/-/e/-/i/ derived from the original speech recording.

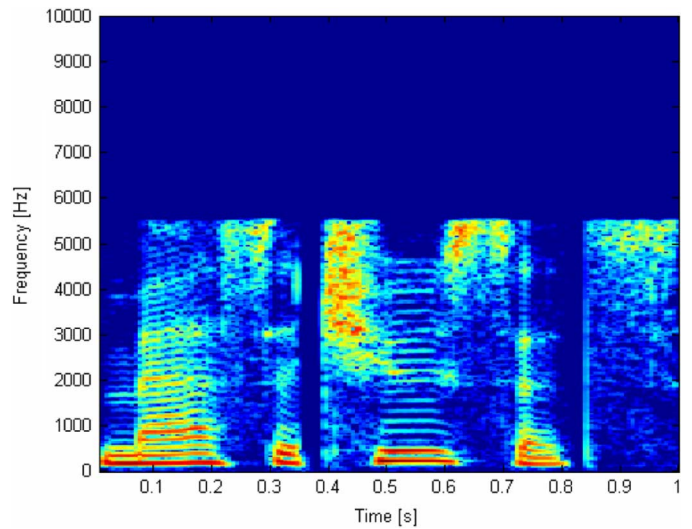


Fig. 12. Spectrogram of a recording of the word “Massachusetts” low-pass filtered at 5.5 kHz.

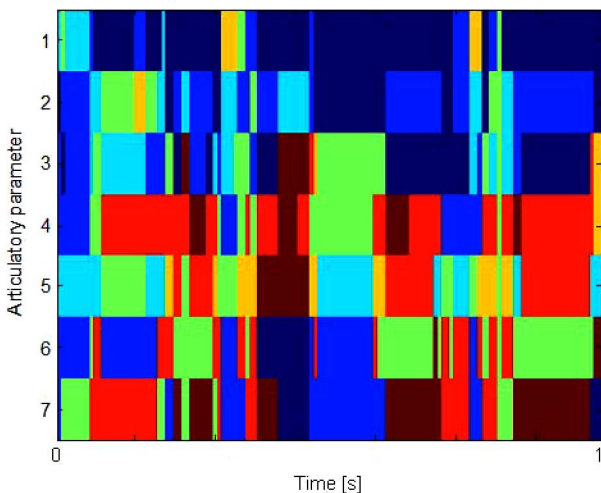


Fig. 10. Articulogram of the word “Massachusetts.”

missing in Fig. 12 were restored by our articulatory silicon vocal tract in Fig. 11. This effect is attributed to the inherent property

of our articulatory silicon vocal tract to synthesize all and only speech signals and, thus, provide high-quality speech even with a compressed articulatory parameter representation.

V. CONCLUSION

We discussed an articulatory silicon vocal tract and its application to hearing and speech prostheses. In particular, we discussed how to exploit the signal-restorative property of the speech-locked loop for improving hearing in noise for cochlear implants and examined the pros and cons of various speech-coding strategies in the context of brain-machine speech prostheses. Our analysis shows that in the context of a BMI-enabled speech prosthesis, coding speech with articulatory parameters is the most efficient and robust strategy. We also described an IC vocal tract that can be used for articulatory synthesis. The IC vocal tract consumes only 275 μ W and can be integrated into an implantable/wearable prosthetic system. We demonstrated the functionality of our system by using a speech-prosthesis simulator, which is able to produce articulograms from audio

recordings. We discussed examples of speech synthesized by our prototype.

