An Analog VLSI Vocal Tract

by

Keng Hoong Wee

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

February 2008

Signature of Author... .....

Department of Electrical Engineering and Computer Science

Jan 15, 2008

Certified by..... .....

Rahul Sarpeshkar

Associate Professor of Electrical Engineering and Computer Science

Thesis Supervisor

Accepted by.....

Terry P. Orlando

Chairman, Committee on Graduate Students

Department of Electrical Engineering and Computer Science

An Analog VLSI Vocal Tract

by

Keng Hoong Wee

Submitted to the Department of Electrical Engineering and Computer Science
on Jan 15 2008, in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

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. A promising approach to improve the naturalness of synthetic speech is to exploit bio-inspired models of speech production with low bit-rate control parameters. In this work, 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 275μW analog vocal tract chip can be used with auditory processors in a feedback *speech 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. Previous attempts that take advantage of the powerful analysis-by-synthesis method employed computationally expensive approaches to articulatory synthesis using digital computation. 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 bionic speech-prosthesis systems.

Thesis Supervisor: Rahul Sarpeshkar

Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgements

An endeavour of this magnitude could not have been accomplished without the help and support of numerous people—my advisor, my Ph.D thesis committee, colleagues, friends, and family—along the way. I would like to acknowledge these people.

I would like to start by acknowledging my thesis committee. I am grateful to my advisor, Professor Rahul Sarpeshkar, for taking me into the Analog VLSI and Biological Systems group at MIT. Rahul is a constant source of inspiration. I have benefited tremendously from his vision and his unrelenting pursuit of truth and real insight. Rahul also instilled in me an attitude to take heart from setbacks and to move on with great enthusiasm. Without doubt, I have acquired knowledge and invaluable research skills from Rahul and I am grateful to him for teaching me to seek out scientific insight.

I am also grateful to Professor Kenneth Stevens and Professor Joel Dawson for serving on my Ph.D thesis committee, and for the numerous discussions and helpful comments. I would also like to thank Joel for being a member of my Ph.D qualification exam committee.

I acknowledge the friends and colleagues, past and present, in the Analog VLSI and Biological Systems group. I am grateful to Lorenzo Turicchia for his invaluable feedback on the thesis. I also benefited greatly from his discerning viewpoints and from our conversations and discussions. I would like to thank Soumyajit Mandal for countless discussions, for spending time to listen and bounce ideas, and for his many helpful suggestions. I am also grateful to Micah O'Halloran and Ji Jon Sit for patiently hearing me out and allowing me to pick their brains on numerous occasions. I thank Michael Baker, Serhii Zhak, Chris Salthouse, Scott Arfin, Ben Rapoport, Woradorn Wattanapanitch, Maziar Tavakoli, Heemin Yang, Tim Lu, and Daniel Kumar for the intellectual and emotional support that I have enjoyed over the years. They have been a major source of inspiration, ideas and help. I have learnt a lot from them and count myself most fortunate to have had the opportunity to work with such a talented group of individuals.

I acknowledge DSO National Laboratories for the opportunity to pursue doctoral research through a study scholarship.

I acknowledge my family who has been most supportive throughout the years. I am grateful to my relatives for pitching in when called upon. Special thanks to my sister, who stepped in for me so willingly and unquestioningly, while I was away from home. To my parents, my gratitude is best portrayed by two verses taken from a beautiful piece of Chinese poetry by the Tang poet Meng Jiao:

誰言寸草心　How could a blade of grass

報得三春暉　Repay the warmth it received from the spring sun

Last but definitely not least, I thank my wife Mei Lin; for her unconditional love and devotion, countless sacrifices and her infinite reserves of patience and understanding. If I may use a circuit analogy, and map her emotional support to a current, and pressure to a voltage, then she is without a doubt an ideal current source. My deepest gratitude to my beloved Mei Lin!

TABLE OF CONTENTS

Acknowledgements 5
Chapter 1 INTRODUCTION 9
1.1 Why a vocal tract: The speech problem 9
1.2 Overview of the human speech production system 11
1.2.1 The subglottal system 13
1.2.2 The supraglottal vocal tract 14
1.2.3 The glottis 17
1.3 Overview of speech synthesis 21
1.3.1 Concatenation synthesis 21
1.3.2 Formant synthesis 23
1.3.3 Linear prediction coding based synthesis 24
1.3.4 Sinusoidal synthesis 25
1.3.5 Articulatory synthesis 26
Chapter 2 ELECTRICAL MODEL OF SPEECH PRODUCTION SYSTEM 29
2.1 Overview of circuit model 29
2.2 Model of the glottis 31
2.2.1 Current source model 31
2.2.2 Variable impedance model 33
2.3 Model of the supraglottal vocal tract 37
2.3.1 Pharyngeal and oral cavities 37
2.3.2 Model of the supraglottal constriction 39
2.3.3 Model of the nasal cavity 42
2.4 Model of sound radiation from the lips and nose 42
2.5 Simulation of speech production 45
2.5.1 Speech production with supraglottal vocal tract 45
2.5.2 Simulation results 46
Chapter 3 DRIVING THE ANALOG VOCAL TRACT 53
3.1 Articulatory representation of speech 53
3.2 The Maeda articulatory model 54
3.3 The synthesis process 56
3.4 Building an articulatory codebook through babbling 56
3.5 Articulatory trajectory optimization 58
3.6 Simulation results 60
Chapter 4 LINEAR OR NONLINEAR MOS RESISTORS 77
4.1 Feedforward biasing technique for electronically tunable MOS resistors 77
4.1.1 Transistor with constant V_{GB} 79
4.1.2 MOS resistor with feedforward biasing technique 81
4.2 Feedback biasing technique for electronically tunable linear or nonlinear resistors using MOS transistors 88
4.3 Linear MOS resistor 92
4.3.1 Circuit description 92
4.3.2 DC characteristics 95

4.3.3 AC characteristics 100
4.3.4 Temperature characteristics 101
4.3.5 Noise analysis and measurements 102
4.4 Nonlinear MOS resistor 109
4.4.1 Circuit description 109
4.4.2 Experimental results 109
Chapter 5 ELECTRONICALLY TUNABLE TWO-PORT π-SECTION 113
5.1 VLSI inductor 113
5.1.1 OTA based second order filter structures 113
5.1.2 OTA based gyrator 116
5.1.3 Operational amplifier based gyrator 118
5.2 VLSI two port equivalent of LC π-section 120
5.2.1 Two port representation of LC π-section 120
5.2.2 Continuously tunable LC π-section 123
5.2.3 Discretely tunable LC π-section 129
5.2.4 The supraglottal vocal tract as a cascade of two-port equivalent π-sections 130
Chapter 6 VLSI IMPLEMETATION OF VOCAL TRACT 133
6.1 Transmission line vocal tract 133
6.2 Subglottal system 136
6.2.1 Current source circuit model of the glottis 136
6.2.2 Nonlinear impedance circuit model of the glottis 137
6.3 Approximate methods for consonant production 138
6.3.1 "Input refer" noise source to glottis 138
6.3.2 Impedance modulation method 143
Chapter 7 CONCLUSIONS 155
7.1 Contributions and accomplishments 155
7.2 Future directions 157
7.2.1 Speech codec 157
7.2.2 Speech recognition via speech locked loop 159
7.2.3 Research tool for speech production 163
BIBLIOGRAPHY 165

Chapter 1 INTRODUCTION

1.1 Why a vocal tract: The speech problem

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. An emerging 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 achieves low bit-rate transmission.

In order to analyze the speech signal and extract its characteristic features, a three dimensional intensity-frequency-time representation known as the spectrogram is commonly used. Spectrogram analysis is a powerful method of speech analysis as it captures and highlights essential features of the speech signal such as frequency content and formant transitions. In fact, the cochlea (inner ear) of the human auditory system performs a similar operation. Input sound waves, converted into mechanical vibrations by the outer and middle ear, impinge on the oval window at the entrance of the cochlea causing the basilar membrane to vibrate at locations as well as at frequencies that reflect the frequency characteristics of the input acoustic wave (e.g., speech formants). Inner hair cells (IHC), distributed along the length of the basilar membrane, sense these vibrations and act as mechanical-to-neural transducers. Vibrations at some point along the basilar membrane activate nerve fibres that innervate the bottom of each IHC at those locations. In this way, the firing activity of the IHC along the membrane provide an indication of the frequency content and formant trajectory of the input acoustic signal.

An important technique employed in speech coding and decoding applications involves analysis-by-synthesis [1]: The speech is analyzed by extracting parameters from it that are used to configure a speech synthesizer to reproduce the speech. More specifically, it is a method of determining the control parameters of speech production—that reproduce desired speech features—in which the consequence of choosing a

particular set of values is evaluated by analyzing the synthesized speech and comparing it to the original input speech signal. The analysis-by-synthesis paradigm employs an active analysis process that is applied to speech synthesized by a generator. The heart of such a system is a speech apparatus that is capable of generating all and only the speech signals of interest.

Fig. 1-1(a) 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 communication systems. The auditory model and controller are analogous to a phase detector and loop filter in a PLL and the vocal tract model is analogous to a voltage-controlled-oscillator (VCO). Fig. 1-1(b) shows a particular embodiment of the SLL employing an analog vocal tract (AVT) and an analog bionic ear processor [2][3] in a feedback configuration.

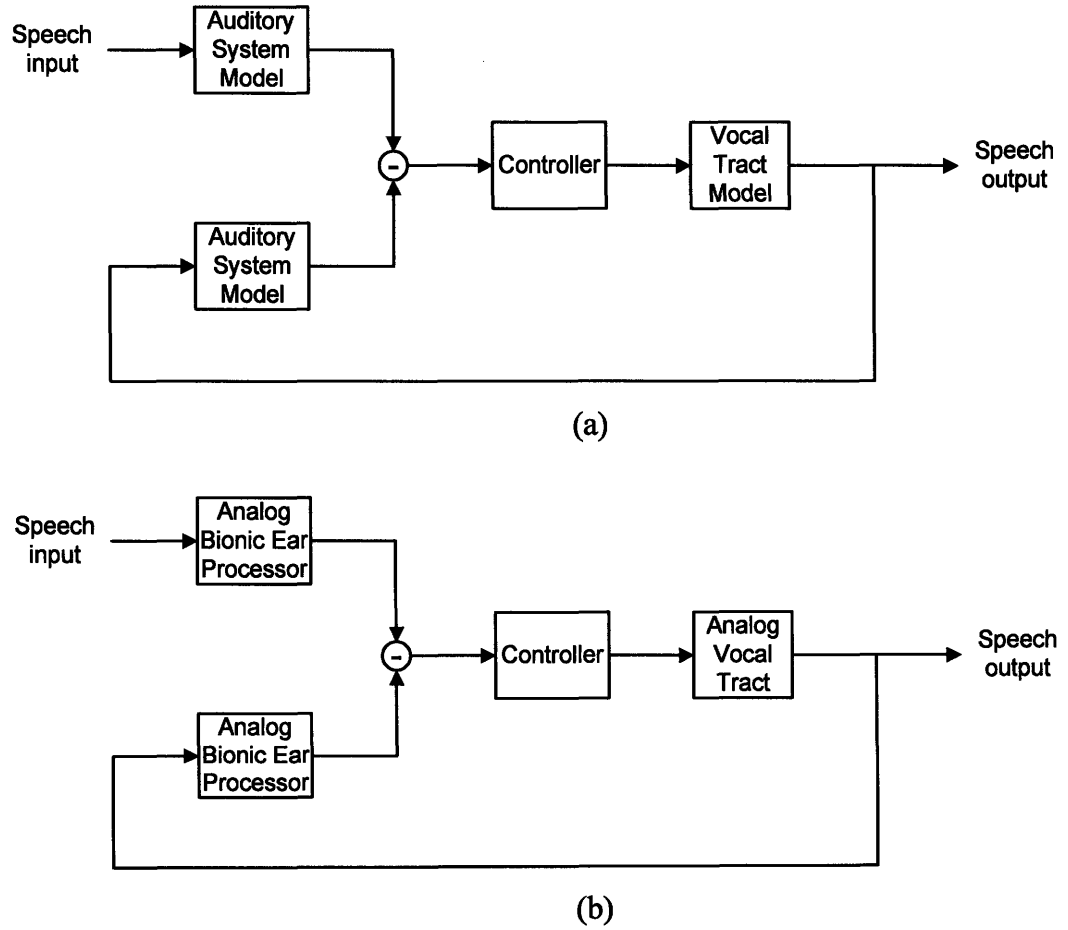


Fig. 1-1: (a) General concept of the speech-locked loop and (b) a particular embodiment.

The speech produced by the AVT is analyzed and compared to that of the input, and a measure of the error is computed. The error is derived from the acoustic difference between the vocal tract output and the target speech, e.g., the L2-norm of the difference in the respective mel-frequency cepstral coefficients, taking into account the articulatory dynamics. Using gradient descent techniques with pre-selected initial conditions that ensure global minimum convergence, 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 optimal vocal tract profile produced by the controller.

In portable systems of moderate complexity, the use of analog processing to reduce power appears to be an emerging technology direction [4]. In particular, circuit models of biology are increasingly being used to improve performance in engineering systems. For example, silicon-cochlea-like models have led to improved speech recognition in noise [5] and low-power cochlear-implant processors for the deaf [2][3]. Silicon models of the retina [6] have been used in machine vision systems and circuit models of the heart have been used to shed insight into cardiac and circulatory malfunction in medicine. In this thesis, we develop an experimental integrated-circuit analog 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. Such silicon vocal tracts 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, and/or find applications in real-time low-power speech production, compression, speaker identification or bionic speech-prosthesis systems.

1.2 Overview of the human speech production system

The human speech production system is illustrated in Fig. 1-2 [7]. It may be broadly classified into a subglottal system of airways, including the lungs and the trachea, extending below the larynx and a supraglottal vocal tract, above the larynx, comprising the pharynx, oral cavity and nasal cavity. The subglottal and supraglottal systems are separated by the glottis, an aperture created within the vocal folds when they open and close. The supraglottal vocal tract extends from the glottis in the throat to the lips in the oral cavity and the nostrils in the nasal cavity. Articulators, namely the soft palate (or

velum), tongue, jaw and lips modify the shape of the vocal tract. The shape in turn determines the transfer function of the vocal tract in response to an excitation. The transfer function is specified by several poles (or formants) and sometimes zeros (or anti-formants). Air pressure produced by the lungs forces air through the vocal folds that when under tension vibrate and produce an airflow which excites the resonances in the vocal and nasal cavities. In general, the excitation can be due to vocal fold vibration at a fundamental frequency, as in the case of voiced speech, and/or from turbulent noise generated at constrictions along the vocal tract, as in the production of unvoiced speech. During speech production, the configuration of the vocal tract varies in a dynamic manner. The source of excitation may also change. It is believed that the state of the articulators is the result of a constant movement towards a sequence of changing targets defined by the phonemes of a given language. Phonemes are basic units of speech produced by a group of vocal tract configurations which are considered to be functionally equivalent in a given language. Two vocal tract configurations represent different phonemes if two words can be found which differ only by the use of these two configurations. Fig. 2-16 shows the classification of phonemes used in standard English.

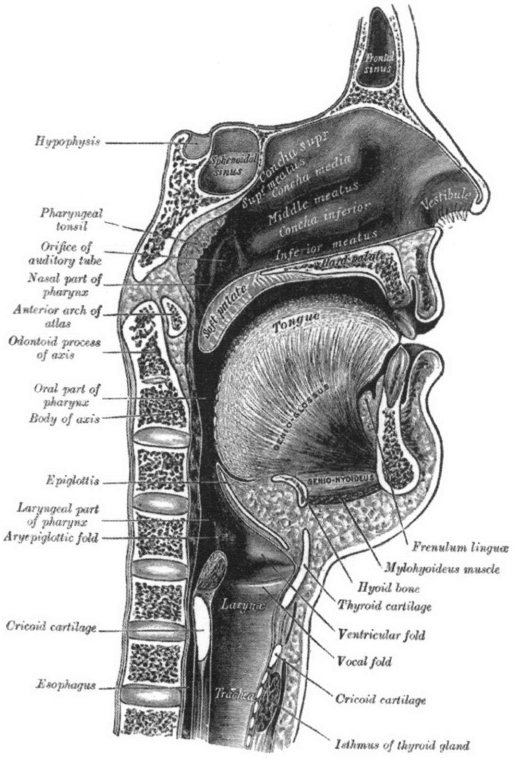
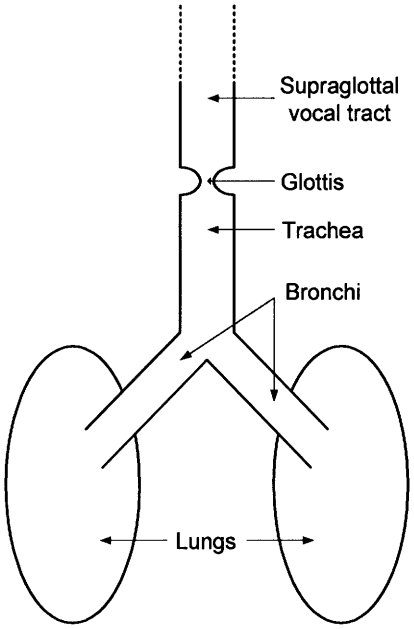


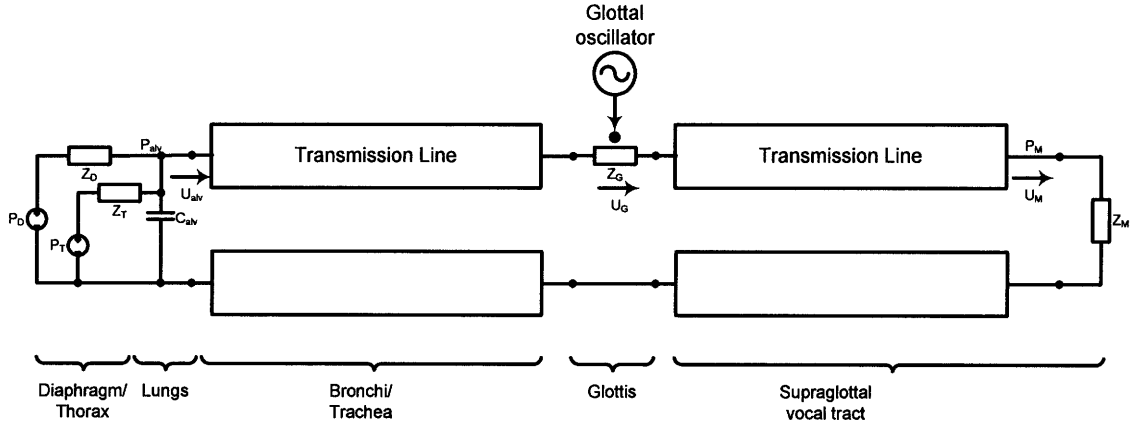
Fig. 1-2: Illustration of the human speech production system (public domain, adapted from [7]).

1.2.1 The subglottal system

The subglottal system is the power source for speech production: it supplies the energy to move air through the trachea, glottis and the supraglottal vocal tract. Fig. 1-3 is a schematic representation of the subglottal system and its equivalent circuit model [8][9].



(a)



(b)

Fig. 1-3: (a) Schematic diagram of subglottal (and also in part the supraglottal) system (adapted from [9]) and (b) its equivalent circuit model.

In this thesis, we adopt the effort-to-voltage (e→V) convention to transform the conjugate power variables of pressure (effort) and volume velocity (flow) in the acoustic energy domain to voltage and current, respectively, in the electric energy domain. Consequently, the lungs are represented by an acoustic compliance C_{alv} and the airways leading to it are represented as a distributed transmission line in Fig. 1-3(b). Changes in airflow, represented by the current U_{alv}, are effected through changes in lung volume, represented by charge stored on C_{alv}. The latter is controlled by a pressure source P_D that raises and lowers the diaphragm and another pressure source P_T that expands and contracts the thorax cavity. The diaphragm and thorax have impedances Z_D and Z_T respectively. The impedance Z_D is comprised of a mass M_D, a compliance C_D, and damping resistance R_D connected in series. Similarly, Z_T consists of a thoracic mass, compliance and damping resistance, M_T, C_T R_T respectively. During inspiration, air is drawn into the lungs: at the end of inspiration, an initial condition of P_{alv}=8 cmH_2O=8000 dynes/cm^2 is set on the alveolar compliance C_{alv} by P_T and P_D. During speech production, the air is expelled through the glottis, via the bronchi and trachea, and passes into the supraglottal vocal tract, which is modeled by another transmission line in Fig. 1-3(b). As air is expelled, C_{alv} is charged through P_D and P_T (through contraction of the thoracic cavity and raising of the diaphragm), thereby maintaining a relatively constant alveolar pressure. During speech production, the supraglottal transmission line varies with the configuration of the supraglottal vocal tract and the pressure in the lungs P_{alv} varies about a nominal value of 8000 dynes/cm^2 [8]. Consequently, the subglottal source may be approximated as a constant or slowly varying pressure source P_{alv}.

1.2.2 The supraglottal vocal tract

Fig. 1-4 [10] illustrates the major regions of the supraglottal vocal tract involved in speech production, namely:

(a) the pharynx

(b) the oral cavity

(c) the nasal cavity

(d) the lips and nostrils

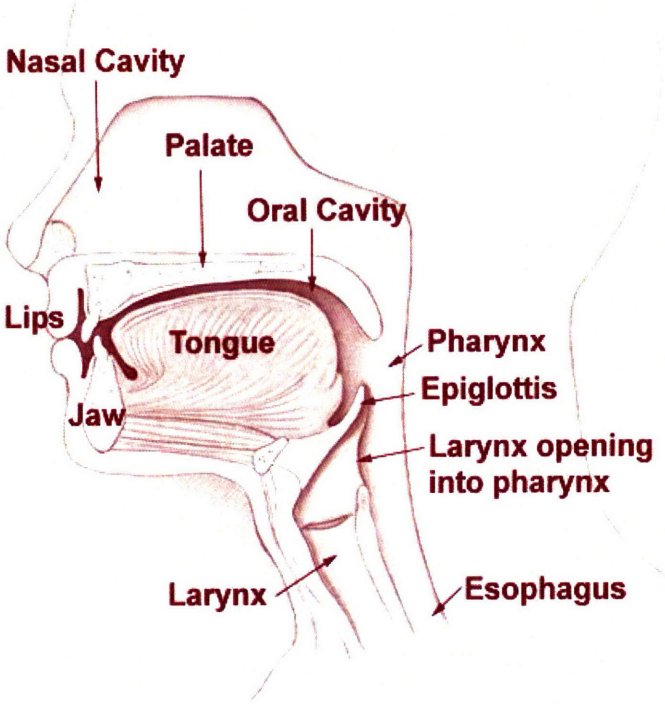


Fig. 1-4: Diagram of the supraglottal vocal tract highlighting the major regions involved in speech production (adapted from [10]).

The physical configuration of the vocal tract is highly variable and is dictated by the positions of the four articulators. The nasal cavity is coupled to the main vocal tract through the velum. 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 well approximated by assuming a circular cross section. The wave equation for planar sound propagation (one dimensional) in a lossless uniform tube of circular cross section can be derived as:

$$-\frac{\partial P}{\partial x} = \frac{\rho}{A}\frac{\partial U}{\partial t} \qquad (1)$$

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

where P is the sound pressure, U is the volume velocity, ρ 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 inertance ρ/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).

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 lossless transmission line can be described by the following coupled partial differential equations:

$$-\frac{\partial V}{\partial x} = L\frac{\partial I}{\partial t} \qquad (2)$$
$$-\frac{\partial I}{\partial x} = C\frac{\partial V}{\partial t}$$

where L and C are the inductance and capacitance per unit length.

The vocal tract is approximately 17.5 cm in length for an average man, which is comparable to the wavelength of sound in air at audible frequencies. Hence, a lumped approximation of the major vocal tract components does not provide an accurate analysis. However, the tube may be discretized in space and the entire tube represented in terms of a concatenation of incremental cylindrical sections. The error introduced by spatial quantization may be kept small if the length, ℓ, of the approximating cylindrical sections are kept short compared to the wavelength of sound corresponding to the maximum frequency of interest.

The electrical analog of a section of a lossy acoustic tube with uniform circular cross sectional area A and length ℓ, is depicted in Fig. 1-5. The series inductance L and the shunt capacitance C represent the discretized acoustic inertance and compliance of the cylindrical section, respectively. The values of L and C are determined by the length ℓ, and cross sectional area A of the section as follows:

$$L = \frac{\rho l}{A} \qquad (3)$$
$$C = \frac{Al}{\rho c^2}$$

Assuming that the flow is laminar, R and G models the energy losses due to viscous friction and heat conduction at the walls respectively. Except for the very smallest of areas ($A<0.01 cm^2$), the series impedance, comprising the sum of R and $j\omega L$, is dominated by the reactive component in the frequencies of interest. In addition to

conductance G, the walls of the vocal tract also have stiffness, mass and damping. These mechanical properties of the vocal tract walls influence sound production and can be modeled as an impedance Z_W in parallel to G, where Z_W is approximated by a compliance C_W, a mass L_W and a resistance R_W connected in series as shown in Fig. 1-5. At low frequencies (100-200Hz), C_W, L_W and R_W, can be assumed to be constant [8].

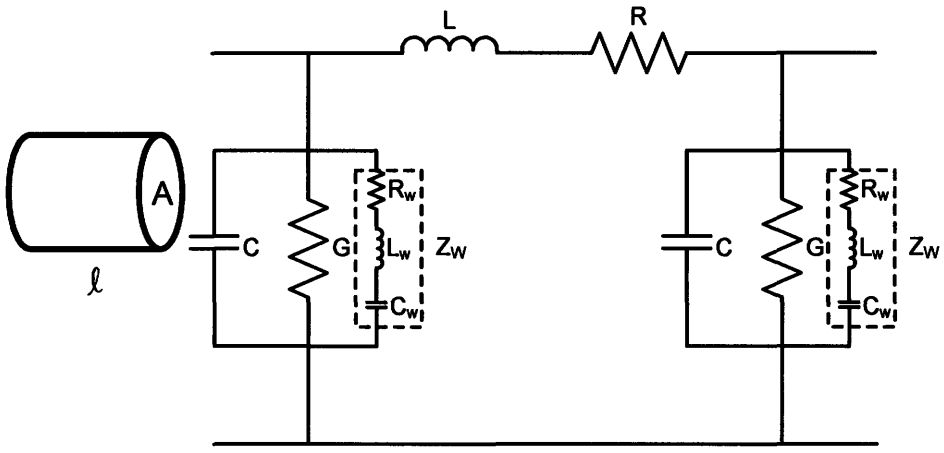


Fig. 1-5: Equivalent π-circuit model of a cylindrical section of acoustic tube with cross sectional area A.

In summary, the following simplifying assumptions were made in the analysis above:

(a) the vocal tract can be straightened out and approximated as a tube with variable cross sectional areas along its length

(b) the wave motion in the tract is planar

(c) the wave equation is linear

Hence, as far as the acoustic properties are concerned, the shape of the vocal tract is completely specified by the cross sectional area of each section. Also, the effects of viscous friction and thermal conduction are accounted for by the appropriate introduction of resistances and conductances.

1.2.3 The glottis

During speech production, the vocal folds vibrate, changing the size of the glottal aperture that connects the subglottal and supraglottal systems. The oscillations of the

vocal folds can be explained in part by a phenomenon known as flow-induced oscillation: a steady stream of air passing by a wall or surface can cause vibrations of that surface. Vocal fold vibrations are the result of aerodynamic forces related to the glottal airflow and mechanical forces associated with the muscular structures. Functionally, during normal phonation, the glottal oscillator can be modeled as shown in Fig. 1-6 where P_{alv} is the alveolar pressure at the lungs, P_S is the subglottal pressure, P_i is the pressure at the glottal end of the supraglottal vocal tract and Z_{VT} is the impedance of the vocal tract load. Z_{SGT} represents the subglottal network of tubes connecting the lungs to the glottis.

In the simplest case, vocal fold vibration can be approximated as simple harmonic motion in response to an aerodynamic force F due to the Bernoulli effect. F acts perpendicularly to the tissue surface of the vocal folds and depends on the mean intra-glottal pressure P over the medial surfaces of the folds. From Bernoulli's energy law, P is approximated by [11]:

$$P \approx \left(1 - \frac{A_{g2}}{A_{g1}}\right)\left(P_S - P_i\right) + P_i \tag{4}$$

where A_{g1} and A_{g2} are the cross sectional areas at the glottal entry and exit respectively.

According to the myoelastic-aerodynamic theory, negative pressure from Bernoulli forces causes the vocal folds to be sucked together, creating a closed airspace below the glottis. Continued air pressure from the lungs builds up underneath the closed

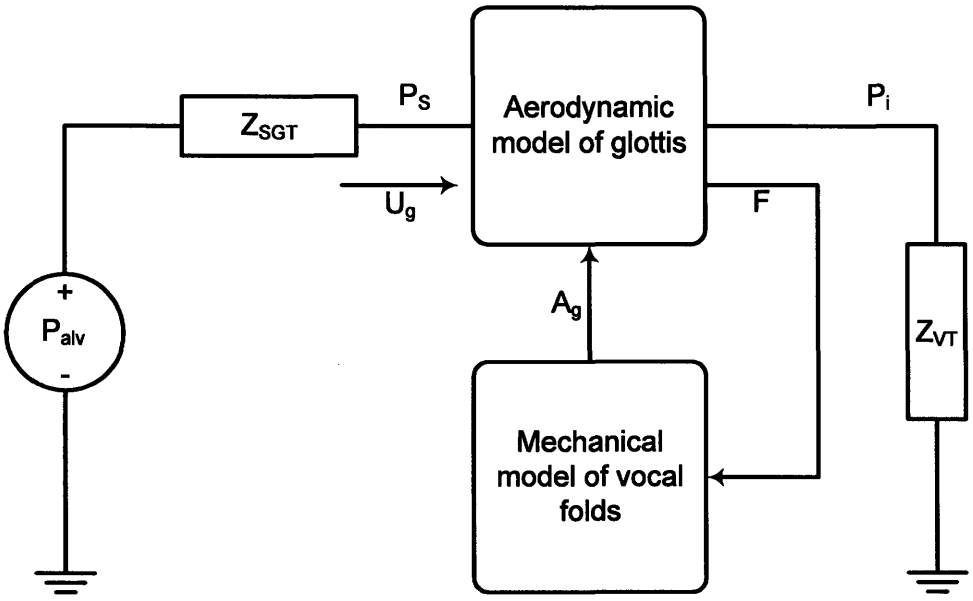


Fig. 1-6: Functional model of glottal oscillator.

folds. Energy is transferred to the tissue by the air and when this pressure becomes sufficiently high, the folds are blown outward, thus opening the glottis and releasing a single pulse of air. The lateral movement of the vocal folds continues until the natural elasticity of the tissue takes over, and the vocal folds move back to their original, closed position. Then, the cycle begins again with each cycle producing a single pulse of air.

It may be shown that Bernoulli forces alone cannot account for continuous energy conversion from air stream to tissue as the oscillations would die out with time. For the vocal folds to sustain oscillation, there must be a negative pressure within the glottis. As pressure from the lungs is always positive, the inertia of the air column in the vocal tract plays an important role in producing negative glottal air pressure [11]. When the glottis is opening and air from the lungs is moving upward, the air column is accelerated by the increasing glottal flow which creates a positive pressure P that drives the folds apart. When the glottis is closing, the airflow begins to decrease, but the air column above the glottis continues to move with the same speed because of inertia. A condition where air is not coming through from the bottom of the glottis as fast as it is leaving above arises, resulting in a suction region of negative pressure just above the vocal folds where the air pressure is reduced. In other words, when the vocal folds are opening, fluid pressure against the walls is greater than when the vocal folds are close together. Oscillation is sustained by the asymmetric driving force [11], giving rise to a two mass model of the glottis, depicted in Fig. 1-7, that incorporates an acoustic tube representing the vocal tract. The force produced by the glottal airflow when it interacts with the acoustic tube of the glottis sustains glottal oscillations.

Experimental observations of the vocal folds show that the bottom of the folds are farther apart than the upper part of the folds at some points in the cycle [8][11]; the glottal path takes on a convergent shape ($A_{g1}<A_{g2}$) with the airflow converging. On the other hand, the airflow diverges when the lowermost parts of the vocal folds are closer together; this is a divergent glottal shape ($A_{g1}>A_{g2}$). In order to model the out of phase motions of the upper and lower parts of the folds, the vocal folds are divided into two masses m1 and m2 [12] coupled by springs. The pressure drop along the glottis depend on the cross-sectional areas A_{g1} and A_{g2}. The changes in these cross-sectional areas are the consequence of the motion of masses m1, m2. The displacement amplitude of the

masses is determined by the subglottal pressure P_S, the initial configuration of the folds described by cross-sectional areas A_{g1}, A_{g2} at the rest position, the masses m1, m2 and the airflow velocity U_g. The driving pressure also depends on the supraglottal pressures [8][11]. Average air pressures within the glottis tend to be larger in the convergent glottal configuration than in the divergent shape, resulting in the asymmetry of air pressures needed to sustain oscillation.

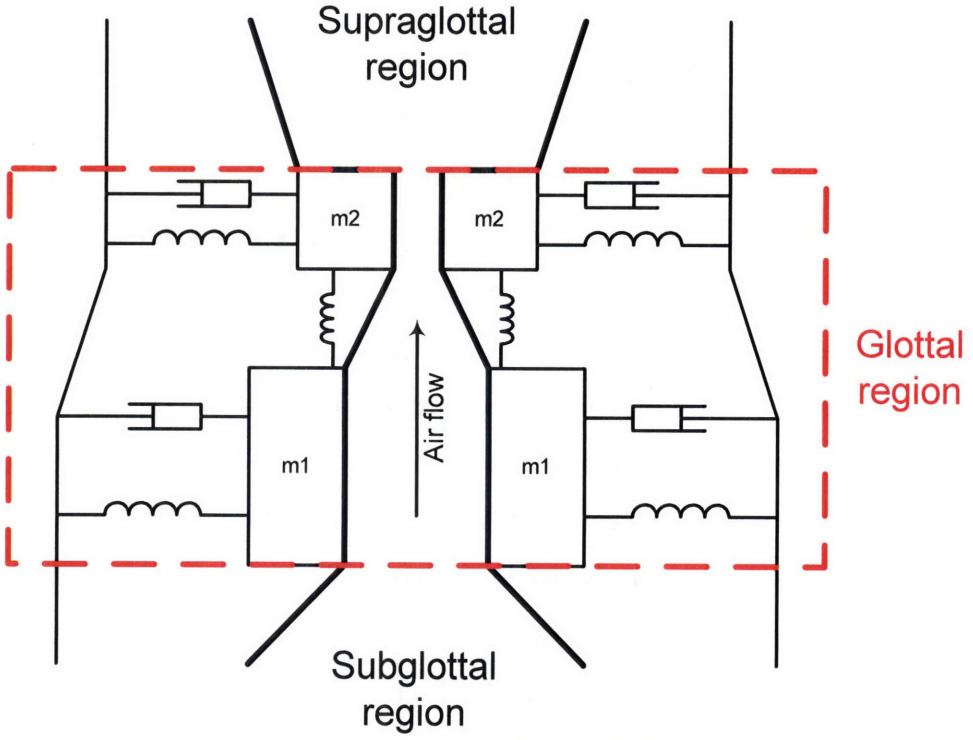


Fig. 1-7: A two mass model of the glottal oscillator.

1.3 Overview of speech synthesis

Synthetic speech is produced by several techniques, which may be broadly classified as follows:

(a) Concatenation synthesis, which uses pre-recorded samples from natural speech;

(b) Formant synthesis, which uses source-filter models to elucidate the transfer function of the vocal tract and pole frequencies of the speech signal;

(c) Linear prediction coding (LPC) based synthesis, which like formant synthesis is based on a source-filter model;

(d) Sinusoidal synthesis, which is based on the assumption that the speech signal is represented by a sum of sinusoids with time varying amplitudes, frequencies and phase;

(e) Articulatory synthesis, which directly models the human vocal tract and speech production system.

The various speech synthesis techniques are described below in more detail together with a discussion of their merits and drawbacks.

1.3.1 Concatenation synthesis

In concatenation synthesis, speech is produced by connecting appropriate units of pre-recorded utterances drawn from an inventory of basic synthesis units. The fundamental frequencies and durations of the synthesis units are altered by signal processing. Special algorithms and rules have been developed to ensure that spectral concatenation discontinuities between units are smoothed away. The choice of synthesis units impacts the quality of the synthesized speech and there is usually a trade-off between unit type, speech quality, and memory. The inventory of basic synthesis units is built by first recording natural speech such that all used units are captured. The units are then segmented and labeled. Finally, the most appropriate units are selected. To produce naturally sounding speech, the chosen prototype units must be concatenated in a natural way. To this end, the Pitch Synchronous Overlap Add (PSOLA) algorithm [16] was developed to enable the pre-recorded units to be concatenated smoothly in time and allows control for pitch and duration. It works by using overlapping segments and

smoothing each segment with a Hanning window such that neighbouring segments blend naturally. The Hanning window is centered around successive time instants called pitch marks. On voiced parts of the speech, the pitch marks are set at a pitch synchronous rate. On unvoiced parts, they are set at a constant rate. The pitch can be manipulated through the time intervals between pitch marks and the duration through repetition or omission of segments. A drawback with PSOLA is the presence of tonal artifacts in unvoiced parts of the speech.

Some common basic synthesis units used to form the inventory of utterances are words, syllables, demi-syllables, diphones, phones and sub-phones. Longer basic synthesis units are advantageous over shorter ones as they have fewer concatenation points and thus preserve continuity over a longer time scale. Moreover, it has been found that longer units have relatively smaller discontinuities at concatenation points. With longer basic synthesis units, the trade-off is a larger inventory for a vocabulary of a given size.

Words can be used as the basic synthesis unit for systems with limited vocabulary. The advantage is that intra-word co-articulation effects are captured within the pre-recorded unit. It is not suited for a large vocabulary as the number of words become prohibitively large. Moreover, pitch and formant discontinuities arising at word boundaries, among other things, also contribute to a difference between words spoken in isolation and in a continuous sentence. The result is unnatural sounding continuous speech. Multiple representations of each unit spoken in various contexts has to be recorded and used in conjunction with sophisticated rules in order for the concatenated speech to sound natural.

Syllables and demi-syllables are shorter synthesis units derived from words. It has been reported [13] that approximately 1000 demi-syllables are sufficient to obtain the estimated 10,000 syllables in the English language, making the memory requirement feasible. However, co-articulation effects between units is a problem.

Diphones are speech units that comprise the last half of a phone followed by the first half of the next phone. They have the property that they preserve the transition between phones. In general, the middle portion of phones is the most spectrally stable region and relatively spectrally consistent across phonetic contexts. Hence, the use of

diphones as the basic unit of synthesis tend to result in small concatenation discontinuities. Research has been conducted to supplement established diphone inventories with longer units to improve the synthesis of highly co-articulated phone sequences. Several diphone based synthesis systems have been reported [14][15] and augmented diphone systems form the basis of many commercial and research text-to-speech systems.

Phone-based speech segments are difficult to connect in a natural manner because they are subject to contextual variations in the acoustic realization of each phoneme. Sub-phone units have also been proposed for use in speech synthesis. Although sub-phones also have a degree of contextual variation, it has been observed that speech becomes more acoustically similar on these time-scales. Hence, the synthesis units can now be represented by a single vector of spectral parameters, making it amenable to state based models such as vector quantization (VQ) and hidden Markov models (HMM).

Concatenation synthesis has fundamental limitations. Data collection and segmentation require much time and effort. As the basic inventory is applicable to only one speaker, the synthetic speech cannot be altered to sound like a specific speaker without incorporating that speaker's voice. There is no straightforward means to obtain the optimal set of prototype units from natural speech and rules have to be developed to concatenate them in a smooth manner.

1.3.2 Formant synthesis

Vocal tract behaviour can be considered in terms of its overall transmission characteristics or its distributed properties. The former is the basis of formant synthesizers while the latter is exploited in articulatory synthesis described later in §1.3.5. Prior to the advent of concatenation synthesizers, formant synthesizers were probably the most widely used. A formant synthesizer is designed to simulate the acoustic output of the human speech process. Generally, three formants are required to produce intelligible speech. A typical formant synthesis system comprises 3-4 resonant circuits each having a frequency range that corresponds to the first three formants of average human speech and a bandwidth of 70-100Hz. A fourth resonantor provides for a nasal formant. Each formant is usually modeled with a two-pole resonator. A larynx pulse generator provides

the input to the resonators. For fricative sounds, a white noise source is used as the input. The first dynamically controlled formant synthesizers consisted of electronic resonators connected in series or in parallel [17][18]. More modern formant synthesizers employ dedicated digital hardware or simulation in place of analog circuitry [18][19].

1.3.3 Linear prediction coding based synthesis

Linear prediction coding (LPC) [20][21] is one of the most effective techniques for speech coding and is widely used in vocoders. A block diagram of a LPC vocoder is shown in Fig. 1-8. LPC performs linear predictive analysis on a finite number of previous speech frames using an all-pole assumption. It deconvolves the contributions of the source and filter by fitting an all-pole filter to the signal. The resulting spectral representation is constrained to a p^{th} order polynomial form, where p is order of the LP analysis. The output of the LP analysis is a vector of coefficients that generates, within an all-pole modeling constraint, a spectrum that best fits the input spectrum over the speech frame. The input speech waveform is also inverse filtered based on estimated filter coefficients to produce an error signal from which pitch is derived. The source is approximated by either a train of impulse-like spikes from the error signal in the case of the LP vocoder or random noise. The main drawback of the LPC method is its all pole assumption which models nasals and nasalized vowels poorly because coupling of the nasal tract introduces zeros into the spectrum. Stop consonants are also modeled poorly mainly because of the short time-scale associated with such events compared to typical frame sizes.

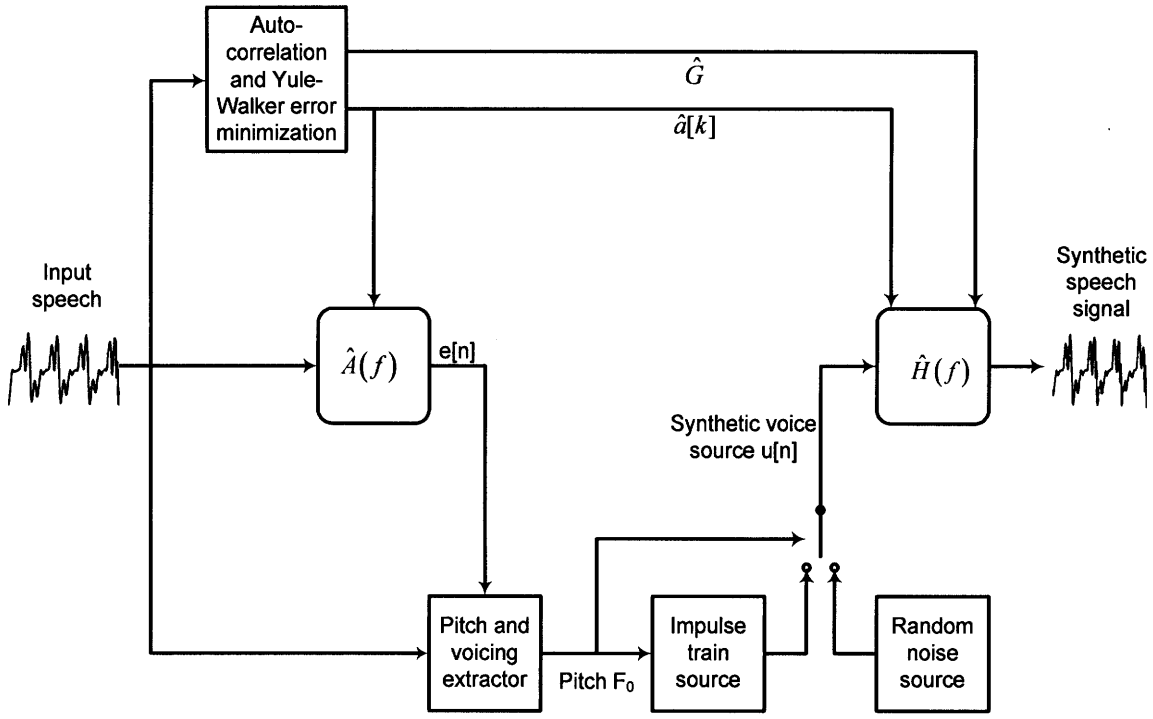


Fig. 1-8: Block diagram of linear predictive vocoder.

1.3.4 Sinusoidal synthesis

A block diagram showing the concept of sinusoidal synthesis is shown in Fig. 1-9. The system may be divided into a speech analysis section and a speech synthesis section. A target speech waveform (input) is first given a sinusoidal representation by using short-time Fourier transform (STFT) to extract phase, frequency and amplitude information [22]. Sine-wave generators then produce sinusoids with the appropriate phase, frequency and amplitude. All the component sinusoids are summed to produce a synthetic output. While the sinusoidal representation is suitable for representing periodic voiced speech signals, it is unclear how well it could approximate unvoiced speech segments with poorly defined or random phase. Some sort of phase randomization may be required for such speech segments.

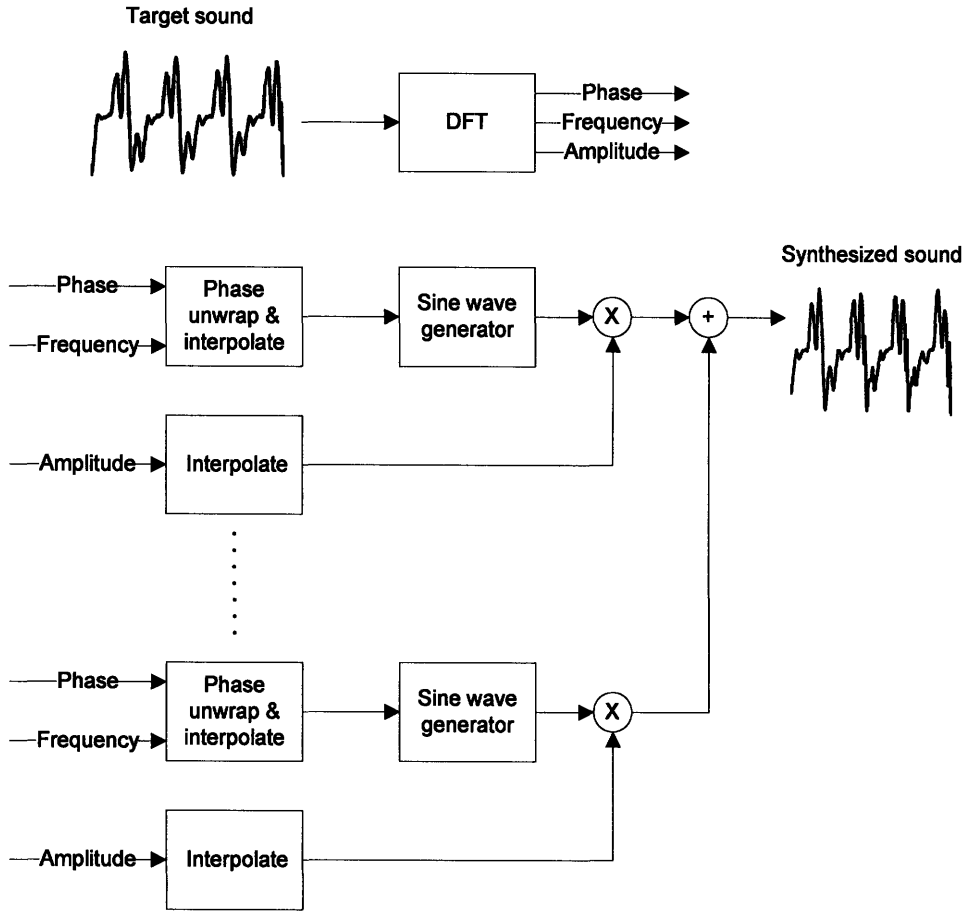


Fig. 1-9: Block diagram of sinusoidal analysis and synthesis system (adapted from [22]).

1.3.5 Articulatory synthesis

Articulatory synthesis is the generation of speech from a bio-physical model of the vocal tract with system parameters that are based on human physiology. It tries to model the human speech production system as closely as possible and hence potentially the most satisfying method for producing natural sounding speech. Potentially, articulatory modeling can also help provide a better understanding of how humans produce speech and thus advance the understanding of speech production. It has also been suggested that incorporating articulatory features in speech parameterization may help to improve recognition. More importantly, an articulatory synthesizer is the crucial link that closes the analysis-by-synthesis feedback loop between auditory processing and speech production [23][24].

In the past, several approaches to articulatory synthesis have been attempted [25][26][27][28][29]. Stevens et al [26] built a static electrical analog of the vocal tract

using discrete elements. A dynamically controllable electrical analog of the vocal tract is described in [27] using vacuum tube technology. Other methods that involve solving discretized partial differential equations [25][28] or a chain matrix approach [29] are computationally intensive and require the use of high-speed computers to perform in a reasonable amount of time.

In this thesis, we develop an experimental integrated-circuit (VLSI) analog vocal tract by exploiting the analogy between sound propagation in tubes and electromagnetic wave propagation in transmission lines. In our silicon VLSI implementation, we map fluid volume velocity to current, fluid pressure to voltage, and linear and nonlinear mechanical impedances to linear and nonlinear electrical impedances. We drive the analog vocal tract using an area function produced by a physiological model that enables speech signals of interest to be produced using parameters that correspond to elementary articulators and hence functionally related to how the biological vocal tract is articulated.

INTENTIONALLY LEFT BLANK

Chapter 2 ELECTRICAL MODEL OF SPEECH PRODUCTION SYSTEM

In this chapter, we describe circuit models for the various components of the speech production system. The circuit models are implemented in Matlab (Simulink) and combined to build a dynamic transmission line model of the vocal tract. The vocal tract model consists of 35 cascaded π-sections whose cross-sectional areas are specified by a vocal tract area profile. As there is a direct mapping between the computational model and analog circuit primitives, an analog integrated circuit implementation is feasible.

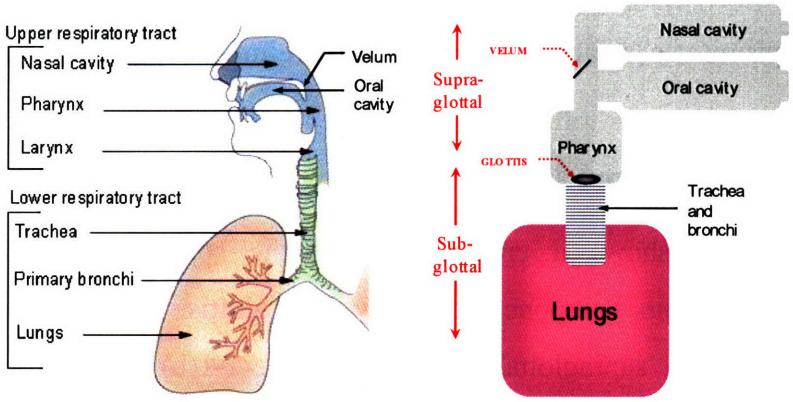
2.1 Overview of circuit model

Fig. 2-1(a) [30] shows a cross-section of the speech production system highlighting the three main components namely, the glottis within the larynx, the subglottal system and the supraglottal vocal tract. Fig. 2-1(b) is the corresponding schematic diagram describing it as an equivalent system of connected tubes and cavities.

Fig. 2-1(b) also shows the supraglottal vocal tract subdivided into three regions: the pharynx, the oral/mouth cavity and the nasal cavity. A structure called the velum separates the oral and nasal cavities. As described in Chapter 1, the lungs and respiratory muscles provide the vocal power supply. Voiced speech is produced by air expelled from the lungs causing the vocal folds to vibrate as a relaxation oscillator. The ejected air stream flows in pulses and is modulated by the vocal tract. In unvoiced speech, sounds are created by passing the stream of air through a narrow constriction in the tract. They can also arise by making a complete closure, building up pressure behind it, and then followed by an abrupt release. In the first case, a turbulent flow is produced while in the second case, a brief transient excitation occurs. The puffs of air are shaped into sound waves of speech and eventually, radiated from the lips and/or nose.

In Chapter 1, we showed that a system of connected acoustic tubes and cavities can be analyzed in the electrical domain using transmission lines. Fig. 2-2 shows our circuit model of the vocal tract. The pharyngeal and oral cavities are modeled as

transmission lines of lengths ℓ_1 and ℓ_2. The nasal tract is modeled as a third transmission line of length ℓ_3 that is coupled to the oral cavity through a velar impedance Z_V. The pharyngeal and oral transmission line parameters vary with time during speech production, corresponding to changes in cross sectional area. With the exception of one or two sections that serve to couple the nasal tract to the main vocal tract, the transmission line parameters of the nasal tract remain relatively fixed and do not change during speech production.



