Elements of Human Voice

This is a self-contained monograph on human voice. It systematically expounds a theory of voice production initiated by Leonhard Euler, through an analysis of large amount of human voice data, especially simultaneously acquired voice signals and electroglossograph signals, as well as temporal variations of pressures directly below and above the vocal folds. Its contents include the physics and physiology of human voice production, parametrical representations of voice signals, and technology applications. Background knowledge on general acoustics and mathematical tools pertinent to quantitative descriptions of human voice are explained in detail.

Readers of this monograph include researchers, practitioners and students in the fields of physiology and medicine, acoustics, computer science, telecommunication, acoustic phonetics, and vocal music.

Prior to joining academia, C. Julian Chen was a research staff member at the Human Languages Technology Department of IBM T. J. Watson Research Center. He invented new algorithms for the recognition of tone languages, giving birth to ViaVoice Mandarin, the first successful Chinese language dictation system. For this achievement, IBM granted him an Outstanding Innovation Award in 1998. Besides research in human voice, he is well known for two monographs, Introduction to Scanning Tunneling Microscopy (Oxford University Press 1993, 2008, 2015), and Physics of Solar Energy (John Wiley and Sons 2011).
Voice communication through speech is a unique ability of human beings, which is probably the most significant feature differentiating humans from all other animals. In addition to speech, since the prehistory time, singing has also been a unique ability of human beings as a means to communicate with fellow human beings, and to express themselves. From the point of view of basic science, voice production is a complex physical-physiological function of human body. After the invention of telephone in 1874 and the invention of the phonograph in 1877, speech science and technology became an important field of research. Since the advance of computers in the 1940s, as the most natural means of man-machine interface, speech recognition and speech synthesis have witnessed explosive development, and are still evolving at a fast pace, as evidenced by the recent advances of Siri, Cortana, and Now. Furthermore, coded speech has always been the dominant mode in telecommunication, including wired, wireless and voice-over-IP.

Due to its importance, the science and technology of human voice has been a perennial subject of research and development since the beginnings of the scientific method and industrial revolution with Galileo and Newton. Among the enlightenment scientists, Leonhard Euler (1707–1783) occupied a unique position. Arguably the most influential mathematician of all time, Euler was also a pioneer in many fields of science, including rigid-body mechanics, fluid mechanics, astronomy, optics, acoustics, music theory, elasticity theory, civil engineering, and articulatory phonetics [47]. Among the 866 publications and communications of Euler, 10 are exclusively on acoustics. In an article on the history of acoustics, six Euler publications on acoustics are cited, more than any other author [56]. In 1727 when he was 20 years old, Euler published a treatise *Dissertatio physica de sono* (A physical dissertation on sound) [26]. It has two chapters: Propagation of Sound and Production of Sound. In Chapter 2, the physics of three categories of music instruments is discussed: string instruments, percussion and wind instruments. In paragraph 23, the physics of human voice production is discussed in analogy to the pipe organ [26]:

Clearly the human voice is produced in the same way; indeed the epiglottis holds in place the seat of the reed tongue in the organ of speech, the vibration of which is maintained by the passage of the air ascending through the base windpipe. Besides, the vibratory motion of the air escaping from the end of the base windpipe is changed in the cavity of the mouth in a number of ways, by which the low and high-pitched tones of the voice can
Following Euler's conjectures, British scientist Robert Willis (1800–1875) conducted an extensive experimental and theoretical study of human voice production, and published a report *On the Vowel Sounds, and on Reed Organ-Pipes* in 1829 [99]. As a talented mechanical engineer, Willis build a number of devices to mimic human voice organs. One of those devices resembles a reed organ-pipe, where the reed tongue mimics the vocal folds, and the pipe mimics the vocal tract, see Fig 1.

Through experimental studies together with the theoretical arguments of Euler [99], Willis found that the timbre of vowel depends on the length of the tube to the right side of the piston $L$, which mimics the vocal tract. By changing its length, vowels $[i]$, $[e]$, $[a]$, $[o]$ and $[u]$ can be produced. The vibration frequency of the reed changes the pitch. However, the pitch has no effect on the timbre of the vowel. The timbre of the voice, which Willis called *mouth tone*, is completely independent of the pitch produced by the vocal folds, which Willis called *larynx tone*.

By quoting Euler's theoretical analysis on a transient resonator [99], see Section 1.3, Willis stated that each pulsation of air generated by the vocal folds triggers a decaying acoustic wave, the waveform of which is determined
Fig. 2. Phonograph traces investigated by Ludimar Hermann. Phonograph traces of six vowels are shown [40]. Apparently, each pitch period starts with a pulsation, and then decays. For different vowels, the spectrum, or the frequency contents of the decaying elementary wave, is different. Hermann coined a term formant for the peak frequencies in the spectrum of the elementary waves [43].

by the vocal tract. The voice of the same vowel with different pitch is determined by the repetition rate of the pulsation of air, generated by the vocal folds. In other words, the voiced sound is produced as a superposition of elementary waves representing the timbre of the vowel, with time intervals determined by the pitch period of the vibration of vocal folds.

In 1877 Thomas Edison invented the phonograph. The waveforms of human voice can be recorded and displayed. German physiologist Ludimar Hermann (1831-1914) further amplified the mechanical grooves optically, recorded the waveforms on photographic plates, and did a large-scale quantitative study [40, 41, 42]. Some examples of the sound wave records are shown in Fig. 2. As shown, each pitch period of the waveform has a clear and consistent internal structure: Each starts with a strong pulsation, and decays within each pitch period. For a given vowel, the waveform in each pitch period is similar. Pitch is the repetition rate of the pulsations that trigger the decaying wave in each pitch period.

Ludimar Hermann observed that the spectrum of the decaying elementary wave of a vowel is peaked at a number of frequencies, characteristic of the vowel. He coined a term formant for those frequencies [43]. The term formant has been then used by voice and speech scientists ever since.

In modern times, waveforms of human voice can be displayed on a computer, with an accuracy far exceeding that of the phonograph. The concept of Willis becomes even more relevant. In Elements of Acoustic Phonetics, Peter Ladefoged (1925 – 2006) shows how the waveforms of his own voice can be straightforwardly described by Willis’s concept [54].

In Chapter 7 of Ladefoged’s book [54], Production of Speech, the vowel
[o] is first discussed, see Fig. 3(A). It has a dominating formant at about 500 Hz. In example (1), with a single glottal event, the elementary wave of the vowel [o] starts with a pulsation at time 0, then decays. It can be approximated by a decaying sinusoidal wave,

\[ x(t) = a e^{-\kappa t} \sin(2\pi ft), \]

where \( a \) is the amplitude, \( \kappa \) is the decay constant, and \( f \) is the formant frequency of the elementary wave representing the vowel.

If the pulsation repeats, as shown in Fig. 3(A) (2) through (4), vowel sounds with different pitch frequencies are generated. In graph (2), the pitch is 100 Hz. In graphs (3) and (4), it is 125 Hz and 150 Hz. The waveform of the same vowel with different pitch frequencies is the superposition of the decaying wave in graph (1) with different repetition rates.