REFERENCES

- [1] R. Sarpeshkar, *Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems*. Cambridge, U.K.: Cambridge Univ. Press, Feb. 2010.
- [2] L. Turicchia, B. Do Valle, J. Bohorquez, W. Sanchez, V. Misra, L. Fay, M. Tavakoli, and R. Sarpeshkar, "Ultra low power electronics for cardiac monitoring," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 9, pp. 2279–2290, Sep. 2010.
- [3] L. Turicchia and R. Sarpeshkar, "A bio-inspired companding strategy for spectral enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 243–253, Mar. 2005.
- [4] L. Turicchia and R. Sarpeshkar, "The silicon cochlea: From biology to bionics," in *The Biophysics of the Cochlea: Molecules to Models*. Singapore: World Scientific, 2003.
- [5] A. Oxenham, A. Simonson, L. Turicchia, and R. Sarpeshkar, "Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing," *J. Acoust. Soc. Amer.*, vol. 121, no. 3, pp. 1709–1716, 2007.
- [6] 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, Music Process.*, vol. 2007, p. 13, 2007.
- [7] S. Mandal, S. Zhak, and R. Sarpeshkar, "A bio-inspired active radio-frequency silicon cochlea," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, pp. 1814–1828, Jun. 2009.
- [8] K. H. Wee, L. Turicchia, and R. Sarpeshkar, "An analog integrated-circuit vocal tract," *IEEE Trans. Biomed. Circuits Syst.*, vol. 2, no. 4, pp. 316–327, Dec. 2008.
- [9] K. H. Wee, L. Turicchia, and R. Sarpeshkar, "An articulatory speech-prosthesis system," in *Proc. IEEE Int. Conf. Body Sens. Netw.*, Jun. 7–9, 2010, pp. 133–138.
- [10] 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.*, vol. 52, no. 4, pp. 711–727, Apr. 2005.
- [11] N. J. Hill, T. N. Lal, M. Schröder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Schölkopf, A. Kübler, and N. Birbaumer, "Classifying EEG and ECoG signals without subject training for fast BCI implementation: Comparison of nonparalyzed and completely paralyzed subjects," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 14, no. 2, pp. 183–186, Jun. 2006.
- [12] N. Weiskopf, K. Mathiak, S. W. Bock, F. Scharnowski, R. Veit, W. Grodd, R. Goebel, and N. Birbaumer, "Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI)," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 966–970, Jun. 2004.
- [13] P. R. Kennedy and R. A. E. Bakay, "Restoration of neural output from a paralyzed patient by a direct brain connection," *Neuroreport*, vol. 9, pp. 1707–1711, 1998.
- [14] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz, "Cortical control of a prosthetic arm for self-feeding," *Nature*, vol. 453, Jun. 19, 2008.
- [15] S. T. Cl, Z. Zohny, M. Velliste, and A. B. Schwartz, "Simultaneous 7-dimensional cortical control of an arm and hand robot via direct brain interface," in *Program No. 494.6/GGG18, 2010 Neuroscience Meeting Planner*, San Diego, CA, 2010.
- [16] D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, no. 3, pp. 971–995, 1980.
- [17] H. Dudley, R. R. Riesz, and S. A. Watkins, "A synthetic speaker," *J. Franklin Inst.*, vol. 227, pp. 739–764, 1939.
- [18] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Prod. Speech Model.*, W. J. Hardcastle and A. Marchal, Eds. Boston, MA: Kluwer, 1990, pp. 131–149.
- [19] K. N. Stevens, *Acoust. Phon.*. Cambridge, MA: The MIT Press, 1998.
- [20] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-35, no. 7, pp. 955–967, Jul. 1987.
- [21] K. H. Wee and R. Sarpeshkar, "An electronically tunable linear or non-linear MOS resistor," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 9, pp. 2573–2583, Oct. 2008.
- [22] F. H. Guenther *et al.*, "A wireless brain-machine interface for real-time speech synthesis," *PLoS ONE*, vol. 4, no. 12, p. e8218, 2009.
- [23] A. Bhattacharya and F.-G. Zeng, "Companding to improve cochlear-implant speech recognition in speech-shaped noise," *J. Acoust. Soc. Amer.*, vol. 122, no. 2, pp. 1079–1089, 2007, Aparajita Bhattacharya and Fan-Gang Zeng.
- [24] P. Loizou, K. Kasturi, L. Turicchia, R. Sarpeshkar, M. Dorman, and T. Spahr, "Evaluation of the companding and other strategies for noise reduction in cochlear implants," presented at the Conf. Implantable Auditory Prostheses, Pacific Grove, CA, 2005.



Keng Hoong Wee (M'08) received the B.S. and M.S. degrees in electrical engineering from Tohoku University, Sendai, Japan, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge.

His work has included circuit modeling of biology, biologically inspired very-large scale integrated systems, low-power integrated circuits for biomedical devices, such as speech and hearing prostheses, and noise robust techniques for audio processing. His research interests include bioelectronics and bioinspired signal processing.



Lorenzo Turicchia is a Research Scientist in the Research Laboratory of Electronics at the Massachusetts Institute of Technology (MIT), Cambridge. He received the Laurea degree in electrical engineering from the University of Padova, Padova, Italy, and the Ph.D. degree in computer science in mathematics and computer science from the University of Udine, Udine, Italy.

In 2002, he joined the Analog VLSI and Biological Systems Group, Massachusetts Institute of Technology, where he completed his doctoral research and is now a Research Scientist. His main research interests are in nonlinear signal processing, especially for audio and biomedical applications, and bioelectronics. His work has included research on cochlear implants for the hearing impaired, visual prostheses for the blind, speech prostheses for individuals with severe communication disabilities, automatic speech recognition in noise, and wearable medical devices. In these areas, he has authored seven patent applications and more than 40 publications. Currently, he is working on robust techniques for the recognition of speech, speaker, and language in noisy environments; bioelectronics for wearable and implantable medical devices; and neural decoding techniques for neural prosthetic devices for the paralyzed. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE (T-ITB) and serves on the program committees of several technical conferences.



Rahul Sarpeshkar (SM'07) received the Bachelor's degrees in electrical engineering and physics from the Massachusetts Institute of Technology (MIT), Cambridge, and the Ph.D. degree from the California Institute of Technology, Pasadena.

After receiving the Ph.D. degree, he joined Bell Labs as a member of the technical staff in the Department of Biological Computation within its Physics division. 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. He holds more than 25 patents and has authored more than 100 publications including one featured on the cover of *Nature*. He has authored the recent text *Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications and Bio-inspired Systems*, which provides a broad and deep treatment of the fields of low-power electronics and bioelectronics. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS and serves on the program committees of several technical conferences.

Dr. Sarpeshkar has received several awards, including the National Science Foundation Career Award, the ONR Young Investigator Award, the Packard Fellows Award, and the Indus Technovator Award for his interdisciplinary bioengineering research.