(a) (b)

Fig. 2-1: (a) Cross-section (adapted from [30]) and (b) schematic diagram of functional components of the speech production system.

The pharyngeal transmission line is terminated at one end by a glottal impedance Z_{GC} modeling the glottal constriction formed by the vocal cords. Z_{GC} is a nonlinear impedance that serves 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. The glottal impedance is modulated by a glottal oscillator to model the opening and closing of the vocal folds. As the bronchial and tracheal tubes are relatively large compared to the glottal constriction, the pressure drop across them is small: The subglottal pressure and the alveolar (lung) pressure P_{alv} are almost the same. Moreover, the subglottal pressure is maintained nearly constant over the duration of several pitch

periods by the low impedance lung reservoir. In our model, the subglottal source is represented as a voltage source denoted by P_{alv}.

For consonant production, a supraglottal constriction Z_{SGC} is included within the oral cavity that comprises a constriction impedance and a turbulent noise source. The oral and nasal cavities are terminated by radiation impedances Z_{rad} and Z_{rad}'. The radiated sound pressures from the lips and nose, i.e., P_{rad} and P_{rad}' in Fig. 2-2, are proportional to the derivative of the current flowing in the respective radiation impedances. The proportionality constant scales inversely with distance from the mouth and nose.

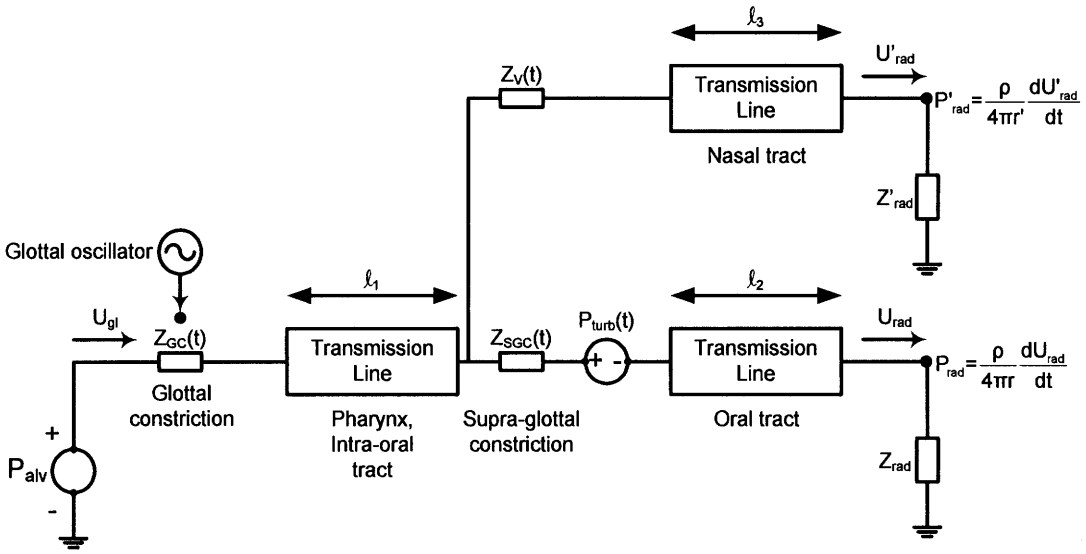


Fig. 2-2: Circuit model of vocal tract for speech synthesis.

2.2 Model of the glottis

2.2.1 Current source model

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. Fig. 2-3 shows an example of one period of a volume velocity (flow) waveform, U_{gl}, and its time derivative, (dU_{gl}/dt) obtained from a typical voiced glottal cycle. Fig. 2-4 shows the spectrum of U_{gl}. From the figure, we observe that the magnitude of the harmonic components decrease with frequency in an

approximately $1/f^2$ fashion. The approximate $1/f^2$ spectral roll-off is attributable to the closing of the glottal opening (as the vocal folds adduct) which produces a step-like change in dU_{gl}/dt. For male voices, the fundamental lies in the range of 100-150 Hz.

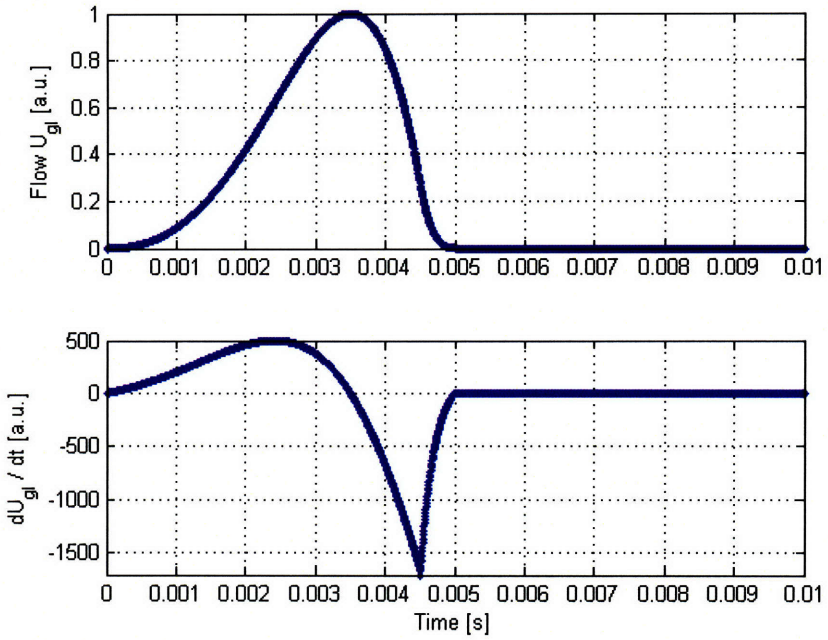


Fig. 2-3: An example of one period of a voiced glottal waveform (with a 10 ms glottal period) and its derivative.

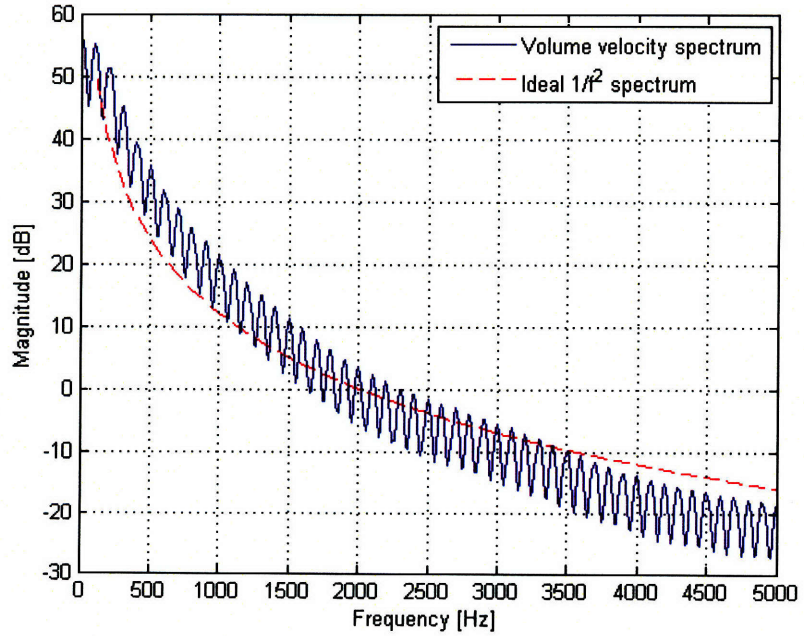


Fig. 2-4: Spectrum of glottal volume velocity waveform.

In the production whispered speech where voicing is absent, the periodic glottal waveform is replaced by turbulent noise. In this case, a random noise current source may be used to excite the vocal tract.

2.2.2 Variable impedance model

Poiseuille's Law states that the pressure difference ΔP across a uniform cylindrical tube of length l and radius r is related to its volume flow rate U as follows:

$$\Delta P = \frac{8\mu l}{\pi r^4} U \tag{5}$$

where μ is the viscosity of the fluid in the tube. The viscous resistance $8\mu l/\pi r^4$ depends linearly with viscosity μ and the length l, but has a fourth power dependence on the radius. Poiseuille's law is found to be in reasonable agreement with experiment for uniform liquids (Newtonian fluids) in cases where there is no appreciable turbulence.

However, when air flows through a constriction, in addition to the viscous component due to Poiseuille's Law, there are also contributions due to losses at the transitions [8]. These losses arise from eddies that form at the vicinity of the contraction and expansion of the tube and are dissipated as heat. Experimentally, it has been found that the pressure drop due to these losses is related to the dynamic pressure in the constricted tube:

$$\Delta P = k_L \frac{\rho}{2}\left(\frac{U}{A}\right)^2 \tag{6}$$

Fig. 2-5 illustrates the relationship between volume velocity U and pressure difference ΔP, with cross sectional area A as a parameter [8]. The dashed lines represent the relation described by (6). The solid lines show the effect of incorporating the losses due to laminar (viscous) resistance, described by (5), with the losses occurring at the expansions and constrictions, described by (6). We observe that viscous losses in the constriction become important at small cross-sectional areas and small airflows where the constriction resistance is larger because the flow is laminar rather than turbulent. Using the electrical analogy V for pressure drop and I for volume velocity, note from the figure that $V = I^2$ in turbulent flow compared with $V = I$ for the laminar case. Thus, the I-V curve looks like a square root in the case of turbulent flow and linear in the laminar case. Hence the turbulent case performs "compressive gain control" at high V's, and reduces

the net current flow compared with the laminar case, but at low V's, the laminar case produces less flow current. In analogy with nonlinear resistor I-V curves, if we think of the turbulent regime as "saturated square root flow" and the laminar regime as "linear flow", and the transition as the saturation voltage, then a small constriction or large resistance has a larger saturation voltage. The I-V curve of an electrical equivalent of a narrow constriction should look like the solid lines in Fig. 2-5. We can think of the electrical equivalent as a linear resistor connected in series with a square-root resistor where the latter is a nonlinear resistor that has an I-V characteristic described by (6).

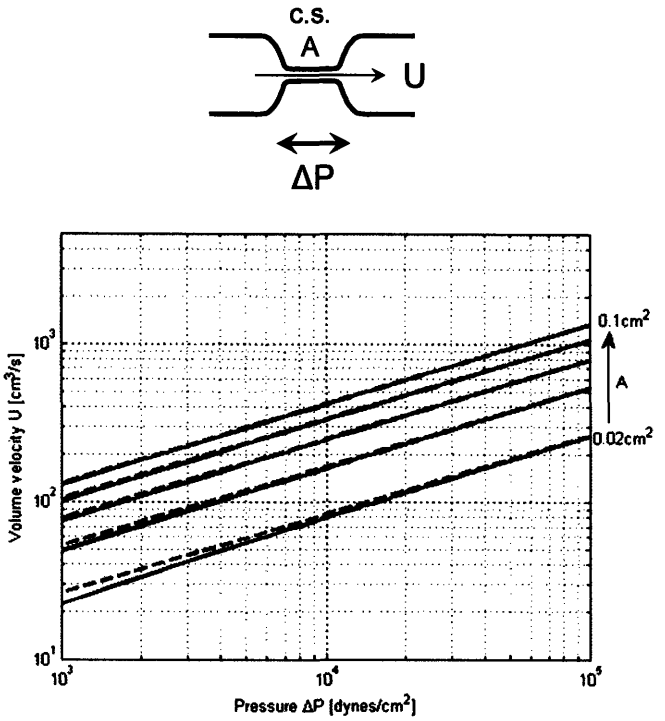


Fig. 2-5: Relation between volume velocity U of air and pressure difference ΔP of constriction [8] with cross sectional area A as a parameter. The dashed lines represent the relation described by (6). The solid lines show the effect of including the viscous resistance. The cross sectional area A of the constriction varies from 0.02 cm^2 to 0.1 cm^2 in 0.02 cm^2 increments.

We approximate a constriction at the glottis as a narrow cylindrical duct. The glottal resistance may then be modeled using a linear resistance in series with a nonlinear resistance as shown in Fig. 2-6. A glottal constriction also has acoustic mass, which may be represented as an inductance that is connected in series to the glottal constriction

resistance described above. As the area of cross section is small, the glottal inertance is dominated by its mass which, according to (3), may be computed as:

$$L_{gl} = \frac{\rho l_{gl}}{A_{gl}} \tag{7}$$

where l_{gl} is the length of the glottal duct. Fig. 2-7 shows that the glottal impedance is dominated by the linear resistance R_{lin} for very small cross sectional areas ($A_{gl} < 0.005$ cm^2) and by the nonlinear resistance R_{nl} for moderate constriction areas ($A_{gl} > 0.005$ cm^2). Fig. 2-7 also shows that the ratio of glottal inertance to the total glottal resistance (given by the sum of R_{lin} and R_{nl}) begins to saturate after $A_{gl} = 0.005$ cm^2. During a typical glottal period, the glottal area varies from zero (or close to zero) to about 0.2 cm^2 (not shown in Fig. 2-7). Hence, the maximum time constant (L_{gl}/R) is on the order of 0.15 ms, assuming a transglottal pressure of 8000 dynes/cm^2. Compared to a typical glottal period of 10ms for an average male voice, the time constant associated with the glottal inductance is small and may be neglected for most practical purposes. Moreover, we observed in Fig. 2-4 that glottal waveforms typically show a $1/f^2$ spectral characteristic: most of the glottal energy is concentrated at low frequencies. A resistor-only glottal model does not take into account the effect of acoustic mass. Nevertheless, we note that for frequencies above 1 kHz, the glottal reactance ωL_{gl} becomes comparable to the total glottal resistance.

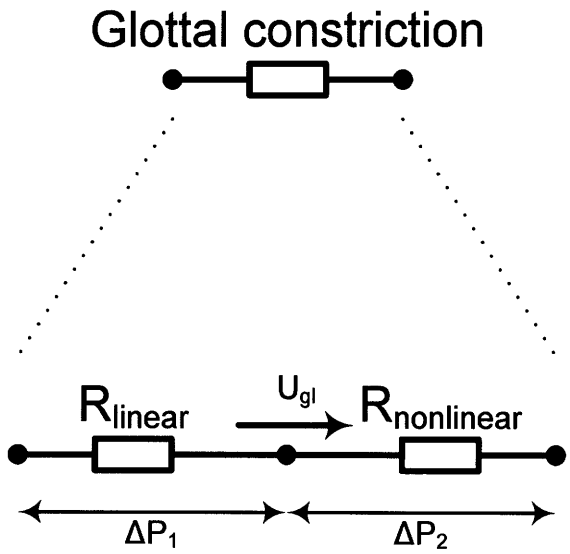


Fig. 2-6: Electrical model of a glottal constriction as a series combination of linear and nonlinear resistors.

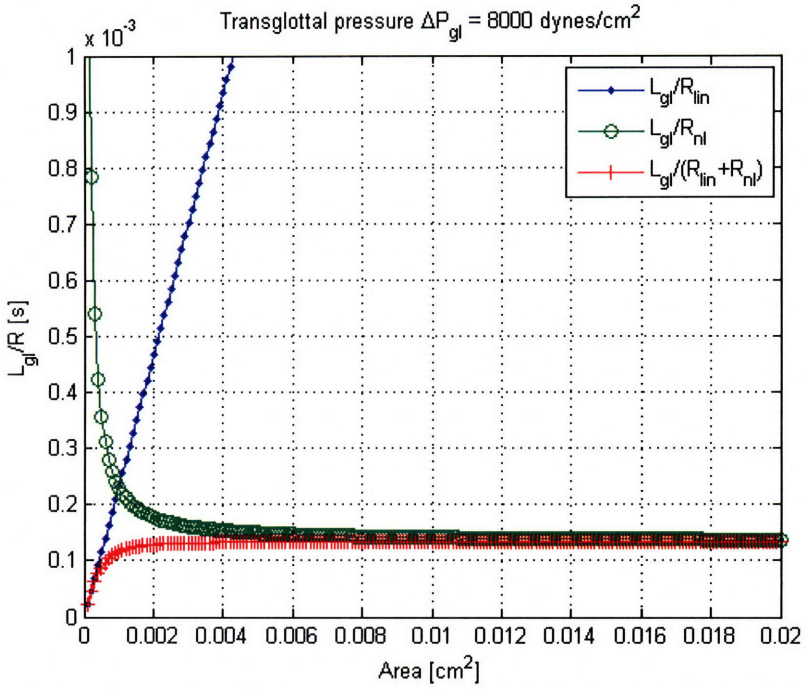


Fig. 2-7: Ratios of glottal inertance (L_{gl}) to linear and nonlinear glottal resistances (R_{lin} and R_{nl}).

Fig. 2-8 shows a model of the glottis consisting of two glottal constrictions connected in series to represent the upper and lower part of the vocal folds. There are two glottal constrictions because the upper and lower folds abduct and adduct with a time lag between them. The displacement of the upper and lower vocal folds with time [8] is shown in the upper part of Fig. 2-8, which clearly illustrates a periodic oscillation with a time lag between the opening and closing of the upper and lower folds. In our model of the glottis, the impedance of each glottal constriction is varied by a glottal oscillator in a corresponding manner. The oscillations in a two mass model of the glottal oscillator described in § 1.2.3 are sustained by a force produced by the glottal airflow when there is interaction with an acoustic tube. In our simple first order approximation of the two mass glottal oscillator, the effect of the glottis-vocal tract interaction is not modeled.

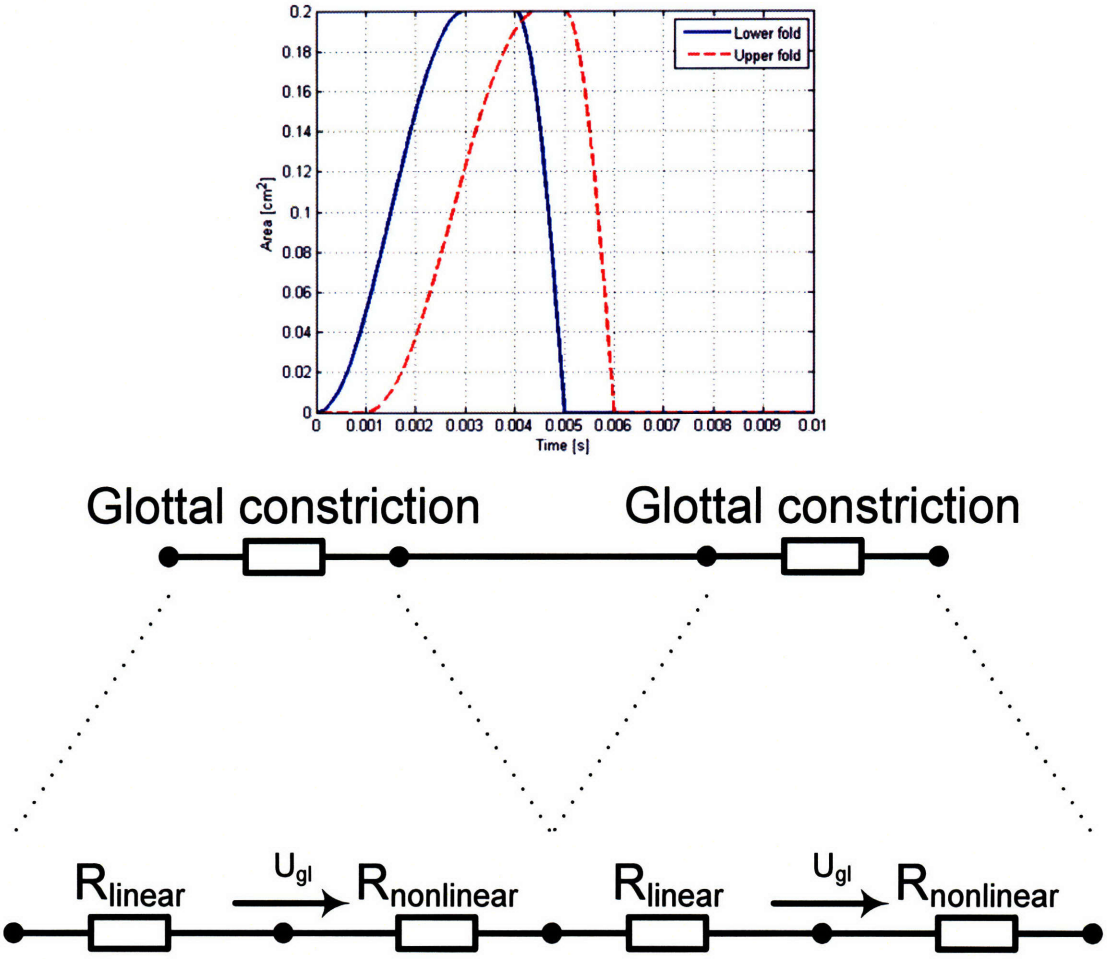


Fig. 2-8: Model of the glottis showing two glottal constrictions connected in series. The displacement of the upper and lower vocal folds with time is shown in the upper part of the figure.

2.3 Model of the supraglottal vocal tract

2.3.1 Pharyngeal and oral cavities

In § 1.2.2, we showed that the electrical analog of an acoustic tube is a transmission line, where voltage along the electrical line corresponds to sound pressure while current is analogous to volume velocity. In the circuit model of Fig. 2-2, the supraglottal vocal tract is represented by a spatially varying transmission line, corresponding to a non-uniform acoustic tube. The distributed transmission line may be approximated by a cascade of sections that forms a ladder network. Fig. 2-9 shows such a discrete representation using a cascade of π-circuit equivalents. Assuming that each

individual π-section represents a uniform acoustic tube of length ℓ and cross sectional area A, the values of L and C in each discrete section are determined from (3). Assuming a lossless line, the characteristic impedance Z at a point on the transmission line and the propagation constant γ is given by:

$$Z = \frac{\rho c}{A} \tag{8}$$

$$\gamma = j\frac{2\pi}{\lambda} = j\frac{\omega}{c}$$

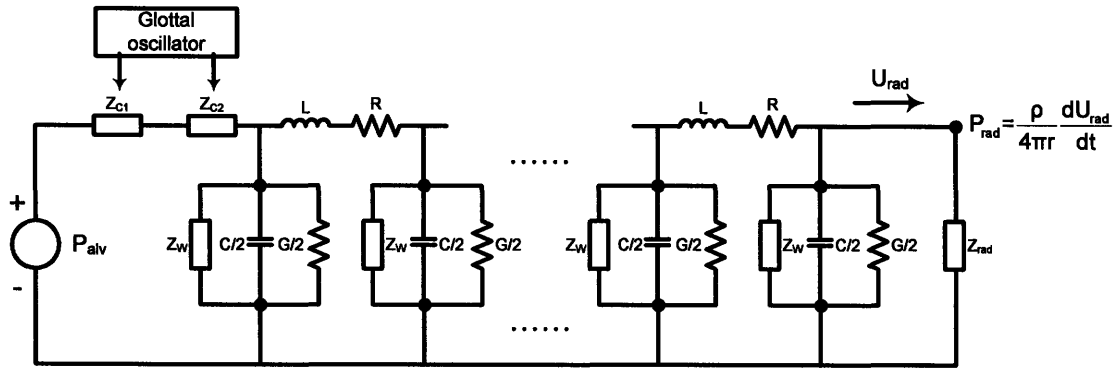


Fig. 2-9: The electrical analog of an acoustic tube implemented with an equivalent ladder network comprising π electrical analogs of cylindrical acoustic tube.

Practical values of electrical capacitance and inductance for the purpose of implementation in state-of-the-art very large scale integration (VLSI) technology, may be obtained by scaling the impedance level Z of the electrical analog by a factor $1/k_S$ with respect to the acoustic network such that the product of the inductance and capacitance per unit length remain unchanged:

$$L_S = \frac{\rho l}{k_S A}, \quad C_S = \frac{k_S A l}{\rho c^2} \tag{9}$$

$$Z_S = \sqrt{\frac{L_S}{C_S}} = \frac{1}{k_S}\sqrt{\frac{L}{C}} = \frac{1}{k_S}\frac{\rho c}{A}$$

$$c = \frac{l}{\sqrt{L_S C_S}} = \frac{l}{\sqrt{LC}},$$

where the L_S, C_S and Z_S denote corresponding inductance, capacitance and impedance parameters that have been scaled by $1/k_S$. The LC product, which determines the propagation speed, is unaffected the impedance scaling. Hence, the frequency characteristics of the scaled transmission line remain unchanged.

The audible frequency range of human speech spans between 100Hz and 7kHz. Consequently, the shortest wavelength of interest is 5cm. To keep the error introduced by spatial discretization small, the length, ℓ, of the approximating π-sections are kept short compared to the wavelength of sound corresponding to the maximum frequency of interest. To model a vocal tract with a length of 17.5cm, we employ a transmission line comprising 35 π-sections (each 0.5cm long) in our model of the vocal tract comprising the pharyngeal and oral cavities.

2.3.2 Model of the supraglottal constriction

A constriction or narrowing of the supraglottal vocal tract occurs during consonant production. A constriction of the vocal tract increases the resistance to air flow and produces a turbulent flow in the vicinity downstream of the narrowing. Turbulence produces random fluctuations in the air pressure. Consequently, a constriction may be modeled by a constriction impedance Z_{CVT} and a turbulent noise voltage source V_{noise}. The noise source is placed downstream of the constriction as illustrated in Fig. 2-10. For simplicity, the nasal tract has been omitted in the illustration of Fig. 2-10. During the production of consonants in standard American English, a constriction is formed in the oral cavity. There is typically only one supraglottal constriction. The constriction may be velar, alveolar or labial, depending on the location where the narrowing occurs. The location of the constriction determines the type of consonant.

We approximate the constriction impedance Z_{CVT} as a resistive component R_{CVT} in series with an inductive component L_{CVT}. The reactive component of Z_{CVT} is determined by the acoustic mass of the corresponding constricted π electrical analog described in previous section. As in the glottal constriction, the resistive component R_{CVT} is comprised of a combination of linear (R_{lin}) and nonlinear (R_{nl}) resistive elements in series, corresponding to viscous and eddy current losses, respectively. Fig. 2-11 shows a plot comparing the linear and nonlinear elements of R_{CVT} and the reactive (ωL_{CVT}) component of Z_{CVT} for a constriction area $A = 0.01$ cm^2. At frequencies above 1 kHz, the reactive component begins to dominate the constriction impedance for most constriction areas of practical interest (0.005 cm^2 $< A <$ 0.2 cm^2).

Fig. 2-12 plots the ratio of the linear and nonlinear resistive components of R_{CVT}. The resistive component of the constriction resistance is dominated by the nonlinear resistance at moderate constriction areas (0.01 cm^2 < A < 0.2 cm^2). For smaller areas R_{lin} and R_{nl} become comparable and eventually, the linear resistive component dominates for areas smaller than 0.0005 cm^2. Unlike the glottal constriction, the cross sectional area of a supraglottal constriction varies from zero (or close to zero) to areas as large as 5 cm^2 when the vocal tract transitions from a consonant configuration to a vowel configuration. As the constricted vocal passage is progressively opened, there comes a point beyond which negligible turbulent noise is generated. Based on the smallest areas for vowel configurations where it is known that the flow is predominantly laminar, we estimate the crossover point to be about 0.3-0.5 cm^2. For these areas, the resistive component of the constriction impedance becomes dominated again by the linear resistance, as there is negligible eddy current loss due to turbulence. At audio frequencies, the linear resistance is negligible compared to the reactive component comprising the acoustic inductance of the cylinder.

In summary, the resistive component R_{CVT} is only relevant for small cross sectional areas and does not interfere with the production of vowels. Moreover, the nonlinear element of R_{CVT} is the dominant component for all but the smallest constriction areas.

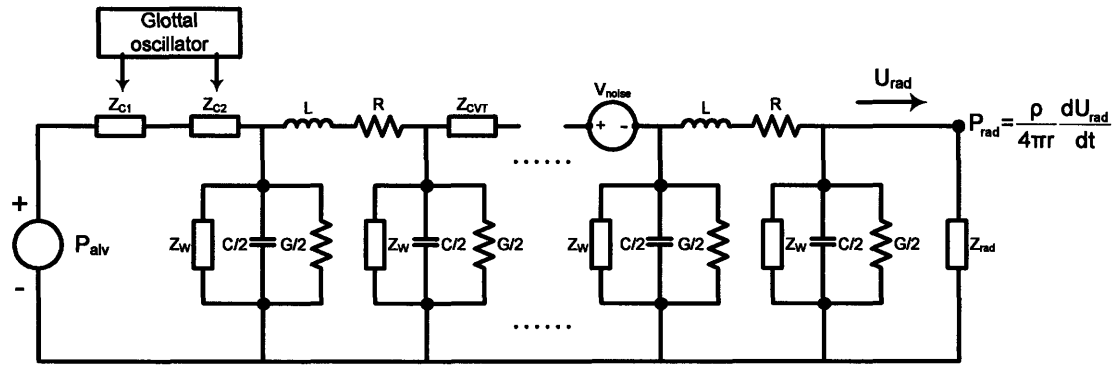


Fig. 2-10: Model of the supraglottal vocal tract with one constriction.

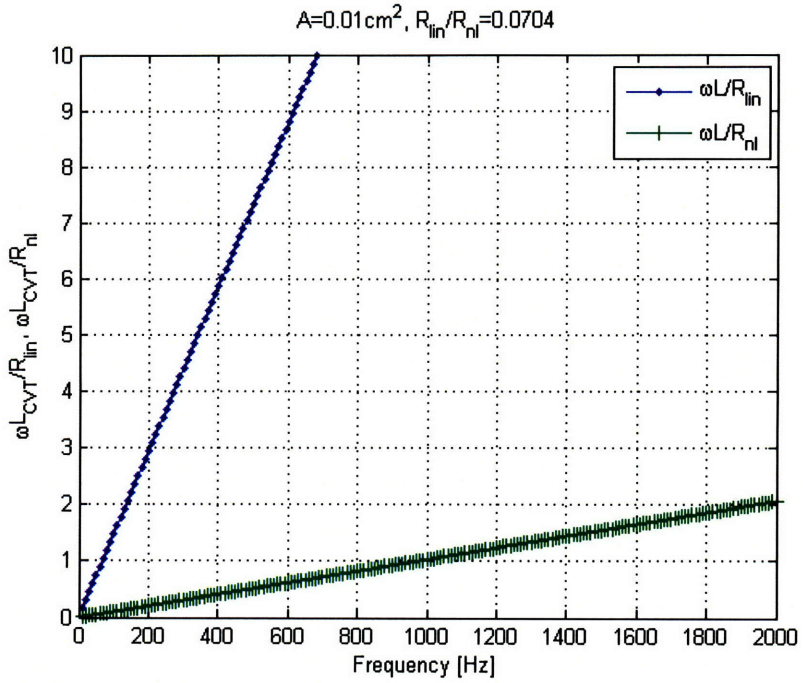


Fig. 2-11: Plot comparing the linear and nonlinear elements of R_{CVT} and the reactive (ωL_{CVT}) component of Z_{CVT} for constriction area A = 0.01 cm^2.

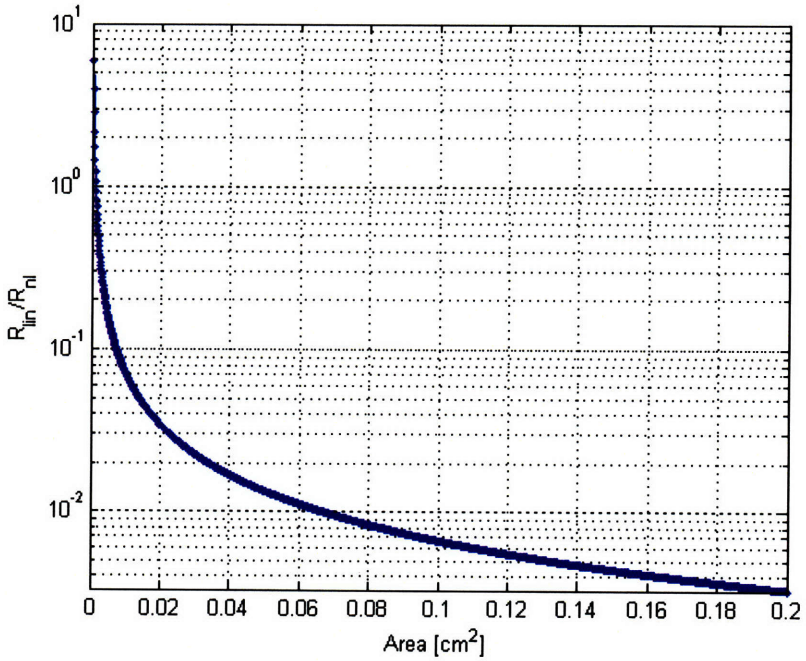


Fig. 2-12: Ratio of the linear and nonlinear resistive components of R_{CVT}.

2.3.3 Model of the nasal cavity

The supraglottal vocal tract divides into two paths at the velum. One path leads to the mouth and lips through the oral cavity. The other path, the nasal cavity, extends from the velopharyngeal opening to the nostrils. The velum acts as a gate; in the raised position it effectively closes the air passage between the oral and nasal cavities. When the velopharyngeal opening is closed, the nasal tract is acoustically decoupled from the main vocal tract comprising the pharyngeal and oral tracts. The velum is lowered during the production of nasal consonants and nasalized vowels.

As in the case of the pharyngeal and oral cavities, the nasal cavity is represented by a transmission line. It is coupled to the main transmission line through a π-section representing the coupling between the nasal and oral tracts. The nasal tract is completely decoupled from the main vocal tract via a switch. The acoustic properties of the nasal tract show considerable variability from one individual to another.

2.4 Model of sound radiation from the lips and nose

The exit of the oral and nasal cavities is connected to free space (atmosphere) where sound is radiated. For the purpose of speech production, it is sufficient to characterize radiation into free space as an acoustic load represented by a radiation impedance that defines the sound pressure and volume velocity relation. The radiation impedance of the nose or mouth opening can be modeled after the radiation impedance of a circular vibrating piston set in a spherical baffle that approximates the effect of the head. Up to a frequency of about 6 kHz, an approximate closed form expression for the radiation impedance is given by [8]:

$$\begin{aligned} Z_{rad} &= \frac{\rho}{c}\pi f^2 K_s(f) + j2\pi f \frac{\rho(0.8r_{rad})}{A_{rad}} \\ &= R_{rad} + j\omega L_{rad} \end{aligned} \tag{10}$$

where A_{rad} is the area of the nose or mouth opening, r_{rad} is the effective radius of the opening and $K_s(f)$ is a dimensionless frequency dependent factor that accounts for the baffling effect of the head. For a man, a spherical baffle with a radius of 9cm may be used as an approximation of the head. The mouth opening for a rounded vowel such as /u/ is on the order of 0.5 cm^2. On the other extreme, a typical mouth opening of an open

vowel, such as /a/, is 5 cm^2. The corresponding radii of circular pistons are 0.4 cm and 1.26 cm, respectively. Hence, the piston-to-sphere radii ratio for the two extreme areas are 0.04 and 0.14, respectively. In the case where the radius of the piston is small compared with that of the sphere, the radiation load approaches that of a piston in an infinite baffle. In this limiting condition, K_s=2. It may be shown that for the above dimensions and in the audio frequency range, up to about 5 kHz, the difference between the radiation impedances of an infinite and spherical baffle is small [9]. The effective radiating area at the nostrils is typically smaller than 5 cm^2. Hence, the radiation load at the lips and nostrils can be approximated as a piston in an infinite baffle.

Fig. 2-13 shows the ratio of the reactive and resistive components of Z_{rad} as a function of frequency for two different mouth openings. From (10), the resistive component R_{rad} is proportional to the square of the frequency, which is difficult to implement. Fortunately, the resistive component is smaller than the reactive component at all frequencies of interest for all but the largest of mouth openings (>5 cm^2). Hence, a simple approximation of the radiation impedance is an inductor connected in series with a linear resistance. Such a model does not take into account the frequency dependent loss of Z_{rad} but should still provide a sufficiently good approximation for frequencies up to about 5 kHz. The Q of the inductor is determined by the linear series resistance, setting a frequency independent loss. Fig. 2-14 shows the transmission line comprising the pharyngeal and oral tracts being terminated by the simple impedance model, denoted by Z_T, at the lip end. For simplicity, the nasal tract and wall impedances have been omitted in Fig. 2-14. A similar impedance may also be used to terminate the transmission line representing the nasal tract.

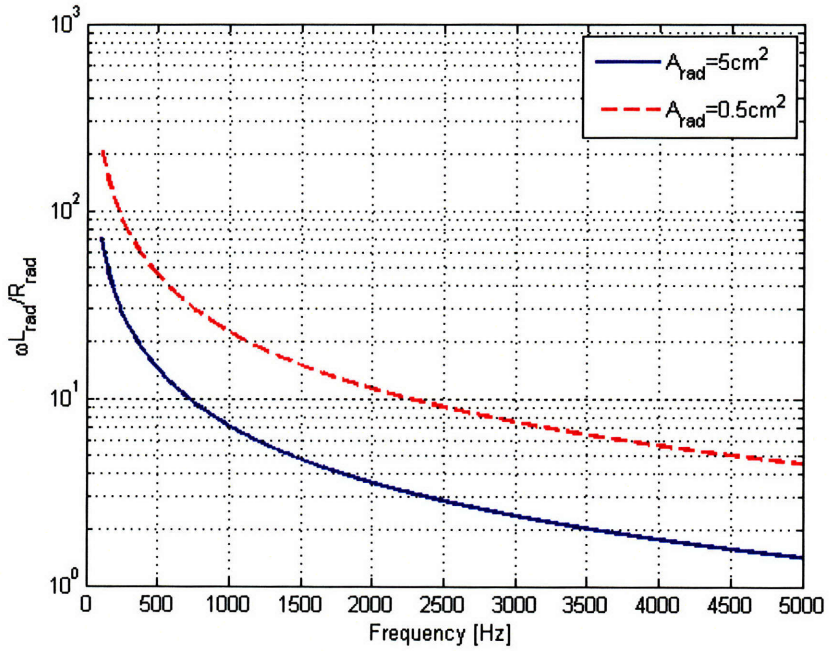


Fig. 2-13: Ratio of radiation reactance and resistance for a piston in an infinite baffle for different cross sectional areas of the piston.

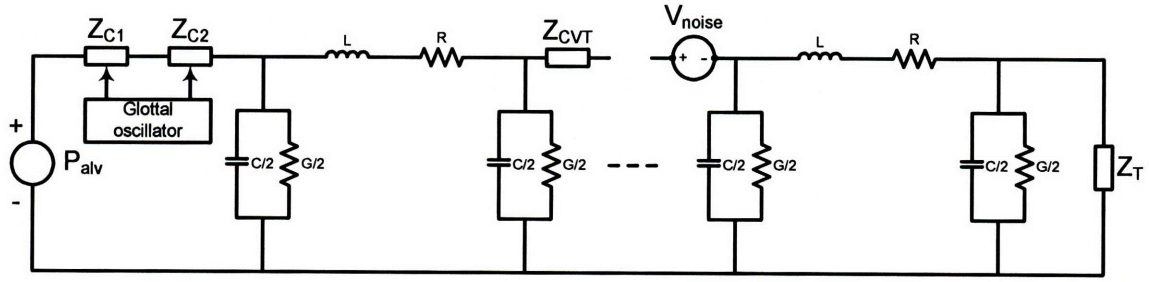


Fig. 2-14: Model of the supraglottal vocal tract (excluding the nasal tract) with termination impedance Z_T at the lips.

Note that an ideal inductor is a high pass element that produces a voltage that is equal to the scaled derivative of the current at the lips. Such a model of the radiation impedance at the lips is equivalent to terminating the transmission line at the lip end by a high Q inductor.

2.5 Simulation of speech production

2.5.1 Speech production with supraglottal vocal tract

Fig. 2-15 is a schematic diagram of the speech production system implemented in Matlab (Simulink). The subglottal source is represented by a voltage source P_{alv}. The glottis is represented by a tunable impedance Z_G. The glottal impedance Z_G is controlled by a glottal oscillator which modulates the value of Z_G in a periodic fashion to produce a volume velocity waveform similar to the one shown in Fig. 2-3. Each rectangular box is a two port representation of an electrical π-section (Fig. 1-5). The boxes, denoted by F, correspond to π-sections of the main vocal tract, comprising the pharyngeal and oral tracts. The main vocal tract is terminated by a radiation impedance Z_M at the mouth. The nasal tract is comprised of sections denoted by F'. The nasal cavity is coupled to the oral cavity through a velar opening represented by a velar impedance Z_V. The nasal tract is terminated by a radiation impedance Z_N at the nostrils.

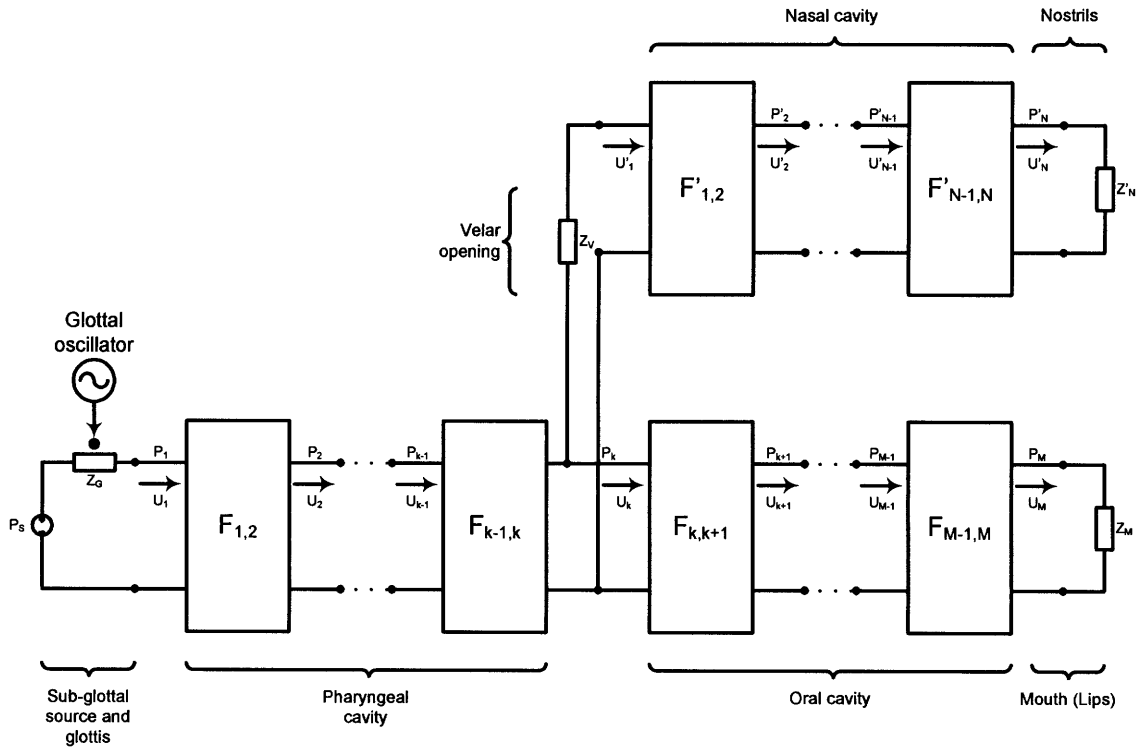


Fig. 2-15: Model of the speech production system.

2.5.2 Simulation results

Fig. 2-16 summarizes the basic units of speech or phonemes in standard English and shows how they are classified in general. Phonemes are divided into two broad categories: vowels and consonants. The distinction is primarily determined by the presence or absence of a narrow constriction located in the oral cavity. Using our Matlab model of the vocal tract, we simulated the production of various speech sounds using vocal tract profiles corresponding to various phonemes. We defer the treatment of non-stationary speech sounds, such as diphthongs, consonant-to-vowel transitions , and words to the next chapter where we describe in detail how we drive the vocal tract. In this section, we present the results pertaining to a stationary vocal tract profile.

Fig. 2-17(a)-Fig. 2-21(a) show the stationary vocal tract profiles [26] used to control the cross-sectional areas of our circuit model of the vocal tract. The corresponding spectrums of the resulting synthesized speech computed from circuit simulation using a Matlab (Simulink) model are shown in Fig. 2-17(b)-Fig. 2-21(b). The variable impedance model of the glottis in conjunction with a glottal oscillator is employed to model the opening and closing of the vocal folds in the simulation.

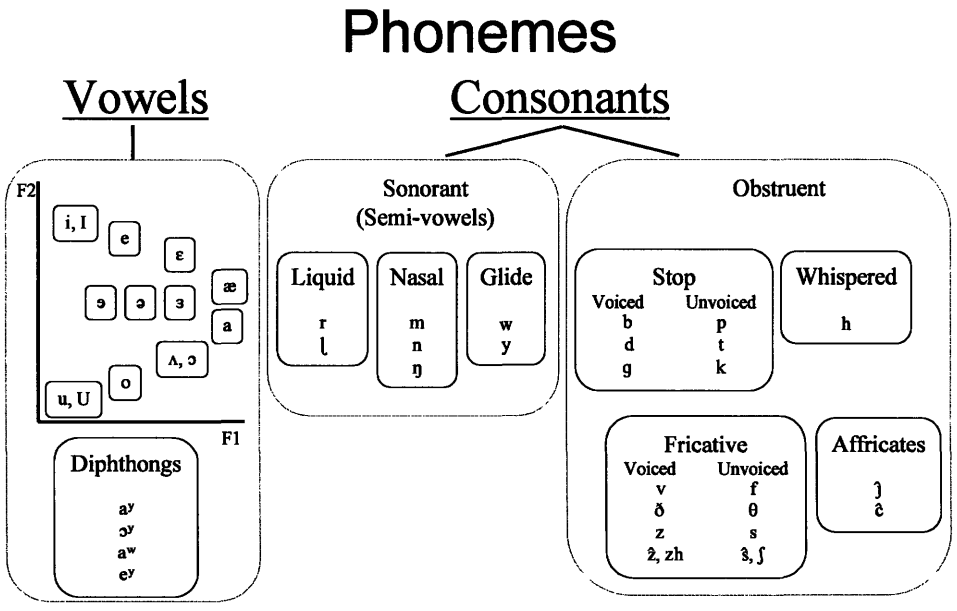
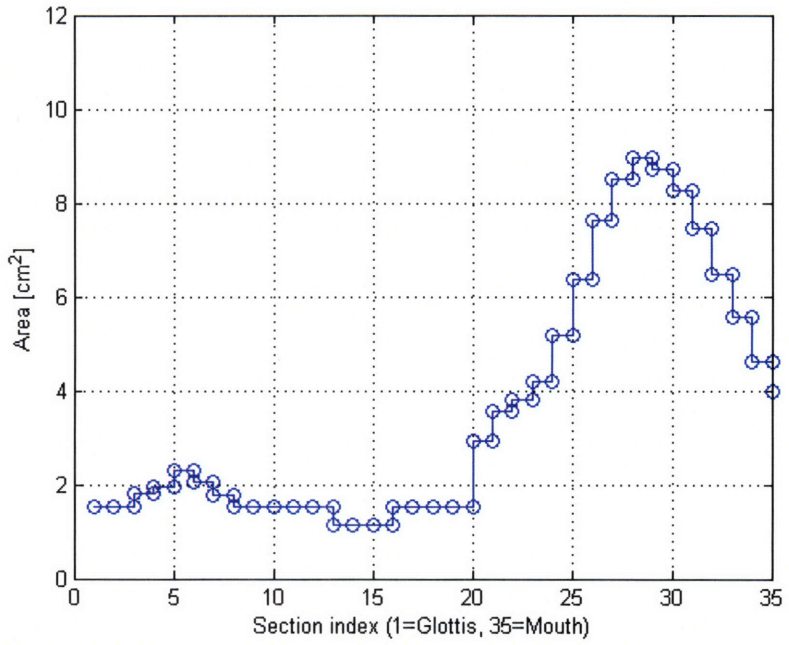
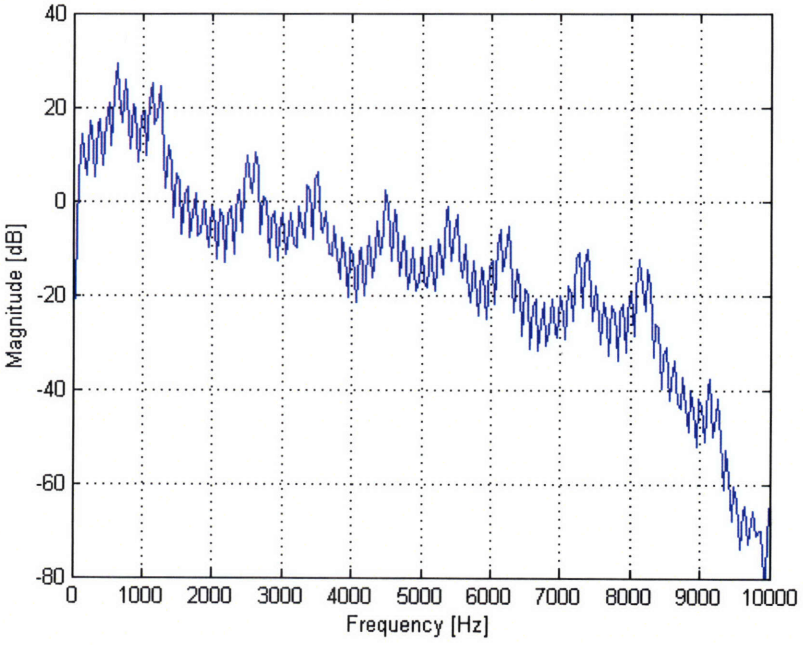


Fig. 2-16: Basic units of speech produced by Matlab (Simulink) model of vocal tract.

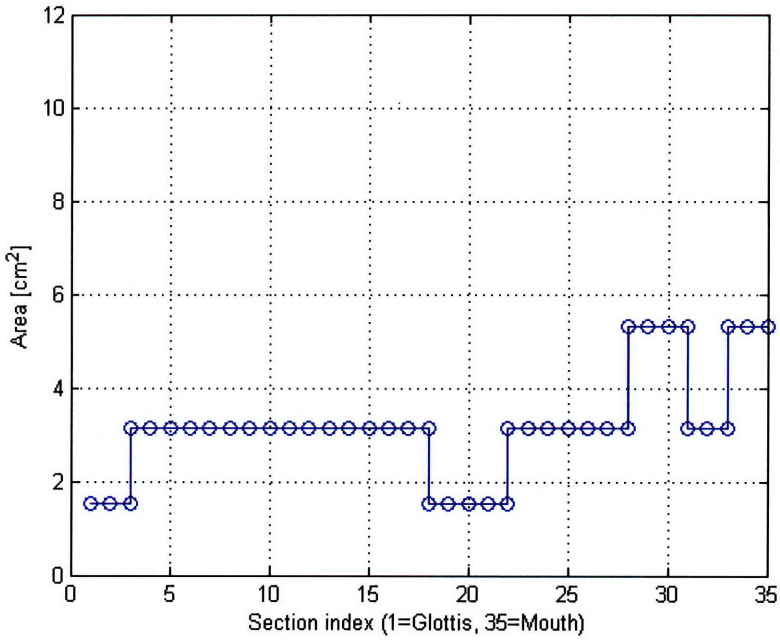


(a)

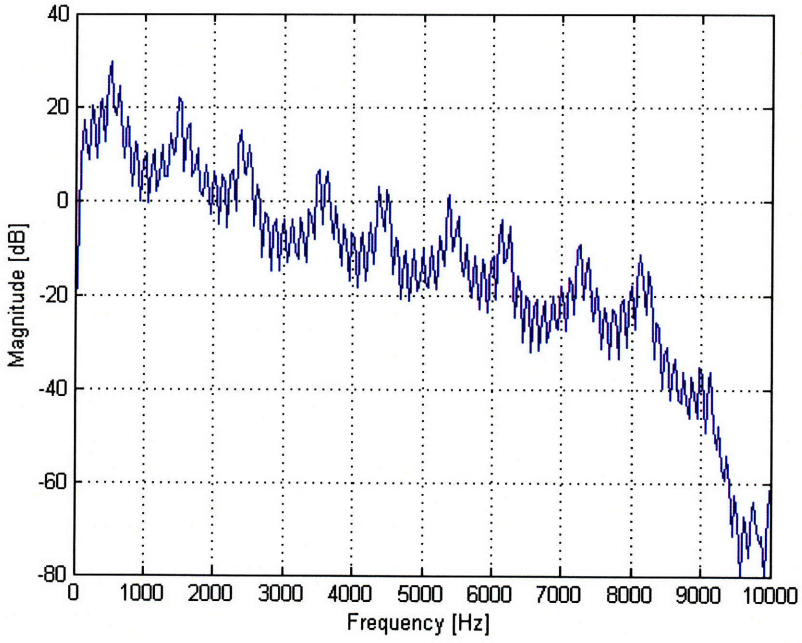


(b)

Fig. 2-17: (a) Vocal tract profile and (b) spectrum of synthesized vowel /a/.