Figure 3(B) shows the waveform of vowel [i] pronounced by him. Again, for each pitch period, the waveform starts with a sharp pulsation, and then decays. The observed waveform contains three decaying waves, starting at the same pulsation instant, with formant frequencies 250 Hz, 2700 Hz and 3500 Hz, respectively. Ladefoged summarized thusly [54]:

If you like to think of it in musical terms, you can say that corresponding to each vowel there is a chord that is characteristic

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
of the vowel. Owing to the pulses from the larynx, this chord is generated many times per second. There is nothing particularly new about this way of looking at speech signals. As long ago as 1829, Robert Willis said: “A given vowel is merely the rapid repetition of its particular note”. This is an oversimplification, because Willis did not realize that vowels are characterized not by one frequency each but by a combination of frequencies; if we alter his remark slightly, however, and say that a given vowel is merely the rapid repetition of its peculiar chord, we have a statement that fits the data very nicely.

This way of looking at human voice should be extremely useful in speech processing, including speech recognition and speech synthesis. If one can decipher the elementary decaying waves that represent the vowels and use the spectral parameters of such elementary waves as the foundation of speech recognition, completely free from the interference of pitch, highly accurate acoustic recognition of the timbre could be achieved. On the other hand, starting from the elementary waves representing individual vowels, speech with desired prosody can be synthesized by making a superposition of those elementary waves according to the prescribed timing and intensity information. Thus synthesized speech will sound natural, because this is exactly the way authentic human speech is produced.

Another important issue regarding the process of human voice production is the timing between the opening and closing of the glottis and the triggering of the elementary waves. Intuitively, one would think that when the glottis is closed, the pressure from the lungs would build up. At a certain point, the air pressure in the trachea becomes high enough to break open the vocal folds and then release a sharp puff of air. That air puff is the pulsation that triggers an elementary wave [23]. However, since the invention of the electroglottograph by French physiologist Philippe Fabre in 1956 [27], a universal experimental fact was found: The speech signal is triggered by the closing of glottis, rather than by its opening, see Fig. 4 and Section 2.2.4. The speech signal is stronger while the glottis is closed, and weaker while the glottis is open. At the glottal opening moment, no pulsation is observed, and the speech signal decays faster. This observation directly contradicts the intuition that the air pressure in the trachea breaks open the vocal folds and then ejects a puff of air to trigger an elementary wave. Another universal experimental fact observed from the aligned voice signal and the EGG signal is that for a given hardware and software system, the sharp peak at the glottal closing moment always has a well-defined polarity. For databases acquired under different conditions, the polarity could either be positive, or be negative. In other words, the voice signal has a
The critical role of glottal closing in the production of voice is generally noticed. In 1992, Robert T. Sataloff made a vivid analogy of human voice production with hand clapping in an article *The Human Voice on Scientific American* [74], and reiterated in 2014 [76]:

Sound is actually produced by the closing of the vocal folds, in a manner similar to the sound generated by hand clapping. Contrary to popular opinion, the vocal folds are not “cords” that vibrate like piano or guitar strings. ... (T)he more frequently they open and close, the higher the pitch.

According to that analogy, each hand clap triggers a decaying elementary acoustic wave. The superposition of such individual decaying waves constitutes a sustaining vowel sound. The pitch frequency of the vowel is the frequency of repetition of hand clapping.

The experimental fact that the elementary acoustic wave of a vowel is triggered by a glottal closing rather than a glottal opening is counterintuitive. To understand human voice production, it is a major mystery to be solved. In order to understand the critical role of glottal closing, the dynamic acoustic process inside the vocal tract immediately after a glottal closing should be studied by solving the time-dependent wave equation. The solution clarifies the acoustic process inside the vocal tract, and also pro-
vides a better understanding of how human voice is produced. A correct and accurate understanding of the production process of human voice enables the design of an accurate mathematical representation of voice and speech signals, which can be applied to improve speech and voice technology. And this is the synopsis of the current book.

Following is a brief summary of why and how I started to be interested in the research of physics and physiology of human voice production as well as the parameterization of voice and speech signals.

The first time I got interested in human voice was at an age of 13 as a private composition student of Professor Chéng Mǎojūn, the composer of the National Anthem of Republic of China. In one of his private lessons, Professor Chéng taught me that when composing vocal music based on a lyric in Mandarin, the melody line should follow the tone of the syllable to make the words clear. During my high school, college and university time, music was my primary hobby. I was a director of Peking University student philharmonic orchestra, and was frequently a conductor or a piano accompanist of choral groups. From time to time I was an arranger and composer. On the other hand, when I entered the Physics Department of Peking University, I already mastered English and Russian. Then I took French, German and Japanese in the respective departments. During the three years I lived in the center of Beijing city, I carefully listened to the native Beijing dialect speakers and made notes. By comparing with my native Shanghai dialect and several foreign languages, I wrote a 250 page monograph about a comparative study of word stresses and tone variations in Chinese languages. The manuscript was destroyed in political turmoil. However, the knowledge I accumulated was enormously helpful when I did a research on Chinese language speech recognition at IBM.

After receiving a Ph.D. in physics from Columbia University under the supervision of Professor Richard M. Osgood, I joined the Physical Sciences Department of IBM T. J. Watson Research Center as a research staff member in 1985 during the golden era of IBM’s basic research. In 1986, Gerd Binnig and Heinrich Rohrer were awarded a Nobel Prize in physics for their design of the scanning tunneling microscope (STM). In 1987, Georg Bednorz and Alexander Müller were awarded another Nobel Prize in physics for their important break-through in the discovery of superconductivity in ceramic materials. I proposed an experimental project to study the mechanism of the high-temperature superconductors by building and running a low-temperature cross-sectional STM, which was fortunately approved and supported by the management. During my research work, I found that the understanding of the imaging mechanism of STM, why and how STM achieved atomic resolution, was missing. Also missing was the theory of
design and calibration of the heart of STM, the tube piezoelectric scanner. I published a series of papers expounding those two problems of physics in STM, which was appreciated by the scientific community worldwide. Based on those papers, I wrote a monograph *Introduction to Scanning Tunneling Microscopy*, published by Oxford University Press in 1993.

As mentioned above, my primary academic interest has always been human voice and languages. In 1993, IBM made a decision to de-emphasize basic research in physics and to re-emphasize software technology. A high-priority research project envisioned by IBM headquarters was speech recognition for Mandarin Chinese. Based on my understanding of the tones of Mandarin Chinese, especially my comparative studies with other languages that I know, I proposed a new algorithm for its recognition. It includes a new phoneme system with tones, and a trigram language model based on mostly multisyllable *words* rather than Chinese characters [15, 16, 17, 18, 19]. My idea was diametrically opposed to the method in Mandarin speech recognition at that time. Because implementing my ideas made it possible to directly map Mandarin into an English speech recognition system, it was immediately approved by IBM’s Human Language Technology Department. In the process of making the first working system, I acted as the test speaker. I also designed a statistical algorithm to automatically segment a large corpus of Chinese text into words, and made a language model from it. The first system achieved an unprecedented accuracy. Owing to the joint effort of many collaborators, the first commercial product of a Chinese language dictation system, ViaVoice Mandarin, was announced in 1997. It was then installed on almost all Chinese language computers sold in China. Because of ViaVoice Mandarin, I received an Outstanding Innovation Award from IBM. The story was documented in a science-history book [8].

In 2000, I was appointed as a technical leader for worldwide speech synthesis technology, and transferred to IBM Speech Systems Division in Boca Raton, in charge of both formant synthesizer and unit-selection synthesizer. Day after day, I stared at the waveforms of speech corpora with simultaneous electroglottograph (EGG) signals of many languages in the world. It was actually a Beagles voyage for me. At that time, the main signal processing method for unit-selection speech synthesis was PSOLA (pitch-synchronous overlap add method). As a physicist, intuition hinted that a good spectral parameterization should be pitch synchronous, rather than the pitch-asynchronous speech parameterization methods, MFCC and LPC. Intuition also hinted that by using an accurate pitch-synchronous spectral representation of human voice, a speech synthesizer with advantages of both formant synthesizer and unit-selection synthesizer could be built. However, before I retired from IBM, no practical method was found.

In January 2004, Professor Roland Wiesendanger invited me to join De-
partment of Physics of Hamburg University, to continue research in scanning tunneling microscopy (STM). During 2004 to 2006, in Hamburg, I continued basic research in STM, and prepared the second edition of *Introduction to Scanning Tunneling Microscopy*. Although concentrated in quantum mechanics of atoms and molecules, I still kept a hobby project of singing synthesizing. Incidentally, I found that the mathematical methods in quantum mechanics fit very well to describe human voice.

In late 2006, at the invitation of Professor Osgood, I joined Department of Applied Physics and Applied Mathematics of Columbia University. To satisfy the popular demand for the science of renewable energy, I prepared a graduate-level course *Physics of Solar Energy*. The lectures were enthusiastically received. My lectures were video-recorded by Columbia Video Network and distributed worldwide. In 2011, a monograph and graduate-level textbook *Physics of Solar Energy* was published by John Wiley and Sons, then a Chinese translation was published in 2012.

After I returned to New York, a start-up company Voice Dream asked me to collaborate in speech synthesis technology, and provided me with software-related support. At that time, the ARCTIC speech databases from Carnegie-Melon University were publicly available. I restarted the search for a better parametrical representation of voice signals. I also met Donald Miller, an opera singer turned to human voice researcher, and found many interests in common. He pointed to me a number of unexplained experimental facts in human voice. By looking into monographs and theoretical papers on human voice from 18th century to current, I found that then popular theory of human voice production, the source-filter model, is neither the first one historically, nor the best one scientifically. A more accurate theory of human voice production could explain all experimental facts, help voice physicians and singers to improve their practice, and enable the invention of more accurate parameterizations of voice and speech signal as the foundation of new methods for voice transformation, speech recognition, speech coding, and speech synthesis [10, 11, 12, 13, 14]. In this book, I wish to share my pleasure of discovery and invention with you.

Samples of speech signals processed by the methods described in this book are posted on my homepage, www.columbia.edu/~jcc2161, under heading Human Voice. The contents are periodically updated.

The book could not be written with the help of many people. First, I thank Academician Hé Zuòxiù for recommending me for participation in an examination to enter Columbia University in 1979 despite my “right-wing-element” history. I also thank Professor T. D. Lee for choosing me as the first graduate student from the People’s Republic of China to enter the Physics Department of Columbia University. His philosophy of scien-
Scientific methodology, that physical intuition and order-of-magnitude estimate are far more important than mathematical development, has guided my research work for years. I sincerely thank Professor Osgood as my thesis adviser at Columbia University. His high moral standard is always my role model. I appreciate numerous discussions with Donald Miller, the founder and president of Voce Vista, on the science of singing and the understanding of human voice in general. I also want to thank Winston Chen, the founder and president of Voice Dream LLC, for giving invaluable support in software technology for my research work on human voice. I am especially grateful to Robert Sataloff for reviewing an early manuscript of this book and sending me valuable comments. I highly value Ronald Baken for his appreciation of the new concepts in this book. I am thankful to Irving P. Herman, the author of Physics of the Human Body, for helpful discussions. I sincerely appreciate the constant encouragement and support of Cevdet Noyan to my research in human voice. Finally, without the spiritual and material support of my wife Liching, the book could never have been written.

At the end of the Preface of *Introduction to Scanning Tunneling Microscopy*, I cited the following lines from the prologue of Faust by Johann Wolfgang von Goethe. My research in human voice took shape after many years of endeavor. Here those lines sound more appropriate:

Oft, wenn es erst durch Jahre durchgedrungen,
Erscheint es in vollendeter Gestalt.
Was glänzt, ist für den Augenblick geboren,
Das Echte bleibt der Nachwelt unverloren.\(^1\)

C. Julian Chen
July 2016, Columbia University
In the City of New York

---

\(^1\)Often, after years of perseverance, it emerges in a completed form. What glitters, is born for the moment. The Genuine lives on to the afterworld. *Faust, Vorspiel auf dem Theater.*

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
## Contents

Preface vii

List of Figures xxiii

List of Tables xxvii

**Part I  Physics and Physiology**  1

**Chapter 1: Acoustic Waves**  7

1.1 Wave Equation in a Tube  7
1.1.1 Particle displacement and perturbation density  7
1.1.2 Particle velocity and equation of continuity  8
1.1.3 Perturbation pressure  8
1.1.4 Wave equation  10

1.2 d’Alembert Solution  11

1.3 Euler’s Transient Resonator  13

1.4 Energy and Power  15
1.4.1 Power of acoustic wave  15
1.4.2 Acoustic energy density  16

1.5 Zero-Particle-Velocity Wavefronts in a Tube  17
1.5.1 Heuristic discussions  17
1.5.2 Laplace-transform solution  19
1.5.3 Energy conversion  20

1.6 Fourier Analysis  21
1.6.1 Amplitude and phase  22
1.6.2 Complex-variable version  23
1.6.3 Fourier transform  23

1.7 Numerical Values  25
1.7.1 Pitch scale  25
1.7.2 Intensity scale  26
Chapter 2: Voice Organs

2.1 Overall Structure ........................................... 29

2.2 Vocal Folds .................................................. 30
  2.2.1 Anatomy ................................................. 31
  2.2.2 Strobovideolaryngoscopy ................................. 32
  2.2.3 Laryngeal electromyography ............................... 33
  2.2.4 Electroglottography ....................................... 34
  2.2.5 Bernoulli force ......................................... 38
  2.2.6 Water-hammer analogy ................................... 44
  2.2.7 Incomplete closures ..................................... 46
  2.2.8 Isolated closures and glottal stops ....................... 49
  2.2.9 Videokymography ....................................... 49
  2.2.10 Closed quotient and voice intensity ..................... 51

2.3 Vocal Tract .................................................. 51
  2.3.1 Plosives ............................................... 52
  2.3.2 Fricatives ............................................. 55
  2.3.3 Vowels ................................................. 57

Chapter 3: Experimental Facts

3.1 Microphones and Voice Signals ............................... 61
  3.1.1 Types and working principles of microphones .......... 61
  3.1.2 Source of voice signals ................................ 64
  3.1.3 Vowels ................................................... 65
  3.1.4 Superposition principle ................................ 68

3.2 Electroglottograph and Voice Data .......................... 69
  3.2.1 Temporal correlation ................................... 70
  3.2.2 Glottal stops .......................................... 72