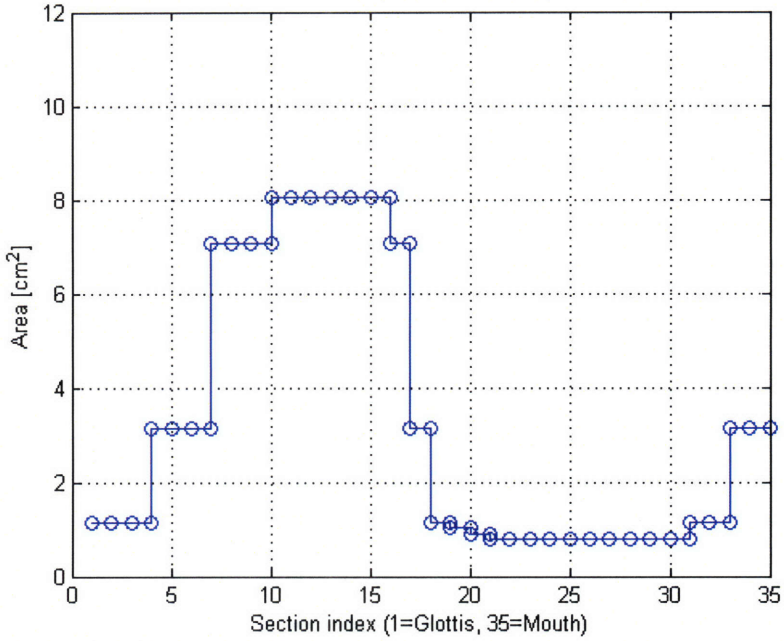


(a)

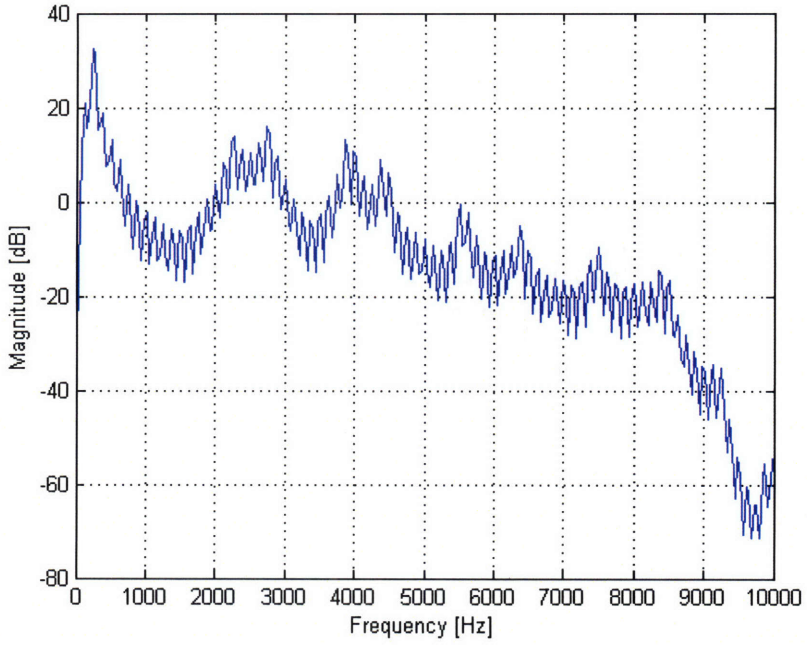


(b)

Fig. 2-18: (a) Vocal tract profile and (b) spectrum of synthesized vowel /e/.

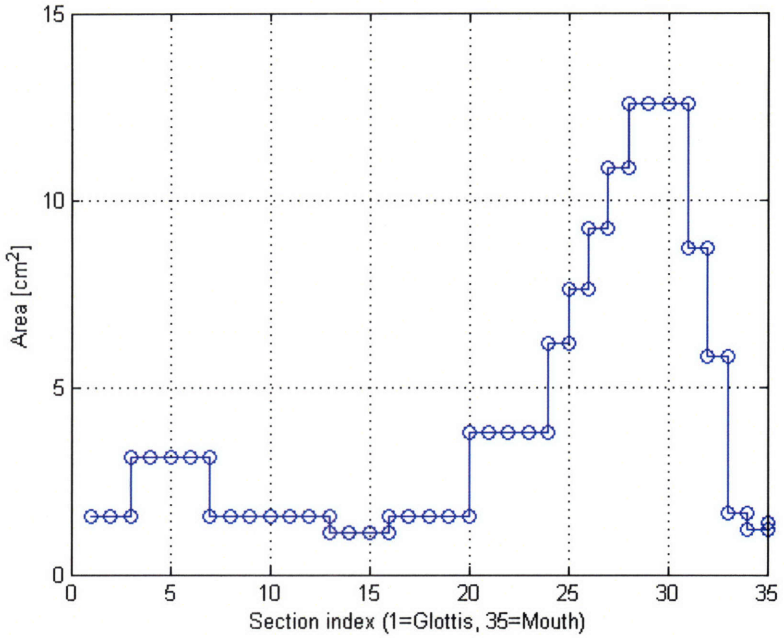


(a)

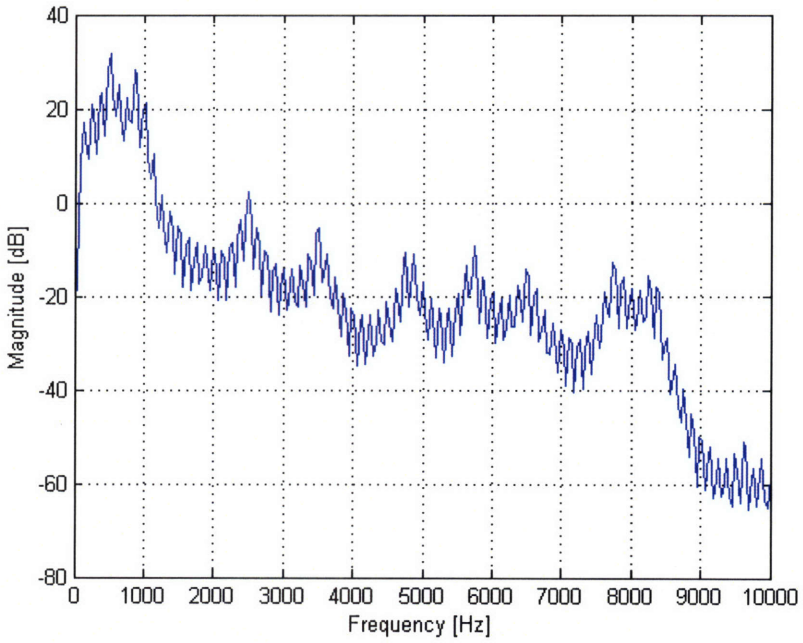


(b)

Fig. 2-19: (a) Vocal tract profile and (b) spectrum of synthesized vowel /i/.

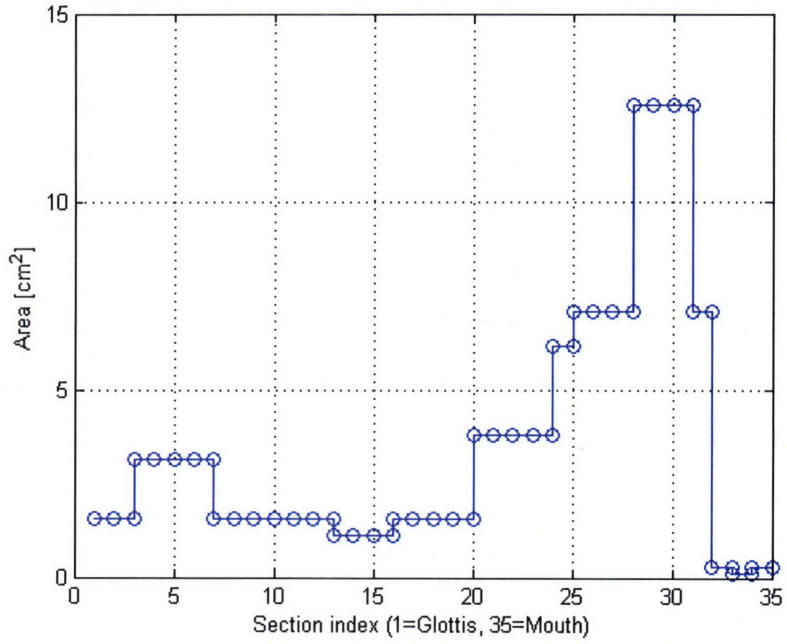


(a)

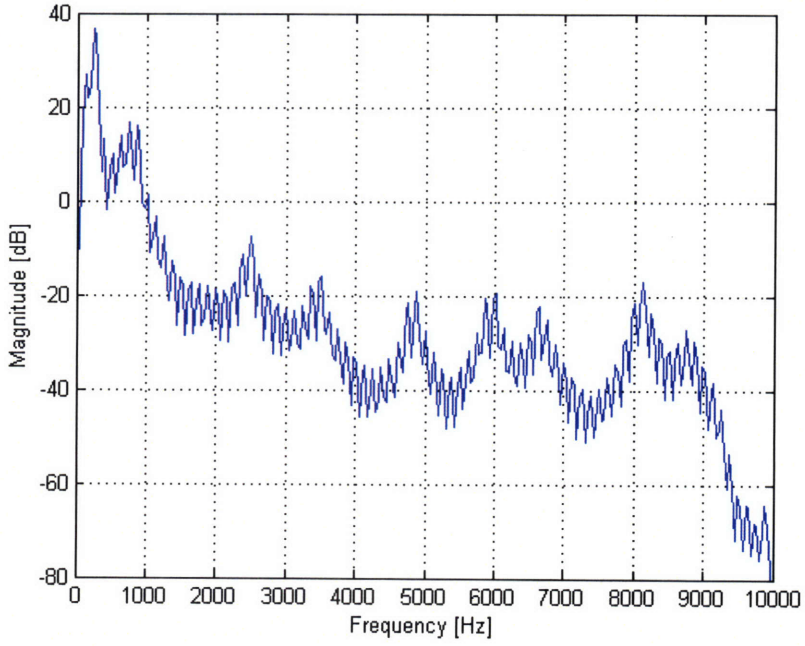


(b)

Fig. 2-20: (a) Vocal tract profile and (b) spectrum of synthesized vowel /o/.



(a)



(b)

Fig. 2-21: (a) Vocal tract profile and (b) spectrum of synthesized vowel /u/.

The Matlab simulation results confirm that the formant positions of the synthesized speech are consistent with what is known about the spectral characteristics of the five vowels /a/, /e/, /i/, /o/, /u/. In particular, the third formant frequency F_3 remains relatively invariant across all the five vowels and is located at approximately 2.5 kHz. In the simulation, the length of each section is 0.5cm, giving a vocal tract length of 17.5 cm. Assuming an uniform tube, as in the case of the neutral vowel, F_3 is given by $5c/4l$, where c is the velocity of sound in air at room temperature, and l is the length of the vocal tract. Using $c = 35000$ cm/s and $l = 17.5$cm, the third formant frequency may be computed as $F_3 = 2.5$kHz; a value that is consistent with what we observe from the spectra shown in Fig. 2-17-Fig. 2-21.

In summary, we showed that the speech production system for vowels may be modeled with an electrical analog. The necessary computational elements such as electronically tunable linear and nonlinear impedances may be realized using circuits comprising MOS transistors in VLSI technology. As there is a direct mapping between the computational model and analog computational primitives, an analog integrated circuit implementation is potentially computationally efficient. The subject of VLSI implementation is the topic of Chapter 4 through Chapter 6. Next, we turn to the topic of driving our circuit model of the vocal tract in the context of producing non-stationary speech.

Chapter 3 DRIVING THE ANALOG VOCAL TRACT

Our circuit model of the vocal tract, presented in the previous chapter 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 this chapter, we introduce a speech coding scheme that tests the efficacy of our analog circuit model of the vocal tract beyond the simple stationary sounds presented in the previous chapter.

3.1 Articulatory representation of speech

The supraglottal vocal tract modulates the glottal excitation signal to produce various linguistic sounds and can be modeled as a discrete acoustic tube with spatially varying and time-varying cross-sections. As shown in previous chapters, the way in which the cross-sectional area varies along the vocal tract determines the formants (resonant frequencies) of the tract and thus the sound that is produced. Hence, the area function of a vocal tract is one of the most important determinants for the production of a given speech signal and essential for a better understanding of the relationship between articulation and the speech acoustic signal. However, the number of degrees of freedom for a discrete acoustic tube approximation of the vocal tract employing quantized sections of 0.5cm-1cm in length is very large. It is also non trivial to impose constraints on an area function representation to ensure that the vocal tract profile is physiological.

In the human vocal tract, the cross-sectional area profile is adjusted by moving the articulators such as the jaw, tongue body, tongue tip, and lips. An additional articulator, the velum, couples the nasal cavity to the oral cavity. Consequently, the speech signal may also be regarded as the output of a time-varying articulatory parametric system that is excited by a glottal source, with articulatory parameters that specify the location of the jaw, tongue body, tongue tip, lips, and velum. The study of X-ray microbeam data reveals that the geometric configuration the vocal tract may be adequately represented by means

of variables specifying the positions of the articulators. An articulatory model allows us to drastically reduce number of degrees of freedom, narrow down the vocal tract area function space and produce realistic vocal tract profiles using a small number of parameters that are functionally related to how we articulate our vocal tract. Moreover, it offers potential benefits in giving a compact, robust, and linearly changing representation that has a closer relationship with the phonetic domain thereby allowing a straightforward treatment of transitions.

3.2 The Maeda articulatory model

The Maeda articulatory model [31] is a statistical anthropomorphic model developed from cineradiographic and labiofilm data of the human vocal tract. Temporal variations of vocal tract profiles during continuous speech were studied using factor analysis to describe the profiles as the sum of seven linear components each corresponding an elementary articulator. The seven articulatory parameters are:

(i) P1: Jaw height; this parameter moves the jaw vertically up or down;
(ii) P2: Tongue body position; this parameter moves the tongue dorsum roughly horizontally from front to back of the oral cavity;
(iii) P3: Tongue body shape; this parameter indicates whether the tongue dorsum is rounded (arched) or unrounded (flat);
(iv) P4: Tongue tip; this parameter deforms the apex part of the tongue by moving it up or down;
(v) P5: Lip height; this parameter affects the area of the mouth opening by moving the lips together or apart;
(vi) P6: Lip protrusion; this parameter extends the vocal tract slightly during the production of rounded vowels;
(vii) P7: Larynx height; this parameter raises or lowers the position of the larynx;

Realistic vocal tract profiles may be represented with reasonable accuracy using the seven articulatory parameters described above and illustrated in Fig. 3-1. Using the

Maeda model, temporal variation of the vocal tract profile may be represented by frame-by-frame samples of the seven articulatory parameters.

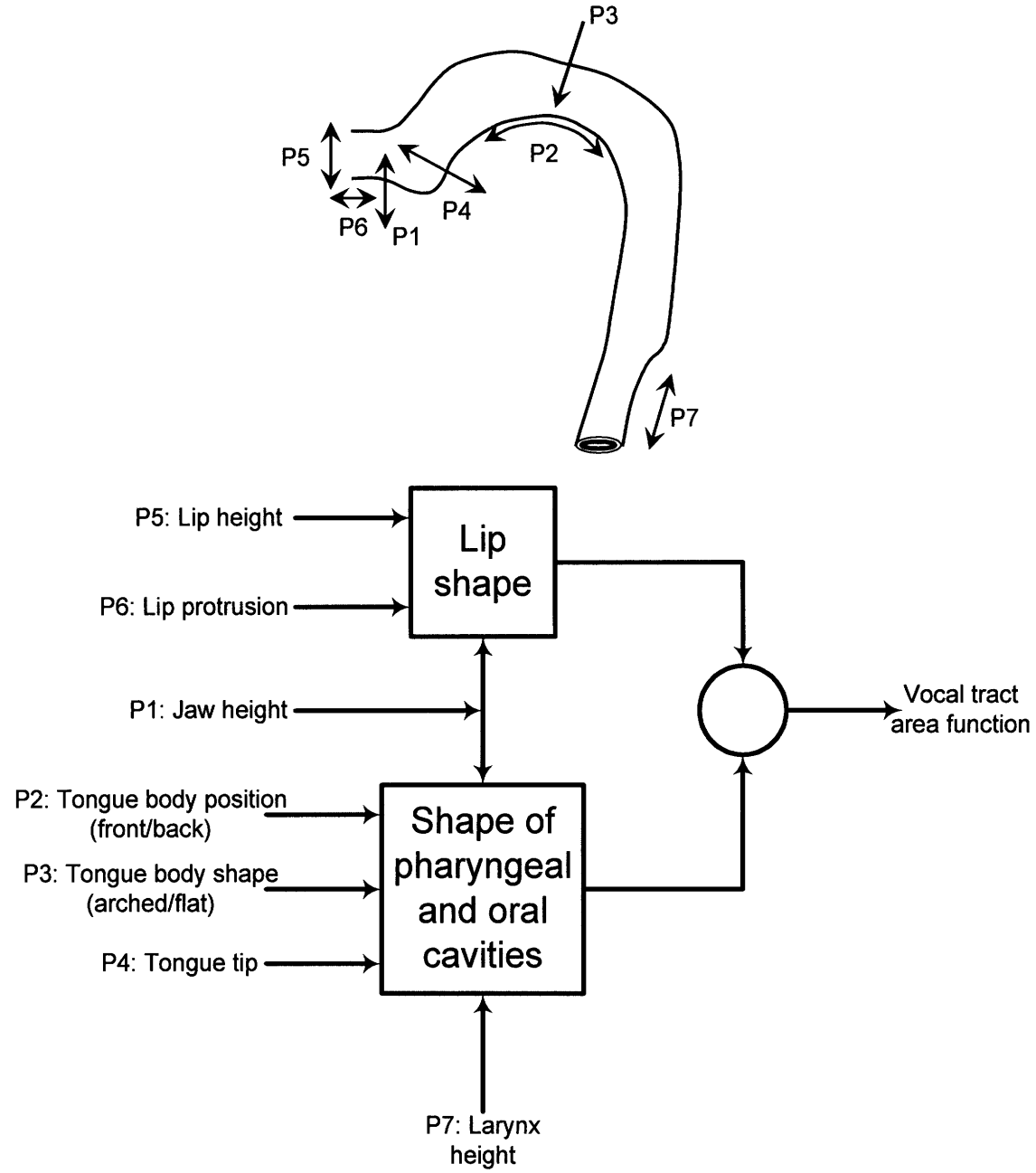


Fig. 3-1: The seven articulatory parameters of the Maeda model.

3.3 The synthesis process

Fig. 3-2 illustrates a synthesis process which employs an articulatory model to specify area functions that control the vocal tract. In the first stage, acoustic features are extracted from input speech. This is followed by an inversion task where the extracted acoustic features are transformed into articulatory parameters and subsequently, the vocal tract area function. Finally, speech is synthesized by a model of the vocal tract using the area function as the control.

The acoustic-to-articulatory transform is by far the most challenging part of the synthesis process. The main difficulty in estimating appropriate parameters from the speech signal for acoustic-to-articulatory inversion is that the mapping between the acoustic and articulatory domains is both nonlinear and one-to-many. As the problem is also under-determined, there exists a multitude of vocal tract shapes that can produce the same speech spectrum. Consequently, constraints that are both sufficiently restrictive and realistic from a phonetic and physiological point of view must be incorporated in order to eliminate false solutions. In order to estimate the articulatory parameters, nonlinear mapping methods such as nonlinear regression, neural networks, or articulatory codebooks have been proposed [32]. We describe a codebook approach to the problem of acoustic-to-articulatory inversion below.

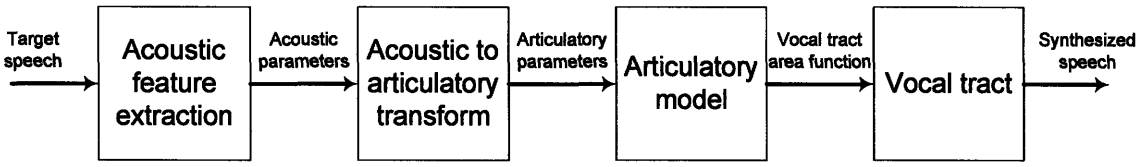


Fig. 3-2: Reconstructing a speech signal.

3.4 Building an articulatory codebook through babbling

In order to reduce the dimensionality of the articulatory description and restrict the area function to reasonable vocal tract shapes, we employ the Maeda articulatory model to specify the vocal tract area function. 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 16,000 entries, specified by 12 mel-frequency cepstral coefficients in the acoustic domain and 7 articulatory parameters. The articulatory space is sampled at intervals shown in Table 3-1. The articulatory parameters are normalized to the maximum and minimum values shown in the table. Fig. 3-3 shows a synthesis process used with an articulatory representation. The input is a target sound to be reconstructed. The mel-frequency spaced (MEL) filter bank, with center frequencies ranging from 130 to 6500Hz, decomposes the target sound into its constituent frequency components which are represented by a set of 30 mel-frequency spaced filter coefficients [33]. A discrete cosine transform (DCT) is applied on the mel-frequency spectrum to generate a set of 12 mel-frequency cepstral coefficients. The mel-frequency cepstrum gives a description of the shape of the mel-frequency spectrum. The set of mel-frequency cepstral coefficients of the target sound are compared against a codebook that contains the mel-frequency cepstral coefficients of babbles synthesized by the vocal tract. The set of mel-frequency 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.

Articulatory parameter	Min value	Max value	Step size
Jaw height	-3	3	1
Tongue body position	-3	3	1
Tongue body shape	-3	3	1
Tongue apex	-2	2	2
Lip height	-1	1	1
Lip protrusion	-2	2	2
Larynx height	-2	2	2

Table 3-1: Sampling the articulatory space.

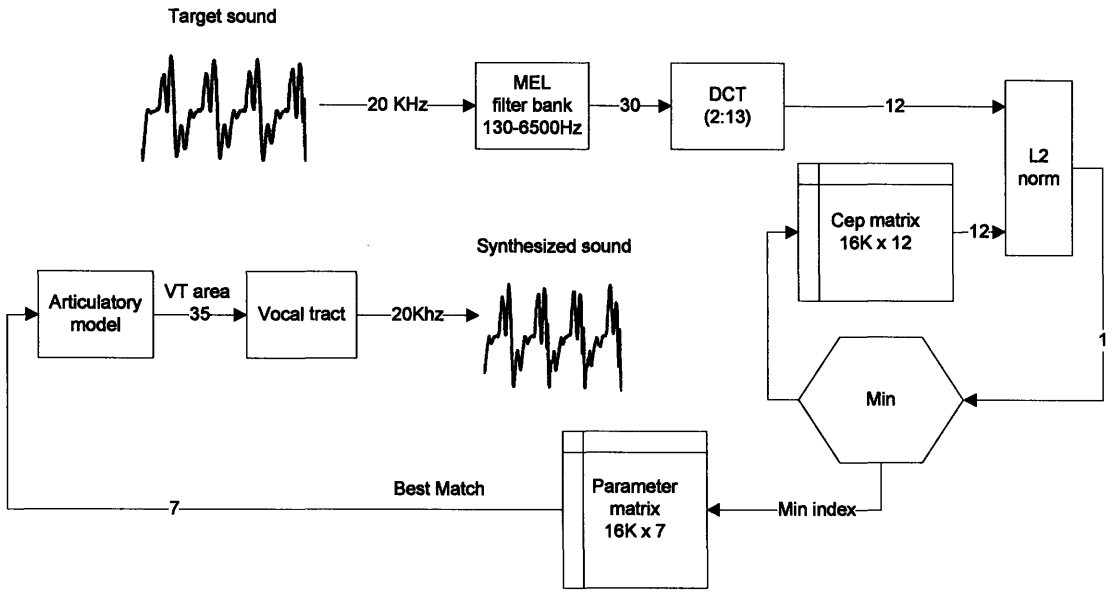


Fig. 3-3: Block diagram illustrating the process of synthesizing a speech signal using an articulatory representation based on the best acoustic match.

3.5 Articulatory trajectory optimization

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, which are not physiological, occurring in the articulatory control. 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 codebook such that the articulatory trajectory is optimized.

In order to evaluate articulatory trajectories between speech frames, a cost c is defined based on the acoustic difference between vocal tract output and target speech, and the articulatory movement between speech frames. The acoustic difference, d(t), between the vocal tract output and the target speech is computed as the L2-norm of the difference in the respective mel-frequency cepstral coefficients (MFCC). The articulatory

movement m(t) may be computed as the variation of the articulators between speech frames or the variation of the area function profile between speech frames. In other words, d(t) is a measure of the acoustic match and m(t) represents the movement of the articulators or area function profile during continuous speech. The latter imposes a continuity constraint on the trajectory of the articulators or area function profile during the optimization process. If we denote the state of the articulators or the area function profile at time t to be A(t), then d(t) and m(t) are given as follows:

$$d(t) = \left| MFCC_{vocaltract}(t) - MFCC_{target}(t) \right|$$
$$m(t) = \left| A(t) - A(t-1) \right|$$

The articulatory trajectory optimization is summarized in Fig. 3-4. Acoustic features are obtained from the incoming speech and compared with entries in the codebook. For every time t, a set of candidate vocal tract profiles for each speech frame corresponding to a given number (N_{best}) of the best acoustic matches is used with the acoustic match derived for each profile. The cost c_{jk} associated with an articulatory transition from vocal tract candidate j to vocal tract candidate k of adjacent speech frames is given by a weighted sum of the acoustic match, d(t), and the articulatory movement, m(t) as follows:

$$c_{j,k}(t) = \alpha d_j(t) + \beta m_{j,k}(t)$$

where the weights α and β are speech dependent and determined empirically. The set of candidate vocal tracts are indexed from 1 to N_{best} and the subscripts j and k denote the j-th and k-th candidate vocal tract shape respectively. Dynamic programming is used to produce a sequence of vocal tract shapes that minimizes function C_{path} given as follows:

$$C_{path} = \sum_{t=0}^{T-1} c_{j(t)\in\{1..Nbest\},k(t)\in\{1..Nbest\}}(t) \quad , \quad \forall k(t) = j(t+1)$$

where the subscript path denotes a sequence of connected moves starting from the first speech frame (t=0) to the last speech frame (t=T-1).

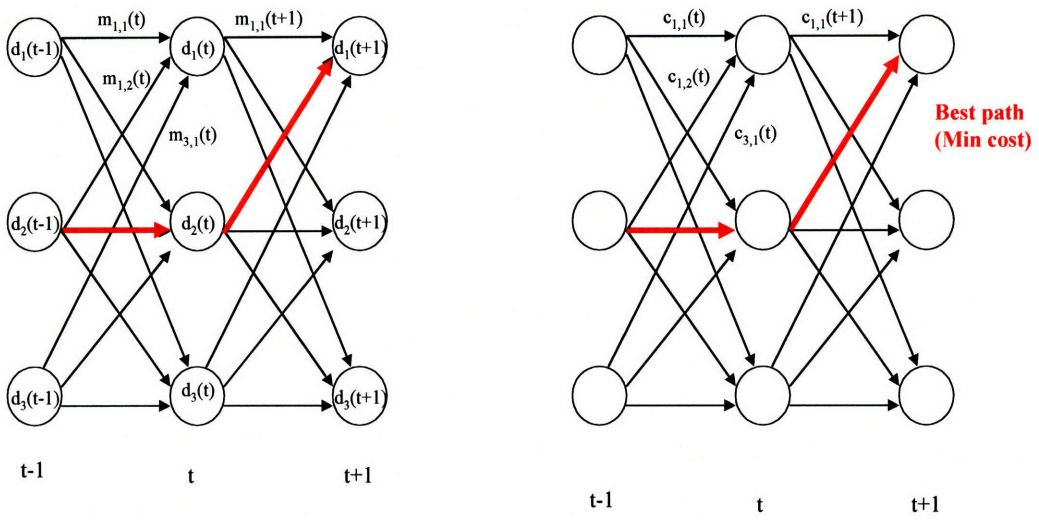
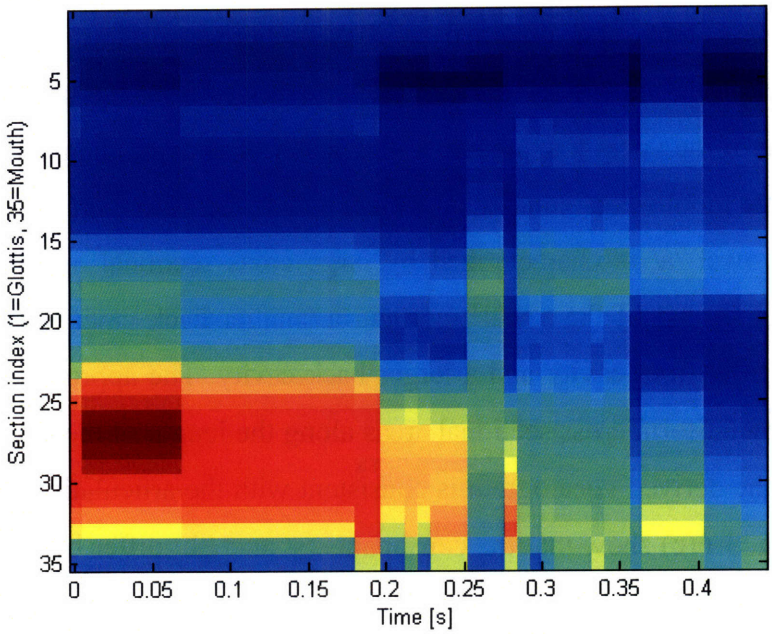


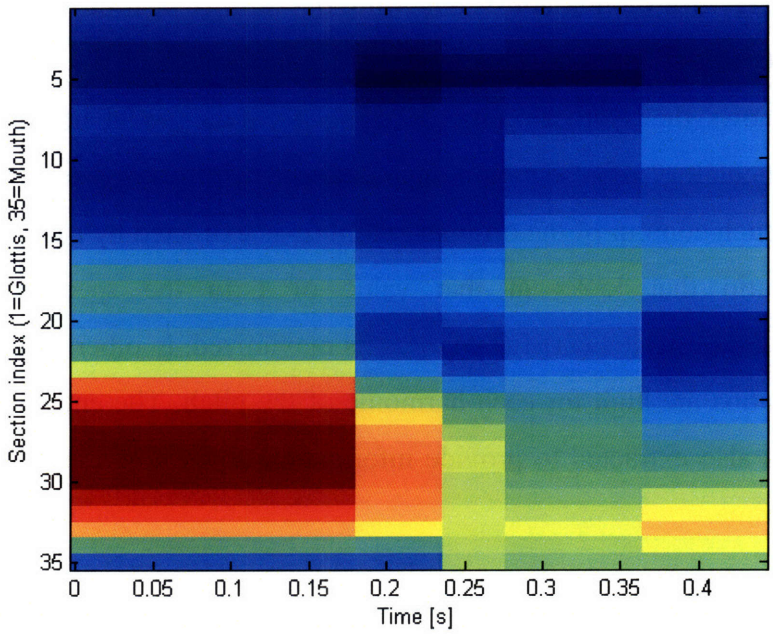
Fig. 3-4: Concept of articulatory trajectory optimization using dynamic programming. The arrows in red indicate the best path i.e. minimum cost.

3.6 Simulation results

Fig. 3-5 compares vocalograms of the diphthong "ae" obtained using an articulatory codebook generated by babbling. A vocalogram is a three dimensional plot that describes the variation of the vocal tract profile with time: 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 cross-sectional areas are color coded; red indicating a large area and blue indicating a small area. The vertical axis denotes the section index; section 1 is the end of the transmission line vocal tract that is closest to the glottis and section 35 is the end closest to the mouth. The horizontal axis denotes time. The vocalogram of Fig. 3-5(a) is derived without articulatory optimization i.e. for every time t, the best articulator configuration at each instant that minimizes the distance d(t) between the synthesized spectrum and the heard spectrum at time t is used. The vocal tract moves in a jittery way since there are no smoothness constraints on its motion. Fig. 3-5(b) shows the vocal tract area profile obtained when articulatory constraints are imposed and the optimal articulatory sequence derived by dynamic programming. The vocal tract movements in Fig. 3-5(b) are clearly smoother.



(a)

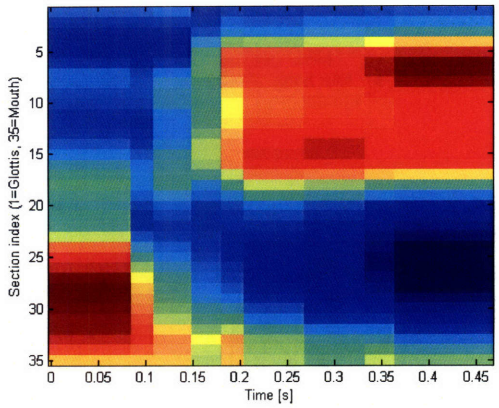


(b)

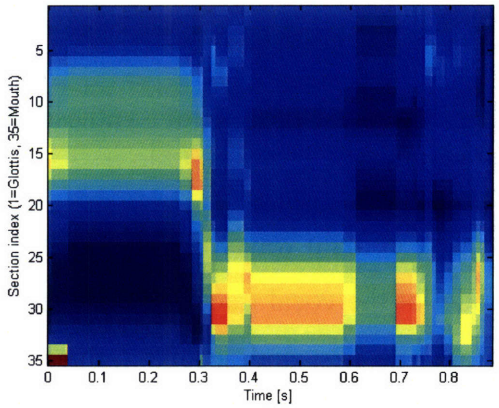
Fig. 3-5: Vocalograms of the diphthong "ae" (a) without articulatory constraints and (b) with articulatory constraints. Regions in red indicate a large cross-sectional area whereas regions in blue indicate a small cross-sectional area.

Fig. 3-6 shows the vocalograms of the vowel transition "aei" and, consonant-to-vowel transitions "sa" and "sha, obtained by dynamic programming with articulatory constraints. For the vowel transition "aei", the tongue moves from the back to the front; initially, narrowing of the vocal tract occurs at the posterior end and then moves forward. Thus, the front part mouth gets smaller and bluer while the back part gets bigger and redder. The low vowel /a/ is characterized by large cross sectional areas in the front cavity near the mouth end of the vocal tract and a smaller back cavity while the high vowel /i/ has a narrow front cavity and a larger back cavity. The approximately neutral vowel /e/ has nearly uniform cross sectional areas along the length of the vocal tract. Fig. 3-6(a) shows that the derived vocalogram is consistent with the articulation of the sound. Fig. 3-6(b) and Fig. 3-6(c) show that the consonants /s/ and /sh/ differ by the location of the supraglottal constriction; with the constriction being slightly more anterior for /s/ than for /sh/. The constriction for /s/ occurs slightly behind the teeth while the constriction for /sh/ is located a little farther back in the oral cavity. Again, the derived vocalograms of Fig. 3-6(b) and Fig. 3-6(c) are consistent with our knowledge of the approximate locations of these constrictions.

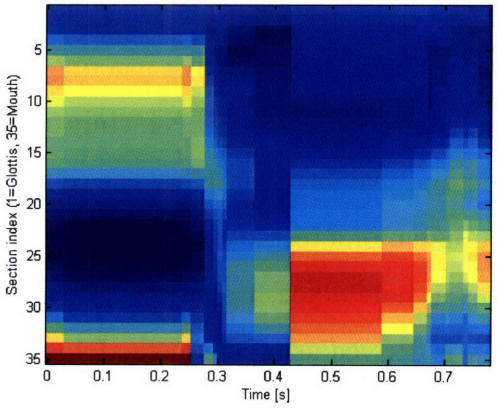
Fig. 3-7 shows the synthesis process using a circuit model of the vocal tract driven with an articulatory representation supplemented with energy and pitch contour information. The pitch contour of the original recording is extracted by using a harmonic-to-subharmonic ratio that looks at that value of f in the spectrum at which the energy sum E(f)-E(1.5f) + E(2f)-E(2.5f) + … is maximized. In other words, the peak-to-valley energy ratios are high for each harmonic of the pitch. The energy envelope of the target sound is also extracted to provide information about how the loudness of the sound varies with time.



(a)



(b)



(c)

Fig. 3-6: Vocalogram of (a) the vowel transition "aei", (b) the consonant-to-vowel transition "sa", and (c) the consonant-to-vowel transition "sha" synthesized by circuit model of vocal tract. Regions in red indicate a large cross-sectional area whereas regions in blue indicate a small cross-sectional area.

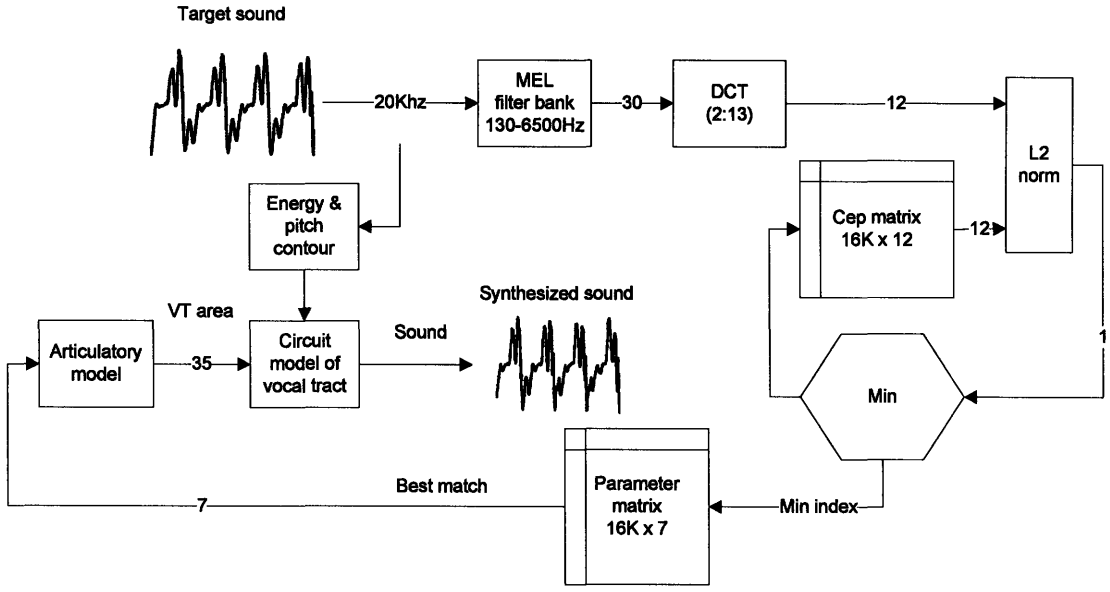


Fig. 3-7: Block diagram illustrating the process of synthesizing a speech signal with energy and pitch contour extraction.

Fig. 3-8(a) shows the time domain waveform and spectrogram of the sound "aei" obtained from our circuit model of the vocal tract implemented in Matlab and driven by the vocalogram of Fig. 3-6(a). The first and second formants (F_1 and F_2) are clearly moving apart as the vowel transitions from /a/ to /e/ to /i/. Fig. 3-9 and Fig. 3-10 show the time domain waveforms and spectrograms of the consonant-to-vowel transitions "sa" and "sha". The fricative energy for "sa" has frequency components that are concentrated above approximately 4kHz whereas the fricative energy for "sha" has frequency components extending down 2kHz. The difference is mainly accounted for by the shorter front cavity (after the constriction) that the fricative /s/ has compared to /sh/.

Fig. 3-11 is the spectrogram of a recording of the word "Boston" lowpass filtered at 5.5kHz. The recording has a male voice. Using the recording as a target sound, the optimal articulatory trajectory is derived using the methodology described in the previous section. The babbles contained in the articulatory codebook are generated using a circuit model of the vocal tract that has a length of 17.5 cm, corresponding to an average adult male. The vocalogram of Fig. 3-12 shows the result of the articulatory trajectory optimization using dynamic programming and it is used to drive our circuit model of the vocal tract to synthesize the sound shown in Fig. 3-13. Comparing the spectrograms of the original recording (Fig. 3-11) and the synthesized sound (Fig. 3-13b), it is evident that

high frequency speech components that were absent in Fig. 3-11 because of the low pass filtering have been re-introduced by the vocal tract. The location and trajectories of the principal formants are also very similar.

The recording of the word "Hello" with a female voice is shown in Fig. 3-14. Using the recording as a target sound, the optimal articulatory trajectory is derived as before, using the same articulatory codebook. The derived vocalogram (Fig. 3-15) drives our circuit model of the vocal tract to synthesize the sound shown in Fig. 3-16. Similarly, Fig. 3-17 depicts the spectrogram of a recording of the word "Technology" lowpass filtered at 5.5 kHz. The recording also has a female voice. The result of the articulatory trajectory optimization is the vocalogram of Fig. 3-18. The time domain waveform and spectrogram of the synthesized sound is shown in Fig. 3-19. In both simulations, the length of the vocal tract was not changed but the extracted pitch was scaled down appropriately for use with a male vocal tract model. Comparison of Fig. 3-14 and Fig. 3-16(b) shows that the trajectories of the principal formants match reasonably well. The same can be said of the spectrograms shown in Fig. 3-17 and Fig. 3-19(b). The principal formant frequencies of the synthesized speech occur at lower frequencies compared to corresponding formants of the recording. For example, the third formant (F_3) of the voiced part of the synthesized speech is located at approximately 2.5 kHz (consistent with the vocal tract length of 17.5cm of our circuit model) compared to about 3 kHz for the recording. The disparity in F_3 location can be explained by the mismatch in vocal tract length between the circuit model and the recording. An estimation of the vocal tract length (L_{VT}) of the female speaker in the recording is given by:

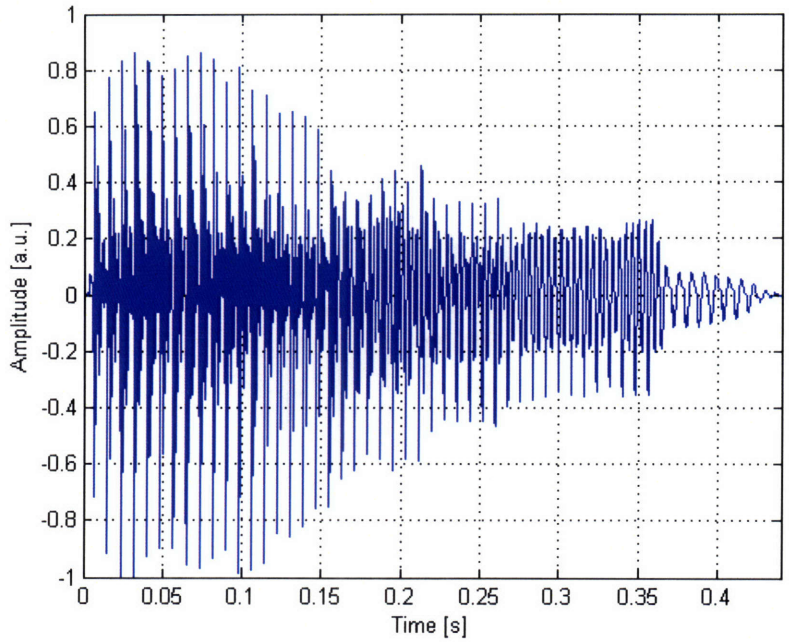
$$F_3 = \frac{5c}{4L_{VT}}$$

$$L_{VT} = \frac{5c}{4F_3}$$

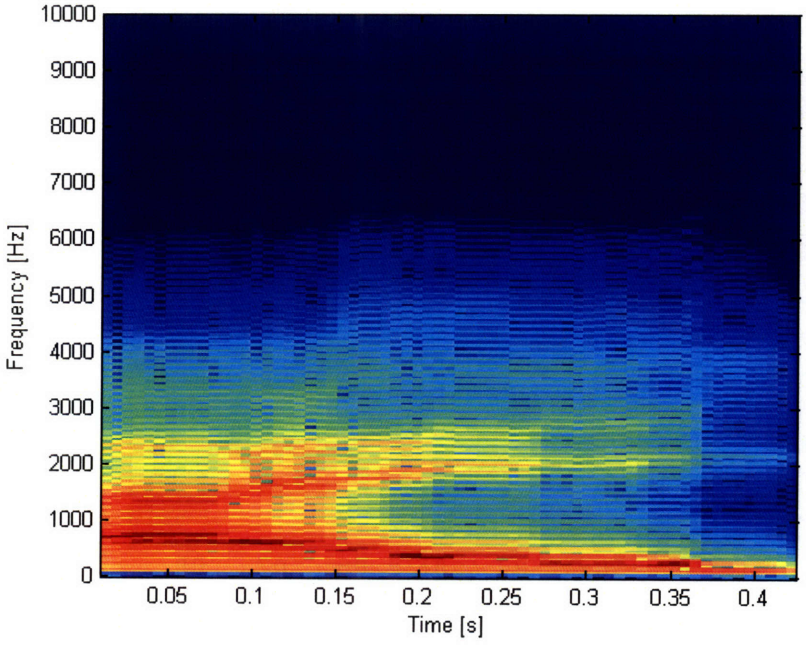
where c denotes the velocity of sound in air and is approximately 35000 cm/s at room temperature. Based on $F_3 \approx 3$ kHz, the female speaker in the recording has a vocal tract length of approximately 14.6cm. The longer vocal tract of the circuit model accounts for the lowering of formant frequencies in the voiced parts of the speech.

The results presented above suggest that our acoustic feature extraction strategy is relatively invariant to pitch scaling, allowing the same articulatory codebook to be used

for recordings of both male and female voices. We will discuss more about varying the length of the vocal tract to accommodate the synthesis of male and female voices in Chapter 6, where a VLSI implementation of the circuit model of the vocal tract is described.

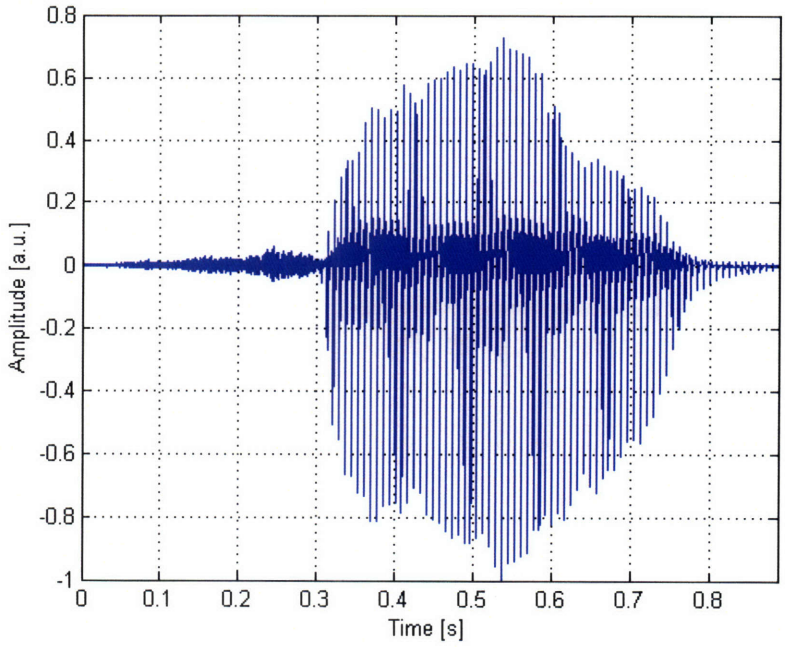


(a)

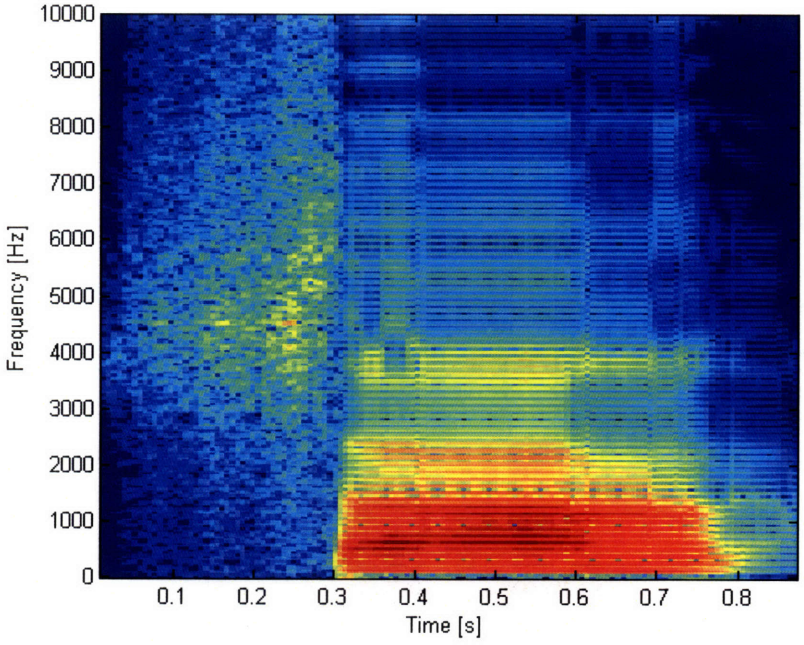


(b)

Fig. 3-8: (a) Time domain waveform and (b) spectrogram of vowel transition "aei" synthesized by circuit model of vocal tract. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

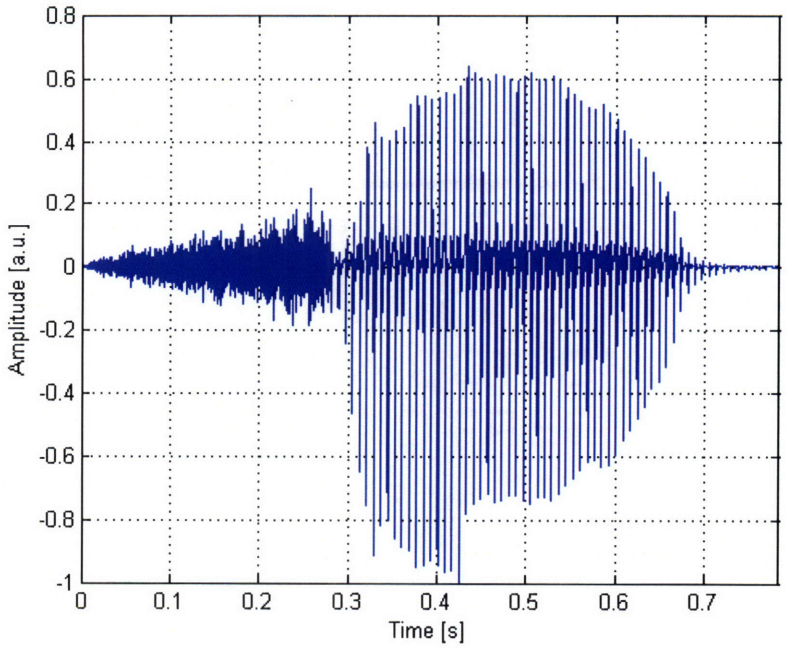


(a)

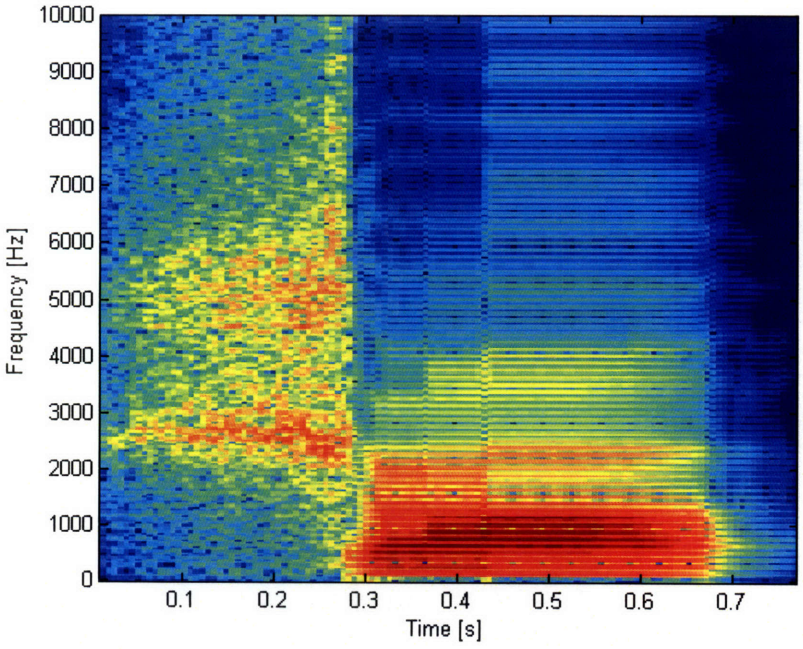


(b)

Fig. 3-9: (a) Time domain waveform and (b) spectrogram of consonant-to-vowel transition "sa". Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.



(a)



(b)

Fig. 3-10: (a) Time domain waveform and (b) spectrogram of consonant-to-vowel transition "sha" synthesized by circuit model of vocal tract. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

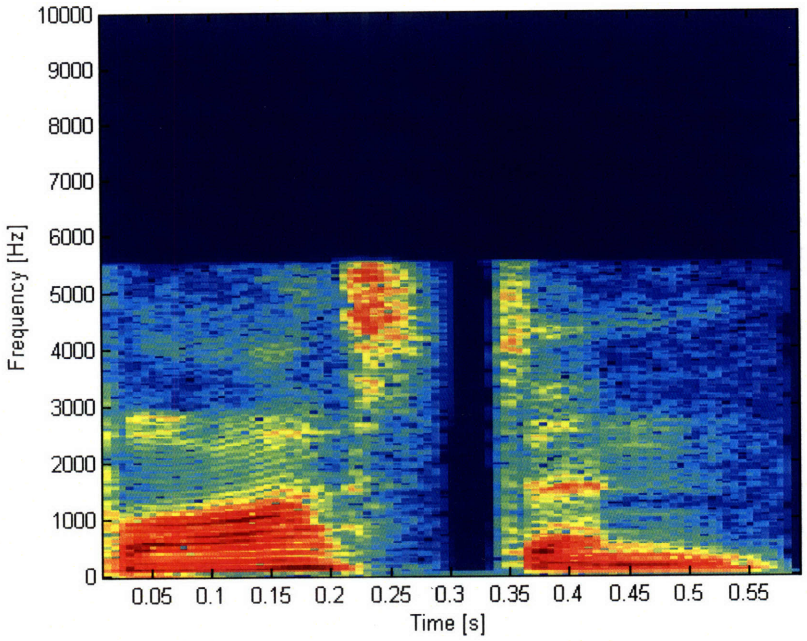


Fig. 3-11: Spectrogram of a recording of the word "Boston" lowpass filtered at 5.5kHz. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

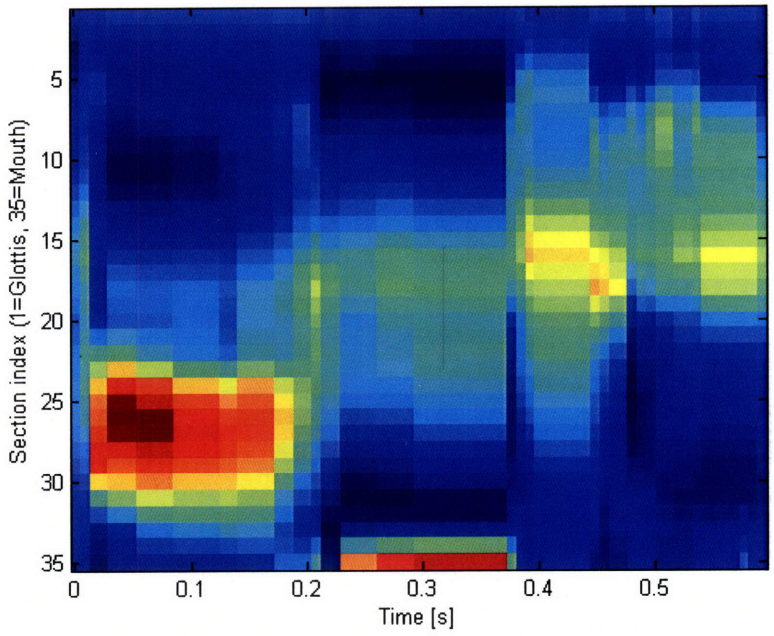
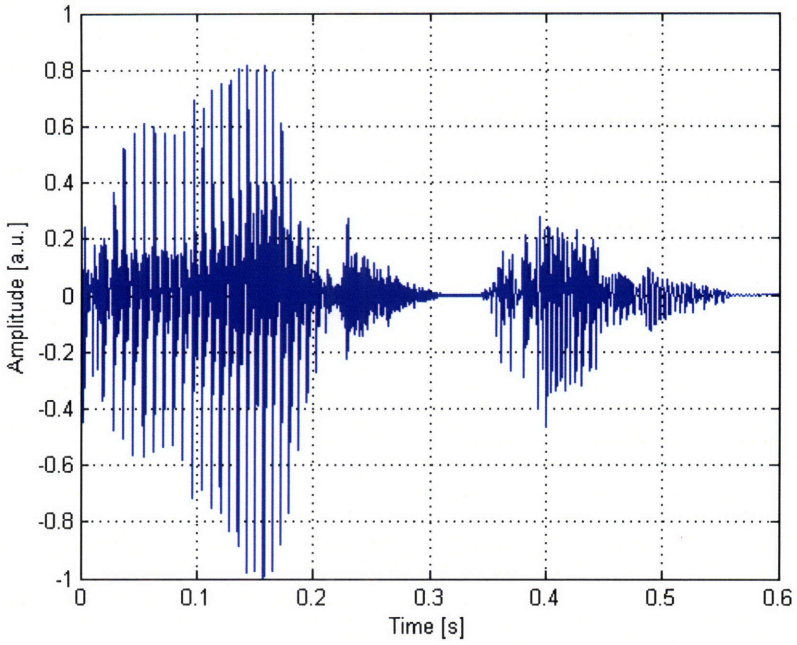
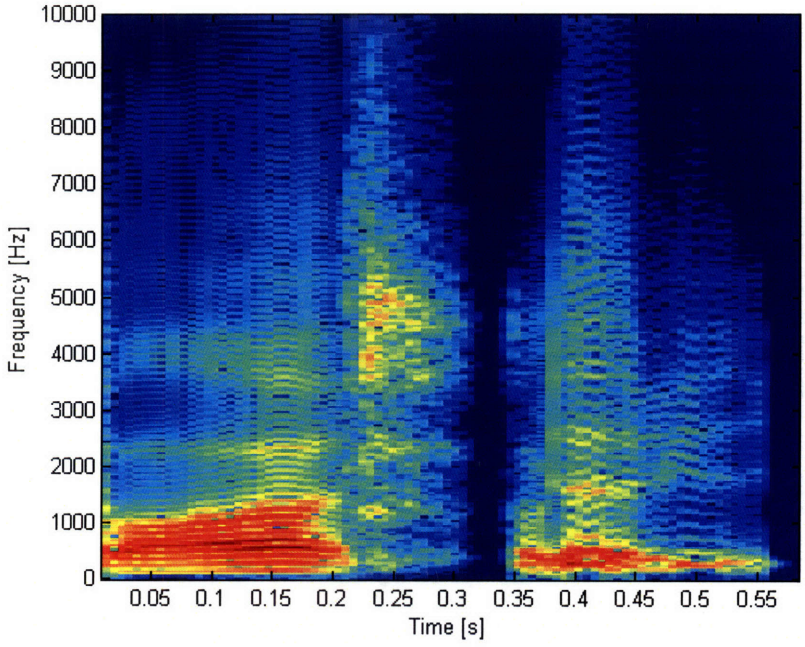


Fig. 3-12: Vocalogram of the word "Boston". Regions in red indicate a large cross-sectional area whereas regions in blue indicate a small cross-sectional area.



(a)



(b)

Fig. 3-13: (a) Time domain waveform and (b) spectrogram of the word "Boston" synthesized by circuit model of vocal tract. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

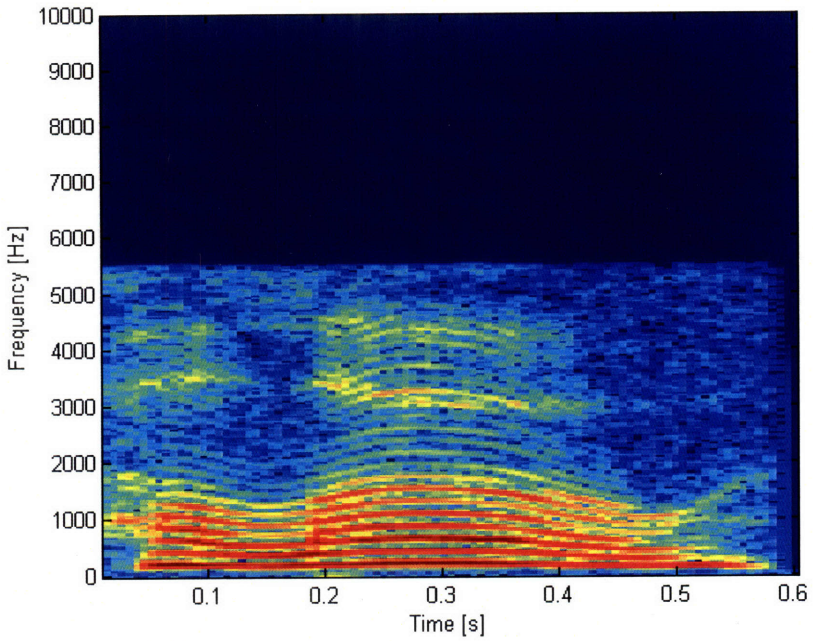


Fig. 3-14: Spectrogram of a recording of the word "Hello" lowpass filtered at 5.5kHz. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

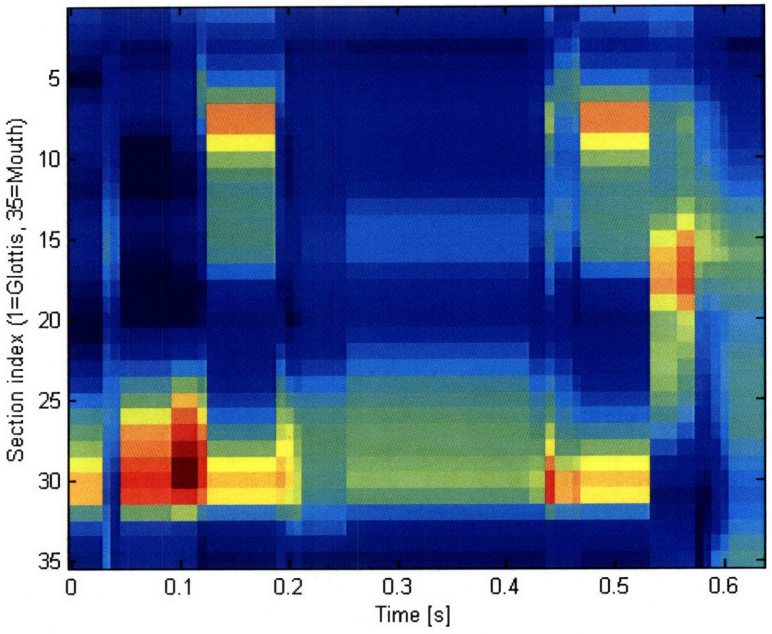
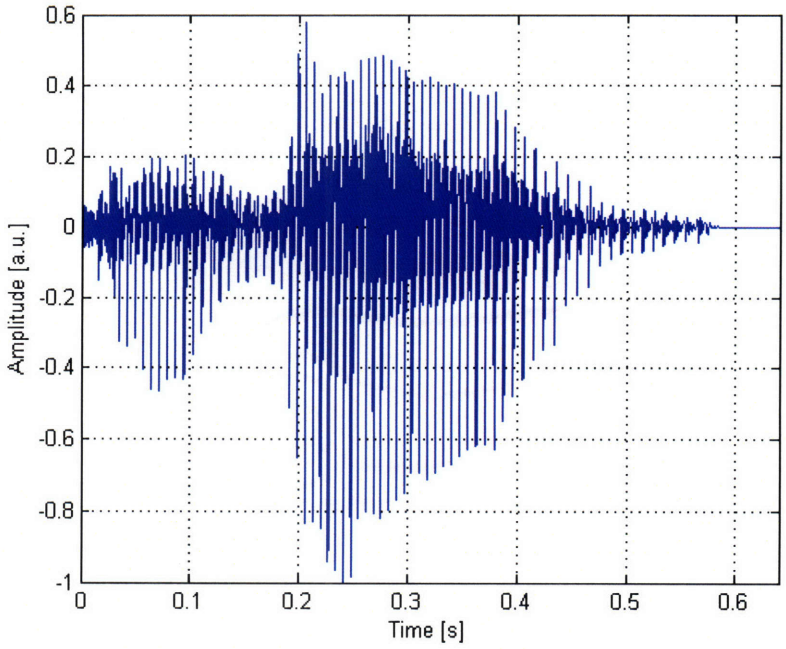
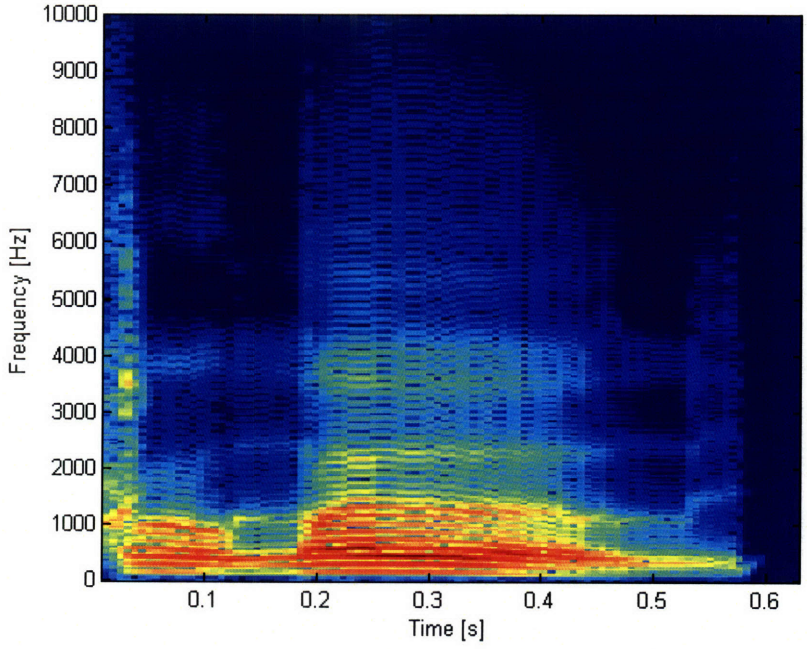


Fig. 3-15: Vocalogram of the word "Hello". Regions in red indicate a large cross-sectional area whereas regions in blue indicate a small cross-sectional area.



(a)



(b)

Fig. 3-16: (a) Time domain waveform and (b) spectrogram of the word "Hello" synthesized by circuit model of vocal tract. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

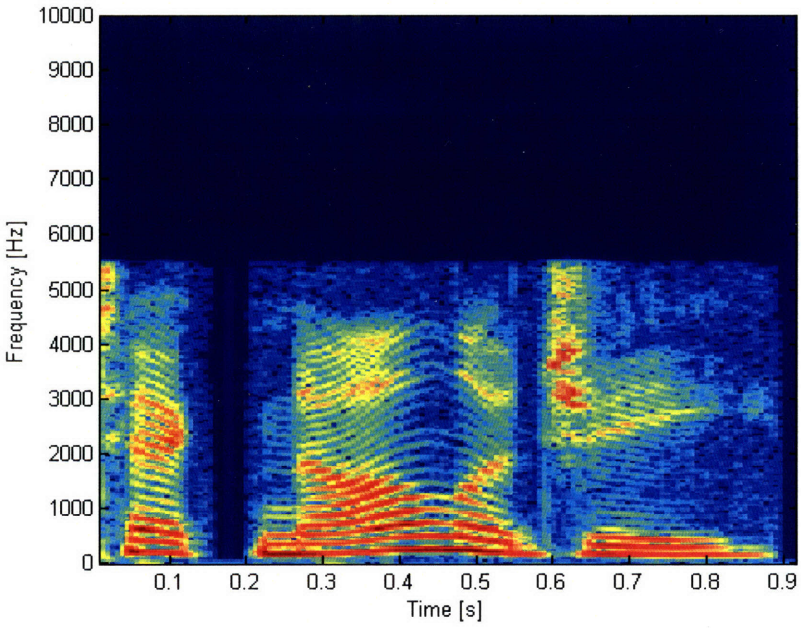


Fig. 3-17: Spectrogram of a recording of the word "Technology" lowpass filtered at 5.5kHz. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

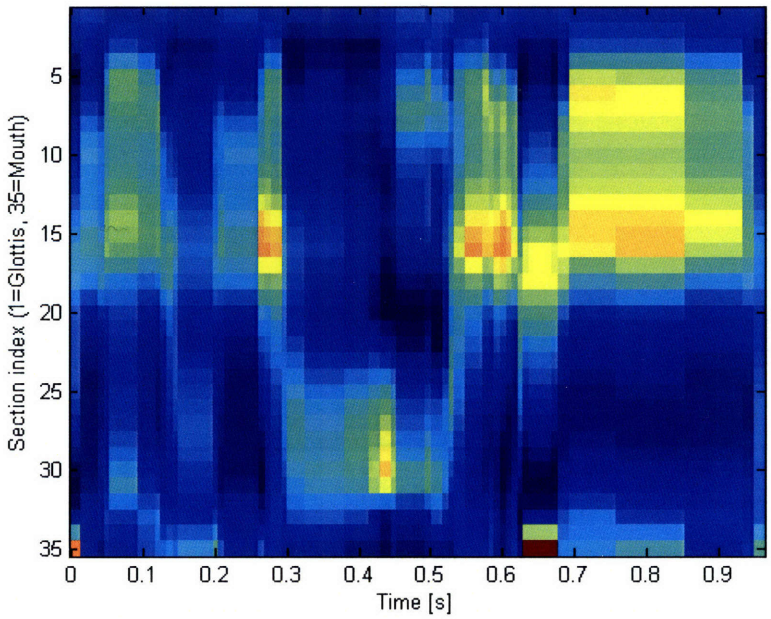
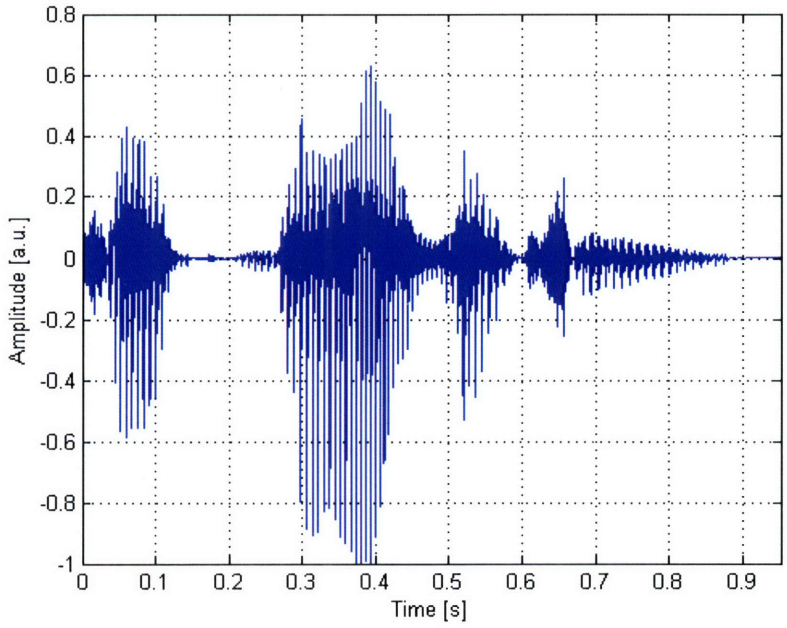
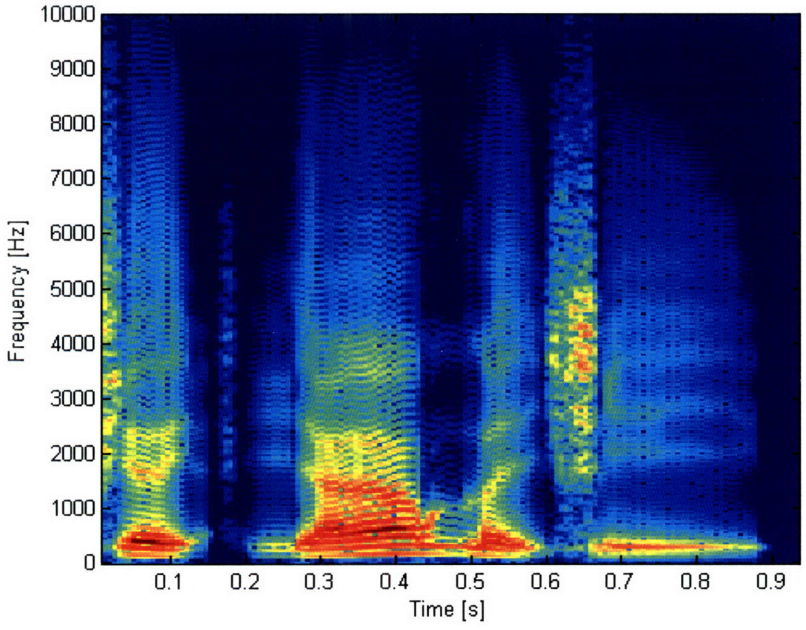


Fig. 3-18: Vocalogram of the word "Technology". Regions in red indicate a large cross-sectional area whereas regions in blue indicate a small cross-sectional area.



(a)



(b)

Fig. 3-19: (a) Time domain waveform and (b) spectrogram of the word "Technology" synthesized by circuit model of vocal tract. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

INTENTIONALLY LEFT BLANK

Chapter 4 LINEAR OR NONLINEAR MOS RESISTORS

In this chapter, we discuss how to implement the glottal constriction resistance as a series combination of linear and nonlinear resistors that model laminar and turbulent flow respectively. Electronically tunable linear resistors are highly versatile circuit elements. They find application in variable gain amplifiers, oscillators, balanced resistive bridges and analog filters. A combination of linear and nonlinear resistances is often useful in creating building blocks in electrical models of physical systems. Electronically tunable resistors may be obtained using MOS transistors. In the past, MOS resistors with approximately linear I-V characteristics were obtained by operating the transistor in the ohmic (triode) region of strong inversion to exploit the resistive nature of the channel. Generally, these approaches were limited by the small ohmic region and its intrinsic non-linearities. Various techniques have been proposed to minimize nonlinear effects associated with operating the MOS transistor in the ohmic strong inversion regime with good results [34][35][36-39][40]. In this chapter, we present a new MOS resistor that does not require triode operation and is valid in weak or strong inversion. In addition, we show that the technique can be applied to produce linear as well as nonlinear resistances.

4.1 Feedforward biasing technique for electronically tunable MOS resistors

In general, resistors with I-V characteristics of the form $I \propto V^n$ may be obtained using devices with an exponential I-V characteristic by re-biasing the exponent such that it becomes a function of the form $\ln(\ )^n$. The concept is illustrated in Fig. 4-1 for a linear I-V relationship with n=1. Bipolar junction transistors (BJT) and MOS transistors operated in weak inversion have I-V characteristics that are exponential in nature. The former cannot produce a bidirectional resistor as it is not a symmetric device with respect to the collector and emitter. On the other hand, MOS transistors have source and drain terminals that are symmetric and gate or bulk voltages that may be varied to provide the

desired characteristic. However, unlike bipolar transistors, regular MOS transistors do not have an ideal exponential I-V characteristic as described below.

In weak inversion, the current I_D through an MOS transistor is dominated by diffusion. We assume an nMOS in the following explanation. The diffusion current is proportional to charge gradient along the length of the device. The charge at the device boundaries, Q_{I0} at the source and Q_{IL} at the drain, are set by the exponential difference in surface potential ψ_s and terminal potentials V_{SB} and V_{DB}. I_D can be expressed explicitly in terms of surface potential as follows:

$$I_D = \mu\phi_t \frac{W}{L}(Q_{I0} - Q_{IL}) \tag{11}$$
$$= \mu\phi_t^2 \frac{W}{L} \frac{\gamma C_{ox}}{2\sqrt{\psi_s}} e^{\frac{\psi_s - 2\phi_F}{\phi_t}} \left(e^{-\frac{V_{SB}}{\phi_t}} - e^{-\frac{V_{DB}}{\phi_t}} \right)$$

where μ is the carrier mobility, ϕ_t is the thermal voltage (kT/q) and C_{ox} is the oxide capacitance per unit gate area. The parameter W/L relates to the aspect ratio of the device.

Note that the voltage on the control terminal V_G is expressed implicitly through $\psi_s(V_{GB})$. To derive the current explicitly in terms of the gate voltage, the bulk-referenced

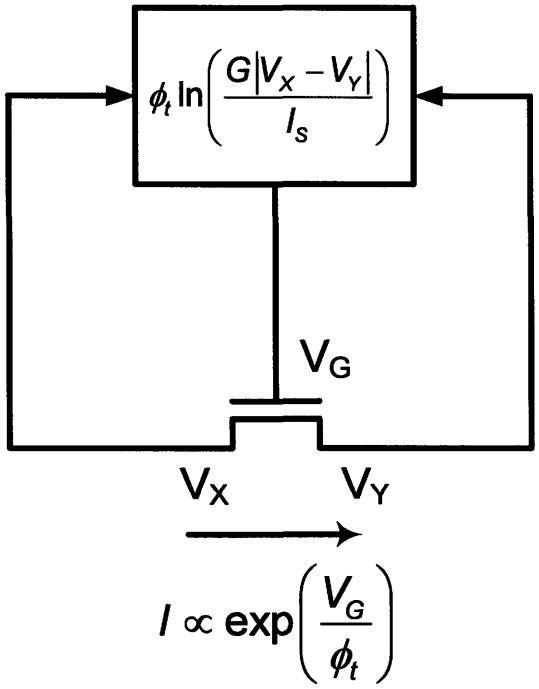


Fig. 4-1: General concept of feedforward biasing technique using an exponential device to derive an arbitrary I-V characteristic.

approximation linearizes the surface potential about an operating point ϕ_0 which is defined as the surface potential when V_G is at the threshold voltage:

$$I_D = I_S e^{\frac{\kappa_0(V_{GB}-V_{T0})}{\phi_t}}\left(e^{-\frac{V_{SB}}{\phi_t}} - e^{-\frac{V_{DB}}{\phi_t}}\right) \tag{12}$$

where I_S is the size-dependent pre-factor ($\mu C_{ox} W/L \phi_t^2 e^{(\phi_0-2\phi_F)/\phi t}$). The parameter ϕ_F is the difference between the intrinsic and Fermi potentials; V_{T0} and κ_0 are the threshold voltage and the subthreshold exponential parameter when $V_{SB}=0$ respectively. The subthreshold exponential parameter κ_0 is defined as the change in surface potential with respect to the gate i.e. the slope at the operating point ϕ_0. Specifically, κ_0 is given by:

$$\kappa_0 = \frac{1}{1+\frac{\gamma}{2\sqrt{\phi_0}}} \tag{13}$$

In the saturation regime of the transistor, $|V_{DB}| >> |V_{SB}|$, hence the current through the transistor I_{sat} is given by:

$$I_{sat} \approx I_S e^{\frac{\kappa_0 V_{GB}}{\phi_t}}\left(e^{-\frac{V_{SB}}{\phi_t}}\right) \tag{14}$$

In order for the transistor to behave as a linear device such that $I_{sat}=G(V_X-V_Y)$, the following relationship must be realized:

$$\frac{\kappa_0 V_{GB} - V_{SB}}{\phi_t} = \ln\left(\frac{G(V_X - V_Y)}{I_S}\right) \tag{15}$$

Hence, the potential V_{GB} on the control terminal must be made as follows:

$$V_{GB} = \frac{V_{SB}}{\kappa_0} + \frac{\phi_t}{\kappa_0}\ln\left(\frac{V_X - V_Y}{\phi_t}\right) \tag{16}$$

4.1.1 Transistor with constant V_{GB}

A MOS transistor with constant gate-to-bulk voltage V_{GB} has an almost ideal exponential I-V characteristic. Such a device may be obtained by inserting a source follower between the gate and bulk terminals of a regular PMOS transistor. The source follower serves to clamp the potential difference V_{GB} across the two terminals. A PMOS transistor is chosen for this purpose because only PMOS devices have floating bulks in a standard CMOS process. Naturally, the principles and techniques described in the

following may be applied in the same manner to nMOS devices in non-standard twin-well CMOS processes.

A PMOS transistor with its gate driving the bulk through source follower action can be modeled as follows. Using the bulk-referenced model of the PMOS transistor, the current I_D through the device is given by:

$$I_D = I_{OP}\exp\left(-\frac{\kappa_0\left(V_{GB}+|V_{T0}|\right)}{\phi_t}\right)\left(\exp\left(\frac{V_{SB}}{\phi_t}\right)-\exp\left(\frac{V_{DB}}{\phi_t}\right)\right) \tag{17}$$

The source follower produces a constant V_{GB} across the gate and bulk of the transistor such that $V_B=V_G+V_{BG}$ where $V_{BG}>0$. When the source follower is biased with a current larger than the maximum current allowed through the transistor, V_B is automatically at a higher potential than the source of the transistor. This condition ensures that all pn junctions are reverse biased. From (17),

$$\begin{aligned} I_D &= I_{OP}\exp\left(\frac{\kappa_0\left(V_{BG}-|V_{T0}|\right)}{\phi_t}\right)\exp\left(-\frac{\left(V_G+V_{BG}\right)}{\phi_t}\right)\left(\exp\left(\frac{V_S}{\phi_t}\right)-\exp\left(\frac{V_D}{\phi_t}\right)\right) \\ &= I_{OP}\exp\left(\frac{(\kappa_0-1)V_{BG}-\kappa_0|V_{T0}|}{\phi_t}\right)\exp\left(-\frac{V_G}{\phi_t}\right)\left(\exp\left(\frac{V_S}{\phi_t}\right)-\exp\left(\frac{V_D}{\phi_t}\right)\right) \\ &= I_{OP}'\exp\left(-\frac{V_G}{\phi_t}\right)\left(\exp\left(\frac{V_S}{\phi_t}\right)-\exp\left(\frac{V_D}{\phi_t}\right)\right) \end{aligned} \tag{18}$$

where I_{OP}' is a pre-exponential constant representing the constant terms in the second equality:

$$I_{OP}' = I_{OP}\exp\left(\frac{(\kappa_0-1)V_{BG}-\kappa_0|V_{T0}|}{\phi_t}\right) \tag{19}$$

Note that in this configuration the device has an ideal exponential current characteristic i.e. the subthreshold exponential parameter is unity. Intuitively, the bulk can be thought of being the *back-gate* of the MOS transistor that has a $1-\kappa_0$ effect on the surface potential as opposed to κ_0 from the gate. Since the gates and the bulks move in tandem through source follower action, the effective subthreshold exponential parameter becomes very close to unity. Fig. 4-2 shows the measured I-V curve of a PMOS transistor whose bulk is driven by its gate through a source follower. The figure shows that the drain current changes by a decade in magnitude for every 61mV change in V_{GS} i.e. the effective subthreshold exponential parameter is very close to unity. This configuration is also

useful for creating current mirrors with tunable gains: Since the pre-exponential factor I_{OP}' is a function of V_{GB}, the current gain of the mirror can be varied through the biasing current I_B of the source follower.

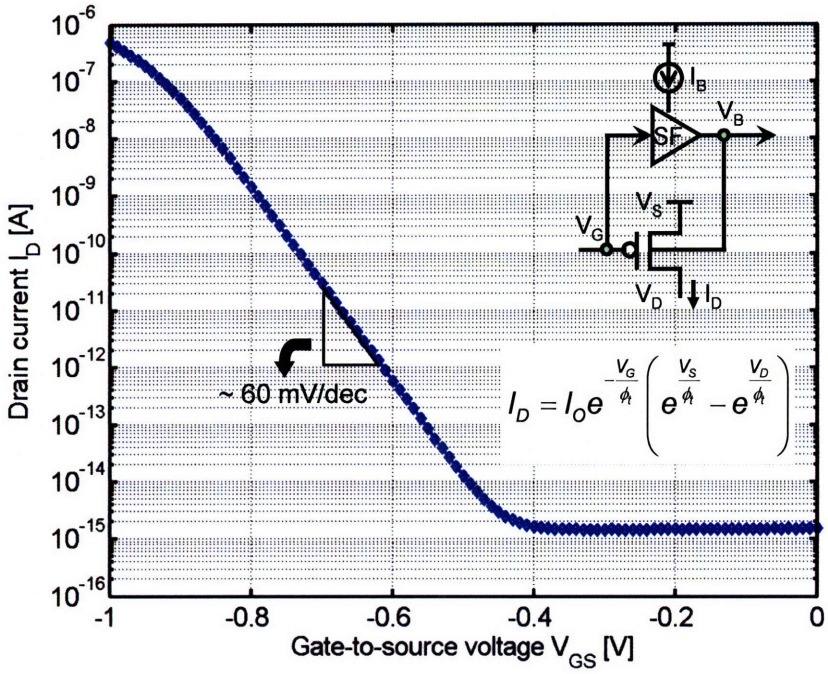


Fig. 4-2: I-V characteristic of a PMOS transistor with its gate driving the bulk through source follower action.

4.1.2 MOS resistor with feedforward biasing technique

A bidirectional resistor may be obtained using a PMOS transistor with a constant V_{GB} as the exponential device in Fig. 4-1. The circuit for a general MOS resistor is shown in Fig. 4-3. It is composed of an absolute value section, a translinear loop that produces a logarithmic voltage output V_1, re-biasing circuitry and a PMOS with constant V_{GB} as the exponential device. In implementation shown in Fig. 4-3, the absolute value circuit consists of two OTAs that operate as transconductance amplifiers. As the two OTAs have same inputs V_X and V_Y connected to their input terminals, the input differential pair may be combined and shared. Hence, the two OTAs are biased by the same current source I_{GM}. The resulting input current I_{in} to the translinear loop is given by:

$$I_{in} = G_M |V_X - V_Y| \tag{20}$$

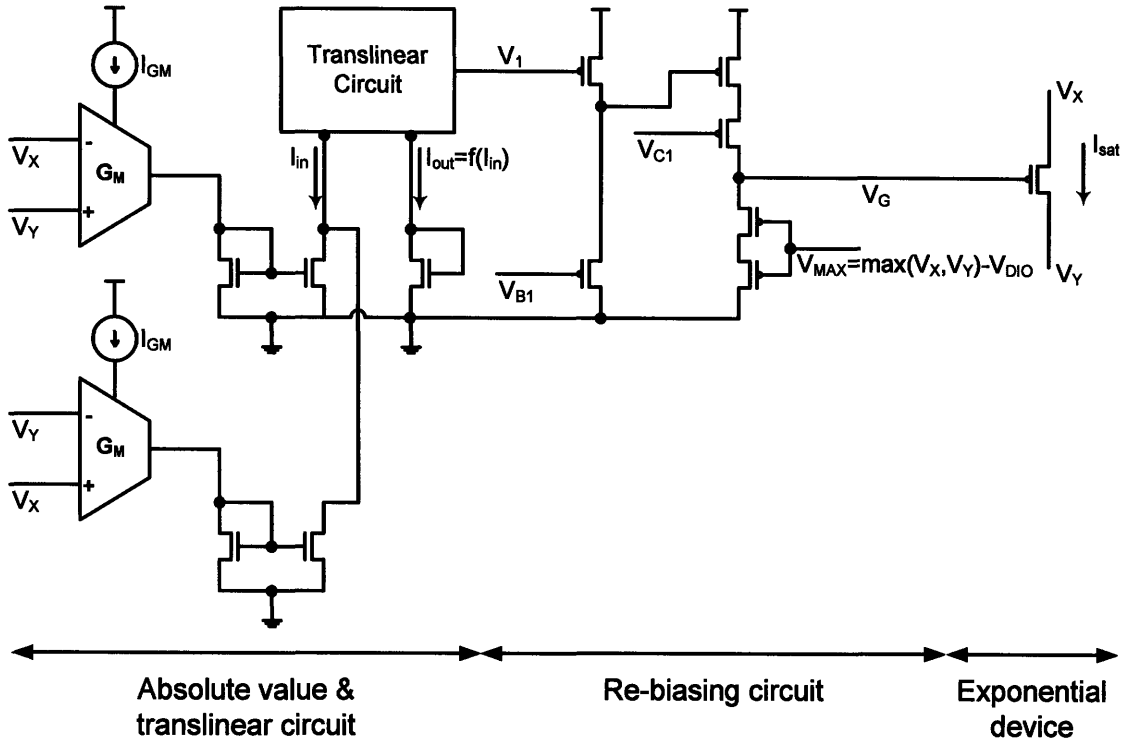


Fig. 4-3: Circuit schematic of MOS resistor using feedforward biasing.

The output current I_{out} of the translinear circuit is a function of its input I_{in}. In the following, a linear MOS resistor is used as an example with no loss of generality as the concept may be easily extended to include nonlinear resistors with compressive (e.g., $I \propto \sqrt{V}$) or expansive (e.g., $I \propto V^2$) characteristics by choosing an appropriate translinear circuit. For a linear MOS resistor, a translinear circuit with the following input-output relationship is used:

$$I_{out} = I_{in} \tag{21}$$

For a resistor with a compressive I-V relationship such as $I \propto \sqrt{V}$, a translinear circuit with the following input-output relationship would be appropriate:

$$I_{out} = \sqrt{I_{REF} I_{in}} \tag{22}$$

where I_{REF} is a reference current.

For the linear MOS resistor, the translinear circuit is a simple current mirror. The logarithmic voltage V_1 produced at the gate of the translinear current mirror is given by:

$$\frac{V_1}{\phi_t} = \ln \frac{I_{out}}{I_{OP}'} = \ln\left(\frac{I_{in}}{I_{OP}'}\right) = \ln\left(\frac{G_M |V_X - V_Y|}{I_{OP}'}\right) \tag{23}$$

Rewriting,

$$\frac{V_1}{\phi_t} = \ln\frac{|V_X - V_Y|}{\phi_t} + \ln\frac{G_M\phi_t}{I_{OP}'} \tag{24}$$

The re-biasing circuit uses V_1 and V_{MAX}=max(V_X,V_Y) as inputs to produce an output V_G that is given by:

$$\begin{aligned}\frac{V_G}{\phi_t} &= \frac{\max(V_X, V_Y) - V_1}{\phi_t} \\ &= \frac{V_{MAX}}{\phi_t} - \ln\frac{|V_X - V_Y|}{\phi_t} - \ln\frac{G_M\phi_t}{I_{OP}'}\end{aligned} \tag{25}$$

Under these biasing conditions, when the exponential device is in saturation, the current through the device I_D=I_{sat} is:

$$\begin{aligned}I_{sat} &= I_{OP}' \exp\left(\frac{V_{MAX} - V_G}{\phi_t}\right) \\ &= G_M|V_X - V_Y|\end{aligned} \tag{26}$$

Hence, we obtain an I-V characteristic that is linear.

The voltage V_{MAX} is generated using a maximum (max) circuit. The schematic of the max biasing circuit is depicted in Fig. 4-4. The max circuit can be viewed as having three parts (all transistors operate in weak inversion): 1) N1, N2, and N3 form a Wilson-mirror-like configuration that forces equalization of the currents through N1 and N2 via negative feedback 2) P1, P2, P3, and P4 form a three-arm differential-pair-like configuration with P4 providing the bias current and diode-connected P3 serving to create the final output voltage of the circuit. When equilibrated, the circuit automatically adjusts itself to have equal current through all arms of the differential-pair-like configuration 3) N4 and N5 serve as a cascode mirror for conveying the current flowing through N2 or N1 to the output transistor P3 such that its output voltage V_{out} is at the maximum of V_{in1} or V_{in2}.

Suppose $V_{in2} < V_{in1}$. Then, the larger current initially flowing through the P2 arm of the differential-pair-like circuit will force P2 to operate in its triode regime as the Wilson mirror functions to equalize the currents through N2 and N1 via negative feedback action on V_A. The current through N2 is then mirrored to the output arm via N4 and N5 such that P3 also conducts a current equal to that flowing through N2 or N1. Since P3 is diode connected, its voltage V_{out} will be very near V_{in1}, the gate voltage of the

other saturated or non-triode transistor conducting the same current as P3. Thus, Vout follows the larger of the two input voltages in this case.

Suppose $V_{in1} < V_{in2}$. Then, the larger current initially flowing through the P1 arm of the differential-pair-like circuit will force P1 and N3 to operate in their triode regimes as the Wilson mirror functions to equalize currents through N2 and N1 via negative feedback action on V_A. The current through N2 is then mirrored to the output arm via N4 and N5 such that P3 also conducts a current equal to that flowing through N2 or N1. Since P3 is diode connected, its voltage V_{out} will be very near V_{in2}, the gate voltage of the other saturated or non-triode transistor conducting the same current as P3. Thus, V_{out} follows the larger of the two input voltages in this case as well. The output characteristic of the max circuit is shown in Fig. 4-5. The output reproduces the higher of the two input voltages faithfully with a maximum error of 5mV when V_{in} is swept over a 5V range.

The DC characteristics of the linear MOS resistor using feedforward biasing is shown in Fig. 4-6. The G_M of the OTA is biased at two different current levels (Ibias=10nA and Ibias=20nA) to produce two different slopes on the linear I-V plot of Fig. 4-6(a), corresponding to two different resistance values. The logarithmic I-V plot of Fig. 4-6(b) show that the resistance is linear when the potential difference $|V_X-V_Y|$ across the exponential MOS device is greater than 100mV. In this regime, the exponential MOS device is operating in the saturation region. As the potential difference across the device falls below 100mV, the MOS device enters the triode region and the slope of the logarithmic I-V curve increases. The change in slope is because when the device is out of saturation, the current I_D through the device becomes a function of both the gate-to-source voltage as well as the gate-to-drain voltage. In other words, assuming $V_X>V_Y$ and V_X-V_Y becomes small such that the MOS exponential device operates in the triode region, the approximation $I_D= I_{sat}$ is no longer valid and I_D is given by:

$$I_D = I_{sat}\left(1-e^{\frac{V_Y-V_X}{\phi_t}}\right) = G_M\left(V_X-V_Y\right)\left(1-e^{\frac{V_Y-V_X}{\phi_t}}\right) \tag{27}$$

Applying the Taylor series expansion on the exponential term gives the following approximation when $V_X-V_Y << \phi_t$:

$$I_D \approx G_M \phi_t (V_X - V_Y)\left[1 - 1 + (V_X - V_Y)\right] \tag{28}$$
$$= G_M \phi_t (V_X - V_Y)^2$$

Fig. 4-6(b) confirms that as V_X-V_Y decreases, the I-V relation departs from being linear and becomes a square (slope of the logarithmic I-V curve approaches 2), as predicted by our analysis above. Hence, the observed DC characteristic is consistent with theory. It is also clear that the linear MOS resistor will introduce cross-over distortion when operated near the origin. The inherent drawback arises because the MOS exponential device is exponential with respect to its gate-to-source voltage only when operated in saturation. In the triode regime, the effect of the gate-to-drain potential on the current becomes significant and sets a limit on the linear range (the range of $|V_X$-$V_Y|$ over which the I-V curve is linear) of the resistor.

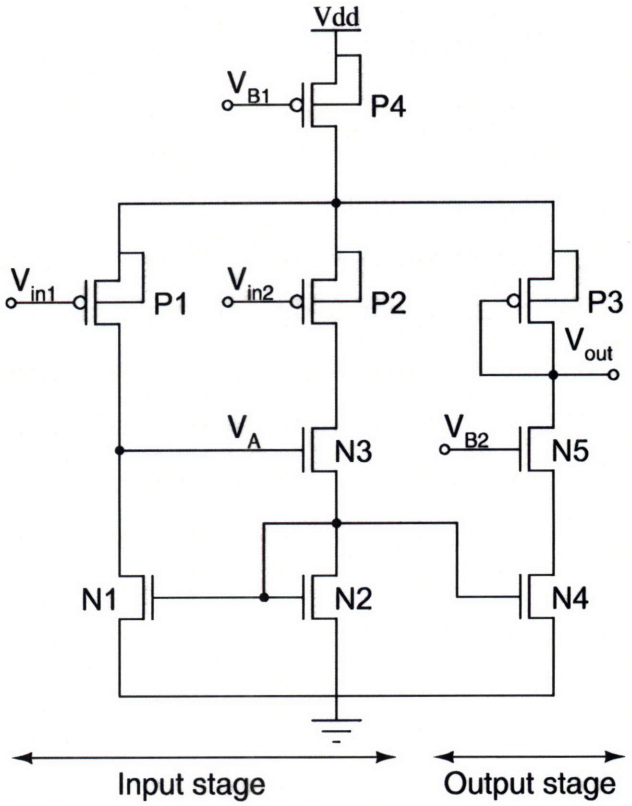


Fig. 4-4: Schematic of maximum circuit that uses a Wilson feedback topology.

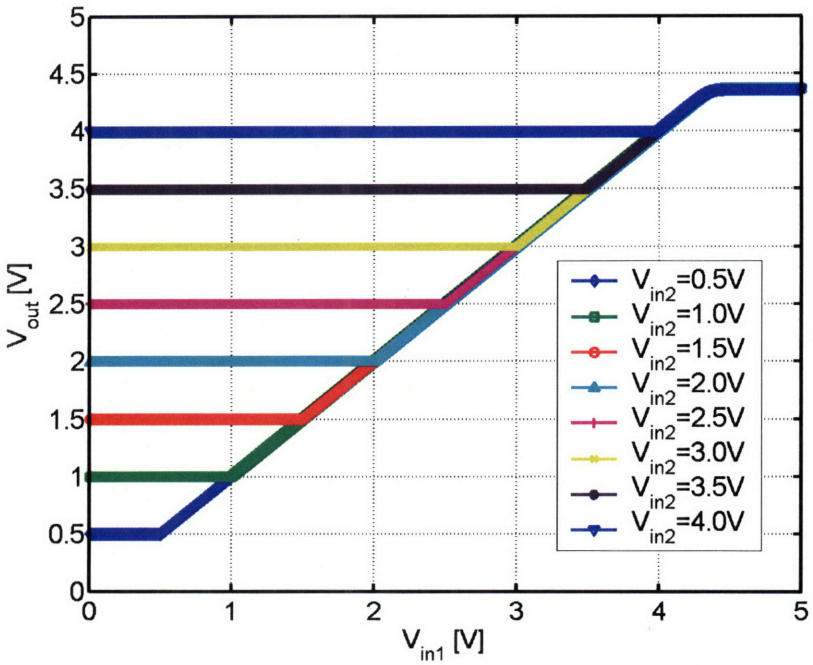
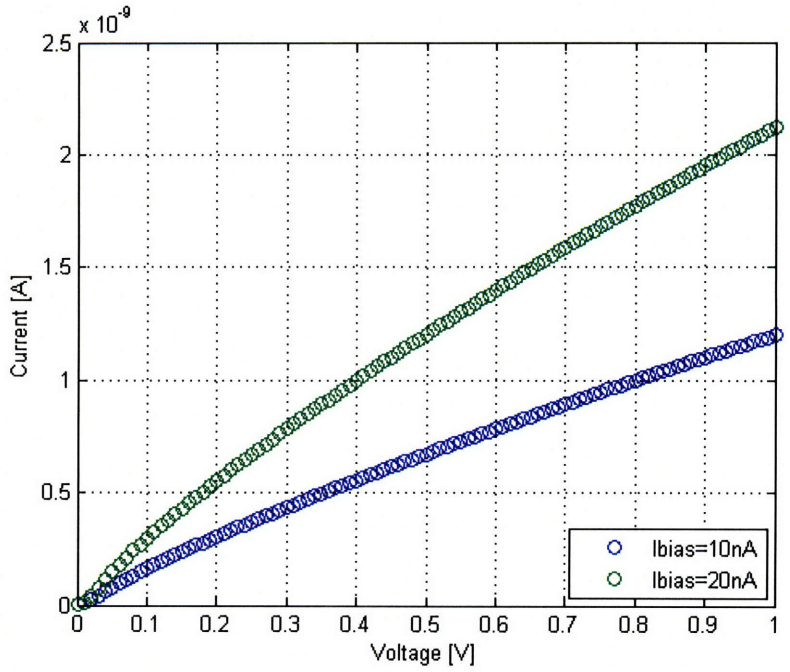
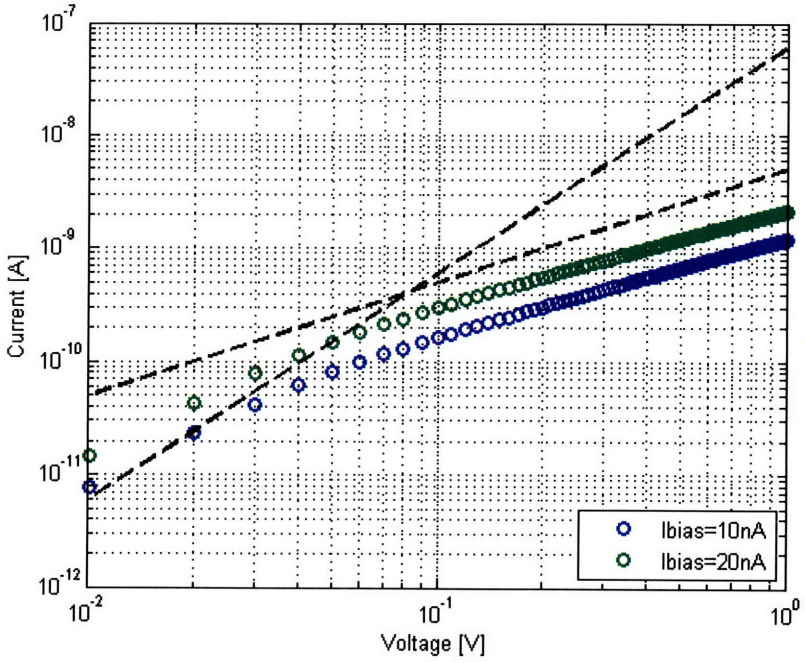


Fig. 4-5: Measured output characteristics of maximum circuit that uses a Wilson-feedback topology.



(a)



(b)

Fig. 4-6: DC characteristics of linear MOS resistor using feedforward biasing plotted on (a) linear and (b) logarithmic axes. The horizontal axis shows the potential difference $|V_X-V_Y|$ across the exponential MOS device. The vertical axis shows the current through the exponential MOS device.

4.2 Feedback biasing technique for electronically tunable linear or nonlinear resistors using MOS transistors

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. 4-7 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. 4-7 (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. 4-7 (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. 4-7 (b). The current I_D through an MOS device may be modeled using the following well-known bulk-referenced expressions:

Weak inversion: (29)

$$I_D = I_O \exp\left(\frac{\kappa_0 (V_G - V_{T0})}{\phi_t}\right)\left(\exp\left(\frac{-V_X}{\phi_t}\right) - \exp\left(\frac{-V_Y}{\phi_t}\right)\right)$$

Strong inversion:

$$I_D = \frac{\kappa_0 \mu C_{ox}}{2}\frac{W}{L}\left[\left(V_G - V_{T0} - \frac{V_X}{\kappa_0}\right)^2 - \left(V_G - V_{T0} - \frac{V_Y}{\kappa_0}\right)^2\right]$$

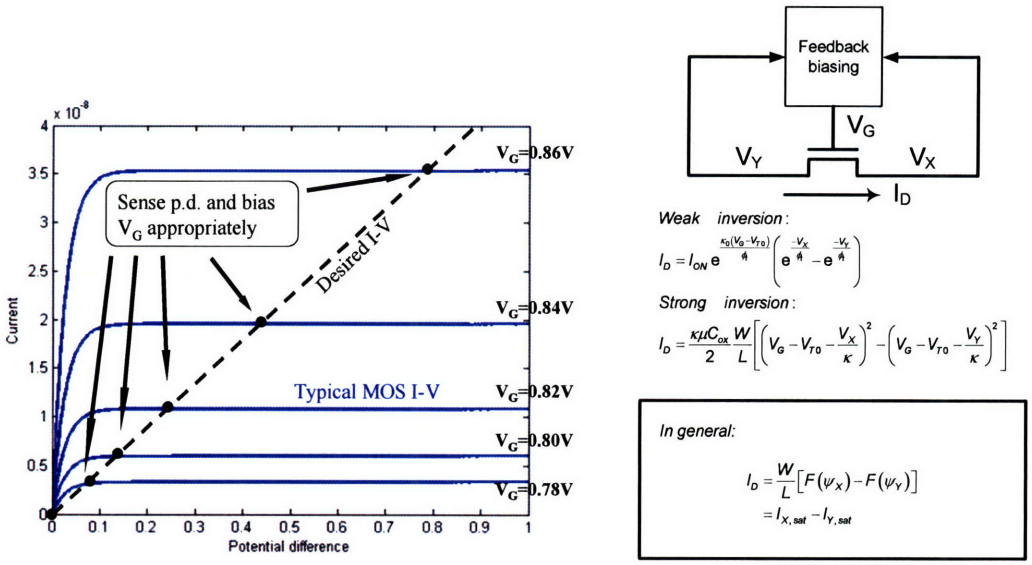
where I_O and ϕ_t are the size-dependent pre-factor and the thermal voltage (kT/q), respectively, and V_{T0} and κ_0 are the threshold voltage and the subthreshold exponential parameter when V_{BS}=0, respectively. Specifically, κ_0 is given by:

$$\kappa_0 = \frac{1}{1 + \frac{\gamma}{2\sqrt{\phi_0}}} \tag{30}$$

where γ is the body effect factor and ϕ_0 corresponds to the surface potential at V_{GB}=V_{T0}. Equation (29) 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 [41]:

$$I_D = I_{X,sat} - I_{Y,sat} \tag{31}$$

where $I_{X,sat}$ and $I_{Y,sat}$ are forward and reverse saturation currents determined by V_{GX} and V_{GY}, the gate-to-source and gate-to-drain potentials respectively.