3.3 In-Vivo Pressure Measurements .............................. 73

Chapter 4: The Physics of Voice Production

4.1 A Brief Summary of Experimental Facts ........................ 79

4.2 The Concept of Timbros ...................................... 80

4.3 Acoustic Waves in the Trachea ................................. 84

4.4 An Analytic Solution of the Wave Equation .................. 85
  4.4.1 Initial conditions and boundary conditions ............. 86
  4.4.2 Acoustic waves in the vocal tract ........................ 87
  4.4.3 Analysis of the solution ................................ 88
  4.4.4 Real vocal tracts ...................................... 89
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.2</td>
<td>Definition of timbre vector</td>
<td>131</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Feature vector of a frame</td>
<td>131</td>
</tr>
<tr>
<td>6.3</td>
<td>Timbre Vector and Amplitude Spectrum</td>
<td>132</td>
</tr>
<tr>
<td>6.4</td>
<td>Timbre Distance</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 7: Waveform Recovery</strong></td>
<td>139</td>
</tr>
<tr>
<td>7.1</td>
<td>Phase Recovery</td>
<td>139</td>
</tr>
<tr>
<td>7.1.1</td>
<td>An analytic model</td>
<td>139</td>
</tr>
<tr>
<td>7.1.2</td>
<td>Dispersion relation: an analytic example</td>
<td>142</td>
</tr>
<tr>
<td>7.1.3</td>
<td>Phase spectrum from speech signals</td>
<td>144</td>
</tr>
<tr>
<td>7.1.4</td>
<td>Dispersion relation: voice data</td>
<td>146</td>
</tr>
<tr>
<td>7.1.5</td>
<td>d’Alembert wave of glottal airflow</td>
<td>147</td>
</tr>
<tr>
<td>7.2</td>
<td>Timbron Recovery</td>
<td>149</td>
</tr>
<tr>
<td>7.3</td>
<td>Unvoiced Consonents</td>
<td>150</td>
</tr>
<tr>
<td>7.4</td>
<td>Recovery of Entire Voice Signal</td>
<td>151</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Breathiness index</td>
<td>151</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Fractional voicedness index</td>
<td>153</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Jitter and shimmer</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td><strong>Chapter 8: Applications</strong></td>
<td>159</td>
</tr>
<tr>
<td>8.1</td>
<td>Voice Transformation</td>
<td>159</td>
</tr>
<tr>
<td>8.1.1</td>
<td>Pitch contour modification</td>
<td>161</td>
</tr>
<tr>
<td>8.1.2</td>
<td>Change of speaker identity</td>
<td>164</td>
</tr>
<tr>
<td>8.1.3</td>
<td>Speed change</td>
<td>165</td>
</tr>
<tr>
<td>8.2</td>
<td>Speech Coding</td>
<td>166</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Basic techniques</td>
<td>166</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Vector quantization</td>
<td>169</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Coding based on timbre vectors</td>
<td>170</td>
</tr>
<tr>
<td>8.3</td>
<td>Speech Recognition</td>
<td>176</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Frontend of a speech recognizer</td>
<td>176</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Acoustic processor with a fixed window size</td>
<td>177</td>
</tr>
<tr>
<td>8.3.3</td>
<td>Speech recognition based on timbre vectors</td>
<td>178</td>
</tr>
<tr>
<td>8.3.4</td>
<td>Concept of phonette</td>
<td>180</td>
</tr>
<tr>
<td>8.4</td>
<td>Speech Synthesis</td>
<td>181</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Formant synthesizer and concatenative synthesizer</td>
<td>181</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Speech synthesizer based on timbre vectors</td>
<td>182</td>
</tr>
<tr>
<td>8.4.3</td>
<td>Timbre fusing</td>
<td>184</td>
</tr>
<tr>
<td>8.4.4</td>
<td>Irreducible units</td>
<td>186</td>
</tr>
</tbody>
</table>

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Willis’ simulation of human voice production</td>
</tr>
<tr>
<td>2</td>
<td>Phonograph traces investigated by Hermann</td>
</tr>
<tr>
<td>3</td>
<td>Voice signals recorded and analyzed by Ladefoged</td>
</tr>
<tr>
<td>4</td>
<td>Temporal correlation of voice signals and EGG signals</td>
</tr>
<tr>
<td>1.1</td>
<td>Air displacement and perturbation density</td>
</tr>
<tr>
<td>1.2</td>
<td>Derivation of wave equation</td>
</tr>
<tr>
<td>1.3</td>
<td>Reflection of acoustic waves</td>
</tr>
<tr>
<td>1.4</td>
<td>Leonhard Euler</td>
</tr>
<tr>
<td>1.5</td>
<td>Euler’s transient resonator</td>
</tr>
<tr>
<td>1.6</td>
<td>Energy transfer of acoustic wave</td>
</tr>
<tr>
<td>1.7</td>
<td>Zero-particle-velocity wavefronts in a tube</td>
</tr>
<tr>
<td>1.8</td>
<td>Music scale</td>
</tr>
<tr>
<td>1.9</td>
<td>Ranges of human voice</td>
</tr>
<tr>
<td>1.10</td>
<td>Intensity scale of human voice</td>
</tr>
<tr>
<td>2.1</td>
<td>Cross-sectional view of voice-production organs</td>
</tr>
<tr>
<td>2.2</td>
<td>Vocal folds</td>
</tr>
<tr>
<td>2.3</td>
<td>Strobovideolaryngoscopic images of vocal folds</td>
</tr>
<tr>
<td>2.4</td>
<td>Needle electrode for laryngeal electromyography</td>
</tr>
<tr>
<td>2.5</td>
<td>An example of laryngeal electromyograph</td>
</tr>
<tr>
<td>2.6</td>
<td>Principle of the electroglossograph</td>
</tr>
<tr>
<td>2.7</td>
<td>Electrodes of the electroglossograph</td>
</tr>
<tr>
<td>2.8</td>
<td>Typical EGG signals of vowels</td>
</tr>
<tr>
<td>2.9</td>
<td>Schematic diagram of vocal-fold vibration</td>
</tr>
<tr>
<td>2.10</td>
<td>Vocal-fold oscillation and Bernoulli force</td>
</tr>
<tr>
<td>2.11</td>
<td>Glottal closing with a glottic angle</td>
</tr>
<tr>
<td>2.12</td>
<td>The water-hammer effect</td>
</tr>
<tr>
<td>2.13</td>
<td>Incomplete closures (1)</td>
</tr>
<tr>
<td>2.14</td>
<td>Incomplete closures (2)</td>
</tr>
<tr>
<td>2.15</td>
<td>Isolated closures</td>
</tr>
<tr>
<td>2.16</td>
<td>Working principle of videokymography</td>
</tr>
<tr>
<td>2.17</td>
<td>Image obtained by videokymography</td>
</tr>
<tr>
<td>2.18</td>
<td>Production of plosive [k]</td>
</tr>
<tr>
<td>2.19</td>
<td>Waveform of plosive [k]</td>
</tr>
<tr>
<td>2.20</td>
<td>Production of plosive [ku]</td>
</tr>
<tr>
<td>2.21</td>
<td>Waveform of plosive [ku]</td>
</tr>
<tr>
<td>2.22</td>
<td>Three resonance waves in a single plosive</td>
</tr>
<tr>
<td>2.23</td>
<td>Oral cavity geometry for fricative [s]</td>
</tr>
</tbody>
</table>
## List of Figures

2.24 Amplitude spectra of fricative [s] ........................................ 56  
2.25 Phase spectrum of fricative [s] ......................................... 57  
2.26 Oral cavity while enunciating [u] ........................................ 58  
2.27 Waveform and amplitude spectrum of [u] .............................. 58  
2.28 Oral cavity while enunciating [a] ........................................ 59  
2.29 Waveform and amplitude spectrum of [a] .............................. 59  

3.1 Condenser microphone ...................................................... 62  
3.2 Working principle of condenser microphone ............................ 62  
3.3 Dynamic microphone ....................................................... 63  
3.4 Working principle of dynamic microphone ............................. 63  
3.5 Electret microphone ....................................................... 64  
3.6 Vowel [a] ................................................................. 65  
3.7 Vowel [i] ................................................................. 66  
3.8 Vowel [u] ................................................................. 66  
3.9 Vowel [e] ................................................................. 67  
3.10 Vowel [o] ................................................................. 67  
3.11 Superposition principle ................................................... 69  
3.12 Microphone signals and electroglossotomograph signals .......... 70  
3.13 Combined voice and EGG signals ....................................... 71  
3.14 Glottal stops ..................................................................... 73  
3.15 In-vivo pressure measurement apparatus ............................... 74  
3.16 Time-sequence of glottal pressures ..................................... 75  
3.17 Pressure measurements for vowel [a] .................................. 76  
3.18 Pressure variation of several vowels .................................... 77  

4.1 Acoustic waves triggered by a glottal closure ......................... 81  
4.2 Acoustic waves in the trachea ............................................ 85  
4.3 Initial conditions for the wave equation in the vocal tract ....... 86  
4.4 Open quotient of an opera singer ........................................ 91  

5.1 Weakly varying section in a pitch period .............................. 103  
5.2 The ends-matching procedure ............................................ 104  
5.3 Glottal closing moment as segmentation point ....................... 105  
5.4 Voiced signals with incomplete closures .............................. 107  
5.5 Asymmetric window for pitch-synchronous segmentation ....... 108  
5.6 Profile function for pitch synchronous frame segmentation .... 109  
5.7 Test the size and polarity of the asymmetric window ............ 110  
5.8 Pitch marks without EGG signals (1) ................................. 111  
5.9 Pitch marks without EGG signals (2) ................................. 112  
5.10 Histogram of intensity variation in a pitch period ................. 113  
5.11 Histogram of intensity variation in a pitch period ................. 114  

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.12</td>
<td>Waveform of a stop</td>
<td>115</td>
</tr>
<tr>
<td>5.13</td>
<td>Stop profile function</td>
<td>116</td>
</tr>
<tr>
<td>5.14</td>
<td>Segmentation including stops</td>
<td>116</td>
</tr>
<tr>
<td>5.15</td>
<td>Pitch-asynchronous spectral analysis</td>
<td>117</td>
</tr>
<tr>
<td>5.16</td>
<td>The sinc function, original and truncated</td>
<td>120</td>
</tr>
<tr>
<td>5.17</td>
<td>Amplitude spectra of vowel [a]</td>
<td>121</td>
</tr>
<tr>
<td>5.18</td>
<td>Amplitude spectra of vowel [i]</td>
<td>122</td>
</tr>
<tr>
<td>5.19</td>
<td>Amplitude spectra of vowel [u]</td>
<td>123</td>
</tr>
<tr>
<td>5.20</td>
<td>Amplitude spectra of consonant [k]</td>
<td>124</td>
</tr>
<tr>
<td>6.1</td>
<td>Mel-scaled triangular windows</td>
<td>126</td>
</tr>
<tr>
<td>6.2</td>
<td>Weight function for Laguerre polynomials</td>
<td>129</td>
</tr>
<tr>
<td>6.3</td>
<td>Waveforms of Laguerre functions</td>
<td>130</td>
</tr>
<tr>
<td>6.4</td>
<td>Spectrogram of a sample sentence</td>
<td>133</td>
</tr>
<tr>
<td>6.5</td>
<td>Spectra and timbre vectors (1)</td>
<td>134</td>
</tr>
<tr>
<td>6.6</td>
<td>Spectra and timbre vectors (2)</td>
<td>134</td>
</tr>
<tr>
<td>6.7</td>
<td>Spectra and timbre vectors (3)</td>
<td>135</td>
</tr>
<tr>
<td>6.8</td>
<td>Spectra and timbre vectors (4)</td>
<td>135</td>
</tr>
<tr>
<td>6.9</td>
<td>Spectra and timbre vectors (5)</td>
<td>136</td>
</tr>
<tr>
<td>7.1</td>
<td>Amplitude spectrum on the entire frequency axis</td>
<td>143</td>
</tr>
<tr>
<td>7.2</td>
<td>Phase recovered from dispersion relations</td>
<td>144</td>
</tr>
<tr>
<td>7.3</td>
<td>Acoustic wave, amplitude spectrum and phase spectrum</td>
<td>145</td>
</tr>
<tr>
<td>7.4</td>
<td>Phase recovered from dispersion relations</td>
<td>146</td>
</tr>
<tr>
<td>7.5</td>
<td>d’Alembert wave of glottal airflow</td>
<td>147</td>
</tr>
<tr>
<td>7.6</td>
<td>Glottal-flow phase correction</td>
<td>148</td>
</tr>
<tr>
<td>7.7</td>
<td>Timbrons recovered from timbre vectors: back vowels</td>
<td>149</td>
</tr>
<tr>
<td>7.8</td>
<td>Timbrons recovered from timbre vectors: front vowels</td>
<td>150</td>
</tr>
<tr>
<td>7.9</td>
<td>Original and recovered spectrograms</td>
<td>151</td>
</tr>
<tr>
<td>7.10</td>
<td>Original and recovered waveforms</td>
<td>152</td>
</tr>
<tr>
<td>7.11</td>
<td>The breathiness index</td>
<td>153</td>
</tr>
<tr>
<td>7.12</td>
<td>Detection of voiced fricatives</td>
<td>154</td>
</tr>
<tr>
<td>7.13</td>
<td>Recovery of voiced fricatives</td>
<td>155</td>
</tr>
<tr>
<td>7.14</td>
<td>Transitional fractional voicedness index</td>
<td>156</td>
</tr>
<tr>
<td>7.15</td>
<td>Jitter: random fluctuation of pitch periods</td>
<td>158</td>
</tr>
<tr>
<td>8.1</td>
<td>Schematics of a voice transformation system</td>
<td>160</td>
</tr>
<tr>
<td>8.2</td>
<td>Voice transformation by changing pitch contours</td>
<td>163</td>
</tr>
<tr>
<td>8.3</td>
<td>Quality vs. bitrate for different speech coding techniques</td>
<td>167</td>
</tr>
<tr>
<td>8.4</td>
<td>Schematics of a parametric coder</td>
<td>168</td>
</tr>
<tr>
<td>8.5</td>
<td>Encoder and decoder with vector quantization</td>
<td>169</td>
</tr>
<tr>
<td>8.6</td>
<td>Encoder based on timbre vectors</td>
<td>171</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.7</td>
<td>Decoder based on timbre vectors</td>
<td>172</td>
</tr>
<tr>
<td>8.8</td>
<td>Data structure of a codec system based on timbre vectors</td>
<td>173</td>
</tr>
<tr>
<td>8.9</td>
<td>An example of vocoded speech</td>
<td>175</td>
</tr>
<tr>
<td>8.10</td>
<td>A schematic diagram of the frontend of a speech recognizer</td>
<td>176</td>
</tr>
<tr>
<td>8.11</td>
<td>Pitch-asynchronous segmentation</td>
<td>177</td>
</tr>
<tr>
<td>8.12</td>
<td>Speech recognition system based on timbre vectors</td>
<td>179</td>
</tr>
<tr>
<td>8.13</td>
<td>Speech synthesis based on timbre vectors</td>
<td>183</td>
</tr>
<tr>
<td>8.14</td>
<td>Principle of timbre fusing</td>
<td>184</td>
</tr>
<tr>
<td>8.15</td>
<td>Effect of timbre fusing</td>
<td>185</td>
</tr>
<tr>
<td>8.16</td>
<td>Effect of timbre fusing</td>
<td>185</td>
</tr>
<tr>
<td>8.17</td>
<td>The IPA vowel chart</td>
<td>187</td>
</tr>
<tr>
<td>8.18</td>
<td>Linear approximation of pitch contours</td>
<td>189</td>
</tr>
<tr>
<td>8.19</td>
<td>Interpolated global pitch contour</td>
<td>190</td>
</tr>
<tr>
<td>8.20</td>
<td>Weight functions of Gegenbauer polynomials</td>
<td>192</td>
</tr>
<tr>
<td>8.21</td>
<td>Computation of expansion coefficients</td>
<td>193</td>
</tr>
<tr>
<td>8.22</td>
<td>Global pitch contours for various phrase types</td>
<td>195</td>
</tr>
<tr>
<td>8.23</td>
<td>Overall algorithm of prosody generation</td>
<td>196</td>
</tr>
<tr>
<td>A.1</td>
<td>Proof of dispersion relations for speech signals</td>
<td>199</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Bernoulli force and elastic force .................. 40
2.2 Incomplete glottal closures .......................... 48
2.3 Dimensions of the vocal tract ......................... 51