(a) (b)

Fig. 4-7: (a) Idea behind MOS resistor and (b) its biasing concept.

For the MOS device to behave like a resistor with an arbitrary I-V characteristic given by

$$I_D = g(V_{XY}) \tag{32}$$

where g() denotes an arbitrary function and the argument V_{XY} denotes the potential difference across the source-drain terminals (V_X-V_Y), an appropriate V_G must be applied to the gate terminal such that:

$$I_D = I_{X,sat} - I_{Y,sat} = g(V_{XY}) \tag{33}$$

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 (33).

Fig. 4-8 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. The potential difference V_X-V_Y across the main MOS device M_R is sensed and converted into a current $I_{OUT,GM}$ using a wide linear range operational transconductance amplifier (WLR OTA) such as that described in [42]. $I_{OUT,GM}$ is linearly related to the sensed input voltages as follows:

$$I_{OUT,GM} = G_M (V_X - V_Y) \tag{34}$$
$$= G_M V_{XY}$$

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

$$G_M = \frac{I_{GM}}{V_L} \tag{35}$$

where I_{GM} and V_L are the biasing current and input linear range of the WLR OTA respectively. Hence, G_M is electronically tunable via I_{GM}. In Fig. 4-8, the two WLR OTAs in conjunction with diode connected transistors M_1 and M_3 produce two half-wave rectified currents that are proportional to $|V_{XY}|$ across the source-drain terminals of M_R with each current being non-zero if and only if $V_{XY} > 0$ or $V_{XY} < 0$ respectively. The rectified output currents are mirrored via M_2 or M_4 to create a full wave rectified current I_{in}. The translinear circuit produces an output current I_{out} that is a function of I_{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. 4-9.

The saturation currents $I_{X,sat}$ and $I_{Y,sat}$ of M_R are proportionally replicated by sensing 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-source terminals of M_X and M_Y. The source followers marked SF in Fig. 4-8 serve as buffers to prevent loading on M_R. Transistors M_7-M_{14} serve to compute $I_{X,sat}$-$I_{Y,sat}$ or $I_{Y,sat}$-$I_{X,sat}$ and transistors M_{15}-M_{20} compare $|I_{X,sat}$-$I_{Y,sat}|$ with a mirrored version of the translinear output current $I_{out} = f(I_{in})$. 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.

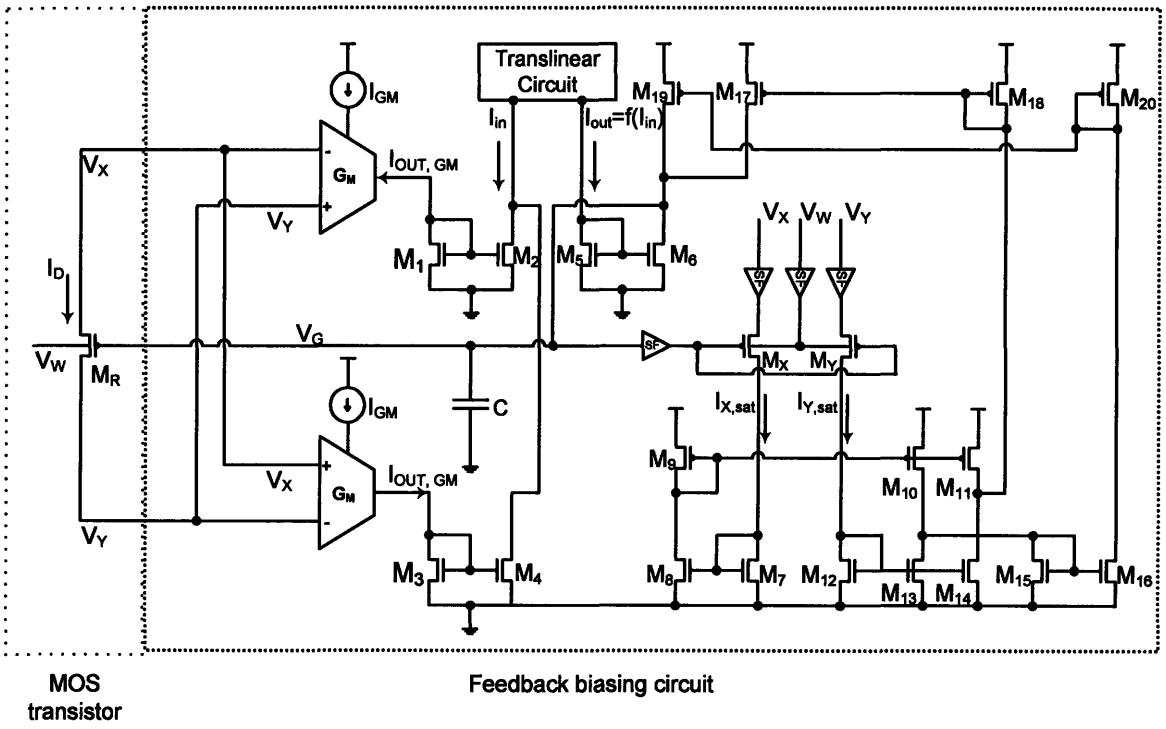


Fig. 4-8: General circuit implementation of MOS resistor.

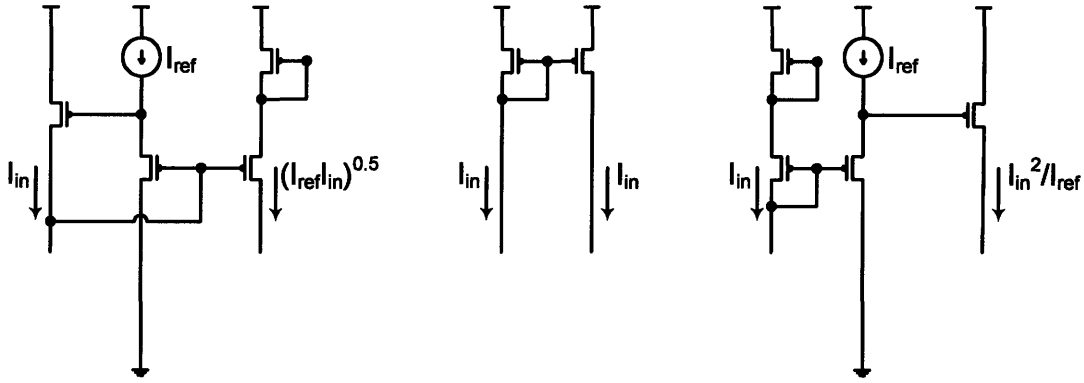


Fig. 4-9: Translinear circuits for MOS resistor with (a) compressive (square-root) (b) linear and (c) expansive (square) I-V characteristics.

4.3 Linear MOS resistor

4.3.1 Circuit description

Fig. 4-10 shows a die micrograph of a testchip fabricated in AMI 1.5μm CMOS technology. The testchip contains a linear and nonlinear MOS resistor. The circuit diagram of an MOS resistor with linear I-V characteristics is shown in Fig. 4-11. Note that the current mirror of Fig. 4-9 (b) is implicit in the circuit implementation. The schematic of the source follower buffer (denoted by SF in Fig. 4-11) is shown in the inset. It comprises a pair of PMOS transistors M_{SF1} and M_{SF2} that together forms a tracking-cascode structure, a pair of current sources I_{BP} and I_{BN}, and an nMOS transistor M_{SF3} that serves as a gain element. The buffer provides a very low output impedance $R_{O,SF}$ given by:

$$R_{O,SF} \approx \frac{1}{\left(g_{m,SF3} r_{o,SF1} g_{mp} r_{o,SF2}\right) g_{mp}} = \frac{1}{A_1 A_2 g_{mp}} \tag{36}$$

where

$$g_{mp} = g_{m,SF1} = g_{m,SF2} = \frac{\kappa_0 I_{BN}}{\phi_t}$$

$$A_1 = g_{m,SF3} r_{o,SF1} \gg 1$$

$$A_2 = g_{mp} r_{o,SF2} \gg 1$$

Note that we have added a tracking-cascode transistor such that the output impedance is even lower than that in topologies that only use M_{SF1} and no M_{SF2} [43]. The source follower buffer also provides a level shift V_{const} that is determined by I_{BN}. The tracking-cascode structure minimizes Early voltage effects by ensuring that the source and drain terminals of the transistor M_{SF1} move in tandem, thereby keeping its V_{DS} relatively constant with input voltage. We ensure that both transistors of the tracking-cascode operate in saturation by biasing them in subthreshold and making the W/L ratio of M_{SF2} larger than M_{SF1}. Fig. 4-12 shows the circuit diagram of a Wilson-mirror version of the WLR OTAs first described in [42] and used to implement the G_M transconductor of Fig. 4-8. 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) bump linearization through B_1 and B_2. The linear range V_L of the WLR OTA may be derived as follows [42]:

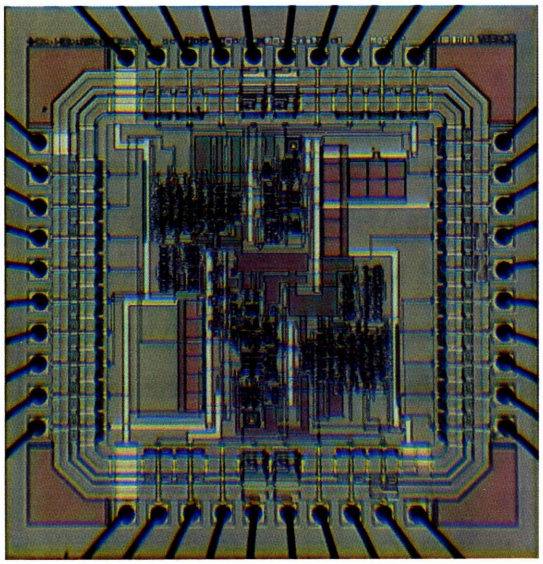


Fig. 4-10: Chip micrograph of MOS resistor fabricated in AMI 1.5μm CMOS technology. The testchip contains a linear and nonlinear MOS resistor.

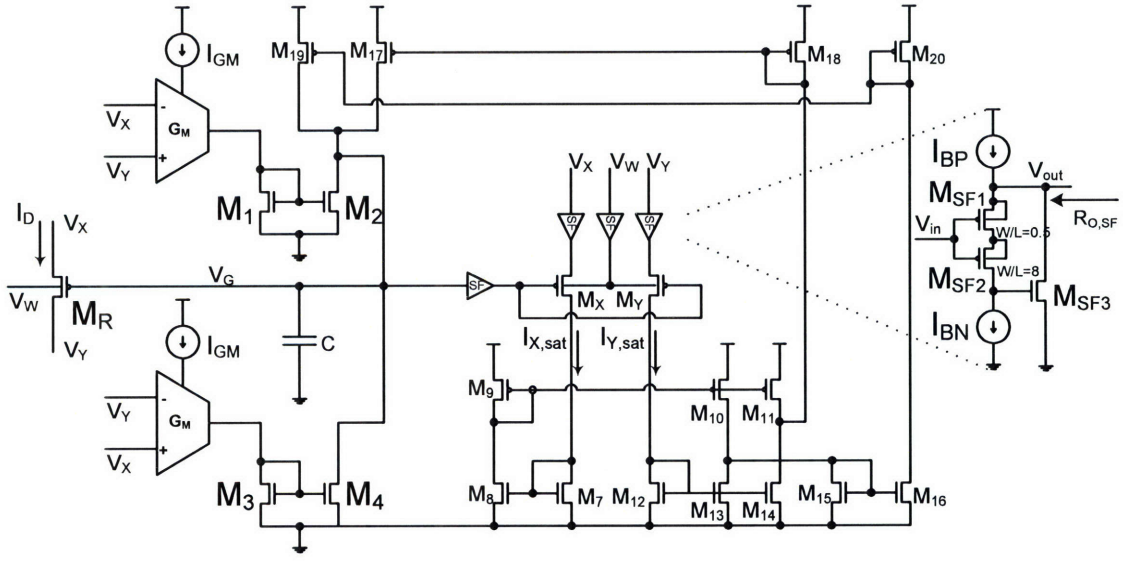


Fig. 4-11: Circuit schematic of linear MOS resistor.

$$V_L = \frac{3kT}{q}\frac{1+\frac{\kappa}{\kappa_N}+\frac{1}{\kappa_P}}{1-\kappa} \tag{37}$$

where κ_P is the subthreshold exponential parameter for transistors M_3 and M_4, κ_N is the subthreshold exponential parameter for transistors M_7 and M_8, and κ is the subthreshold exponential parameter for the input pair M_1 and M_2. The current sources I_{OC}^+ and I_{OC}^- serve to compensate for current offsets that may arise due to device mismatch. The current at the output of the WLR OTA is given by (34) and hence the desired linear I-V characteristic is:

$$I_D = g(V_{XY}) = G_M V_{XY} \tag{38}$$

In this manner, the conductance G of the linear MOS resistor may be determined by the transconductance G_M which in turn is electronically controlled by the bias current I_{GM}.

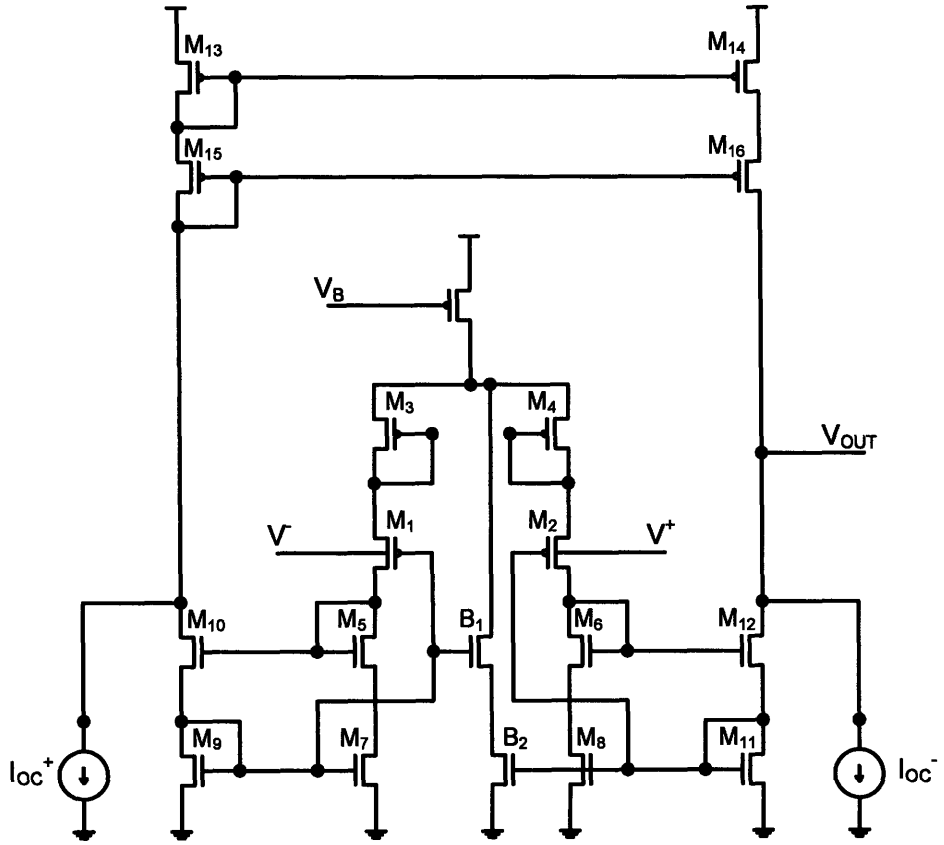


Fig. 4-12: Circuit diagram of WLR OTA.

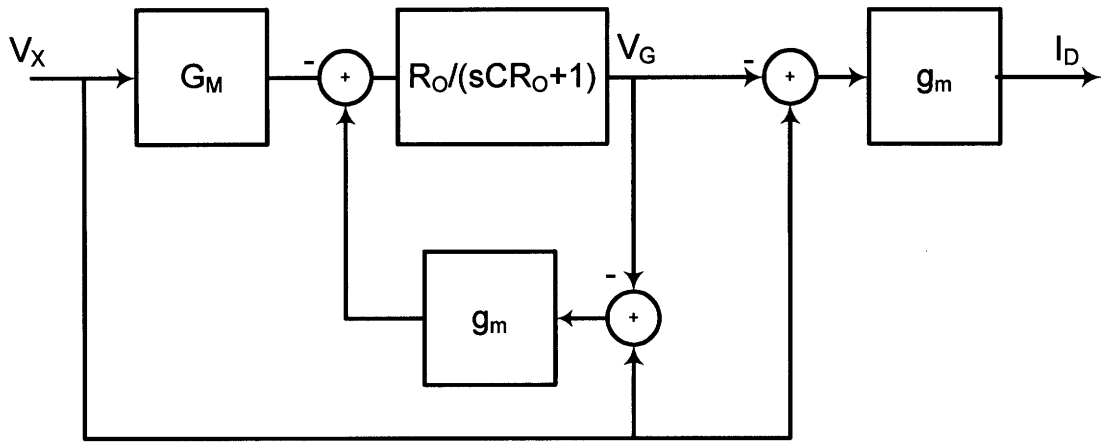
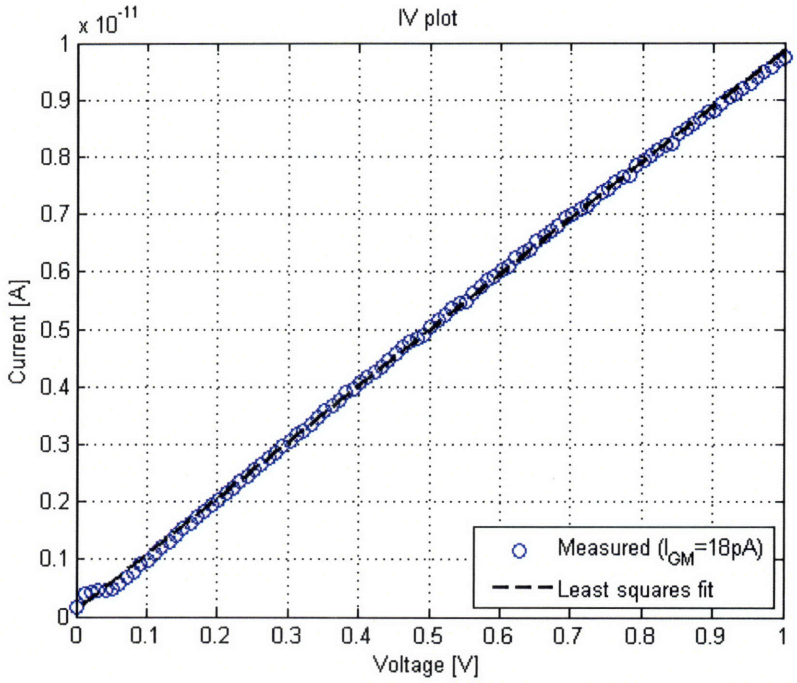


Fig. 4-13: Block diagram of MOS resistor with linear I-V characteristic.

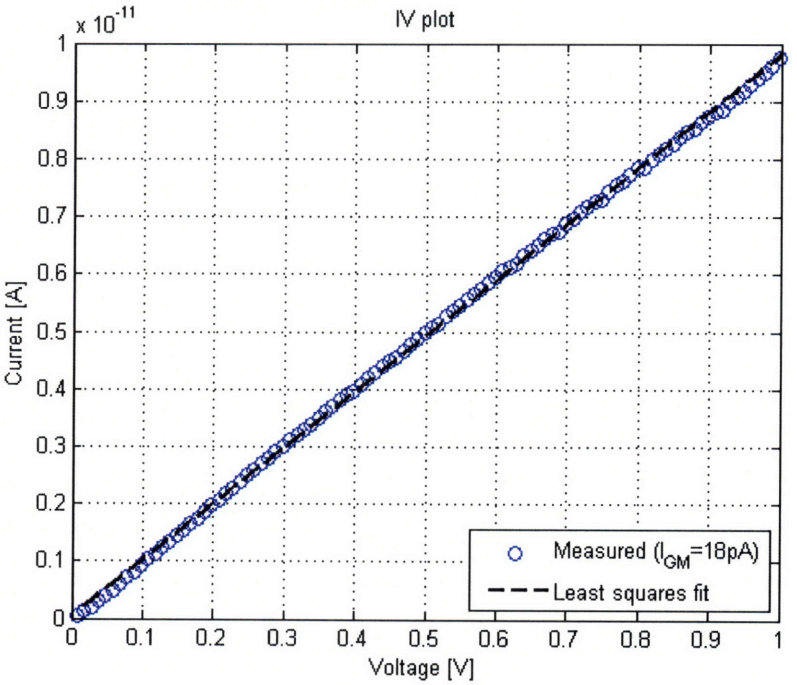
Fig. 4-13 shows a block diagram representation of the circuit depicted in Fig. 4-11. As the circuit is symmetrical, we may arbitrarily assume that V_X (or V_Y) is the signal variable and V_Y (or V_X) is grounded. The negative feedback loop servos V_G to maintain the equality of (33). G_M denotes the transconductance of the OTA as given in (34) while g_m denotes the small signal transconductances of transistors M_R and M_X (or M_Y) in Fig. 4-11. The dominant small-signal time constant at the gate terminal of M_R is given by R_OC, where $R_O = r_{O,M2} \| r_{O,M4} \| r_{O,M17} \| r_{O,M19}$ (r_O represents the small signal Early-voltage resistances of the respective transistors) and C is the total capacitance at the node.

4.3.2 DC characteristics

Fig. 4-14 shows the measured I-V characteristic of our linear MOS resistor electronically configured to have a resistance of 100GΩ. The tiny currents flowing through the MOS resistor are accurately sensed and measured using an on-chip current integration technique [44]. The potential difference V_{XY} across its source-drain terminals is varied in 10mV increments. The plot in Fig. 4-14(a) shows the I-V data without offset compensation. In this case, the slope of the I-V curve changes near the origin. The slope deviation can be attributed to offsets arising from: (a) the WLR OTAs and (b) current subtraction and mirroring operation by transistors M_7-M_{19}. Close to the origin, the current through the MOS resistor is comparable to the offset currents. Offset compensation is



(a)



(b)

Fig. 4-14: Measured I-V characteristics of (a) uncompensated and (b) offset compensated linear MOS resistor.

performed by tuning the current injected through I_{OC}^{+} or I_{OC}^{-} of the WLR OTAs of Fig. 4-12. Fig. 4-14(b) shows the I-V plot with offset compensation.

Fig. 4-15 shows the measured I-V characteristics for various values of WLR OTA biasing current I_{GM}. The slope of the I-V characteristic i.e. the conductance is determined by I_{GM}. Fig. 4-16 shows a plot of conductance G with I_{GM}. G varies linearly with I_{GM} 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_{GM} in the subthreshold regime. As I_{GM} is increased, the WLR OTA begins to transition into moderate inversion. The change in G with I_{GM} gradually departs from being linear and eventually becomes square-root when the WLR OTA operates above threshold.

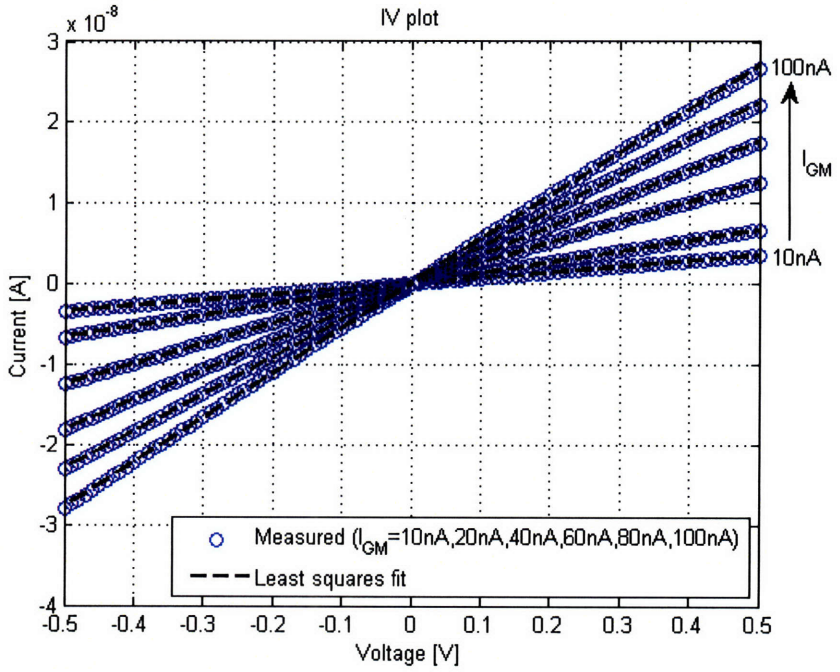


Fig. 4-15: Measured I-V characteristics of linear MOS resistor with varying biasing current I_{GM}.

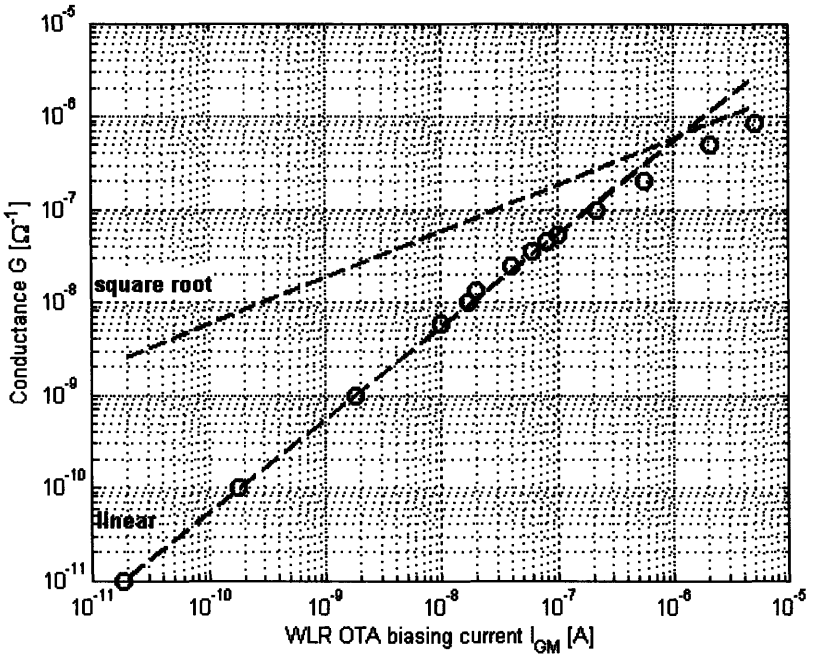
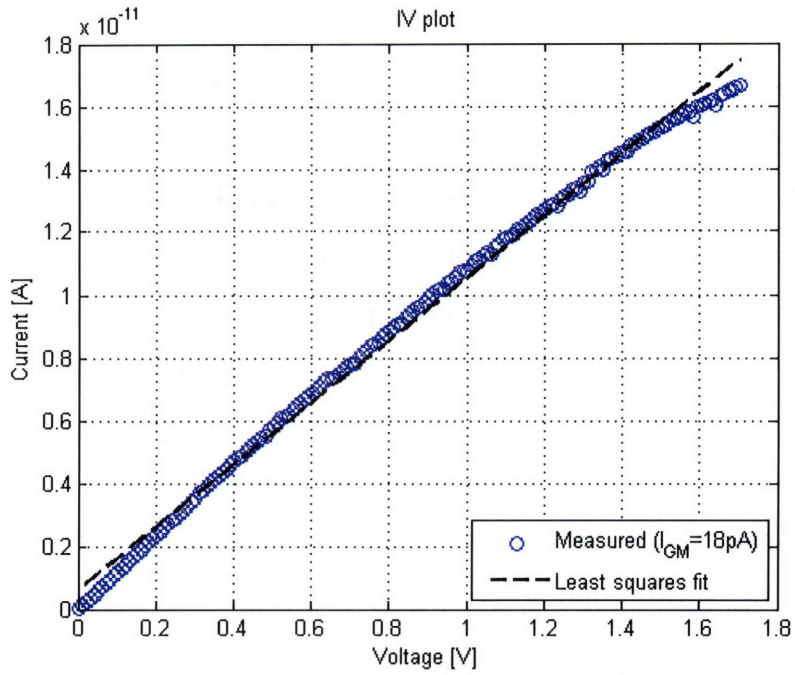
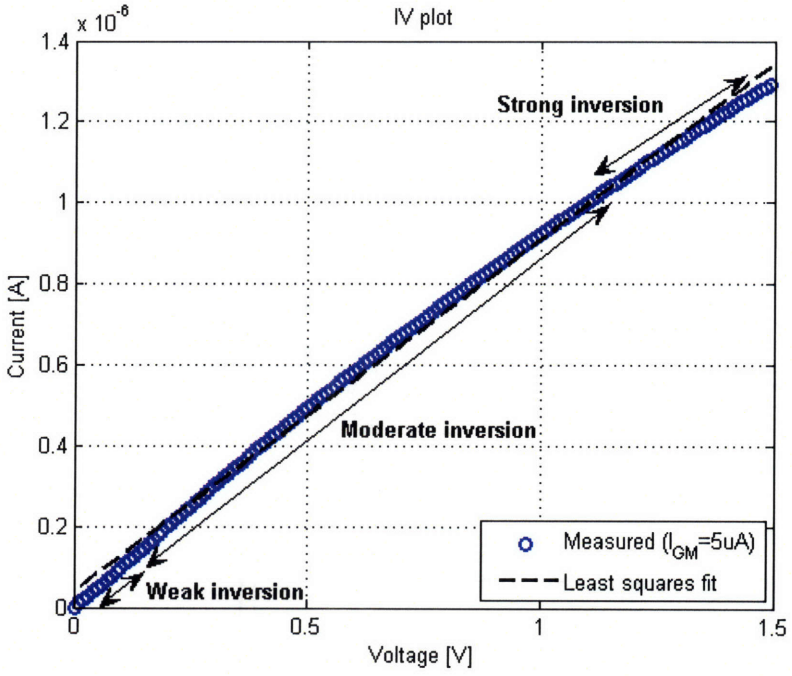


Fig. 4-16: Change in conductance G with biasing current I_{GM}.

The linear range V_L of the WLR OTA determines the linear range of the MOS resistor. A theoretical estimate of V_L may be computed from (37) to be 1.7V. The I-V data obtained by varying V_{XY} over a range of V_L=1.7V is shown in Fig. 4-17(a). The slight curvature in the I-V characteristic may be attributed to κ variation of the input pair of the WLR OTA. As V_X or V_Y is varied, the gate-to-bulk and source-to-bulk voltages of the input pair changes, giving rise to depletion width modulation which causes κ and hence the transconductance to vary slightly. Fig. 4-17(b) shows the I-V characteristic obtained when I_{GM} is biased at 5 uA. As V_{XY} is increased, the main MOS device M_R goes from weak inversion to strong inversion as indicated in the figure. The W/L ratio of M_R is 2. The above-threshold operation of M_R is limited by the WLR OTA. In our present implementation of the WLR OTA, the input transistors (M_1 and M_2 of Fig. 4-12) begin to come out of saturation when I_{GM} is increased above 5 uA.



(a)



(b)

Fig. 4-17: (a) I-V plot of linear MOS resistor taken over the theoretical linear range of WLR OTA. (b) I-V plot of linear MOS resistor showing operation of main MOS device in weak, moderate and strong inversion.

4.3.3 *AC characteristics*

Fig. 4-18 shows the measured AC characteristics of the linear MOS resistor. The experimental setup used to make the measurements and the parameters are also shown. The device under test (DUT) is hooked up to a sense amplifier comprising a resistor R_f and an operational amplifier to form an inverting amplifier configuration. In this measurement, R_f is 25MΩ and the DUT is configured to give an inverting gain of 7. The input signal V_{IN} is centered at 2.8V with an amplitude of 200mVpp and a frequency of 250Hz. V_{REF} at the non-inverting input of the operational amplifier is set at 2.5V. The measured total harmonic distortion (THD) of V_{OUT} is 0.56%.

The experiment was repeated with V_{IN} centered at various offsets from V_{REF}. Fig. 4-19 is a plot of signal distortion at the output with respect to offset at two different signal frequencies, namely 250Hz and 1kHz. The higher distortion measured at the origin may be attributed to: (a) slope mismatch at crossover, (b) residual offset from WLR OTA, current subtraction and mirroring operations and (c) current rectification dead-zone.

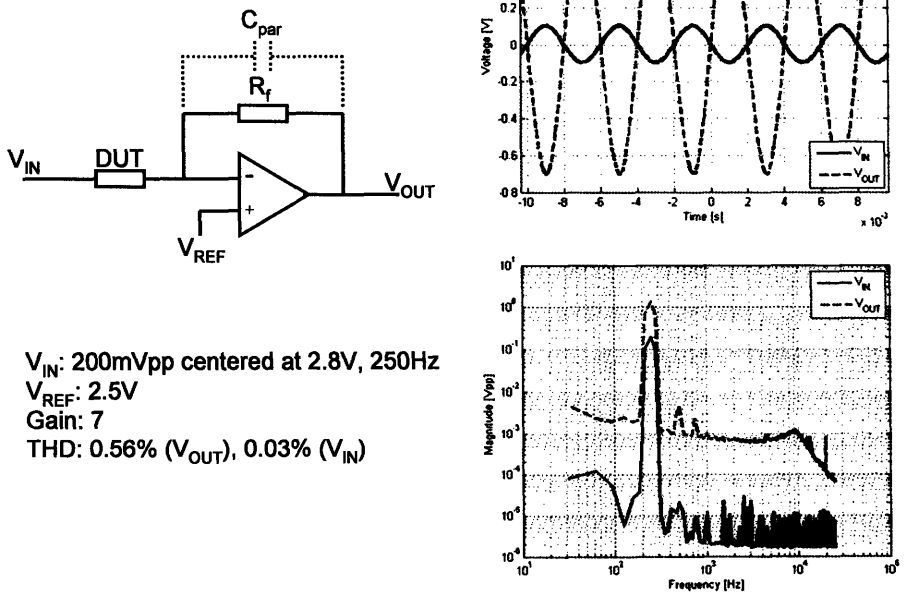


Fig. 4-18: Measured AC characteristics.

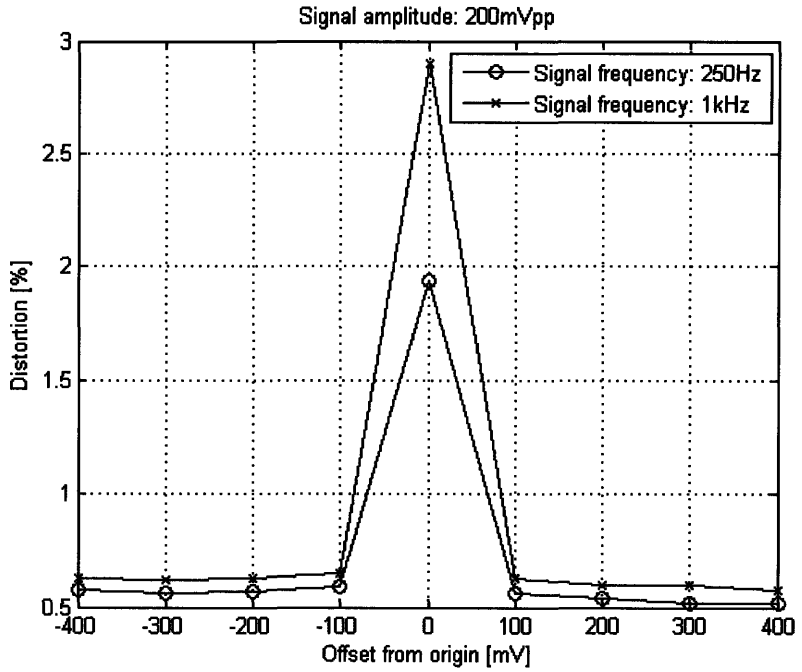


Fig. 4-19: Distortion characteristics.

4.3.4 Temperature characteristics

As Fig. 4-15 shows, the resistance R of the MOS resistor may be varied through the biasing current I_{GM} of the WLR OTA. Specifically,

$$R = \frac{V_L}{I_{GM}} \tag{39}$$

Substituting V_L from (37) in (39), the change in resistance with respect to temperature $\partial R/\partial T$ can be written as:

$$\frac{\partial R}{\partial T} = \frac{\frac{3k}{q}\frac{1+\frac{\kappa}{\kappa_N}+\frac{1}{\kappa_P}}{1-\kappa}}{I_{GM}} \tag{40}$$

Fig. 4-20 shows the measured variation of resistance with temperature. In the experiment, the nominal resistance was set by I_{GM} from a temperature invariant current source. The temperature was varied between 6 and 46°C at 2°C intervals. The resistance at each temperature was measured by computing the slope of the I-V plot taken after the temperature has stabilized to the set value. The subthreshold slopes of a PMOS and NMOS transistor with V_{BS}=0 were measured to be

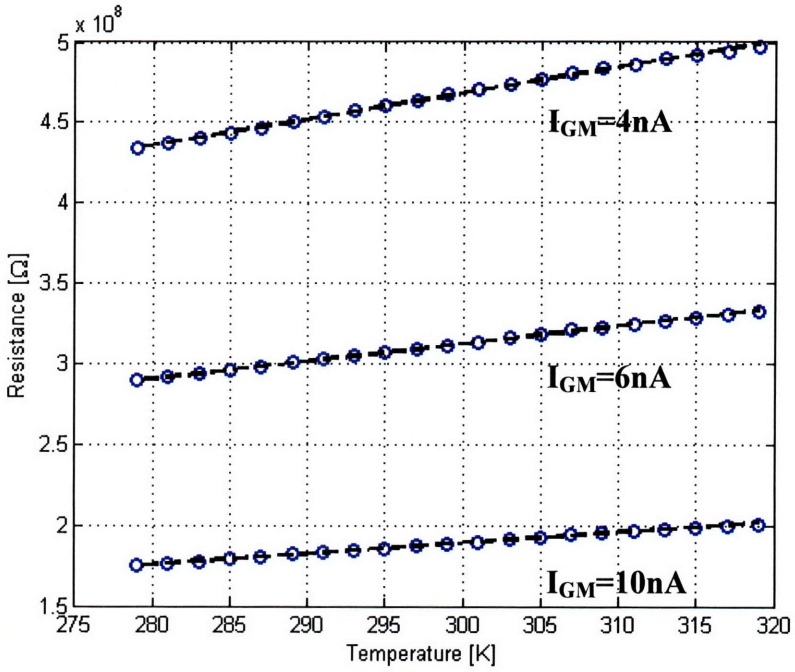


Fig. 4-20: Measured temperature characteristics.

S_P=78mV/dec and S_N=96mV/dec respectively. The corresponding subthreshold exponential parameters are $\kappa_P=(\phi_t/S_P)\ln 10=0.76$ and $\kappa_N=(\phi_t/S_N)\ln 10=0.6$. The subthreshold exponential parameter of the input pair was estimated to be $\kappa=0.85$ by accounting for its non-zero V_{BS}. The measured and theoretical values of $\partial R/\partial T$ are tabulated in Table 4-1. We see that there is good agreement between the measured and theoretical values.

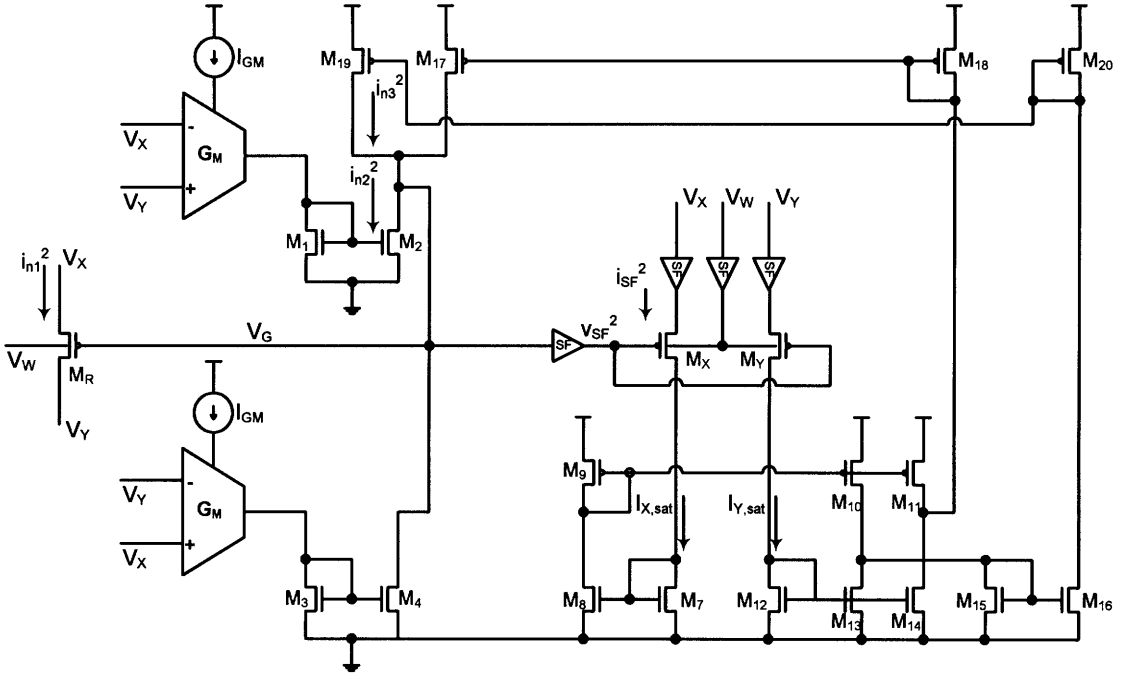
I_{GM} [nA]	$\partial R/\partial T$ (Measured) [MΩ/K]	$\partial R/\partial T$ (Theoretical) [MΩ/K]
4	1.65	1.62
6	1.1	1.08
10	0.668	0.646

Table 4-1: Measured and theoretical values of $\partial R/\partial T$ for various WLR OTA biasing currents.

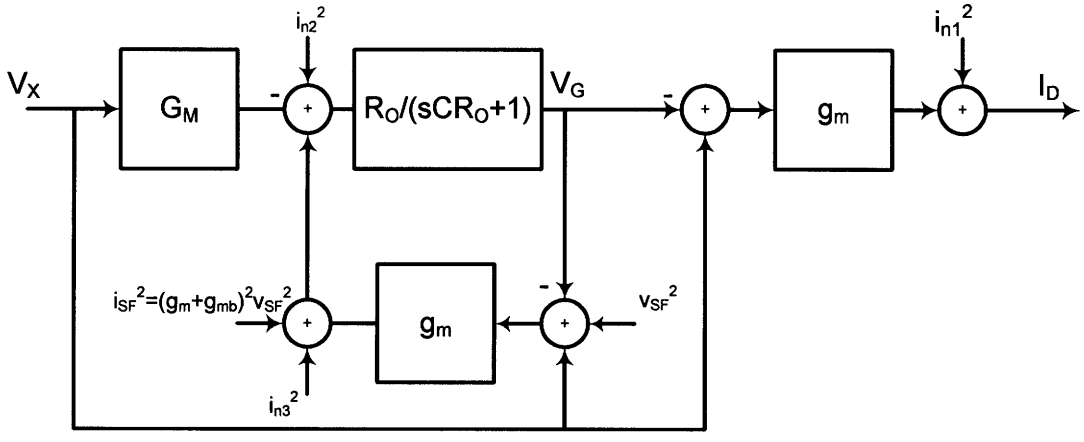
4.3.5 Noise analysis and measurements

The noise sources associated with the MOS resistor are depicted in Fig. 4-21. Since the MOS resistor circuit is bidirectional and symmetric, we assume $V_X>V_Y$ in the following analysis of the half-circuit with no loss in generality. The noise component of

the main transistor M_R is denoted by i_{n1}^2. The combined output noise of the active WLR OTA and the noise from the current mirror formed by M_1 and M_2 is denoted by i_{n2}^2. The noise contribution of transistors M_X, M_Y and M_7-M_{19} is denoted by i_{n3}^2. The output voltage and current noise of the source-follower buffers with input V_G and input V_X (or V_Y) are denoted by v_{SF}^2 and i_{SF}^2 respectively. When configured as a resistor of



(a)



(b)

Fig. 4-21: (a) Circuit and (b) block diagram of MOS resistor showing the dominant noise sources.

conductance G carrying a current I_D ($V_{XY} \neq 0$) such that $I_D \approx I_{X,sat}$ ($I_{X,sat} >> I_{Y,sat}$), the noise contribution i_{n1}^2 from transistor M_R may be derived as follows:

$$i_{n1}^2 = 2qI_D = 2qG_M V_{XY} = N_1 kTG_M = N_1 kTG \tag{41}$$

where G is servoed by feedback to equal G_M as revealed by the feedback block diagram of Fig. 4-13, and N_1 is defined to be

$$N_1 = \frac{2V_{XY}}{kT/q} \tag{42}$$

The output current noise $i_{n,GM}^2$ of the WLR OTA may be derived as [42]:

$$i_{n,GM}^2 = N_{GM}\left(2q\frac{I_{GM}}{3}\right) \tag{43}$$

where N_{GM}, the effective number of noise sources in the WLR OTA, has a value of 3.8. Hence, i_{n2}^2 may be written as follows:

$$\begin{aligned} i_{n2}^2 &= i_{n,GM}^2 + N_2(2qI_D) \\ &= 3.8\left(2q\frac{I_{GM}}{3}\right) + N_2(2qI_D) \end{aligned} \tag{44}$$

where N_2 is the number of noise sources in the current mirror formed by M_1 and M_2. Applying (41) and (37) in (44), we get

$$\begin{aligned} i_{n2}^2 &= 3.8\left(2q\frac{G_M V_L}{3}\right) + N_2(N_1 kTG) \\ &= 7.6\left(\frac{1+\frac{\kappa}{\kappa_N}+\frac{1}{\kappa_P}}{1-\kappa}\right)kTG + N_2(N_1 kTG) \end{aligned} \tag{45}$$

Since the current through M_X is $I_{X,sat} \approx I_D$,

$$i_{n3}^2 = N_3(2qI_D) = N_3 N_1 kTG \tag{46}$$

where N_3 is the number of noise sources originating from the current subtraction and mirroring operation performed by M_X, M_Y and M_7-M_{19}. From the inset of Fig. 4-11 and using the techniques described in [42], the voltage noise v_{SF}^2 at the output of the source follower driving the gates of M_X and M_Y may be derived as:

$$\begin{aligned} v_{SF}^2 = \big[&\alpha_{SF1}^2(2qI_{BN}) + \alpha_{I_{BN}}^2(2qI_{BN}) + \alpha_{I_{BP}}^2(2qI_{BP}) \\ &+ \alpha_{SF3}^2 2q(I_{BP} - I_{BN})\big]R_{O,SF}^2 \end{aligned} \tag{47}$$

where α_{SF1}, α_{IBN}, α_{SF3}, α_{IBP} are the current noise transfer functions from M_{SF1}, I_{BN}, M_{SF3}, I_{BP} to the gate terminal of M_X, respectively. Using $A_1=g_{m,SF3}r_{O,SF1}$ and $A_2=g_{mp}r_{O,SF2}$ given in (36), the noise transfer functions α_{SF1}, α_{IBN}, α_{SF3} and α_{IBP} are given by:

$$\begin{aligned}
\alpha_{SF1} &= \left|A_1\left(1+A_2\right)\right| \approx A_1A_2 \\
\alpha_{I_{BN}} &\approx \left|1-A_1\left(1+A_2\right)\right| \approx \alpha_{SF1} \\
\alpha_{SF3} &= 1 \ll \alpha_{SF1} \\
\alpha_{I_{BP}} &= 1 \ll \alpha_{SF1}
\end{aligned} \tag{48}$$

Hence, the noise contribution from v_{SF}^2 is:

$$\begin{aligned}
v_{SF}^2 &\approx \Big[A_1^2A_2^2\left(2qI_{BN}\right)+\left(A_1^2A_2^2+2A_1\left(1+A_2\right)+1\right)\left(2qI_{BN}\right) \\
&\quad +2qI_{BP}+2q\left(I_{BP}-I_{BN}\right)\Big]R_{O,SF}^2 \\
&\approx 4q\left(A_1^2A_2^2I_{BN}+I_{BP}\right)R_{O,SF}^2 \\
&\approx 4q\left(A_1^2A_2^2I_{BN}+I_{BP}\right)\left(\frac{1}{A_1A_2g_{mp}}\right)^2 \\
&\approx \frac{4qI_{BN}}{g_{mp}^2}
\end{aligned} \tag{49}$$

If $g_{m,X}$ and $g_{mb,X}$ are the transconductance and back-gate transconductance of M_X respectively, the output current noise i_{SF}^2 of the source follower driving the source terminal of M_X may be derived as:

$$\begin{aligned}
i_{SF}^2 &= \left(g_{m,X}+g_{mb,X}\right)^2 v_{SF}^2 \\
&\approx \left(\frac{g_{m,X}+g_{mb,X}}{g_{mp}}\right)^2 4qI_{BN} \\
&= \left(\frac{I_D}{\kappa_P I_{BN}}\right)^2 4qI_{BN} \\
&= \left(\frac{I_D}{\kappa_P^2 I_{BN}}\right)4qI_D
\end{aligned} \tag{50}$$

From the block diagram of Fig. 4-21(b), the total noise at the output is given by:

$$i_{n,tot}^{\ 2} = i_{n1}^{\ 2} + \left(\frac{g_m R_O / 1 + g_m R_O}{s \frac{CR_O}{1+g_m R_O} + 1} \right)^2 \left(i_{n2}^{\ 2} + i_{n3}^{\ 2} + i_{SF}^{\ 2} + g_{m,X}^{\ 2} v_{SF}^{\ 2} \right) \tag{51}$$

$$\approx i_{n1}^{\ 2} + \left(\frac{1}{s \frac{C}{g_m} + 1} \right)^2 \left(i_{n2}^{\ 2} + i_{n3}^{\ 2} + i_{SF}^{\ 2} + g_{m,X}^{\ 2} v_{SF}^{\ 2} \right)$$

$$\approx i_{n1}^{\ 2} + \left(\frac{1}{s \frac{C}{g_m} + 1} \right)^2 \left[i_{n2}^{\ 2} + i_{n3}^{\ 2} + \left(\frac{I_D}{\kappa_P^{\ 2} I_{BN}} \right) 4qI_D + g_{m,X}^{\ 2} \frac{4qI_{BN}}{g_{mp}^{\ 2}} \right]$$

$$= i_{n1}^{\ 2} + \left(\frac{1}{s \frac{C}{g_m} + 1} \right)^2 \left(i_{n2}^{\ 2} + i_{n3}^{\ 2} + \frac{2}{\kappa_P^{\ 2}} \frac{I_D}{I_{BN}} (2qI_D) + 2 \frac{I_D}{I_{BN}} (2qI_D) \right)$$

Now, by applying (41), (45) and (46) in (51), we get

$$i_{n,tot}^{\ 2} \approx N_1 kTG + \tag{52}$$

$$\left(\frac{1}{s \frac{C}{g_m} + 1} \right)^2 \left[\left(N_2 + N_3 + \frac{2}{\kappa_P^{\ 2}} \frac{I_D}{I_{BN}} + 2 \frac{I_D}{I_{BN}} \right) N_1 + 7.6 \left(\frac{1 + \frac{\kappa}{\kappa_N} + \frac{1}{\kappa_P}}{1-\kappa} \right) \right] kTG$$

In essence, the current noise from the WLR OTA's and current mirrors is low pass filtered by the integrating feedback loop and adds to the intrinsic noise of the main transistor M_R.

The experimental setup used to measure noise of the MOS resistor is depicted in Fig. 4-22(a). The inverting amplifier configuration allows the potential difference across the DUT to be varied through V_{DC} and V_{REF}. The noise at the output of the amplifier is given by:

$$v_{n,out}^{\ 2} = \left(\frac{R_f}{R_i} \right)^2 v_{n,Ri}^{\ 2} + v_{n,Rf}^{\ 2} + \left(1 + \frac{R_f}{R_i} \right)^2 v_{n,Amp}^{\ 2} \tag{53}$$

where R_i is the resistance of the DUT and R_f is the resistance of the feedback resistor. $v_{n,Amp}^{\ 2}$ is the input referred voltage noise of the operational amplifier (LF356) and has a value of $15nV/\sqrt{}$ Hz. In the measurement, a real resistance of R_f=25MΩ was used as the

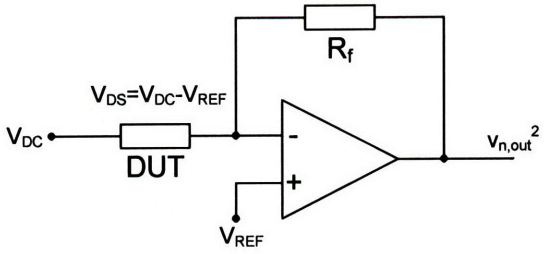
feedback resistor and the DUT was configured such that R_i=5MΩ, giving an inverting gain of 5. The measured noise power spectral densities (PSD) at the output of the amplifier are plotted in Fig. 4-22(b) for various source-to-drain potentials V_{DS} (=V_{XY}) of the MOS resistor. As a reference, the noise PSD of a real 5MΩ resistor is also shown in the same figure. The theoretical value of $v_{n,out}^2$ for a real 5MΩ resistor may be computed to be 1.5μV/√ Hz. From Fig. 4-22(b), the measured value is 1.63μV/√ Hz. Using the noise estimate given in (51) at low frequencies ($\omega << C/g_m$), the theoretical value of $v_{n,out}^2$ for an MOS resistor may be computed as:

$$v_{n,out}^2 = v_{n,Rf}^2 + \left(1+\frac{R_f}{R_i}\right)^2 v_{n,Amp}^2 + \left(\frac{R_f}{R_i}\right)^2 \left[\left(1+N_2+N_3+\frac{2}{\kappa_P^2}\frac{I_D}{I_{BN}}+2\frac{I_D}{I_{BN}}\right)N_1+7.6\left(\frac{1+\frac{\kappa}{\kappa_N}+\frac{1}{\kappa_P}}{1-\kappa}\right)\right]kTR_i \tag{54}$$

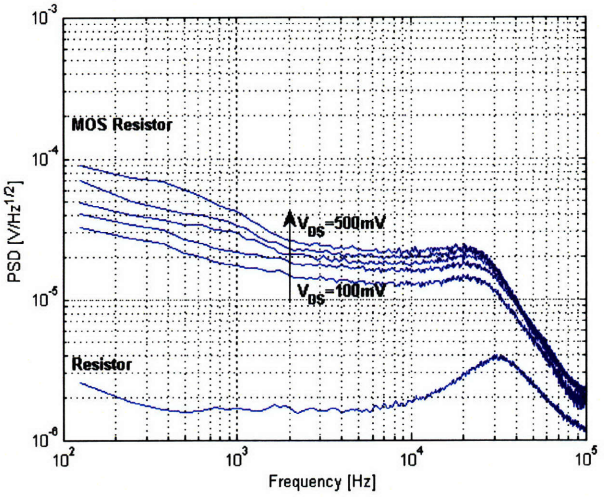
As N_1 and I_D are functions of V_{DS}, the MOS resistor's noise varies with the potential difference across its terminals, unlike a real resistor. The measured output noise $v_{n,out}^2$ and its theoretical values for an MOS resistor with an equivalent resistance R_i=5MΩ are plotted as a function of V_{DS} in Fig. 4-22(c); we see that there is good agreement of the measured noise and that predicted by theory. Also, compared to the 4kTG noise spectral density of a real resistor, the MOS resistor has an excess noise factor N_e given by:

$$N_e = \frac{\left[\left(N_2+N_3+1+\frac{2}{\kappa_P^2}\frac{I_D}{I_{BN}}+2\frac{I_D}{I_{BN}}\right)N_1+7.6\left(\frac{1+\frac{\kappa}{\kappa_N}+\frac{1}{\kappa_P}}{1-\kappa}\right)\right]kTG}{4kTG} \tag{55}$$

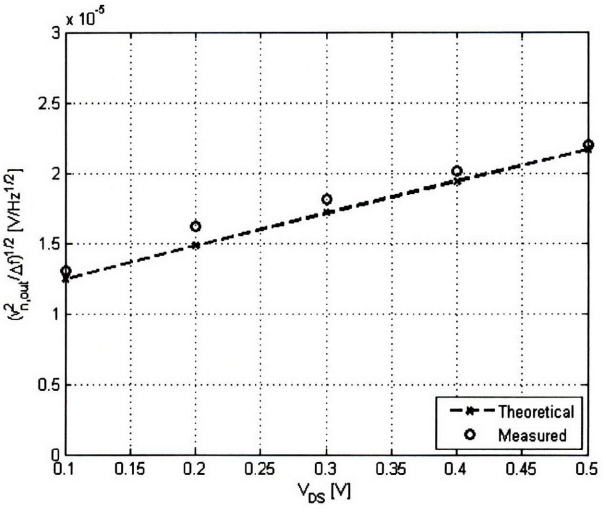
$$= \frac{1}{4}\left[\left(N_2+N_3+1+\frac{2}{\kappa_P^2}\frac{I_D}{I_{BN}}+2\frac{I_D}{I_{BN}}\right)N_1+7.6\left(\frac{1+\frac{\kappa}{\kappa_N}+\frac{1}{\kappa_P}}{1-\kappa}\right)\right]$$



(a)



(b)



(c)

Fig. 4-22: (a) Experimental setup. (b) Measured noise power spectral density of MOS resistor (5MΩ) with varying potential difference across its terminals. The noise PSD of a real 5MΩ resistor is also shown for reference. (c) Plot of measured output noise $v_{n,out}^2$ and its theoretical values as a function of V_{DS} for an MOS resistor configured as an equivalent resistance R_i=5MΩ.

4.4 Nonlinear MOS resistor

4.4.1 Circuit description

In this section, we describe two electronically tunable nonlinear resistors: the first has a compressive I-V characteristic such that $I=K\sqrt{V}$ while the second has an expansive I-V characteristic such that $I=KV^2$ where K is an electronically controlled scale factor with the appropriate dimensions. Both nonlinear resistors are implemented using the general circuit architecture depicted in Fig. 4-8. The compressive resistor having a square-root I-V characteristic employs the translinear circuit of Fig. 4-9(a). The output current of the WLR OTA given by $G_M V_{XY}$ is compressed by the translinear circuit in a square-root manner to produce the desired I-V relation:

$$I_D = \sqrt{I_{ref} G_M V_{XY}} \tag{56}$$

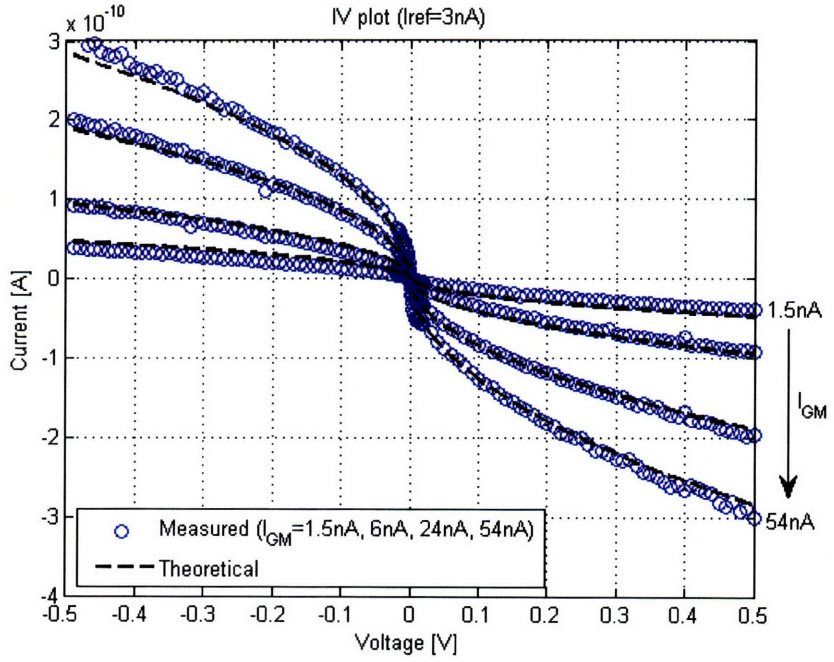
The expansive resistor employs the translinear circuit of Fig. 4-9 (c) to expand $G_M V_{XY}$ and produce the desired I-V relation given by:

$$I_D = \frac{(G_M V_{XY})^2}{I_{ref}} \tag{57}$$

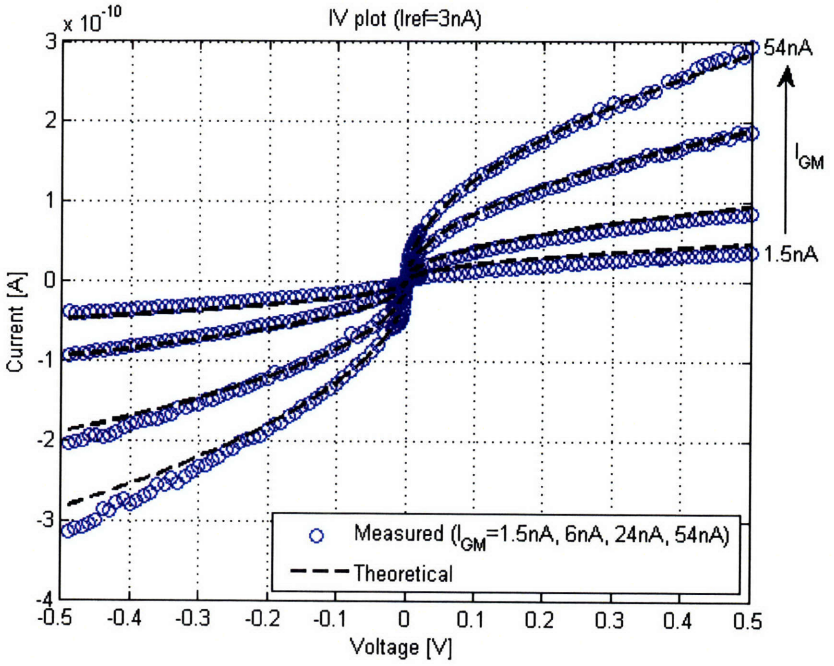
In either case, the negative feedback loop servos V_G such that the difference in saturation currents $I_{X,sat}$-$I_{Y,sat}$ become equal to I_D given by (56) or (57).

4.4.2 Experimental results

Fig. 4-23 (a) shows the measured I-V data for the compressive resistor having an I-V relation given by (56). The theoretical I-V curve is also plotted in dashed lines for comparison. The measurement was repeated with V_X and V_Y interchanged in Fig. 4-23(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_{GM} of the OTA. The same effect may also be achieved by varying I_{ref} in the translinear circuit. Fig. 4-24 shows the measured and theoretical I-V curves of the expansive resistor having an I-V relation given by (57). A logarithmic plot of the measured and theoretical data in Fig. 4-25 shows that they are in good agreement.



(a)



(b)

Fig. 4-23: Measured I-V characteristics of compressive MOS resistor ($I=K\sqrt{V}$).

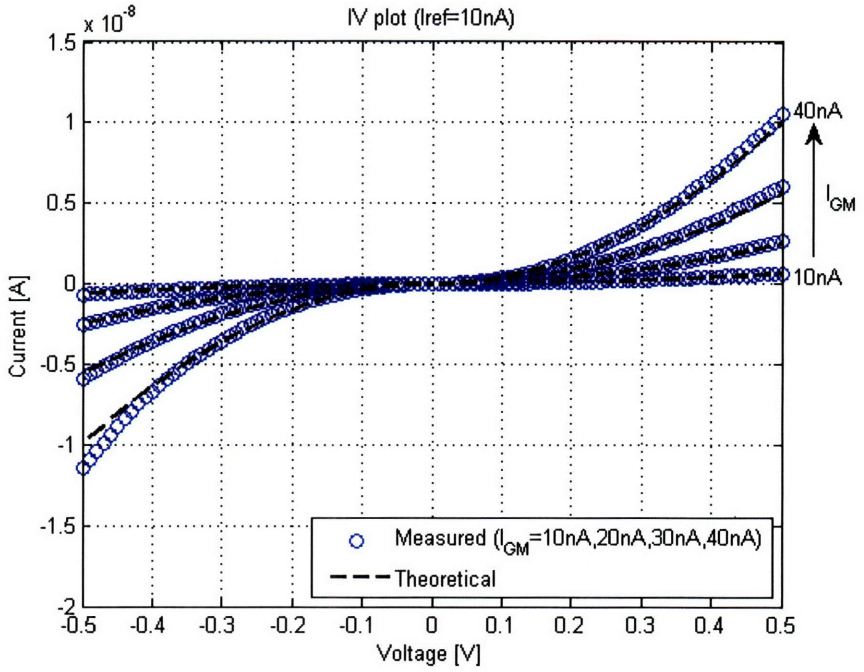


Fig. 4-24: Measured I-V characteristics of expansive MOS resistor ($I=KV^2$).

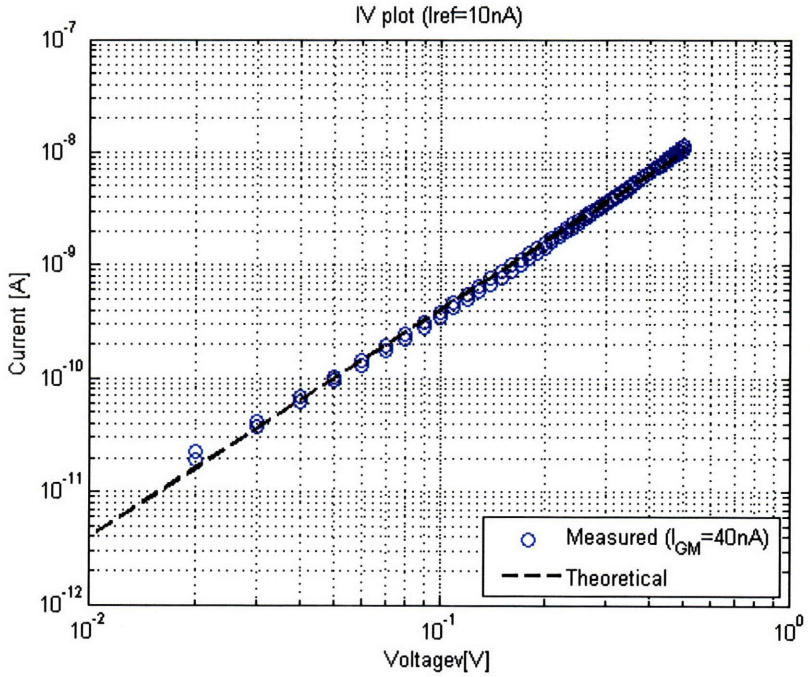


Fig. 4-25: Logarithmic plot of measured I-V characteristics of expansive MOS resistor ($I=KV^2$).

INTENTIONALLY LEFT BLANK

Chapter 5 ELECTRONICALLY TUNABLE TWO-PORT π-SECTION

In this chapter, we describe circuit topologies for realizing a tunable transmission line vocal tract which simulates the variations in the acoustic characteristics of the human vocal tract that occur as a result of changes in the cross-sectional areas of the tract at various points along its length. The transmission line analog vocal tract comprises a cascade of tunable two port sections each of which is electrically equivalent to a LC π-section having an inductor as the series component and a capacitor as a shunt component. Both the series and shunt components are readily tunable via electronic control signals.

5.1 VLSI inductor

5.1.1 OTA based second order filter structures

Fig. 5-1 shows the topology of a second order filter section employing two operational transconductance amplifiers (OTA), G_1 and G_2, connected in a feedback configuration. The values of the transconductances G_1 and G_2 are tunable via the biasing currents of the OTAs. When an input signal V_{in} is connected to the non-inverting input of OTA G_1, the circuit serves as a second order low-pass filter whose output V_o is given by:

$$V_o = \frac{1}{\tau_1\tau_2 s^2 + \frac{\tau_1}{Q}s + 1} V_{in} \tag{58}$$

$$\tau_1 = \frac{C_1}{G_1}$$

$$\tau_2 = \frac{C_2}{G_2}$$

$$Q = \sqrt{\frac{\tau_2}{\tau_1}}$$

If a DC reference voltage V_{DC} is applied to the non-inverting input of the first OTA G_1, as shown in Fig. 5-2, the impedance looking into the output of OTA G_2, i.e., V_o, may be computed as follows. First, consider that the output node is driven in voltage

mode by a voltage V_o. OTA G_1 produces a current proportional to V_o that is integrated on capacitor C_1 resulting in a voltage V_{c1}. OTA G_2 produces a current I_{in} that is proportional to the difference between V_{c1} and V_o. The overall effect is to produce a current I_{in} that is the integral of the voltage V_o with respect to the DC voltage V_{DC}. Hence the impedance at V_o looks inductive with respect to V_{DC} (ac ground). The block diagram representation of such a voltage mode operation is depicted in Fig. 5-3(a). Assuming $R_{O1} >> 1/j\omega C_1$, the impedance at V_o may be computed from the block diagram as:

$$I_{in} = -\left(-\frac{G_1}{sC_1+1}-1\right)G_2V_o = \left(\frac{G_1}{sC_1+1}+1\right)G_2V_o \tag{59}$$

$$\frac{V_o}{I_{in}} = \frac{s\frac{C_1}{G_1G_2}}{s\frac{C_1}{G_1}+1}$$

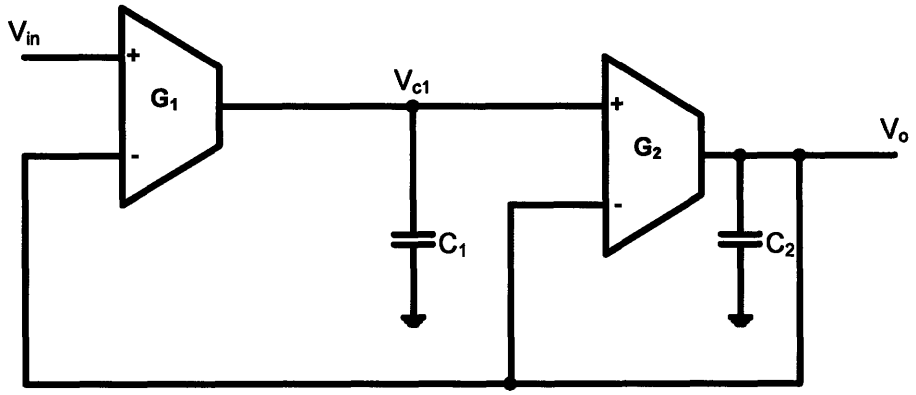


Fig. 5-1: Second order filter section.

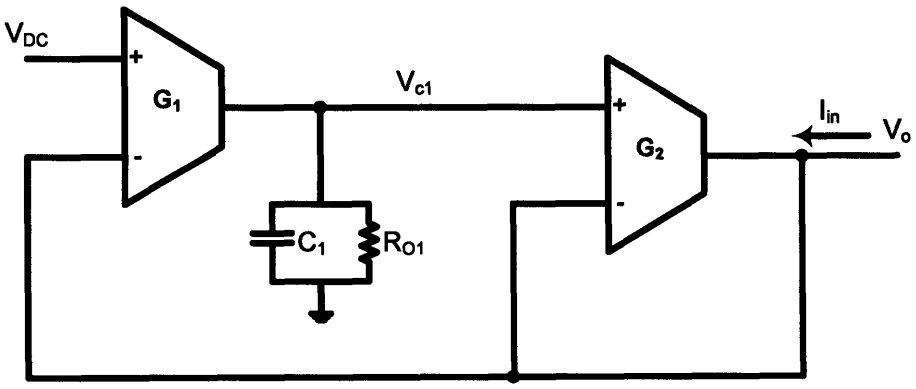


Fig. 5-2: Unidirectional tunable inductance.

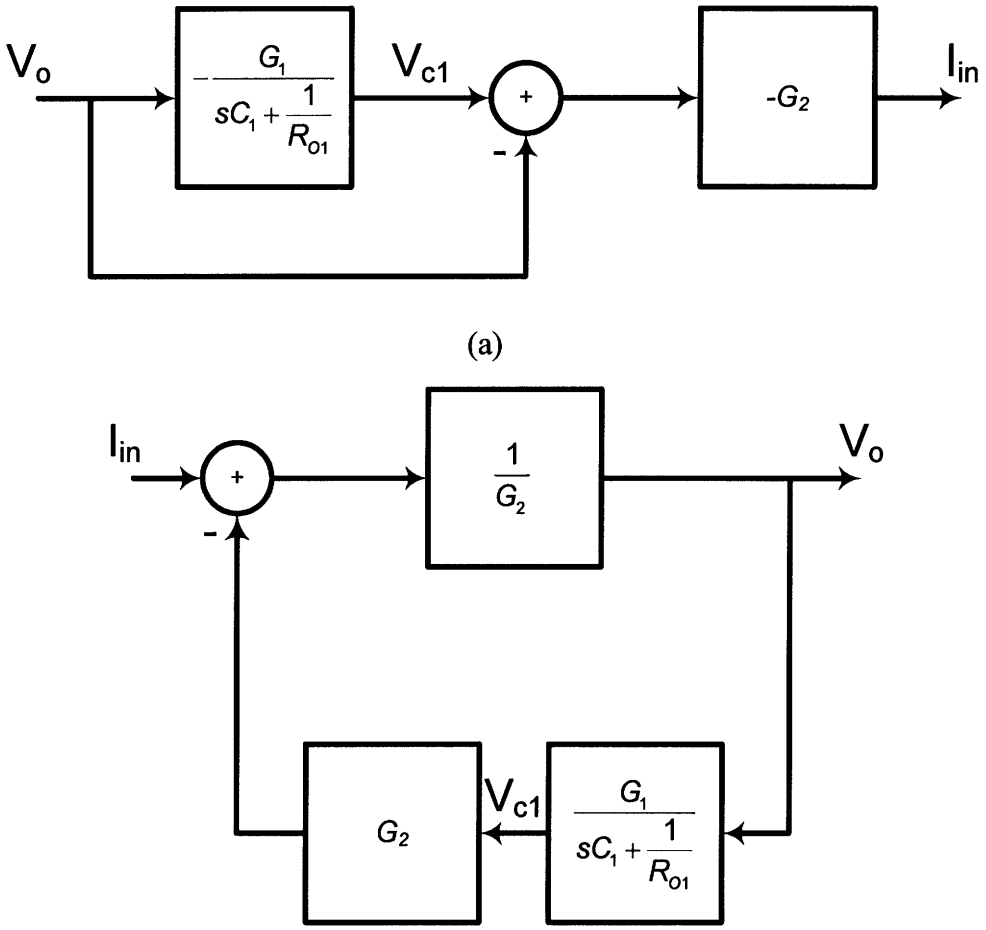


Fig. 5-3: Block diagram representation of Fig. 5-2 for (a) voltage mode operation and (b) current mode operation.

Alternatively, consider that the output node is driven in current mode by a current I_{in}. Fig. 5-3(b) shows the block diagram of such a current mode operation. Analyzing the the feedback loop (assuming $R_{O1} >> 1/j\omega C_1$) for the impedance gives the following result:

$$V_o = \frac{\frac{1}{G_2}}{\frac{G_1}{sC_1}+1} I_{in} \tag{60}$$

$$\frac{V_o}{I_{in}} = \frac{s\frac{C_1}{G_1 G_2}}{s\frac{C_1}{G_1}+1}$$

The impedance V_o/I_{in} obtained in (59) and (60) are identical. At frequencies $\omega << G_1/C_1$ the impedance at V_o with respect to V_{DC} (ac ground) is given by:

$$Z_{in} = sL_{eq}, \quad L_{eq} = \frac{C_1}{G_1 G_2} \tag{61}$$

In other words, the impedance looking into the output of OTA G_2 is inductive and the value of the equivalent inductance with respect to V_{DC} is given by the ratio C_1/G_1G_2.

It is noteworthy that the block diagram of Fig. 5-3(a) is completely feedforward even though OTAs G_1 and G_2 are connected in a feedback configuration. Intuitively, the feedback loop from the output of OTA G_2 (i.e., V_o in Fig. 5-2) to the inverting input of OTA G_1 is inactive because V_o is driven by a low impedance source in voltage mode operation. In contrast, the feedback loop is active for current mode operation, as the output of OTA G_2 is driven by a high impedance current source. The block diagram of Fig. 5-3(b) clearly shows the feedback action.

5.1.2 OTA based gyrator

The gyrator architecture is a classic approach for realizing active inductors. The topology consists two OTAs connected in a feedback configuration as shown in Fig. 5-4. Fig. 5-5 shows a block diagram analysis of the OTA based gyrator under current mode operation. Assuming that $R_{O1} \gg 1/j\omega C_1$ and $R_{O2} \ll 1/j\omega C_2$, the impedance may be computed from the block diagram as:

$$V_o = \frac{R_{O2}}{\frac{G_1 G_2 R_{O2}}{sC_1} + 1} I_{in} \tag{62}$$

$$\frac{V_o}{I_{in}} = \frac{s\frac{C_1}{G_1 G_2}}{s\frac{C_1}{G_1 G_2 R_{O2}} + 1}$$

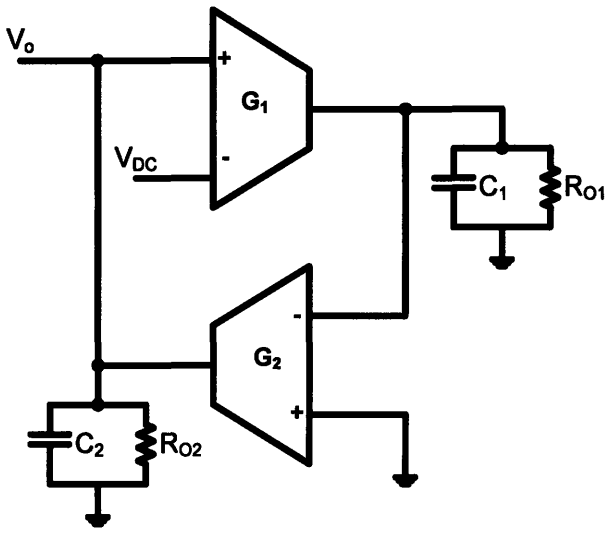


Fig. 5-4: OTA based gyrator.

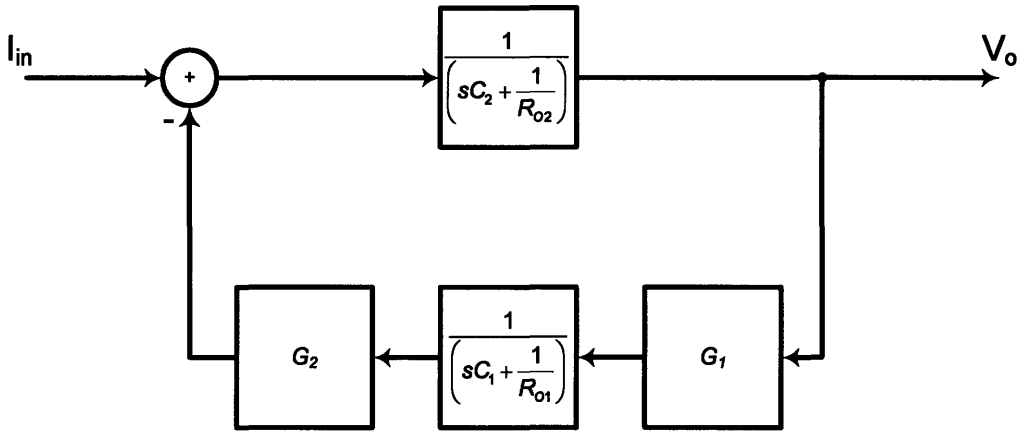


Fig. 5-5: Block diagram representation of OTA based gyrator.

Hence, at frequencies $\omega << G_2R_{O2}(G_1/C_1)$ the impedance at V_o with respect to V_{DC} (ac ground) is given by:

$$Z_{in} = sL_{eq}, \quad L_{eq} = \frac{C_1}{G_1 G_2} \tag{63}$$

As in the case of the second order filter structure, the impedance Z_{in} at the output of OTA G_2 is given by the ratio of the integrating capacitance C_1 and the product of the transconductances G_1 and G_2. Note that although the values of the equivalent inductance L_{eq} are identical in (61) and (63), the self resonant frequency is higher by a factor of G_2R_{O2} in the OTA based gyrator.

5.1.3 *Operational amplifier based gyrator*

Fig. 5-6 shows an operational amplifier A_1 with its output driving a capacitor C_C added in a feedback loop to turn it into a current derivative circuit. A block diagram analysis of the circuit is shown in Fig. 5-7.

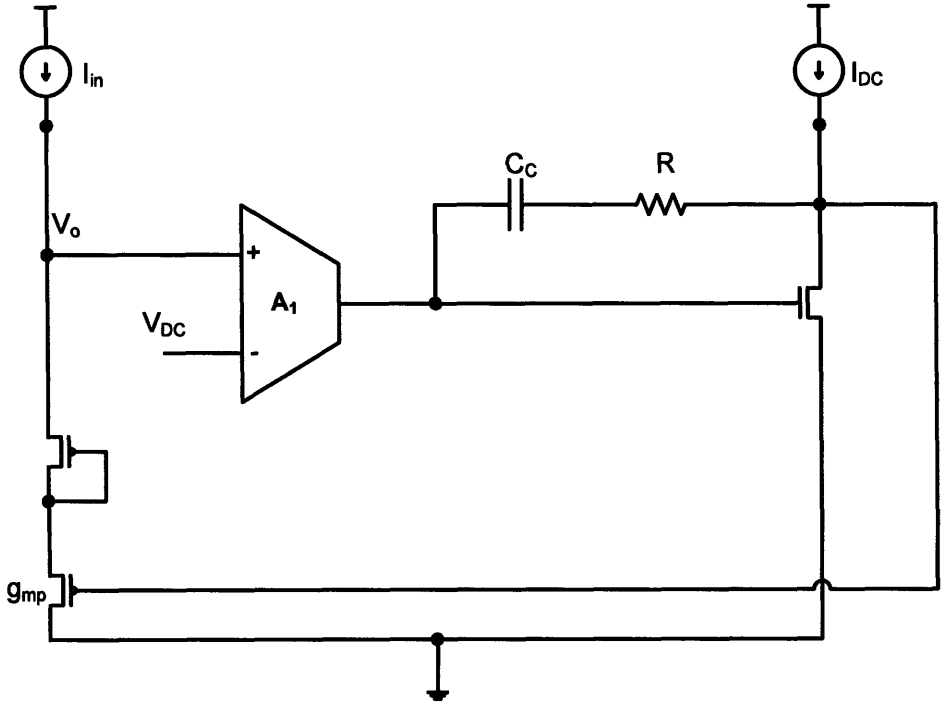


Fig. 5-6: Operational amplifier based current derivative circuit.

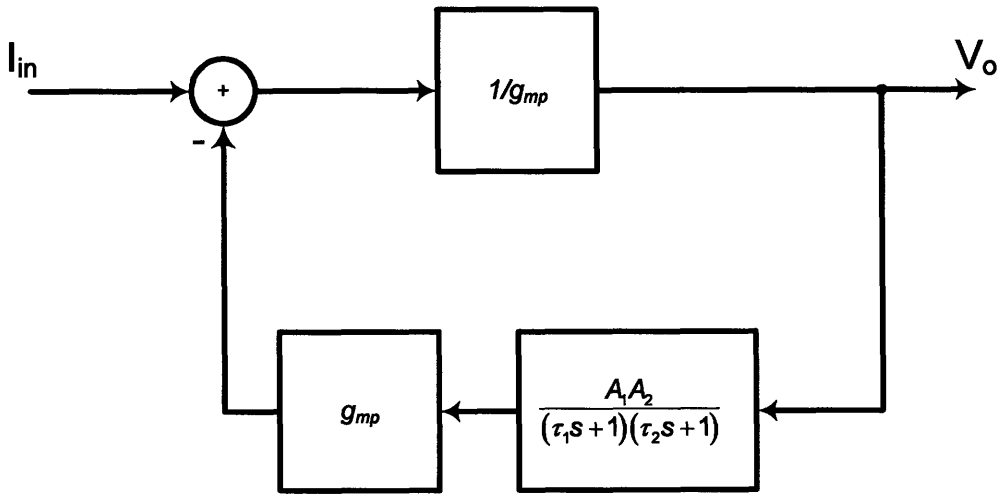


Fig. 5-7: Block diagram of current derivative circuit.

Analysis of the circuit shows that:

$$\tau_1 = A_2 C_C R_{O1} \qquad (64)$$
$$\tau_2 = C_{par} R_{O2}$$
$$\tau_1 >> \tau_2$$

$$\frac{V_o}{I_{in}} = \frac{1}{g_{mp} A_1 A_2} \frac{s\tau_1 + 1}{\frac{\tau_1 s}{A_1 A_2} + 1}$$
$$\approx \frac{\frac{sC_C R_{O1}}{g_{mp} A_1}}{\frac{C_C R_{O1} s}{A_1} + 1}$$

Fig. 5-8 shows the simulated magnitude and phase plots of the current to voltage transfer function of the circuit. The input current I_{in} is injected as shown in Fig. 5-6 and the output voltage V_o is observed at the same node. From the bode plot, the output voltage is the time derivative of the input current. Hence, the impedance looking into the node is inductive.

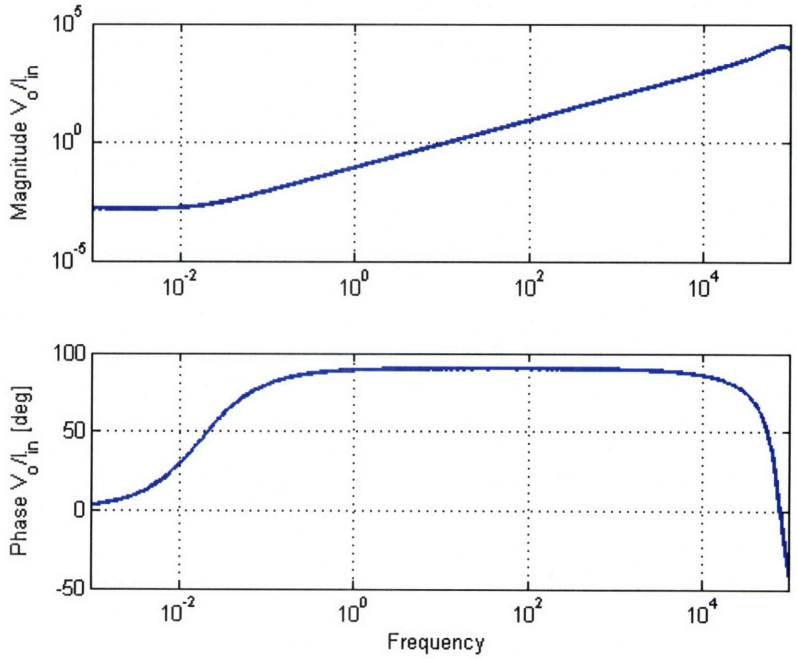


Fig. 5-8: Simulated bode plot of the current to voltage transfer function of the current derivative circuit depicted in Fig. 5-6.

Fig. 5-9 shows a bode plot of the open loop gain of the circuit for stability analysis.

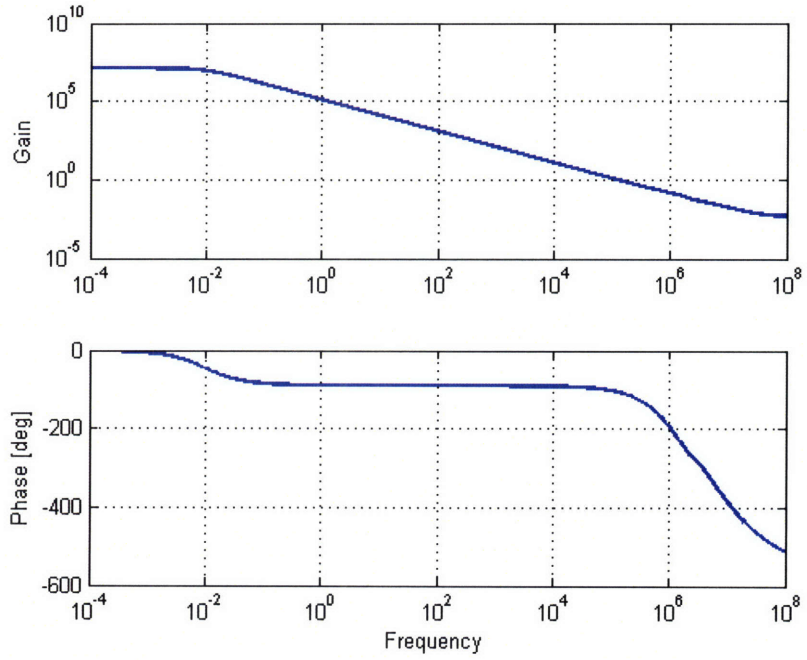


Fig. 5-9: Bode plot of the open loop gain.

5.2 VLSI two port equivalent of LC π-section

5.2.1 Two port representation of LC π-section

Fig. 5-10 depicts a passive LC π-section comprising resistors, inductors and capacitors. The series resistor R and shunt conductance G determine the loss associated with the LC circuit. The π-section may be represented by a two-port network as follows:

$$\begin{bmatrix} P_2 \\ U_2 \end{bmatrix} = F \begin{bmatrix} P_1 \\ U_1 \end{bmatrix} \tag{65}$$

where F is the 2x2 transmission (cascade) matrix given by:

$$F = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{66}$$

$$= \begin{bmatrix} \dfrac{(G+sC)(R+sL)}{2}+1 & -(R+sL) \\ \left\{\dfrac{(G+sC)(R+sL)}{2}+1\right\}\left\{\dfrac{(G+sC)(R+sL)+2}{2(R+sL)}\right\}-\dfrac{1}{R+sL} & -\dfrac{(G+sC)(R+sL)}{2}-1 \end{bmatrix}$$

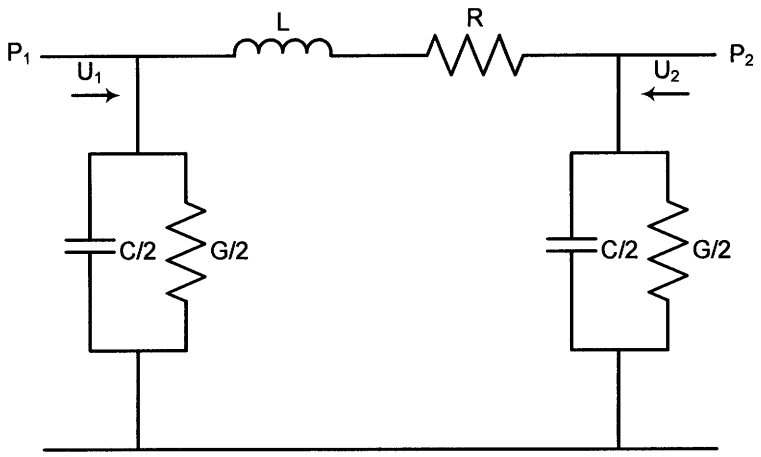


Fig. 5-10: π equivalent circuit of lossy LC section.

An example of a topology that is an active equivalent of the above network is shown in Fig. 5-11. It comprises two second order filter structures connected in parallel to serve as a bidirectional inductance. OTA G_1 and capacitance C_1 form a lossy current integrator whose time constant is determined by the output resistance R_O of OTA G_1 and capacitance C_1. It is this lossy current integration that eventually produces the lossy inductive effect corresponding to a series combination of a passive R and L. OTA G_2 together with capacitor C_2 form a shunt impedance to ground that corresponds to the G and C of the passive π-section.

In this topology, P_2 may be expressed in terms of P_1 and U_1 as follows:

$$P_2 = \left[\left(\frac{1}{G_2 G_1 R_O} + \frac{sC_1}{G_2 G_1}\right)(G_2 + sC_2) + 1\right] P_1 - \left(\frac{1}{G_2 G_1 R_O} + \frac{sC_1}{G_2 G_1}\right) U_1 \tag{67}$$

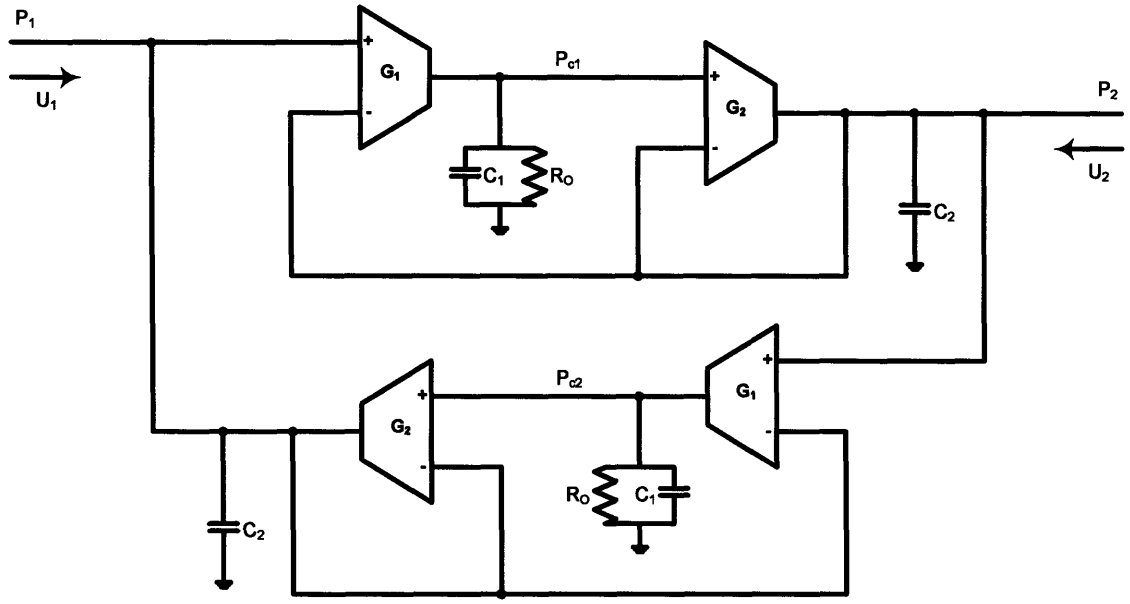


Fig. 5-11: An equivalent active implementation of Fig. 5-10 using two second order filter structures connected in parallel.

Comparing (67) with that obtained from (65) and (66), we can define an equivalent R and L for the active circuit as follows:

$$R = \frac{1}{G_2 G_1 R_O} \tag{68}$$

$$L = \frac{C_1}{G_2 G_1}$$

Using (67) and (68), U_2 may be expressed in terms of P_1 and U_1 as follows:

$$\begin{aligned} U_2 &= \left(G_2 + sC_2 + \frac{1}{R+sL}\right)P_2 - \frac{1}{R+sL}P_1 \\ &= \left(G_2 + sC_2 + \frac{1}{R+sL}\right)\{[(R+sL)(G_2+sC_2)+1]P_1 - (R+sL)U_1\} - \frac{1}{R+sL}P_1 \\ &= \left\{\left(G_2 + sC_2 + \frac{1}{R+sL}\right)[(R+sL)(G_2+sC_2)+1] - \frac{1}{R+sL}\right\}P_1 - \left(G_2 + sC_2 + \frac{1}{R+sL}\right)(R+sL)U_1 \end{aligned} \tag{69}$$

Comparing (69) with that obtained from (65) and (66) of the passive circuit, we observe that the transconductance G_2 and capacitance C_2 correspond to the shunt conductance G/2 and shunt capacitance C/2 of Fig. 5-10.

Note that OTA G_1 of both second order filter structures may be implemented with the same input differential pair.

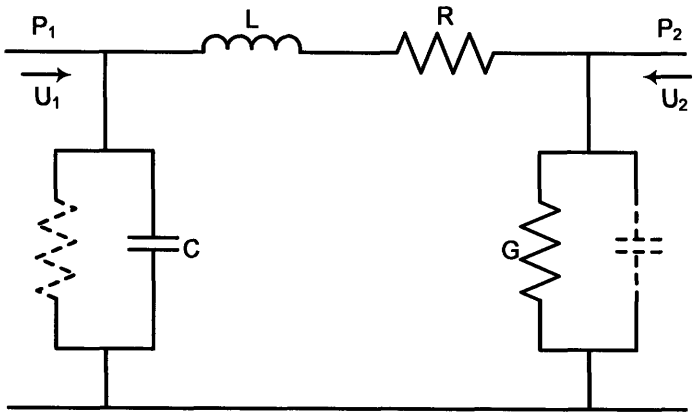
5.2.2 Continuously tunable LC π-section

Fig. 5-12 shows a circuit diagram of 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. A chain of basic two-port π-sections, concatenated end to end, yields a complete transmission line. In Fig. 5-12(b), wide linear range operational transconductance amplifiers G_1, G_{2A}, G_{2B} and capacitor C_1 form a tunable bidirectional gyrated inductance given by $L=C_1/G_1G_2$. A unidirectional gyrated inductance $Z_C=j\omega L_C$ with OTAs G_{2A} and G_3 implement a tunable shunt capacitance given by $C=G_2G_3L_C$. The series resistance R is given by $1/G_1G_2R_O$ where R_O 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 this topology, L and C may be controlled through a common transconductance G_2 by means of current. As C is proportional to G_2 and L is inversely proportional to G_2, the L and C of each section may be varied electronically to produce variations in the ratio L/C while maintaining the LC product constant. Fig. 5-12(b) lists the algebraic relationships that define the mapping of the circuit of Fig. 5-12(a) to the equivalent circuit of Fig. 5-12(b).

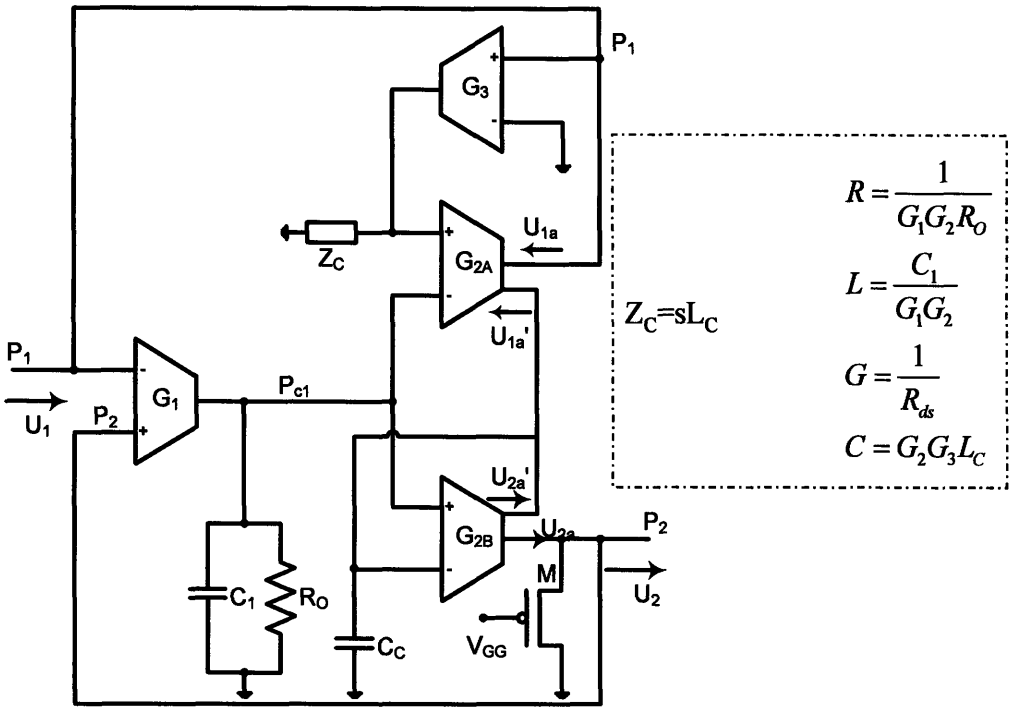
The values of G_{2A} and G_{2B} are designed to be equal since they are intended to represent a symmetrical L and R in series. However, unlike a real series L and R, the d.c. current from P_1 to P_2 is not zero when P_1 and P_2 are equal because of circuit offsets. Such mismatch leads to a large d.c. offset in each stage. In order to mitigate such mismatch, mirrored copies of currents U_{1a} and U_{2a} (U_{1a}' and U_{2a}') are compared and the difference integrated to generate an offset compensation bias voltage on capacitor C_C.

The length of each section may also be varied by adjusting G_1 and G_3 respectively. The series inductance L and the shunt capacitance C of each section are proportional to its length. Thus, a longer section has a larger L, C and vice versa. To increase the length of a section, we decrease G_1 and increase G_3 proportionately, such that both L and C increase proportionately in the algebraic expressions of Fig. 5-12. The control of the length of each section may be implemented globally such that all sections

are affected equally when the overall length of the vocal tract needs to be varied, as for example, when modeling female versus male speakers. In this case, the number of control signals are minimized. Alternatively, it may be implemented for every section, requiring separate control signals for each section. A good compromise is to implement fine control



(a)



(b)

Fig. 5-12: (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 the π-circuit model shown in (a).

for sections in crucial sections of the vocal tract such as the sections in the oral cavity close to the mouth. In particular, vowels such as /u/ and /o/ are often articulated with lip protrusion and we may increase the length of the two-port π-sections near the lips to model this effect.

Fig. 5-13 and Fig. 5-14 show experimental results obtained from a 16-stage cascade of two-port π-sections fabricated in a 1.5μm AMI CMOS process. The input to the cascade is a current source implemented on chip using a WLR OTA. The output current of the OTA is produced by setting the inverting input terminal to a fixed reference voltage V_{REF} and applying a sinusoidal a.c. voltage centered about V_{REF} to the non-inverting input terminal. The amplitude of the sinusoidal voltage is 0.5V. The biasing current of the OTA is kept constant. The sinusoidal current produced by the WLR OTA serves as the input to the 16-stage cascade of two-port π-sections under test. The output of the 16-stage cascade is terminated by a radiation impedance Z_{rad} (to ground). The radiation impedance is well modeled by an inductance L_{rad} implemented by gyrating a capacitor. Thus, the output is proportional to the voltage across $Z_{rad}=sL_{rad}$. The output signal is the voltage measured across $Z_{rad}=sL_{rad}$ via an on-chip buffer circuit.

Fig. 5-13 shows the measured signal and noise characteristics at the output of a 16-stage cascade of two-port π-sections. In this measurement, the cascade of two-port π-sections are configured electronically to form an equivalent uniform acoustic tube corresponding to the voiced phoneme /ə/. The signal frequency response is obtained by sweeping the frequency of the sinusoidal a.c. voltage source from 100Hz to 8kHz. The noise characteristic is obtained when the a.c. source is set to zero. The measured SNR is 64dB, 66dB, and 63dB for the first three formant resonances (F1, F2, and F3) of /ə/.