3.1 Sources of recorded voice and EGG signals ............ 64

5.1 Comparison of PMs and GCIs .......................... 111
5.2 Formants of vowel [a] by four speakers ............... 121
5.3 Formants of vowel [i] by four speakers ............... 122
5.4 Formants of vowel [u] by four speakers ............... 123

8.1 Change of speaker identity ............................ 164
8.2 Mean opinion score (MOS) scale ....................... 167
Part I

Physics and Physiology
Part I: Physics and Physiology

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

On the Method of Theoretical Physics
Albert Einstein
The Herbert Spencer Lecture
Oxford, June 10, 1933

A good understanding of human voice production is the starting point of improving voice and developing algorithms for voice and speech technology. In Part I, a theory of human voice production is presented, which also serves as the scientific foundation of Part II, Mathematical Representations and the applications in speech and voice technology.

Chapter 1 presents background theory of acoustic waves. For simplicity, only the one-dimensional wave equation in a uniform tube is presented. It is sufficient for the understanding of the entire book.

Chapter 2 presents the basic anatomy and physiology of the voice-producing organs, including vocal folds and vocal tract. Instruments for probing and measuring the functions of the voice organs are presented. Special emphasis is directed to the non-invasive probing methods, the electroglottograph (EGG) and miniature pressure sensors, which can be applied simultaneously with the microphone during normal voicing.

Chapter 3 presents the experimental facts of human voice. First, the superposition principle formulated by Edward W. Scripture [84] is illustrated by numerous examples of voice signals. Next, the universal temporal correlation of the voice signal with the electroglottograph signal, the subglottal and the supraglottal pressures is presented. The temporal correlation strongly implies the critical role of glottal closings in voice production.

In Chapter 4, inferred from the experimental facts presented in Chapter 3, a theory of human voice production, especially for vowels, is presented. Briefly, the theory is as follows. Immediately before a glottal closure, there is a steady airflow in the vocal tract. A glottal closing abruptly blocks the airflow from the trachea into the vocal tract, triggers a zero-particle-velocity d’Alembert wavefront, which propagates and resonates in the vocal tract to form a decaying acoustic wave. Kinetic energy of the airflow in the vocal tract immediately before a closure is converted into acoustic energy. Linear superposition of these elementary resonance waves constitutes voice. The
acoustic wave in the vocal tract triggered by a glottal closing is determined by the geometry of the vocal tract at that moment, thus representing the instantaneous timbre. It is reasonable to term the decaying acoustic wave triggered by a glottal closing a “timbron”. The timbrons are literally the elements of human voice. The production mechanism of consonants is then presented, which is relatively straightforward.

In the history of the theory of human voice, especially for vowels, there are two schools of thought, analogous to the centuries-long controversy of the theory of light: the particle theory of Isaac Newton and the wave theory of Christiaan Huygens [31]. The first school of human voice, the transient theory or inharmonic theory, was proposed by British scientist Robert Willis (1800-1875) in 1829 [99]. Motivated by the similarity between human voice organ and pipe organ proposed by Leonhard Euler (1707-1783), Willis designed a series of mechanical models to artificially imitate human voice production. By following Euler’s theoretical analysis, he showed that vowel sounds are composed of a series of decaying acoustic waves excited by pulsations emitted from the vocal folds. After the invention of phonograph by Thomas Edison in 1877, speech waveforms could be recorded and displayed. Ludimar Hermann, a German physiologist (1838-1914), using an optical amplification system to record the speech waveform on photographic plates, then verified Willis’s theory with extensive data [40, 41, 42]. In 1902, American physiologist Edward Wheeler Scripture (1864-1945) published a monograph *The Elements of Experimental Phonetics* [83], systematically expounding the transient theory of human voice.

However, the early transient theories conjectured that the source of excitation is the air puff coming through the glottis after being pushed open by the pressure in the trachea. After the invention of electrolaryngograph by French physiologist Philippe Fabre in 1956 [27], a universal experimental fact was found: The speech signal is triggered by the closing of glottis, rather than by its opening. The waveform of the air puff during the open phase of glottis has little effect on the voice. In order to elucidate that experimental fact, in this book, the acoustic process inside the vocal tract immediately after a glottal closing is studied by solving the time-dependent wave equation. The solution, a dynamic acoustic process inside the vocal tract, is a quantitative representation of the transient sound wave.

An alternative theory of human voice production, the overtone-resonance theory or source-filter theory, was proposed in 1837 by Sir Charles Wheatstone (1802-1875) in a comment on Willis’s paper [98]. Wheatstone agreed in every respect with Willis’s theory, but added an alternative view in terms of overtones and resonance [73]. Wheatstone’s view was elaborated by Hermann von Helmholtz (1821-1894) in *Sensation of Tone* [38].

Wheatstone and Helmholtz assumed that the vibration of vocal folds is
truly periodic with a frequency $f_0$. A periodic function can be treated as a Fourier series, which consists of a fundamental component with frequency $f_0$, and the overtones with frequencies $2f_0$, $3f_0$, and so on. The vocal tract can be treated as a Helmholtz resonator with resonance frequencies $F_1$, $F_2$, $F_3$, and so on, which are called formant frequencies. (An interesting historical fact is that the term “formant” was coined by Ludmir Hermann [31].) If the frequency of an overtone $nf_0$ is equal or very close to a formant frequency $F_m$, it is reinforced. Therefore, the intensity envelope of the overtones on a frequency scale exhibits the spectrum of formants, or the resonance frequencies of the vocal tract. Experimentally, for sustained vowels with a fixed fundamental frequency, that theory is valid.

In late 19th century, the controversy between the Euler-Willis transient theory and the Wheatstone-Helmholtz overtone-resonance theory ran red hot. John William Strutt, also known as Lord Rayleigh (1842-1919), described both theories in detail in his monograph Theory of Sound [72]. Here are the subsection titles on human voice in the Table of Contents:

Willis’ theory of vowel sounds. Artificial imitation. Helmholtz’s form of the theory. No real inconsistency. Relative pitch characteristic, versus fixed pitch characteristic. Auerbach’s results. Evidence of phonograph. Hermann’s conclusions. His analysis of A. Comparison of results by various writers. ...

According to Lord Rayleigh, the acoustic treatment of the subject of vowel production dated from a “remarkable memoir by Willis”, and spent two full pages to quote Willis’s original paper, including Euler’s theoretical analysis of a transient resonator [72]. Then, after describing Wheatstone and Helmholtz’s overtone-resonance theory, Lord Rayleigh argued that for an infinite array of pulsations with equal time interval, the only observable waves are the overtones of the fundamental frequency of the vocal cord vibration. In this case, the result of the Euler-Willis theory becomes identical to that of the Wheatstone-Helmholtz theory. Lord Rayleigh concluded, “From these considerations it will be seen that both ways of regarding the subject are legitimate and not inconsistent with one another.”

Nevertheless, such a consistency is a one-way street. By definition, the subject matter of the overtone-resonance theory or source-filter theory is truly periodic signals over a sufficiently long stretch of time. For mechanical devices or music instruments, that condition is easy to fulfill. However, humans never produce a voice with truly periodic pitch. The pitch period, or the time interval between two consecutive glottal closures, varies constantly. Even if a person intentionally makes a voice of constant pitch, jitter (random variation of pitch periods), shimmer (random variation of intensity),
and vibrato (in singing) always present. Synthesized voice without jitter, shimmer, and vibrato sounds buzzy, boring, and unnatural. In addition, in normal speech, pitch varies constantly to convey prosody. Within the time interval of a single vowel, pitch often varies by more than 6 semitones. More than 60% of the world’s languages are tone languages [103], where pitch variations in individual syllables distinguish lexical or grammatical meaning. Isolated glottal closures and isolated decaying acoustic waves at the beginning or the end of a vowel are indispensable elements of speech. In speech, vocal fry is not unusual, where the pitch is lower than the average pitch and somewhat irregular. Among professional narrators and young women, vocal fry near the end of phrases is intentionally practiced to make their speech stylish and attractive. A startling fact is that the vowels in the vocal-fry sections of speech can be clearly perceived. It indicates that a single decaying acoustic wave triggered by a single glottal closure contains sufficient timbre information of the vowel. From the point of view of the transient theory, such phenomena are part of genuine human voice that can be treated naturally and straightforwardly.

On the other hand, by applying the transient theory to a truly periodic train of pulsations, all the phenomena predicted by the overtone-resonance theory can be derived. Because any experimental datum treatable by the overtone-resonance theory can also be treated equally well by the transient theory, and a substantial portion of experimental data can be treated only by the transient theory but not by the overtone-resonance theory, according to Einstein’s criterion of a good theory, the transient theory, especially the concept of timbron introduced in Chapter 4, is presented as the sole irreducible basic element of voice production throughout the book.
Chapter 1
Acoustic Waves

Acoustic waves in air are the carrier of human voice. The production, transmission, and receiving of human voice follow the laws of acoustics. The theory of acoustic waves is a very mature branch of physics. There are excellent monographs and textbooks on this subject \[63, 72\]. In this Chapter, we will review the basic theory and facts of acoustic waves in air as the background information for the understanding of human voice.

1.1 Wave Equation in a Tube

An acoustic wave in a tube of uniform cross section is a simple case, but worthy thorough examination. Much of the conceptual understanding of the human voice can be obtained by considering this simple case. To help us understand the acoustic process of human voice production, the derivation is made as transparent as possible, and illustrated by a number of figures.

1.1.1 Particle displacement and perturbation density

Figure 1.1 shows a tube with a uniform cross section $A$. Air is a compressible fluid. A plane in the air column can only move in the $x$-direction. The displacement $\xi$ of a plane in the air is a function of original location $x$ of the plane, and time $t$. Suppose at a time $t$, the displacement of an air particle originally at $x$ is $\xi(x, t)$, and the displacement of an air particle originally at $x + \Delta x$ is $\xi(x + \Delta x, t)$, the volume of the air mass originally in a volume...
between \( x \) and \( x + \Delta x \), \( V = A\Delta x \), becomes
\[
A\Delta x \rightarrow A \left[ (x + \Delta x + \xi(x + \Delta x, t) - (x + \xi(x, t)) \right] \\
\approx A\Delta x \left[ 1 + \frac{\partial \xi(x, t)}{\partial x} \right].
\tag{1.1}
\]
Assuming the mass of the air package is originally
\[
m = A\Delta x \rho_0,
\tag{1.2}
\]
where \( \rho_0 \) is the unperturbed air density; as a consequence of conservation of mass, as the volume changes, the density is changed to
\[
\frac{\rho_0 A\Delta x}{A\Delta x \left[ 1 + \frac{\partial \xi(x, t)}{\partial x} \right]} \approx \rho_0 - \rho_0 \frac{\partial \xi(x, t)}{\partial x}.
\tag{1.3}
\]
The additional term in the right-hand side is the perturbation density of the air particle, denoted as \( \rho(x, t) \),
\[
\rho(x, t) = -\rho_0 \frac{\partial \xi(x, t)}{\partial x}.
\tag{1.4}
\]
The perturbation density varies with \( x \) and \( t \). If the particle displacement \( \xi(x, t) \) increases with coordinate \( x \), the volume of the air package is expanding, and the perturbation density is negative, as expected.

1.1.2 Particle velocity and equation of continuity

The velocity of a particle originally at location \( x \) and time \( t \) is defined as
\[
u(x, t) \equiv \frac{\partial \xi(x, t)}{\partial t}.
\tag{1.5}
\]
By differentiating Eq. 1.4 with respect to \( t \), using Eq. 1.5, we find
\[
\frac{\partial \rho(x, t)}{\partial t} = -\rho_0 \frac{\partial \nu(x, t)}{\partial x}.
\tag{1.6}
\]
Equation 1.6 is the equation of continuity, which is a direct consequence of conservation of mass.

1.1.3 Perturbation pressure

In all applications regarding human voice, air can be treated as an ideal gas. According to the law of ideal gas, the unperturbed air pressure \( p_0 \) is related to the unperturbed air density \( \rho_0 \) by
\[
p_0 = \rho_0 RT,
\tag{1.7}
\]
1.1 Wave Equation in a Tube

where \( R \) is the gas constant, 8.31 J/(mol·K), and \( T \) is the absolute temperature. Under a constant temperature, pressure is directly proportional to density. When the density is changed from \( \rho_0 \) to \( \rho_0 + \rho(x, t) \), the pressure is changed from \( p_0 \) to \( p_0 + p(x, t) \). The perturbation pressure \( p(x, t) \) is

\[
\frac{p(x, t)}{p_0} = \frac{\rho(x, t)}{\rho_0}.
\] (1.8)

When an air particle is compressed or decompressed, the temperature changes due to the work done to it. As shown in Eq. 1.1, the variation of the particle displacement \( \xi(x, t) \) with \( x \) results in a change of volume,

\[
\delta V = A\Delta x \frac{\partial \xi(x, t)}{\partial x}.
\] (1.9)

Because of the pressure \( p_0 \), an amount of work \( \delta W \) is done to the air particle,

\[
\delta W = -p_0 \delta V = A\Delta x p_0 \frac{\partial \xi(x, t)}{\partial x} = -A\Delta x RT \rho(x, t),
\] (1.10)

which changes its internal energy by \( \delta W \). If the specific heat of air is \( c_v \), the heat capacity of the air particle is \( c_v A\Delta x \). Temperature changes by

\[
\delta T = \frac{-p_0 \delta V}{c_v A\Delta x} = \frac{RT}{c_v} \rho(x, t).
\] (1.11)

By differentiating a logarithmic version of Eq. 1.7, we find

\[
\frac{p(x, t)}{p_0} = \frac{\rho(x, t)}{\rho_0} + \frac{\delta T}{T} = \left(1 + \frac{R}{c_v}\right) \frac{\rho(x, t)}{\rho_0}.
\] (1.12)