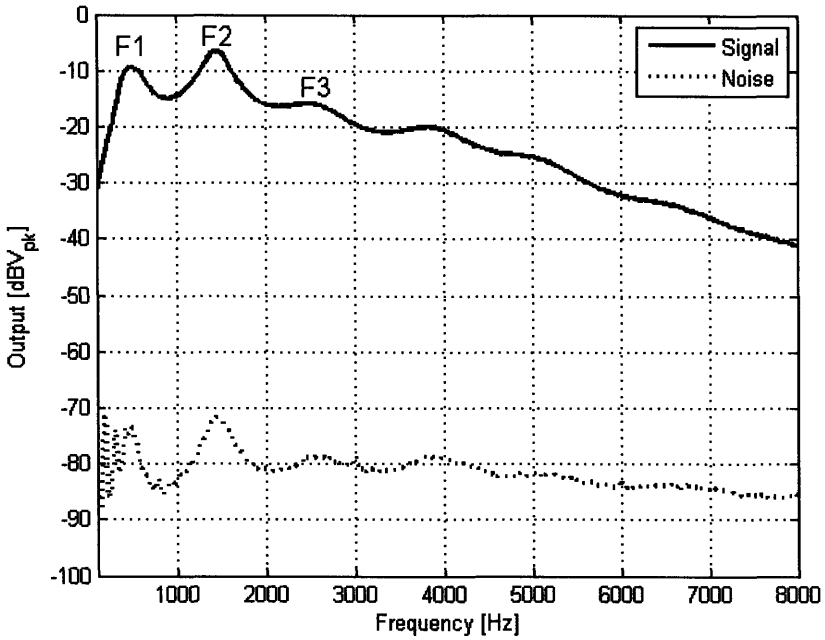
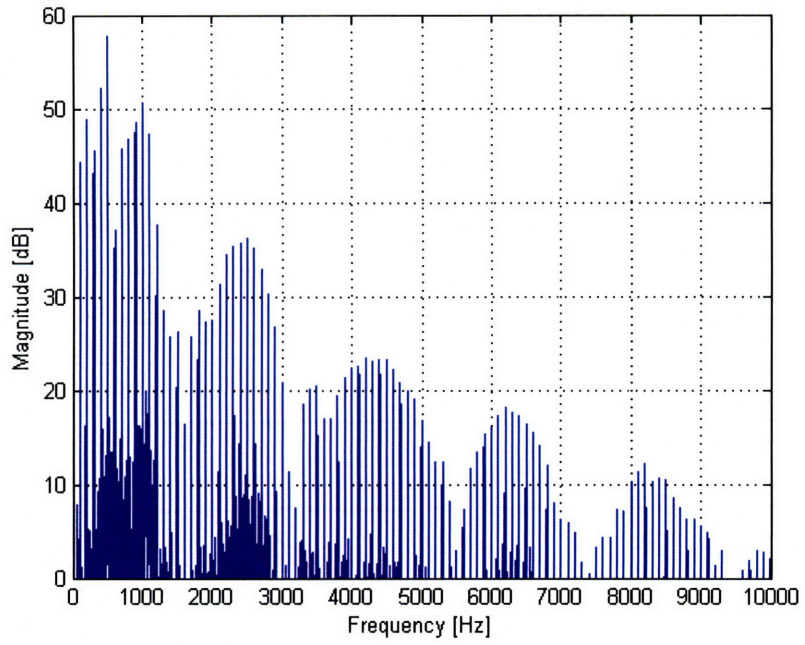


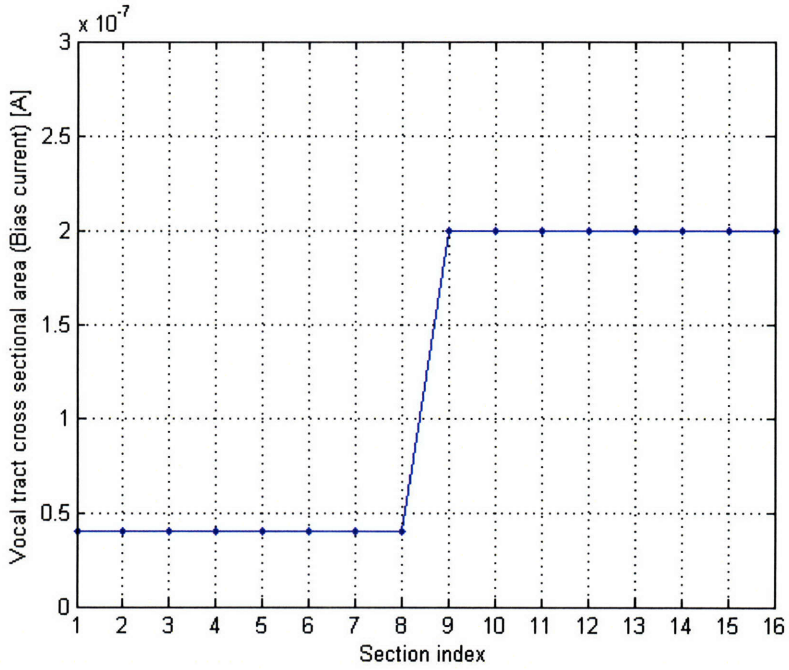
Fig. 5-13: Measured signal and noise characteristics as a function of sinusoidal input frequency.

Fig. 5-14(a) shows the measured output spectrum when the two-port π-sections are configured to form an acoustic tube with an area profile corresponding to /a/ and depicted in Fig. 5-14(b). In this measurement, the input is a LF [45] glottal pulse train with a pitch period of 10ms. The harmonics of the periodic glottal pulse train are clearly illustrated in Fig. 5-14(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).

Fig. 5-15(a) shows an equivalent π-circuit model of a cylindrical section of acoustic tube with non-rigid walls. It has a shunt impedance element Z_W comprising a series combination of R_W, L_W, and C_W that is connected in parallel with capacitance C and conductance G to model the mass and compressibility of the non-rigid walls. Fig. 5-15(b) is a circuit diagram of the tunable two-port π-section of Fig. 5-12 modified to incorporate the effect of Z_W. To this end, an additional OTA G_4 is employed to gyrate an impedance Z_{RLC} comprising a parallel combination of R_W, L_W and C_W.

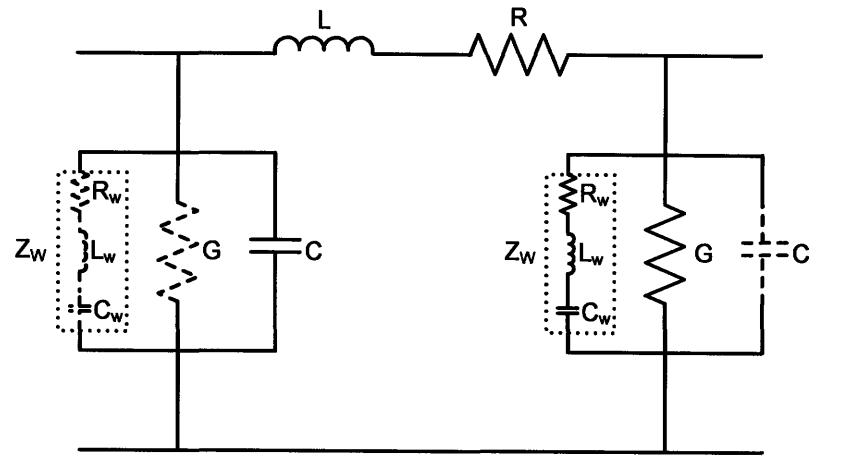


(a)

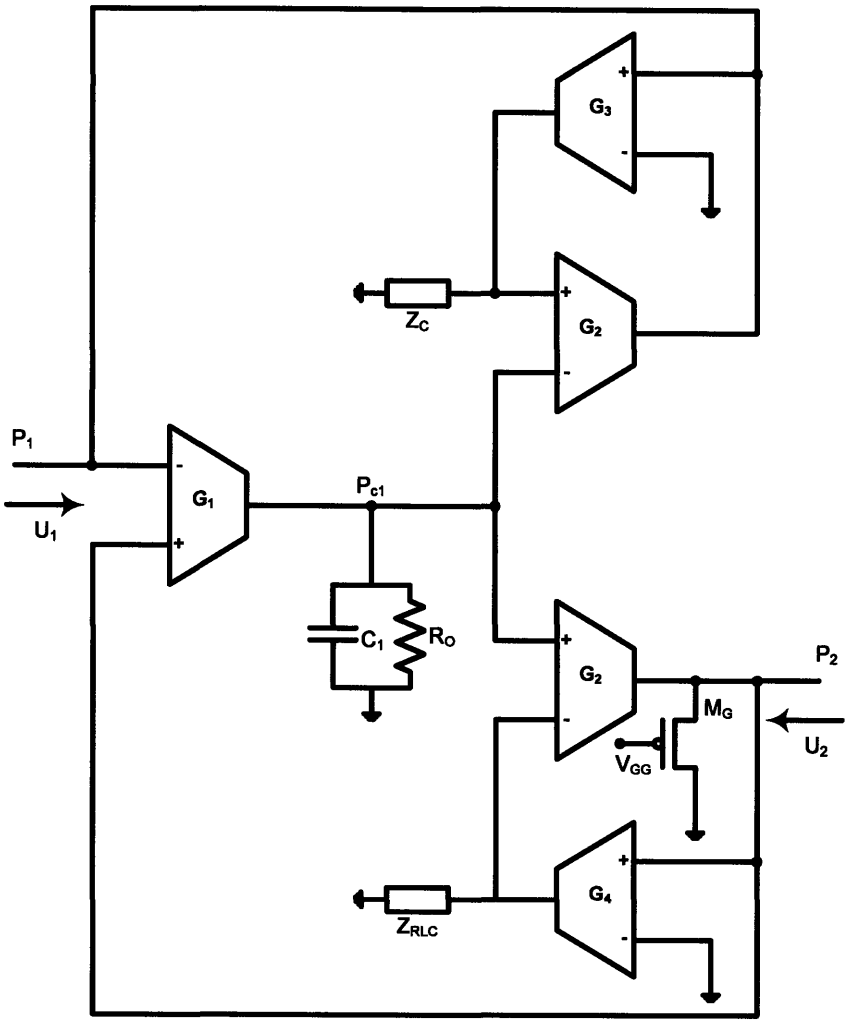


(b)

Fig. 5-14: (a) Measured spectrum at output of 16-stage cascade of two-port π-sections when the π-sections are configured to form an acoustic tube with vocal tract area profile shown in (b).



(a)



(b)

Fig. 5-15: (a) Equivalent π-circuit model of a cylindrical section of acoustic tube with non-rigid walls. (b) Circuit diagram of tunable two-port π-section modified to include the effect of non-rigid wall impedance.

5.2.3 *Discretely tunable LC π-section*

A tunable bidirectional gyrated inductance given by $L=C_1/G_1G_2$ is constructed as shown in Fig. 5-16 using wide linear range operational transconductance amplifiers G_1, G_2 and capacitor C_1. For capacitive tunability, the capacitor C_2 may be implemented using an array of binary weighted capacitors in conjunction with a switching network. Note that the input terminals of OTAs G_2 share the same signals i.e. P_{c1} and AC ground. Hence, they may be implemented using the same input differential pair. For the same input differential voltage, the respective output currents should ideally be equal and opposite. Fig. 5-17 shows an equivalent circuit implementation of Fig. 5-16 using a fully differential input-output WLR OTA G_2 in place of the two single-ended WLR OTAs G_2 of Fig. 5-16. The fully differential circuit implementation of OTA G_2 shown in Fig. 5-17 eliminates potential offsets due to device mismatch that may otherwise occur in the single ended implementation of Fig. 5-16.

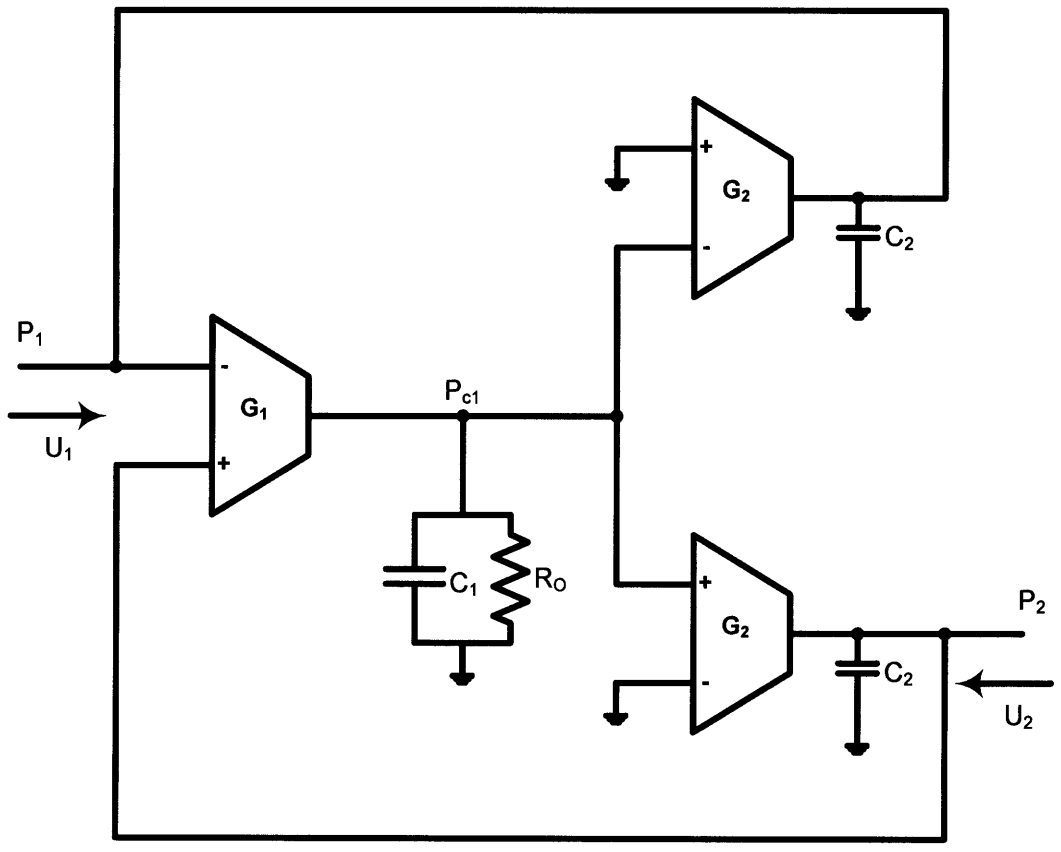


Fig. 5-16: Circuit diagram of tunable two-port π-section that is electrically equivalent to a π-circuit model of a cylindrical section of acoustic tube assuming rigid walls. Capacitor C_2 is implemented using a binary weighted capacitor array for tunability.

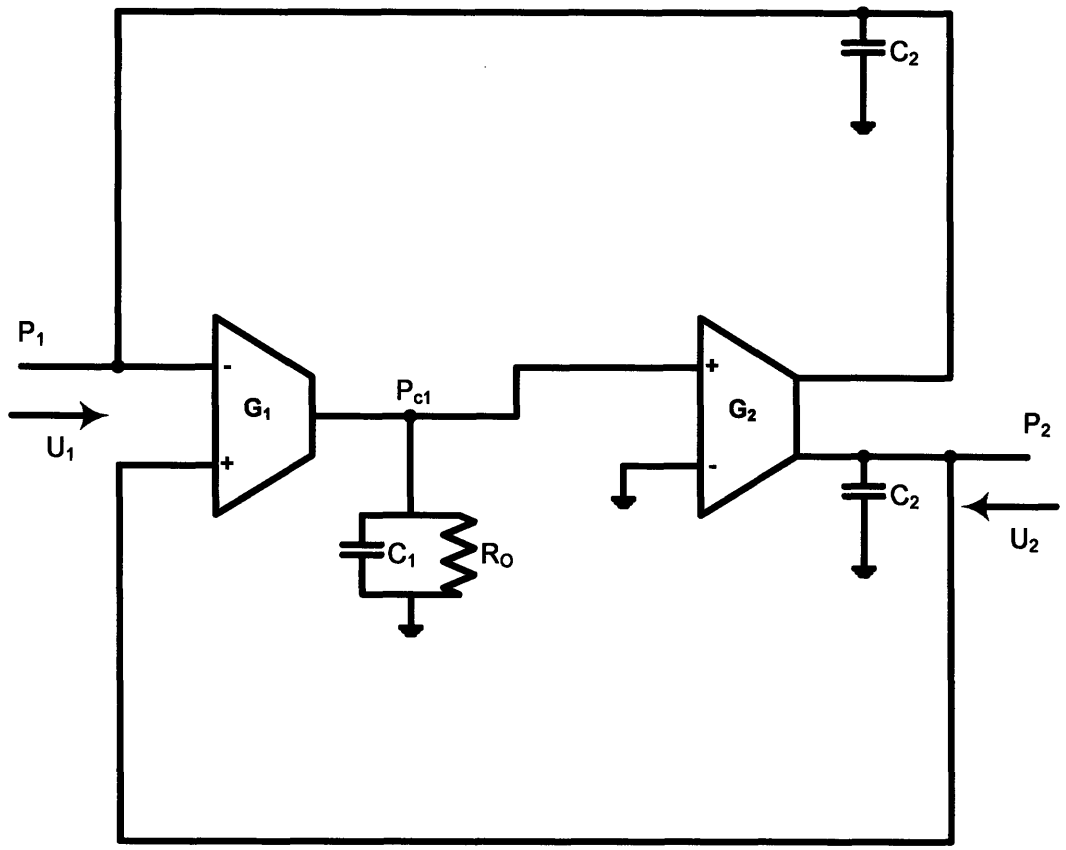


Fig. 5-17: An equivalent implementation of Fig. 5-16 using a fully differential input-output WLR OTA G_2.

In the implementation of Fig. 5-17, the inductance $L=C_1/G_1G_2$ is tuned discretely using a binary weighted current source array to bias OTA G_2, such that the same digital signals used to control the tunable capacitance C_2 may be used to control the current source array. In this manner, the control of each π-section is greatly simplified as only one set of digital signals are required to ensure that cross-sectional area variations produce a change in the L/C ratio but not the LC product. To implement cross-sectional area variations of 0.1 cm^2 to 10 cm^2, a seven bit digital signal is required. Consequently, a seven bit capacitor and current source array are required for each π-section.

5.2.4 The supraglottal vocal tract as a cascade of two-port equivalent π-sections

Fig. 5-18 shows a schematic diagram of the supraglottal vocal tract using our discrete transmission line model where each passive LC section is represented as a two-port equivalent and implemented in VLSI technology using the topologies described in

§5.2.2 and §5.2.3. Multiple two-port equivalents are cascaded to form an active transmission line. The transmission line models the pharyngeal, oral and nasal cavities. Each active two-port network is defined by a transmission matrix $F_{i,j}$ corresponding to the passive LC section. At the source end (beginning of pharyngeal tube), the transmission line is connected to a sub-glottal pressure source P_S through a glottal impedance Z_G representing the constriction at the glottal opening. The transmission line is terminated at the nostrils and mouth by radiation impedances Z_N and Z_M. During the production of nasal sounds, the oral and nasal cavities are coupled through the velar opening which is represented by an impedance Z_V. For non-nasal sounds, Z_V has infinite impedance as the velar opening is closed and the oral and nasal cavities are completely decoupled.

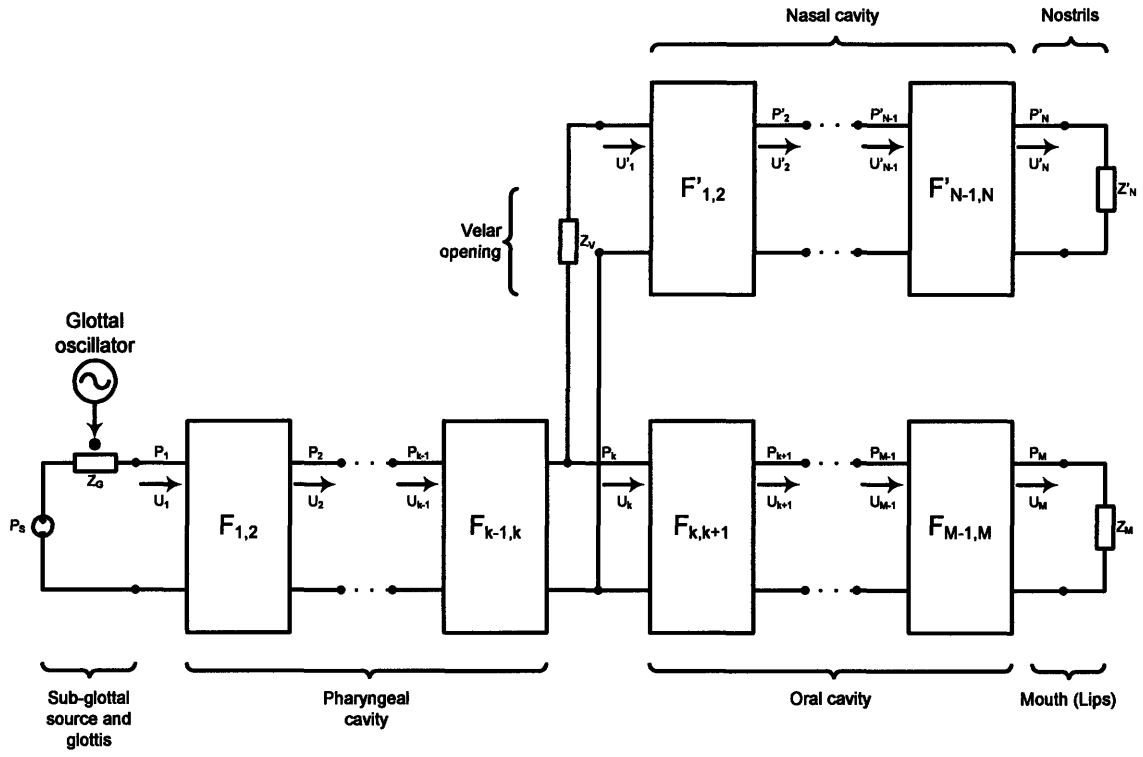


Fig. 5-18: Discrete transmission line representation of vocal tract using cascade of two-port networks.

INTENTIONALLY LEFT BLANK

Chapter 6 VLSI IMPLEMETATION OF VOCAL TRACT

In this chapter, we present an experimental integrated-circuit analog 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. Such silicon vocal tracts 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, and/or find applications in real-time low-power speech production, compression, speaker identification or bionic speech-prosthesis systems. 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 [1][29]. Below, we describe how our chip is architected and demonstrate some of its potential applications. The 275μW analog vocal tract chip can be used with auditory processors in a feedback speech locked loop—analogous to a phase locked loop—to implement speech recognition that is potentially robust in noise. It is also useful for low-power, real-time speech production, speech compression and bionic speech-prosthesis systems.

6.1 Transmission line vocal tract

A uniform cylindrical tube section of the vocal tract was shown to be analogous to a discrete section of an electrical transmission line. Specifically, sound pressure P corresponds to electrical voltage V, volume velocity U corresponds to electrical current I, acoustic inertance ρ/A is analogous to electrical inductance L and acoustic compliance $A/\rho c^2$ is analogous to electrical capacitance C. Our transmission line vocal tract consists of a cascade of two-port equivalent LC π-sections representing a concatenated tube approximation with abutment of short cylindrical acoustic tubes end-to-end. The electrical parameters L and C of each section of the line are determined by the cross-

sectional area of the corresponding acoustic tube. These parameters vary dynamically as the vocal tract configuration is varied to produce different sounds.

Fig. 6-1 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. The error introduced by spatial quantization is kept small by making ℓ short compared to the wavelength of sound corresponding to the maximum frequency of interest. Each two-port is an electrical equivalent of a 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:

(i) a periodic source at the glottis and/or

(ii) a turbulent noise source P_{turb} at some point along the tube.

In Fig. 6-1, the glottal source is represented by a voltage source P_{alv} (corresponding to the alveolar pressure source) with variable source impedance Z_{GC}, that is modulated by a glottal oscillator or an AC current source U_{gl} (corresponding to a volume velocity source). In the former, 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 AC current source U_{gl} is implemented using a WLR OTA with a cascoded output stage to obtain high output impedance.

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}.

Fig. 6-2 shows a die photo of our AVT fabricated in a 1.5μm AMI CMOS process. The AVT is comprised of a cascade of 16 tunable two-port π-sections each representing a uniform tube of adjustable length. The chip consumes less than 275μW of

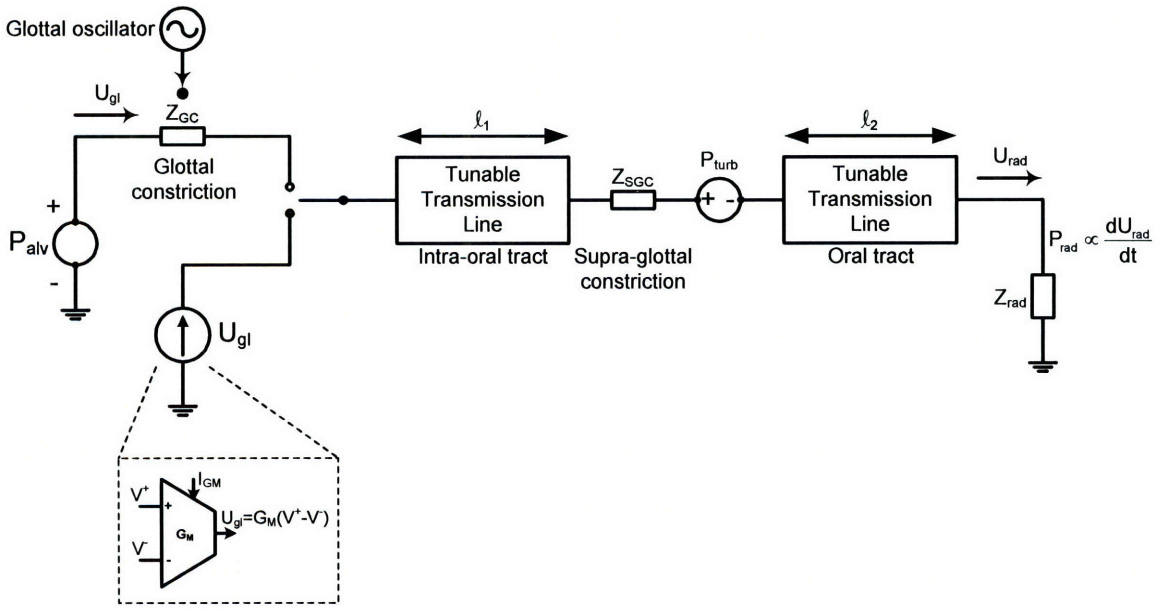


Fig. 6-1: Schematic diagram of transmission line vocal tract.

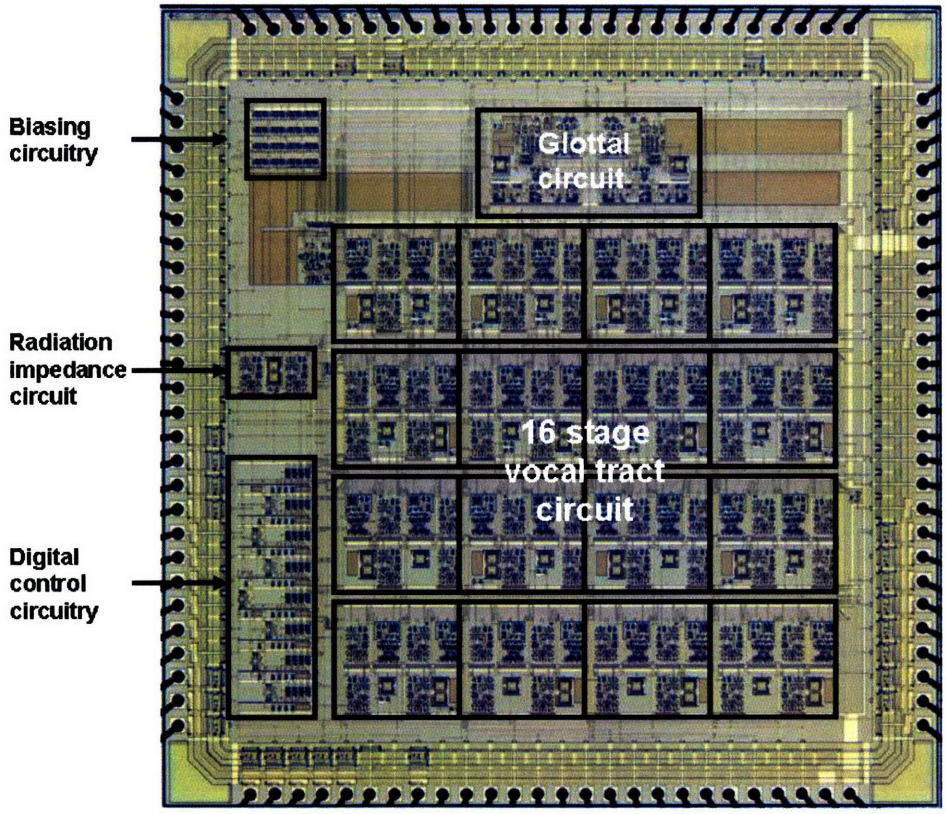


Fig. 6-2: Chip micrograph of 16-stage analog vocal tract fabricated in a 1.5μm AMI CMOS process.

power when operated with a 5V power supply. 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/.

6.2 Subglottal system

6.2.1 Current source circuit model of the glottis

The subglottal system is a network of tubes including the trachea and bronchi, that extend to the lungs. The subglottal system is the power source for speech production whereas the supra-glottal vocal tract is responsible for modulating the airflow to produce speech. The subglottal system comprises the respiratory system (lungs), the subglottal cavity and the vocal folds at the glottis. During speech production, air ejected from the lungs flow in pulses. A typical glottal pulse was described in Fig. 2-3 as a volume velocity waveform and its derivative. As described in § 2.2.1, the glottal source has a high acoustic impedance compared to the driving point impedance of the vocal tract at most frequencies of interest. Consequently, a current source may be used to provide the necessary periodic excitation of the supraglottal vocal tract. A simple current source may be constructed using a WLR OTA such as the one shown in Fig. 6-3. The output current of the WLR OTA is proportional to the voltage difference across its two input terminals:

$$I_{out} = G_M \left(V^+ - V^- \right) \tag{70}$$

$$G_M = \frac{I_{GM}}{V_L}$$

where G_M and V_L are the transconductance and linear range of the WLR OTA respectively.

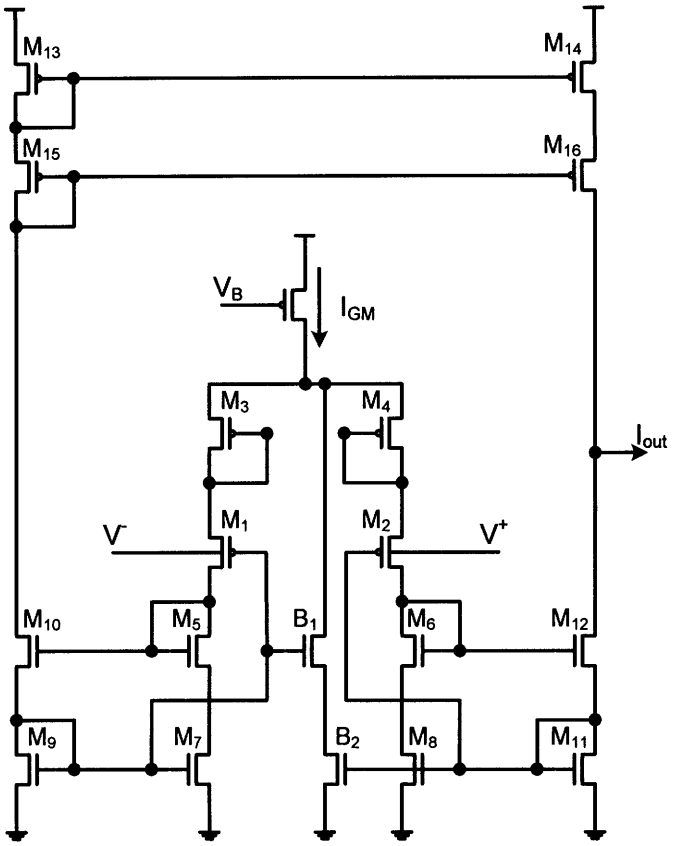


Fig. 6-3: Current source circuit that implements volume velocity source at the glottis.

6.2.2 Nonlinear impedance circuit model of the glottis

During speech production, air is expelled from the lungs causing the vocal folds to vibrate as a relaxation oscillator. The lungs provide a source of alveolar pressure and consequently may be represented by a voltage source P_{alv}. Vocal fold vibration produces a periodic interruption of the air flow from the lungs to supraglottal vocal tract. The ejected air stream is modulated by the glottal constrictions formed when the vocal folds open and close.

We model a constriction at the glottis using a linear resistance in series with a nonlinear resistance. During steady laminar flow, the volume velocity (or analogously current) U_{gl} is linearly related to pressure drop (or analogously voltage drop) ΔP_1. Hence, the linear resistance models the viscous dissipation associated with laminar flow. It is proportional to the viscosity of air μ, the length of the glottal constriction l_{gl}, and

inversely proportional to the square of the cross sectional area of the constriction and has a linear I-V characteristic given by:

$$\Delta P_1 = \frac{8\pi\mu l_{gl}}{A_{gl}^{\ 2}} U_{gl} \tag{71}$$

In addition to the viscous resistance, there is also a nonlinear kinetic resistance due to eddy current losses. The additional energy losses in the flow arise in the vicinity of the entrance to the constriction, where there is a transition from a wide tube to a narrow section, and again in the vicinity of the exit, where there is an expansion of cross sectional area. Empirically, it is found that the pressure drop due to eddy current losses is proportional to the square of the volume velocity in the constriction. In other words, the current U_{gl} is proportional to the square root of the voltage ΔP_2 across its terminals. In circuit terminology, the nonlinear resistance has a square-root I-V characteristic:

$$\Delta P_2 = \frac{\rho}{2} \frac{U_{gl}^{\ 2}}{A_{gl}^{\ 2}} \tag{72}$$

Our circuit model of the glottis consists of two glottal constrictions in series to represent the upper and lower part of the vocal folds. As the upper and lower folds open and close in a periodic fashion (phase-shifted with respect to one another), the impedance of each glottal constriction is varied by a glottal oscillator in a corresponding manner.

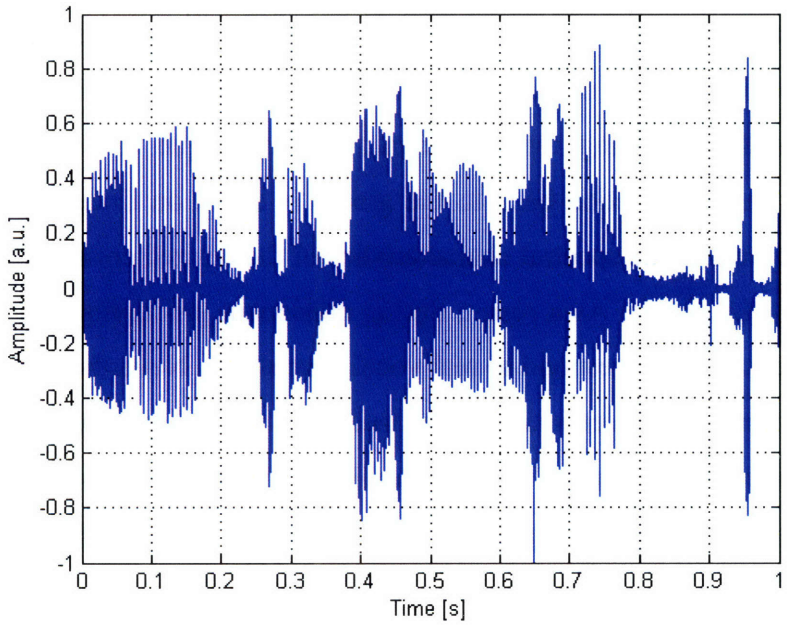
6.3 Approximate methods for consonant production

6.3.1 "Input refer" noise source to glottis

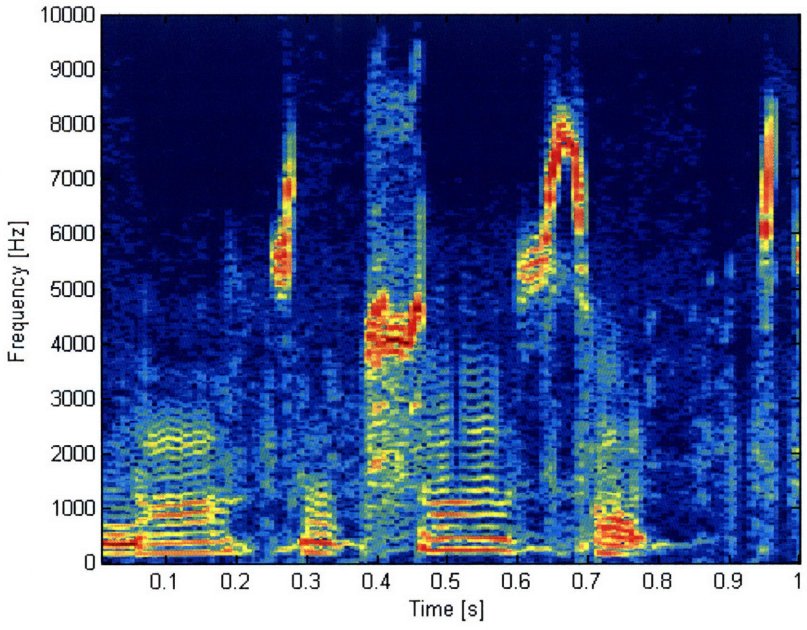
A periodic glottal source provides the stimulus to produce voiced speech, which includes vowels and voiced consonants. The source of excitation for consonant sounds originates in the turbulent flow of air through a constriction located in the oral cavity of the supraglottal vocal tract. For unvoiced consonants such as fricatives, noise resulting from the turbulent flow located a little downstream of the constriction provides the source of fricative energy. For unvoiced stop consonants, a transient source caused by the sudden pressure release following a closure somewhere in the oral cavity immediately precedes the frication noise due to turbulence. In the case of voiced consonants, both the periodic glottal stimulus and the turbulent noise generation occur simultaneously.

As illustrated in Fig. 6-1, the source of turbulence downstream of a constriction may be modeled as a bandlimited white noise source P_{turb} inserted between two adjacent π-sections. The transfer function of the noise source to the output (lips) is determined by the vocal tract profile producing the consonant. The technique to "input refer" the turbulent source at the constriction back to the glottis is based on finding a vocal tract profile that best reproduces the spectral characteristics of the desired consonant when that profile is excited at the glottis. As a result, during consonant production the vocal tract profile does not necessarily correspond to real vocal tract shape. Nevertheless, vocal tract profiles of vowels continue to correspond to the true vocal tract shape. Input referring the turbulent noise source to the glottis consolidates the various excitation sources to the "input" of the vocal tract, i.e., the glottis, thereby making VLSI implementation and subsequent control simple and easy to manage. On the other hand, the use of non-physiological vocal tracts for consonant production makes interpolation between consonant-vowel and vowel-consonant transitions non-trivial as it is not clear that they are linearly interpolatable. As a result, there is a risk of introducing undesirable components during these transitions.

Fig. 6-4 and Fig. 6-6 show the results of speech produced by the AVT by input referring the noise source to the glottis. Fig. 6-5 and Fig. 6-7 show the vocalograms used to generate the speech. The spectrograms of the voice recordings used to derive the vocalograms are depicted in Fig. 6-9 ("Massachusetts") and Fig. 6-14 ("Technology"). Fig. 6-4 shows the time domain waveform and spectrogram of the word "Massachusetts" synthesized by the AVT. The intensity information on the spectrogram is color coded with blue representing low intensity and red representing high intensity. The vocalogram used to control the AVT is shown in Fig. 6-5. The cross-sectional area is colour coded with blue representing small areas and red representing large areas. Fig. 6-6 shows the time domain waveform and spectrogram of the word "Technology" synthesized by the AVT. The vocalogram used to control the AVT is shown in Fig. 6-7.



(a)



(b)

Fig. 6-4: (a) Time domain waveform and (b) spectrogram of the word "Massachusetts" synthesized by the AVT by input referring turbulent noise source to the glottis.

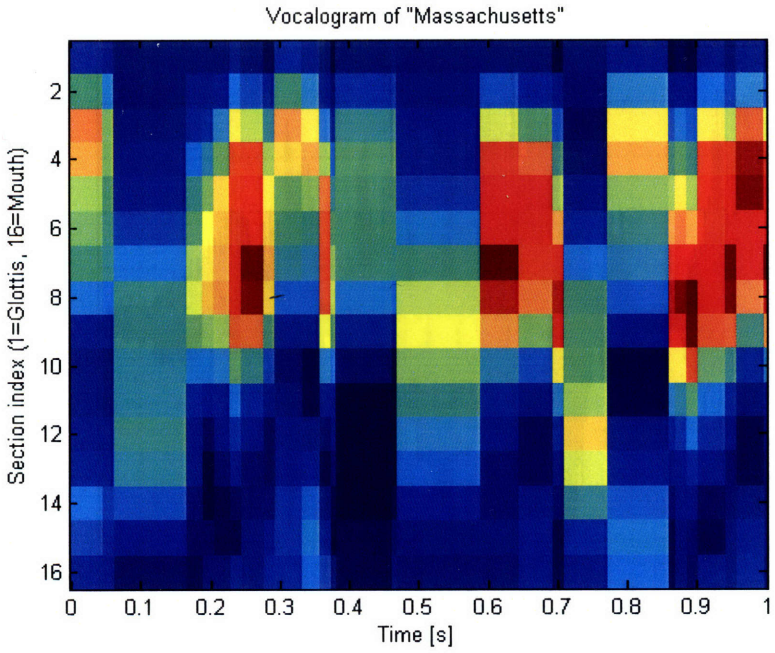
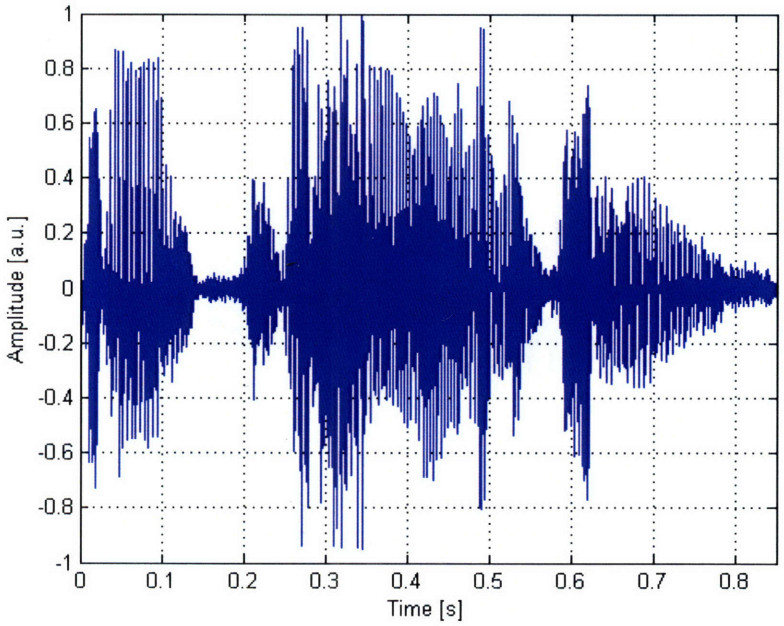
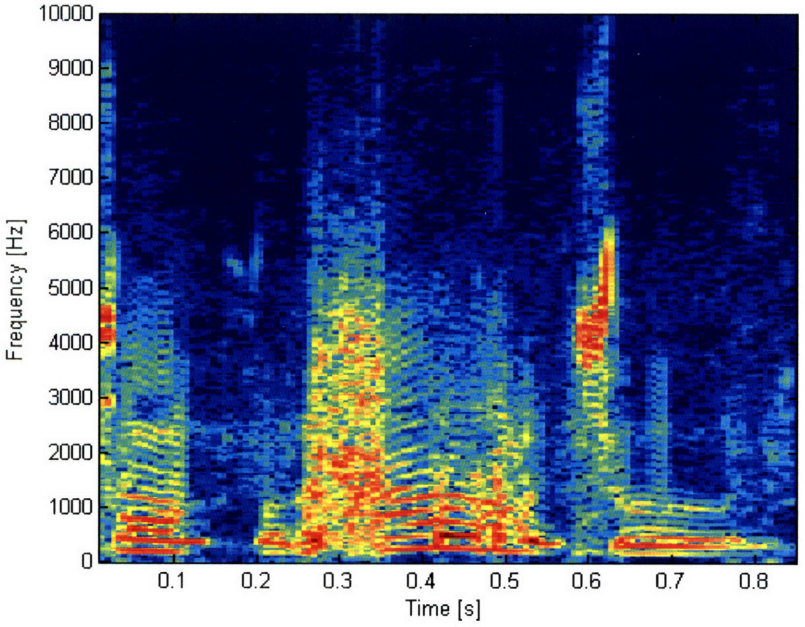


Fig. 6-5: Vocalogram of the word "Massachusetts" synthesized by input referring turbulent noise source to the glottis.



(a)



(b)

Fig. 6-6: (a) Time domain waveform and (b) spectrogram of the word "Technology" synthesized by the AVT by input referring turbulent noise source to the glottis.

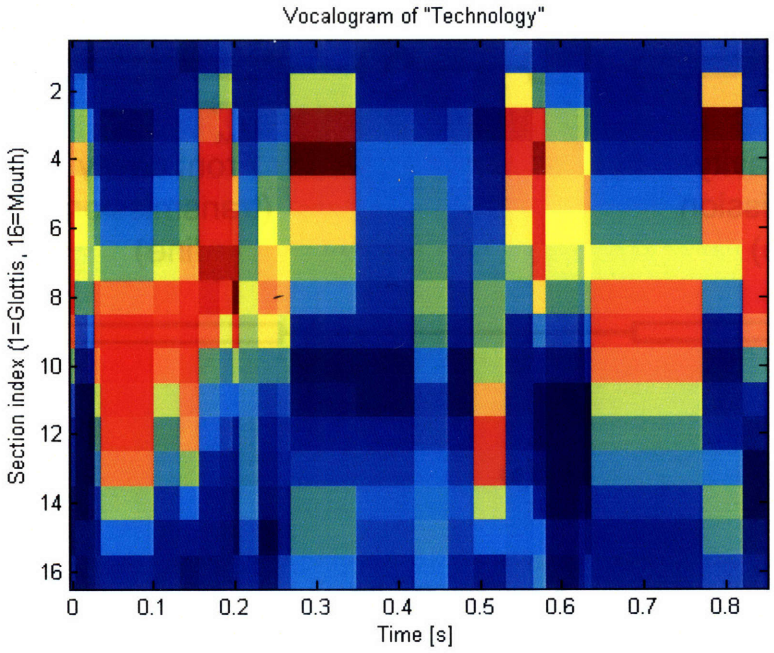
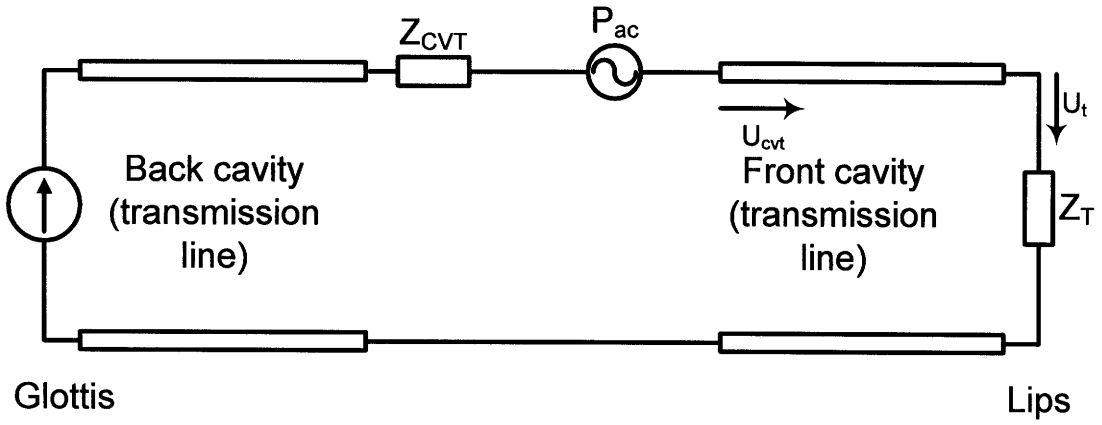


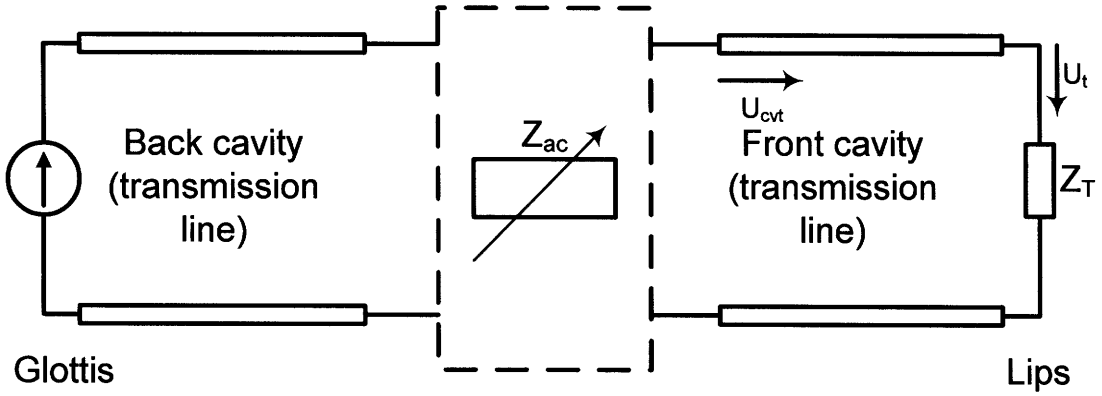
Fig. 6-7: Vocalogram of the word "Technology" synthesized by input referring turbulent noise source to the glottis.

6.3.2 Impedance modulation method

The impedance modulation technique is another method to approximate the turbulent noise source at the constriction that is convenient for circuit design. Fig. 6-8(a) is a circuit diagram showing the circuit equivalent of a vocal tract profile that has a constriction located somewhere in the oral cavity. In Fig. 6-8(a), we represent the constriction with an impedance Z_{CVT} and the turbulence generated in the vicinity of the constriction with a turbulent AC source P_{ac}. The location of P_{ac} is downstream of the constriction location. The front and back cavities, corresponding to portions of the vocal tract located anterior and posterior to the constriction respectively, are represented as transmission lines appropriately terminated at the glottis and at the mouth. The general idea of the impedance modulation technique is illustrated in Fig. 6-8(b). The technique involves generating noise downstream of the constriction by modulating the area of the section in front of the constriction, i.e., we approximate the noise generated by P_{ac} with a signal produced by modulating the cross sectional area of a two-port section downstream of the constriction in a noisy fashion.



(a)



(b)

Fig. 6-8: An approximate method to produce turbulence at the constriction by modulating the impedance downstream of the constriction in a noisy fashion.

Fig. 6-9 depicts the spectrogram of a recording of the word "Massachusetts" lowpass filtered at 5.5kHz. The recording has a female voice. During the training phase, the AVT undergoes the babbling process described in Chapter 3 to produce an articulatory codebook. Using the recording as a target sound, an optimal articulatory trajectory, given by the vocalogram of Fig. 6-10, is derived through dynamic programming. The motor-domain vocalogram is used to drive the AVT, in conjunction with the impedance modulation technique to produce the synthesized speech shown in Fig. 6-11. 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. 6-9) and the synthesized sound (Fig. 6-11(b)), 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. 6-9 have been re-introduced by the AVT in Fig. 6-11(b). 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.

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. 6-12 show 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. 6-11(b), the resulting speech has a lower third formant, F_3, in the voiced segments, which is consistent with a longer vocal tract. Note that with only pitch scaling an unrealistic voice is obtained.

Fig. 6-13 show the synthesis results when the female vocal tract of Fig. 6-11 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 of the synthesized speech has a third formant that is located at approximately the same frequency as the one in Fig. 6-11. The result is not surprising as the speech in Fig. 6-11 and Fig. 6-13 are produced by vocal tracts of the same length albeit with different pitch periods.

Fig. 6-14 shows the spectrogram of a recording of the word "Technology" lowpass filtered at 5.5kHz. The vocalogram used to control the AVT for the synthesis of the word is shown in Fig. 6-15. Fig. 6-16 shows the time domain waveform and spectrogram of the word synthesized by the AVT.

Fig. 6-17(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. 6-17(b) shows the spectrogram of the same word re-synthesized by our AVT using the scheme illustrated in Fig. 6-18. In Fig. 6-17(b), it is

evident that high frequency speech components that were absent in Fig. 6-17(a) have been introduced by the AVT. It is also noteworthy that the noise added to the recording of Fig. 6-17(a) is reduced in Fig. 6-17(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.

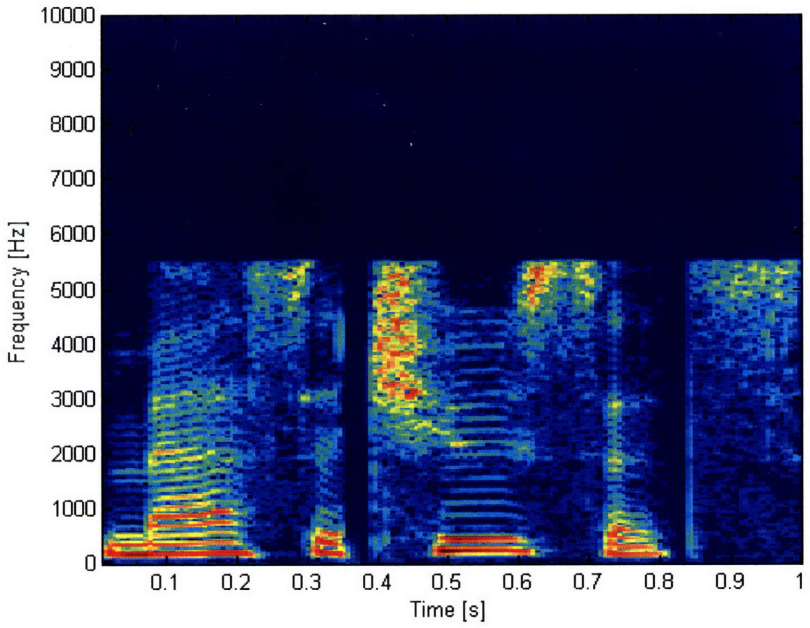


Fig. 6-9: Spectrogram of a recording of the word "Massachusetts" lowpass filtered at 5.5kHz. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

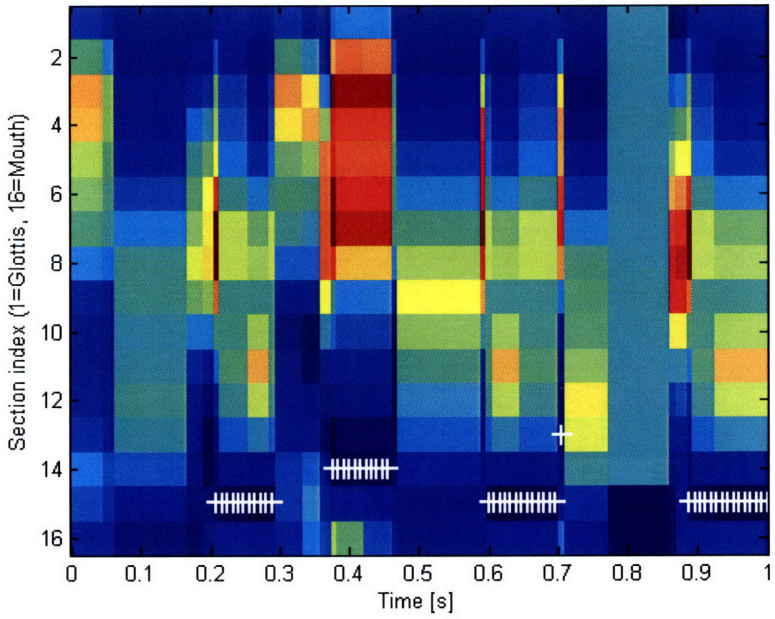
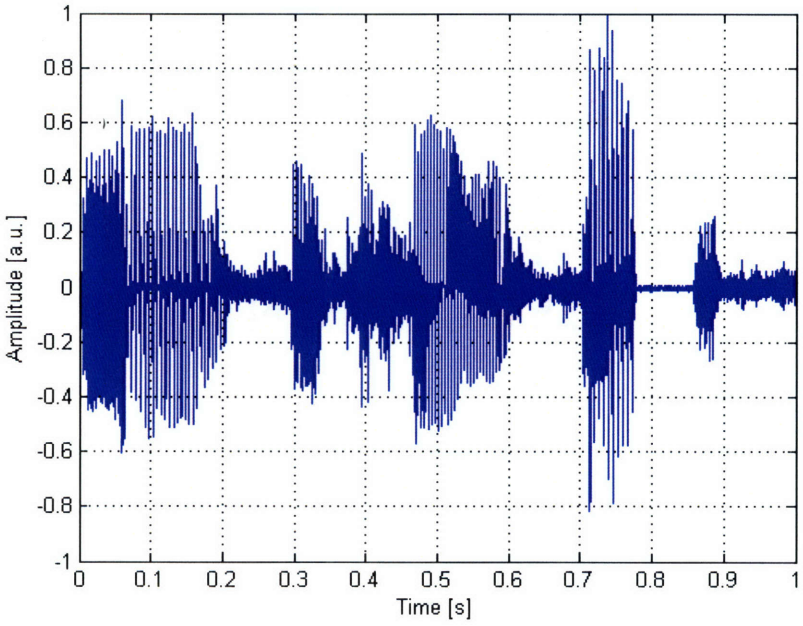
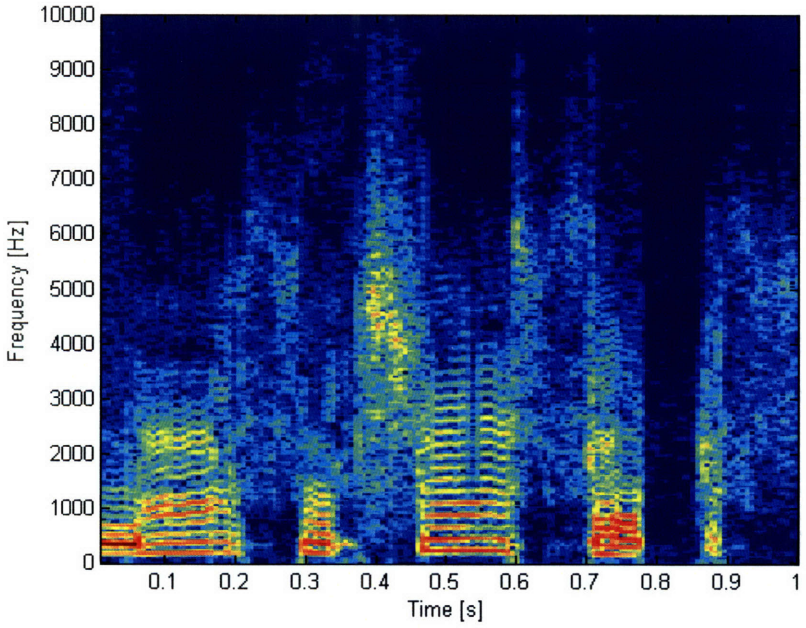


Fig. 6-10: Vocalogram of the word "Massachusetts". The white plus sign markers on the vocalogram indicate the position of the constriction when the speech segment is a consonant.

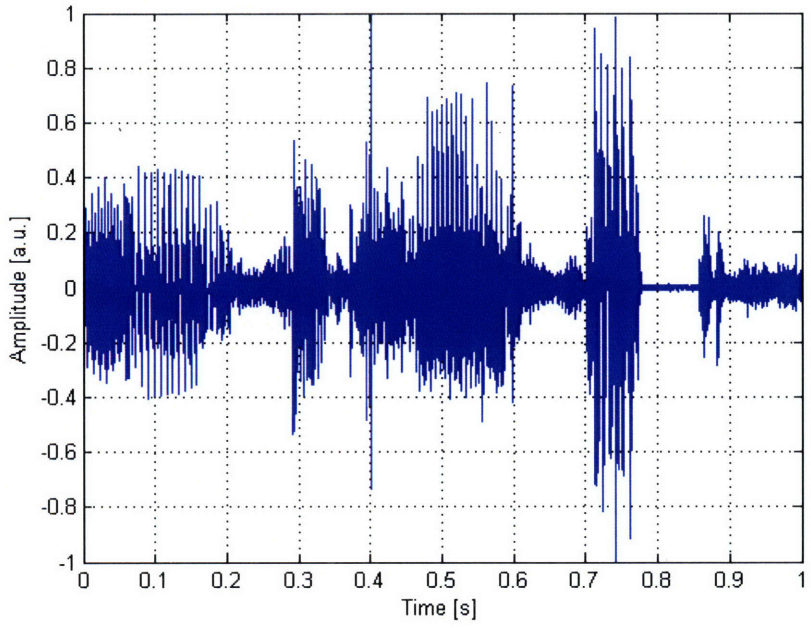


(a)

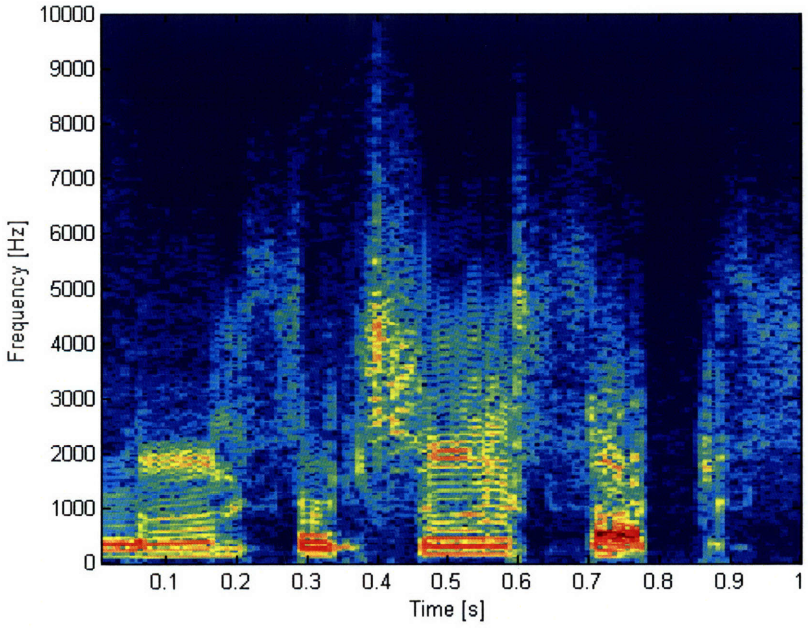


(b)

Fig. 6-11: (a) Time domain waveform and (b) spectrogram of the word "Massachusetts" synthesized by the AVT using the impedance modulation technique for consonants. A female vocal tract and the original extracted pitch contour are used.

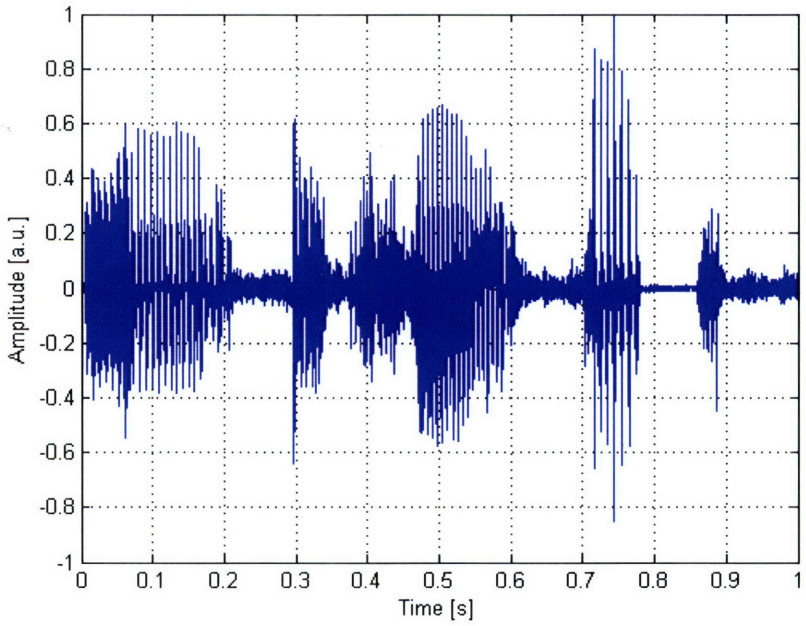


(a)

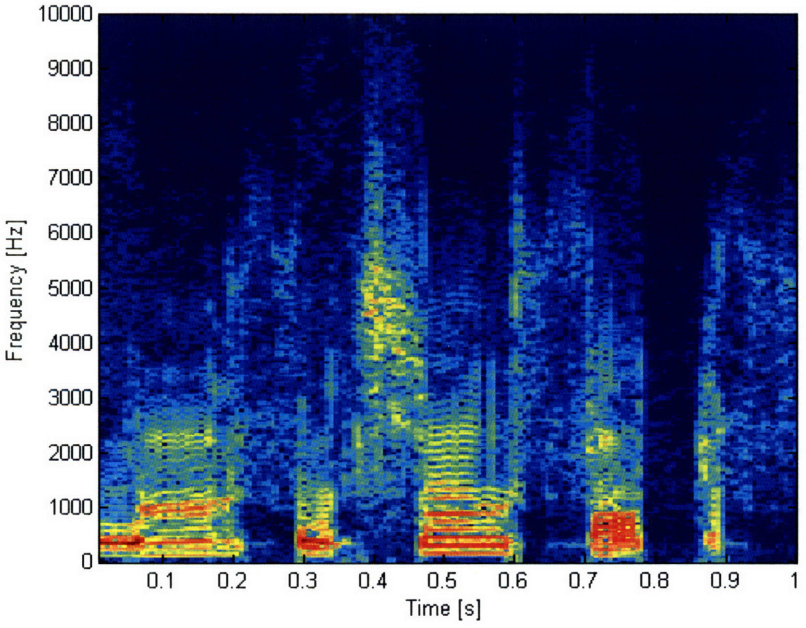


(b)

Fig. 6-12: (a) Time domain waveform and (b) spectrogram of the word "Massachusetts" synthesized by the AVT using the impedance modulation technique for consonants. A male vocal tract is used and the extracted pitch contour is scaled down by a factor of 1.5.



(a)



(b)

Fig. 6-13: (a) Time domain waveform and (b) spectrogram of the word "Massachusetts" synthesized by the AVT using the impedance modulation technique for consonants. A female vocal tract is used and the extracted pitch contour is scaled down by a factor of 1.5.

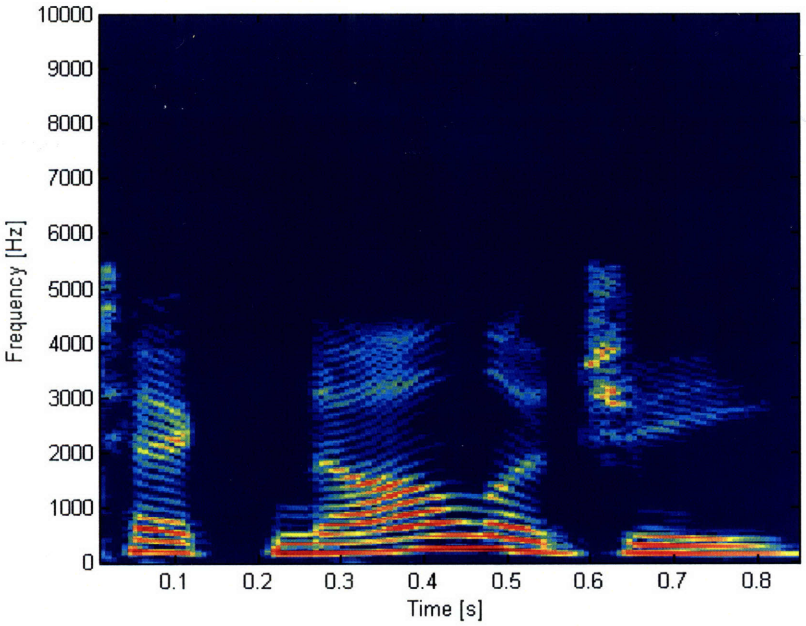


Fig. 6-14: Spectrogram of a recording of the word "Technology" lowpass filtered at 5.5kHz. Regions in red indicate the presence of high intensity frequency components whereas regions in blue indicate low intensity.

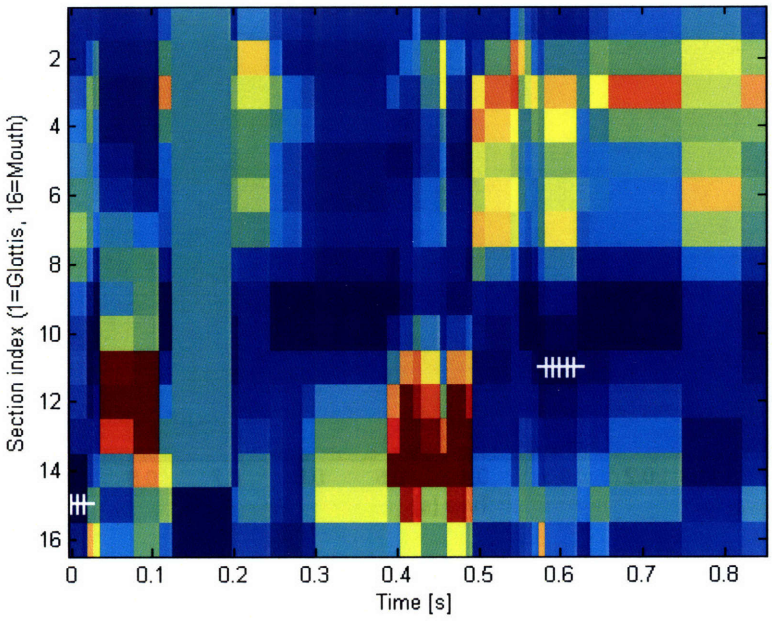
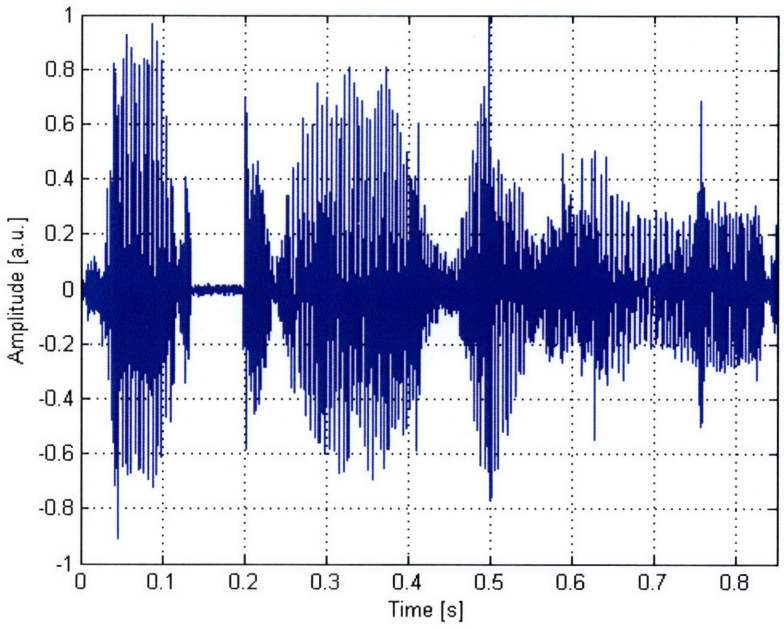
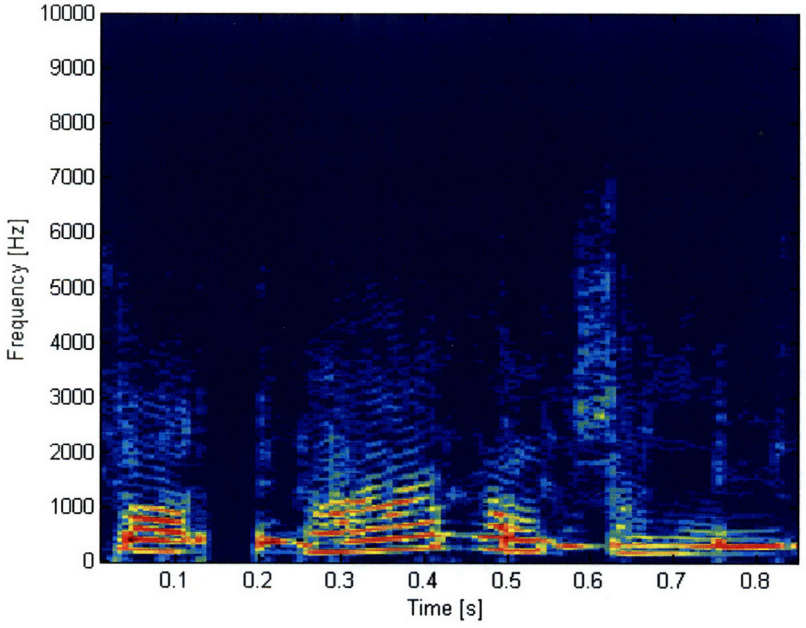


Fig. 6-15: Vocalogram of the word "Technology". The white plus sign markers on the vocalogram indicate the position of the constriction when the speech segment is a consonant.

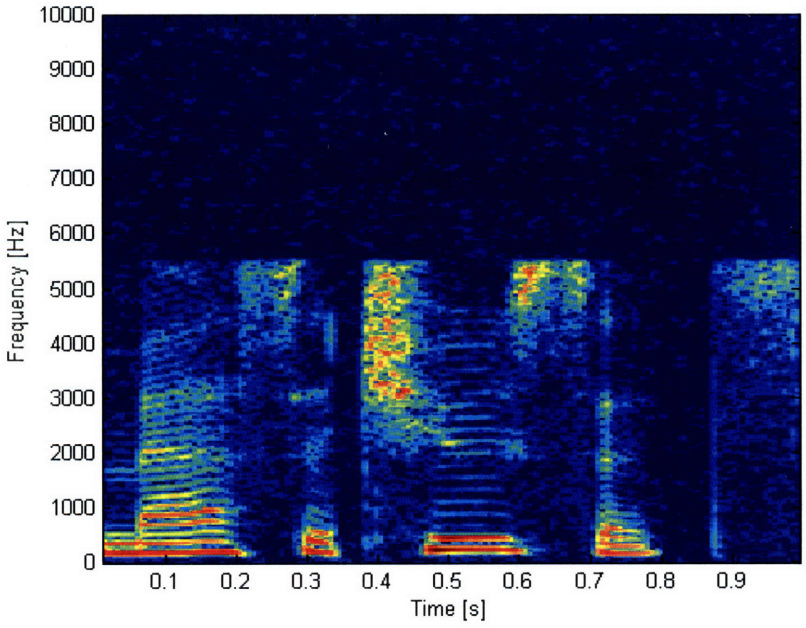


(a)

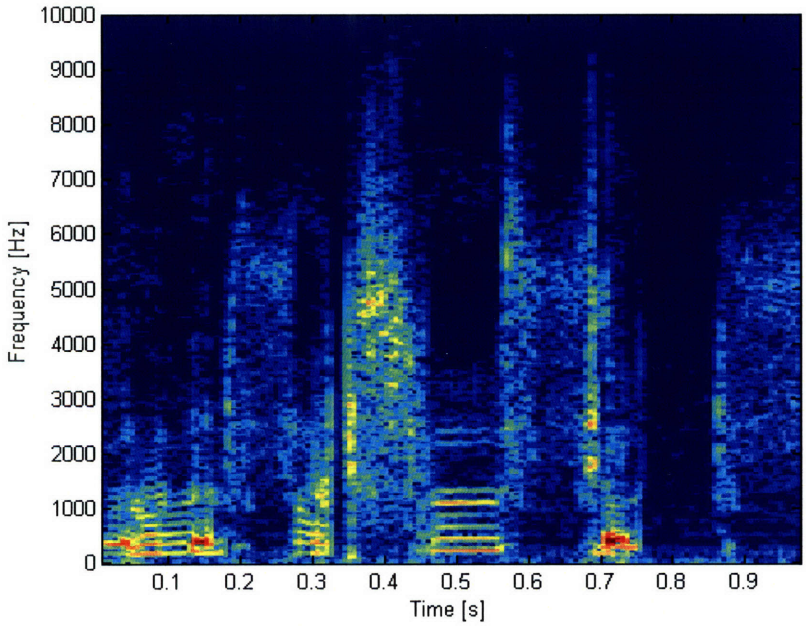


(b)

Fig. 6-16: (a) Time domain waveform and (b) spectrogram of the word "Technology" synthesized by the AVT using the impedance modulation technique for consonants.



(a)



(b)

Fig. 6-17: Spectrogram of (a) "Massachusetts" recording and (b) "Massachusetts" re-synthesized by the AVT.

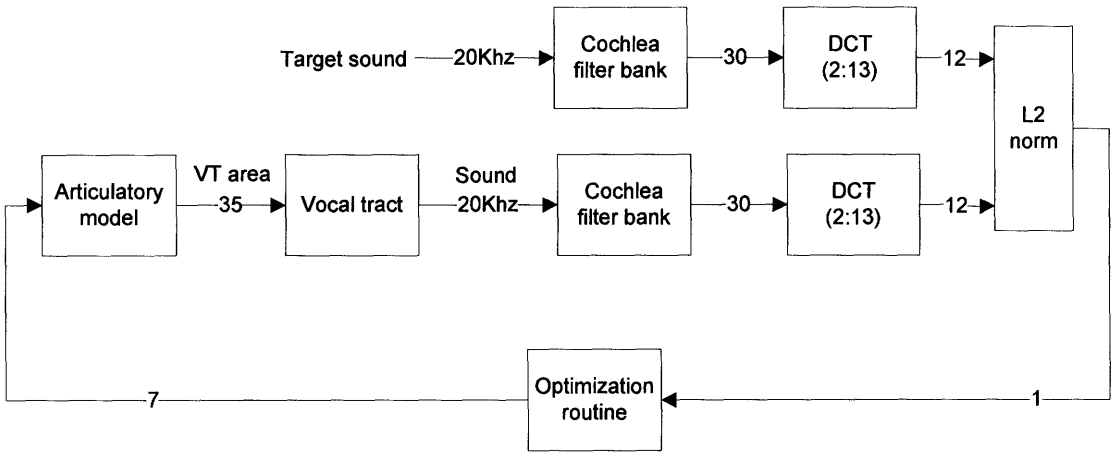


Fig. 6-18: Speech locked loop.

Chapter 7 CONCLUSIONS

In this chapter, we highlight and summarize the major contributions of the thesis. We also suggest potential areas of future work.

7.1 Contributions and accomplishments

Previous attempts to build speech apparatus based on the powerful analysis-by-synthesis method employed computationally expensive approaches to articulatory synthesis using digital computation [29]. 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. Our approach uses bio-inspired models of speech production that facilitates low bit-rate transmission and improves the naturalness of synthetic speech. Our use of a physiological model of the human vocal tract enables our analog vocal tract to synthesize speech signals of interest, using articulatory parameters that are intrinsically compact, robust, and linearly interpolatable. In the preceding chapters, the following were accomplished:

In the second chapter, we developed and analyzed a circuit model of the vocal tract comprising the subglottal system, glottis, and supraglottal vocal tract using elementary circuit elements. We implemented the circuit model successfully in Matlab and verified its functionality through simulations.

In the third chapter, we presented a dynamic programming technique to derive an optimal sequence of articulatory parameters that can be used to drive the vocal tract and verified its utility through simulations with our circuit model.

In the fourth chapter, we presented a new bidirectional electronically tunable linear and nonlinear MOS resistor implemented in CMOS technology 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. The linear MOS resistor does not require triode operation and operates in weak or strong inversion. Our MOS

resistor exploits the symmetry of an MOS device and has inherently zero d.c. offset. Our negative feedback biasing architecture enables the resistor to have arbitrary linear and nonlinear I-V characteristics. We presented experimental results of MOS resistors having linear, compressive and expansive I-V relations. DC measurements show that our linear MOS resistor in its current implementation has a tunable resistance range spanning 1MΩ to 100GΩ. We theoretically analyzed and experimentally verified the temperature dependence of the linear MOS resistor to be proportional to absolute temperature. AC measurements showed that the resistor has a distortion of 0.7% when the signal is centered away from the origin ($|V_{DS}|>100mV$) and 3% when centered at the origin ($V_{DS}=0$). We presented noise measurements of the linear MOS resistor that agreed well with theory and showed that, unlike a real resistor, the linear MOS resistor's noise is a function of V_{DS}.

In the fifth chapter, we developed electronically tunable two-port equivalents of LC π-sections that are used as building blocks for our transmission line vocal tract. We presented a two-port topology that produced 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 vocal tract walls into the circuit topology.

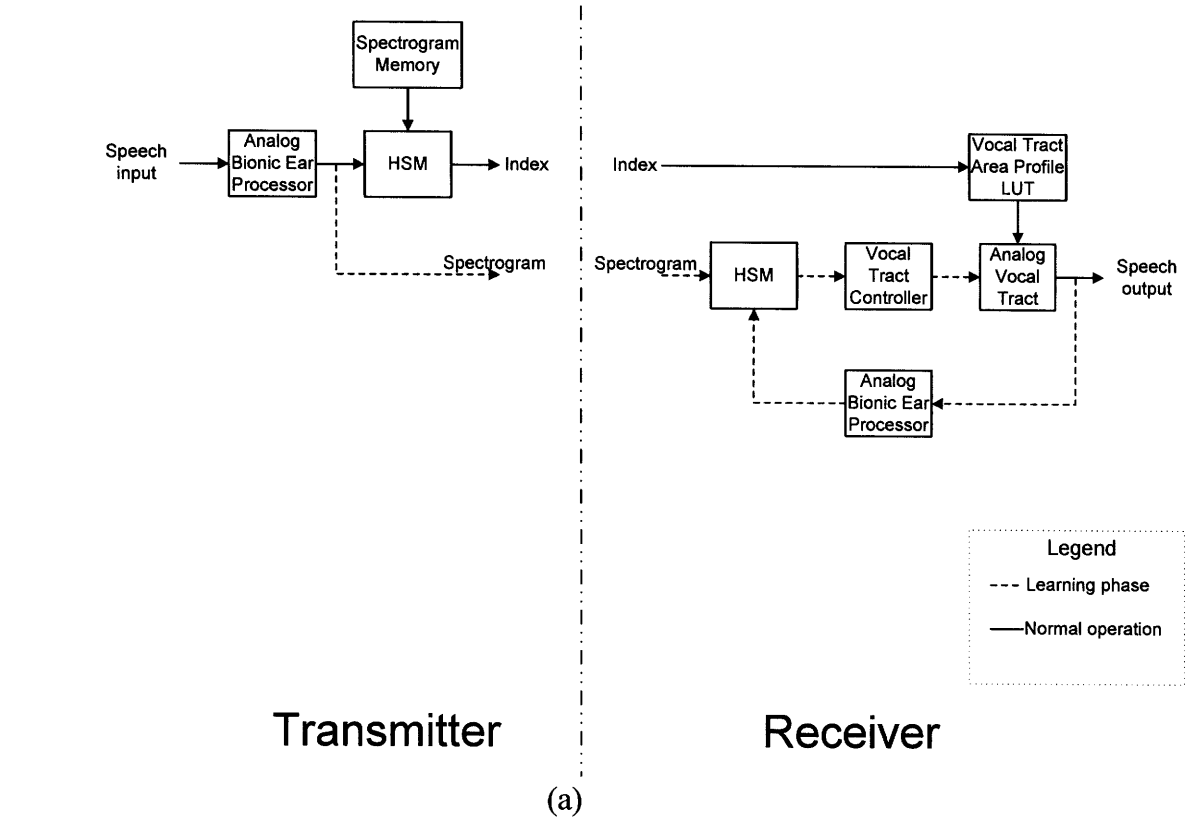
In the sixth chapter, we presented the first experimental integrated-circuit vocal tract. The analog VLSI vocal tract comprises a cascade of 16 tunable two-port π-sections each representing a uniform tube of adjustable length. We also developed two new techniques for producing consonants and presented experimental results that demonstrated their feasibility. The first method input-refers the turbulent noise source to the glottis. The second method modulates the impedance of a two-port π-section downstream of the constriction in a noisy manner. The analog vocal tract was fabricated in a 1.5μm AMI CMOS process. The chip consumes less than 275μW of power when operated with a 5V power supply. 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/.

7.2 Future directions

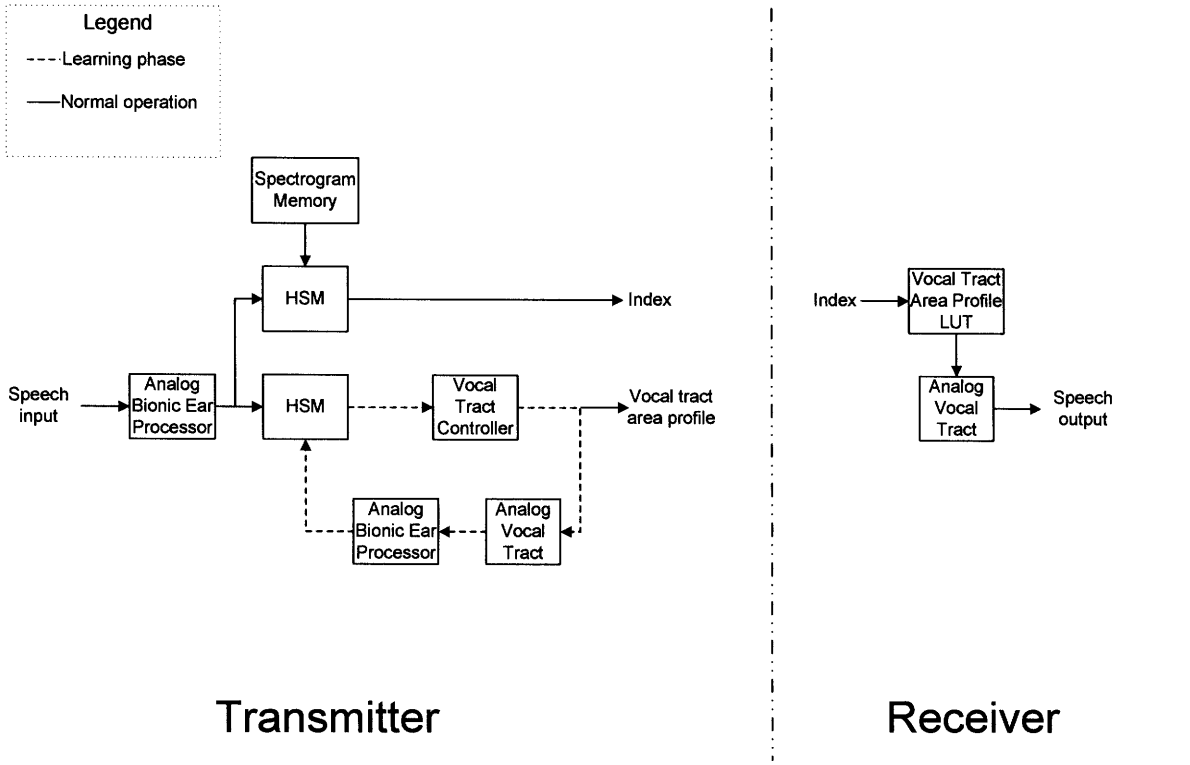
7.2.1 Speech codec

Fig. 7-1 shows two possible architectures that illustrate a promising technology direction for low power speech coding and decoding applications involving a hybrid analog-digital analysis-by-synthesis scheme. The architecture employs a speech locked loop that is applied to speech generated by an analog vocal tract. In Fig. 7-1(a), an analog bionic ear processor performs spectral analysis on the input to the transmitter. A hybrid state machine (HSM) [46] compares the analysis results of speech synthesized by the AVT and the input to produce an error signal. A hybrid state machine—analogous to a digital finite state machine—operates in a hybrid analog-digital domain where analog and digital dynamical systems interact in a feedback fashion such that analog spike-time codes trigger state transitions in the digital dynamical system which in turn produce binary vectors to reconfigure the analog dynamical system. During the training phase, different sounds are generated until one is found that produces the least error, at which time the speech locked loop learns the vocal tract profile and stores it in look-up-table (LUT) on the receiver. The corresponding spectrograms are stored in memory on the transmitter. In normal operation, a HSM compares the input spectrogram with learned spectrograms and produces an index corresponding to the best acoustic match. The AVT synthesizes speech using a vocal tract area profile corresponding to the transmitted index.

Fig. 7-1(b) is an alternative architecture that has a speech locked loop on the transmitter side instead of the receiver side. During the learning phase, optimal vocal tract profiles of various sounds are generated and stored in a LUT on the receiver side. The corresponding spectrograms are stored in memory on the transmitter. In normal operation, a HSM compares the input spectrogram with learned spectrograms and produces an index corresponding to the best acoustic match. The AVT on the receiver synthesizes speech using a vocal tract area profile corresponding to the transmitted index.



(a)



(b)

Fig. 7-1: Hybrid analog-digital analysis-by-synthesis architecture for low power speech codecs.

7.2.2 *Speech recognition via speech locked loop*

Fig. 7-2 is a schematic block diagram illustrating the general concept of speech recognition via analysis by synthesis using a speech locked loop. It comprises a multitude of speech generators (SG) and noise generators (NG), two signal analyzers (SA) that extract salient signal attributes, an apparatus D that computes the differences in signal attributes, and a controller C that controls the speech and noise generators using the output of D in a feedback loop.

The speech generator produces speech signals and has a set of control parameters (driven by C) that shape its output. The analog vocal tract controlled using an articulatory model is a particular embodiment of the speech generator. The noise generator produces non-speech (noise) signals. It has a set of control parameters (driven by C) that shape its output. An example of the noise generator is a model of vehicle noise. A pre-recorded or real-time feed of the desired signal and/or noise is included as additional speech generators and/or noise generators.

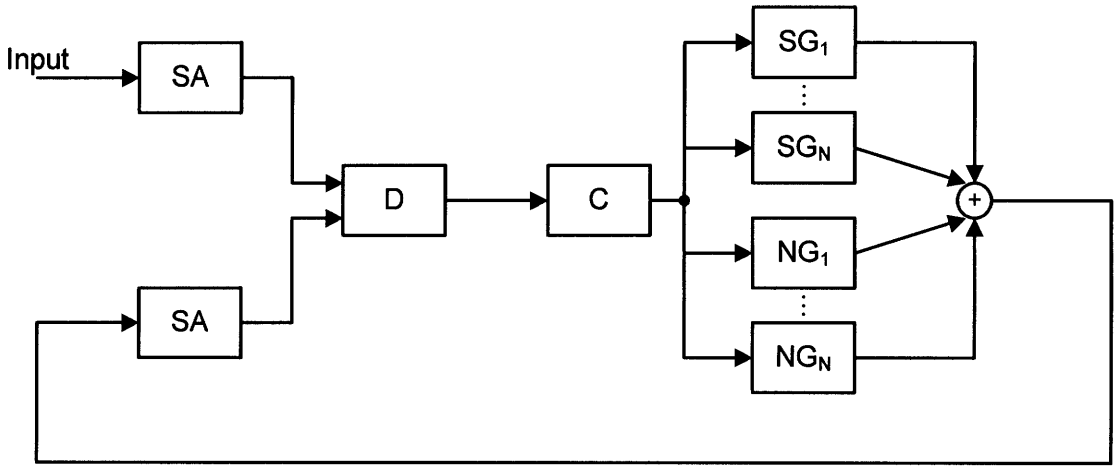


Fig. 7-2: Schematic block diagram illustrating speech recognition via speech locked loop.

The signal analyzer analyses speech/noise signals and extracts salient characteristics of the signals. An example of the signal analyzer is a frequency analysis system such as a perceptually shaped filter bank or cochlea-like apparatus. The extracted characteristics of the input signal and feedback signal are compared by D to produce an error signal. An example of D is an apparatus which computes the L2-norm. The output of D is processed by the controller C to generate a control signal which drives the speech

and noise generators such that the error signal is eventually minimized through feedback action. In this way the output signal is locked to the input signal. In the speech locked condition the parameters characterizing the speech and noise generators provide the optimal description of the input sound.

Multiple speech locked loops functioning in parallel is used to process the input. In this case, the input signal is shared or propagated through a delay line as depicted in Fig. 7-3. In such a parallel structure, the individual C controllers are influenced by one another through a distributed network of interconnections or a central controller that force the parameters to behave in a manner that is consistent with what is observed in natural speech waveforms. In order to drive the speech and noise generators such that the error is minimized, the C controllers use an acoustical distance between the generated sound and the input, and a value related to the control parameter dynamics for every speech and noise generator. The collective behavior of the various C controllers attempts to minimize a cost comprising a linear or nonlinear combination of the acoustic distance and control parameter dynamics. For example, in estimated high-noise conditions, acoustic distance contributions are reduced in favor of contributions from control parameter dynamics (rely more on dynamic/articulatory constraints than on acoustic similarities). Control parameter dynamics vary according to a priori knowledge and an estimation of the input (vowel to consonant, stops, grammar, etc)

Different strategies may be used to set the initial condition of the C controllers. For example, they are learned a priori in a way that guarantees minimum error. This may be done by trying all the possible initial conditions and input signals, and finding the minimal set of initial conditions that will guarantee convergence to the global minimum by the feedback loop. As arrival at global minimum is assured, a fully parallel architecture with multiple feedback loops starting from the minimum set is useful to speed up the convergence process. Otherwise, one or multiple initial conditions of the minimum set are processed serially.

In order to generate an optimal control signal which drives the speech and noise generators such that the error signal is eventually minimized through feedback action, a perturbation-method-based model (or other models that correlate the error signal to the control parameters) is employed. In such an embodiment, the signal analyzers are mel-

frequency spaced filter banks whose outputs are subtracted to produce a vector representing the spectral error. The spectral error vector is used with perturbation methods to vary the control parameters in the feedback loop.

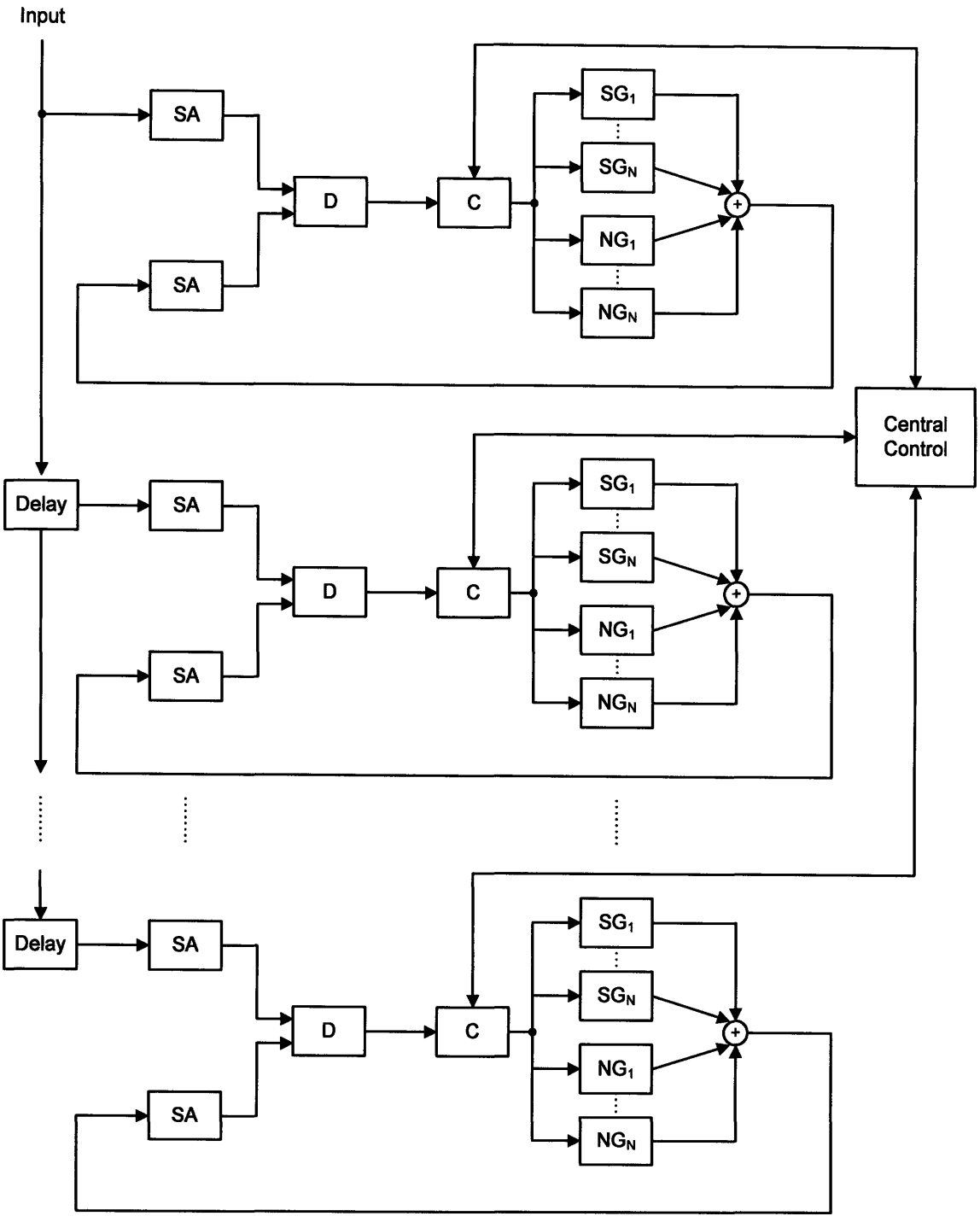


Fig. 7-3: Schematic block diagram illustrating multiple speech locked loops functioning in parallel.

7.2.3 Research tool for speech production

In its present form the analog vocal tract serves as a basic tool for speech production. To have a more complete tool for speech science education and speech research, improvements can be made in the following areas:

(a) Turbulent noise source for consonant production

(b) Subglottal and nasal tracts

(c) Glottal oscillator

The consonant production methods described in §6.3 are approximations that are convenient for circuit design. Improved circuit implementations employing more sophisticated models of the turbulent noise source (e.g., distributed noise sources as opposed to a lump approximation) could be incorporated in the future to better approximate the noise production mechanism at a constriction. In order to include the effects of subglottal resonances, a subglottal tract comprising a cascade of two-port elements that models the trachea and bronchi should be incorporated. The addition of such a subglottal transmission line would capture the effect of the subglottal pole-zero pairs observed in realistic speech spectra. Similarly, the pole-zero pairs due to the nasal tract are introduced by a cascade of two-port elements modeling the nasal passages and sinuses. A glottal oscillator that takes into account the mechano-aerodynamic nature of the vocal fold vibrations would also be useful.

INTENTIONALLY LEFT BLANK

BIBLIOGRAPHY

[1] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens and A. S. House, "Reduction speech spectra by analysis by synthesis techniques," *J. Acoust. Soc. Am.,* vol. 33, pp. 1725-1736, 12. 1961.

[2] R. Sarpeshkar, M. W. Baker, C. D. Salthouse, J. -. Sit, L. Turicchia and S. M. Zhak, "An analog bionic ear processor with zero-crossing detection," in *2005 IEEE International Solid-State Circuits Conference,* 2005, pp. 78-9.

[3] R. Sarpeshkar, C. Salthouse, Ji-Jon Sit, M. W. Baker, S. M. Zhak, T. K. -. Lu, L. Turicchia and S. Balster, "An ultra-low-power programmable analog bionic ear processor," *IEEE Transactions on Biomedical Engineering,* vol. 52, pp. 711-27, 04. 2005.

[4] F. Serra-Graells, L. Gomez and O. Farres, "A true 1 V CMOS log-domain analog hearing-aid-on-a-chip," in *Proceedings of the 27th European Solid-State Circuits Conference,* 2001, pp. 420-3.

[5] B. Raj, L. Turicchia, B. Schmidt-Nielsen and R. Sarpeshkar, "An FFT-based Companding Front End for Noise-Robust Automatic Speech Recognition," *J. Audio, Speech, and Music Process.,* 2007.

[6] C. Mead, *Analog VLSI and Neural System.* Addison-Wesley, 1989,

[7] H. Gray, *Anatomy of the Human Body.* Philadelphia: Lea & Febiger, 1918,

[8] K. N. Stevens, *Acoustic Phonetics.* , vol. 30, Cambridge, Mass.: MIT Press, 1998, pp. 607.

[9] J. L. Flanagan, *Speech Analysis; Synthesis and Perception.* ,2nd ed.Berlin, New York: Springer-Verlag, 1972, pp. 444.

[10] Public domain. Image:Illu01_head_neck. Available: http://en.wikipedia.org/wiki/Image:Illu01_head_neck.jpg; http://training.seer.cancer.gov/ss_module06_head_neck/unit02_sec02_anatomy.html

[11] I. R. Titze, *Principles of Voice Production.* National Center for Voice and Speech, 2000,

[12] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Technical Journal,* vol. 51, pp. 1233-68, 07. 1972.

[13] J. B. Lovins, M. J. Macchi and O. Fujimura, "A demisyllable inventory for speech synthesis," *J. Acoust. Soc. Am.,* vol. 65, pp. 130-131, 1979.

[14] J. -. Courbon and F. Emerard, "SPARTE: A text-to-speech machine using synthesis by diphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1982, pp. 1597-600.

[15] F. J. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *86CH2243-4),* 1986, pp. 2015-18.

[16] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.,* vol. 9, pp. 453-467, 1990.

[17] B. Gold and L. R. Rabiner, "Analysis of digital and analog formant synthesizers," in *1967 Conference on Speech Communication and Processing,* 1968, pp. 81-94.

[18] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.,* vol. 82, pp. 737-93, 1987.

[19] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.,* vol. 87, pp. 820-57, 02. 1990.

[20] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology,* vol. 1, pp. 40-9, 04. 1982.

[21] J. P. Campbell Jr. and T. E. Tremain, "Voiced/unvoiced classification of speech with applications to the US government LPC-10E algorithm," in *86CH2243-4),* 1986, pp. 473-6.

[22] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. ASSP-34, pp. 744-54, 08. 1986.

[23] F. H. Guenther, S. S. Ghosh and A. Nieto-Castanon, "A neural model of speech production," in 2003, pp. 85-90.

[24] D. E. Callan, R. D. Kent, F. H. Guenther and H. K. Varperian, "An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system," *Journal of Speech, Language, and Hearing Research,* vol. 43, pp. 721-36, 06. 2000.

[25] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Commun.,* vol. 1, pp. 199-229, 1982.

[26] K. N. Stevens, S. Kasowski and C. G. M. Fant, "An electrical analog of the vocal tract," *J. Acoust. Soc. Am.,* vol. 25, pp. 734-742, 07. 1953.

[27] G. Rosen, "Dynamic Analog Speech Synthesizer," *J. Acoust. Soc. Am.,* vol. 30, pp. 201-209, 03. 1958.

[28] J. L. Flanagan, K. Ishizaka and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell System Technical Journal,* vol. 54, pp. 485-505, 03. 1975.

[29] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. ASSP-35, pp. 955-67, 07. 1987.

[30] Public domain. Image:Illu_conducting_passages. Available: http://en.wikipedia.org/wiki/Image:Illu_conducting_passages.jpg; http://training.seer.cancer.gov/module_anatomy/images/illu_conducting_passages.jpg

[31] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model." in *Speech Production and Speech Modelling* W. J. Hardcastle and A. Marchal, Eds. Dordrecht, Netherlands; Boston: Kluwer Academic Publishers, 1990, pp. 131-149.

[32] J. Schroeter, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing,* vol. 2, pp. 133-150, 1994.

[33] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, N.J.: PTR Prentice Hall, 1993, pp. 507.

[34] K. Nay and A. Budak, "A voltage-controlled resistance with wide dynamic range and low distortion," *IEEE Transactions on Circuits and Systems,* vol. CAS-30, pp. 770-2, 10. 1983.

[35] K. Nagaraj, "New CMOS floating voltage-controlled resistor," *Electron. Lett.,* vol. 22, pp. 667-8, 06/05. 1986.

[36] Y. Tsividis, M. Banu and J. Khoury, "Continuous-time MOSFET-C filters in VLSI," *IEEE Transactions on Circuits and Systems,* vol. CAS-33, pp. 125-140, 1986.

[37] Y. Tsividis and K. Vavelidis, "Linear, electronically tunable resistor," *Electron. Lett.,* vol. 28, pp. 2303-5, 12/03. 1992.

[38] K. Vavelidis and Y. Tsividis, "Design considerations for a highly linear electronically tunable resistor," in *1993 IEEE International Symposium on Circuits and Systems,* 1993, pp. 1180-3.

[39] K. Vavelidis, Y. P. Tsividis, F. Op't Eynde and Y. Papananos, "Six-terminal MOSFET's: Modeling and applications in highly linear, electronically tunable resistors," *IEEE J Solid State Circuits,* vol. 32, pp. 4-12, 1997.

[40] J. Ramirez-Angulo, M. S. Sawant, R. G. Carvajal and A. Lopez-Martin, "Linearisation of MOS resistors using capacitive gate voltage averaging," *Electron. Lett.,* vol. 41, pp. 511-512, 2005.

[41] Y. Tsividis, *Operation and Modeling of the MOS Transistor.* ,2nd ed.Boston: WCB/McGraw-Hill, 1998, pp. 620.

[42] R. Sarpeshkar, R. E. Lyon and C. Mead, "A low-power wide-linear-range transconductance amplifier," *Analog Integr. Cir. Signal Proc.,* vol. 13, pp. 123-51, 05. 1997.

[43] P. R. Gray, *Analysis and Design of Analog Integrated Circuits.* ,4th ed.New York: Wiley, 2001, pp. 875.

[44] M. O'Halloran and R. Sarpeshkar, "A 10-nW 12-bit accurate analog storage cell with 10-aA leakage," *IEEE J Solid State Circuits,* vol. 39, pp. 1985-96, 11. 2004.

[45] C. G. M. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow," Tech. Rep. STL-QPSR 4/1985, 1985.

[46] R. Sarpeshkar and M. O'Halloran, "Scalable hybrid computation with spikes," *Neural Comput.,* vol. 14, pp. 2003-38, 2002.