The dimensionless constant

\[
\gamma = 1 + \frac{R}{c_v}
\] (1.13)

is a thermodynamic constant of the gas, which is the ratio between the constant-pressure specific heat and the constant-volume specific heat. For dry air, \( \gamma = 1.40 \). Finally we have

\[
\frac{p(x, t)}{p_0} = \gamma \frac{\rho(x, t)}{\rho_0}.
\] (1.14)

In terms of perturbation pressure, Eq. 1.6 becomes

\[
\frac{\partial p(x, t)}{\partial t} = -\gamma p_0 \frac{\partial u(x, t)}{\partial x}.
\] (1.15)

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
1.1.4 Wave equation

The wave equation can be obtained by applying Newton’s law on an air particle, see Fig. 1.2. An air particle with volume $A\Delta x$ is experiencing the force by the pressure difference of the two sides. Newton’s equation is

$$\rho_0 \frac{\partial u(x,t)}{\partial t} A\Delta x = -\frac{\partial p(x,t)}{\partial x} A\Delta x.$$  (1.16)

Differentiating both sides with regard to $t$, we obtain

$$\rho_0 \frac{\partial^2 u(x,t)}{\partial t^2} = -\frac{\partial^2 p(x,t)}{\partial t \partial x}.  \quad (1.17)$$

On the other hand, differentiating Eq. 1.15 with regard to $x$ yields

$$\frac{\partial^2 p(x,t)}{\partial x \partial t} = -\gamma \rho_0 \frac{\partial^2 u(x,t)}{\partial x^2}. \quad (1.18)$$

The differentiations should be independent of their order. Therefore, we obtain a second-order differential equation for particle velocity $u(x,t)$,

$$\rho_0 \frac{\partial^2 u(x,t)}{\partial t^2} = \gamma \rho_0 \frac{\partial^2 u(x,t)}{\partial x^2}. \quad (1.19)$$

Defining the velocity of sound $c$ by

$$c^2 = \frac{\gamma \rho_0}{\rho_0} = \gamma RT, \quad (1.20)$$

we obtain the wave equation in air,

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u(x,t)}{\partial t^2}. \quad (1.21)$$

Because of the linear relation between velocity and perturbation pressure, the same type of equation is valid for the perturbation pressure,

$$\frac{\partial^2 p(x,t)}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 p(x,t)}{\partial t^2}. \quad (1.22)$$

At $0^\circ$C and one atmosphere pressure, the velocity of sound is $c = 330$ m/s. It is proportional to the square root of the absolute temperature. At $20^\circ$C, $c = 342$ m/s; and at $37^\circ$C, $c = 352$ m/s.

Fig. 1.2. Derivation of wave equation. The wave equation is a direct consequence of Newton’s equation, the conservation of mass, and the state equation of the ideal gas.
1.2 d’Alembert Solution

For an infinitely long tube of uniform cross section, Eq. 1.21 has a simple and general solution, obtained by French mathematician and physicist Jean le Rond d’Alembert in 1747,

\[ u(x, t) = F(x - ct) + G(x + ct), \]

(1.23)

where \( F(x) \) and \( G(x) \) are two independent, arbitrary functions. The solution can be proved by direct verification. On one hand,

\[ \frac{\partial u(x, t)}{\partial t} = c F'(x - ct) - c G'(x + ct), \]

(1.24)

thus

\[ \frac{\partial^2 u(x, t)}{\partial t^2} = c^2 F''(x - ct) + c^2 G''(x + ct). \]

(1.25)

On the other hand,

\[ \frac{\partial^2 u(x, t)}{\partial x^2} = F''(x - ct) + G''(x + ct). \]

(1.26)

Therefore, the d’Alembert solution satisfies the wave equation, Eq. 1.22. It is a combination of a wave \( F(x - ct) \) propagating in the positive \( x \) direction at velocity \( c \), and a wave \( G(x + ct) \) propagating in the negative \( x \) direction at velocity \( -c \).

As an application of d’Alembert solution, the reflection of a wave pulsation at the end of a tube is examined, see Fig. 1.4. First, consider the case of a rigid wall. At \( x = L \), the particle velocity is always zero. Let a pulsation \( F(x - ct) \) propagate from the negative side into the positive \( x \) direction. After the pulsation hits the wall, another wave \( G(x + ct) \) propagating in the negative \( x \) direction to compensate the incoming wave such that at \( x = L \), the velocity is always zero. The condition means

\[ F(L - ct) + G(L + ct) = 0. \]

(1.27)

Therefore, the form of the reflected wave should be

\[ G(x) = -F(2L - x). \]

(1.28)

Figure 1.3(A) shows the evolution of the reflection event. As shown, the wave reflected at a rigid wall has a particle velocity opposite to that of the incoming wave.

If the end of the tube is open, the perturbation pressure should always be zero. According to Eq. 1.15, the condition at \( x = L \) is

\[ \frac{\partial u(x, t)}{\partial x} \bigg|_{x=L} = 0, \]

(1.29)

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
which means that at the open end of the tube, the derivative of velocity with respect to \( x \) is zero. It requires that

\[
F'(L - ct) + G'(L + ct) = 0. \tag{1.30}
\]

Following a similar argument, we have

\[
G'(x) = -F'(2L - x). \tag{1.31}
\]

By making an integration, notice the negative sign in \( x \), we obtain

\[
G(x) = F(2L - x). \tag{1.32}
\]

As shown in Fig. 1.3(B), the particle velocity of the pulsation reflected at the open end has the same direction as the incoming pulsation.

For a wave pulsation in perturbation pressure or equivalently, perturbation density, following the same argument, similar results can be obtained. The wave pulsation reflected at a solid wall has a perturbation density of the same polarity as the incoming wave pulsation, and the wave reflected at an open end has a perturbation density of the opposite polarity as the incoming wave pulsation, see also Fig. 1.3.

---

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
1.3 Euler’s Transient Resonator

In a monograph *Tentamen novae theoriar musicae*, Leonhard Euler laid down a mathematical theory of music, including the physics of many musical instruments. For woodwind instruments, he derived a formula for its pitch, or frequency. He showed that the pitch depends only on the length of the tube. For the process of sound generation, Euler wrote thusly [99]:

If a single pulsation be excited at the bottom of a tube closed at one end, it will travel to the mouth of this tube with the velocity of sound. Here an echo of the pulsation will be formed which will run back again, be reflected from the bottom of the tube, and again present itself at the mouth where a new echo will be produced, and so on in succession till the motion is destroyed by friction and imperfect reflection. If it be a compressed pulsation that is echoed from the open end of a tube, the echo will be a rarefied one and *vice versa*, but the direction of the particle velocity will be the same. On the other hand when the reflection takes place from the stopped end, the pulsation retains its density, but the propagation from the mouth of the tube of a succession of equidistant pulsations alternatively compressed and rarefied, at intervals corresponding to the time required for the pulse to travel down the tube and back again; that is to say, a short burst of the musical note corresponding to a stopped pipe of the length in question, will be produced.

---

*Fig. 1.4. Leonhard Euler.* Swiss mathematician and physicist (1707-1783), who created much of the terminology, notations, and a large number of formulas in modern mathematics, for example, Euler’s formula $e^{ix} = \cos x + i \sin x$. A special case, $e^{i\pi} + 1 = 0$, is arguably the most beautiful formula in mathematics. The constant $e = 2.71828\ldots$ is called *Euler’s number.* He also made substantial contributions to physics and engineering. His portrait was printed on a Swiss bank note, a rare honor for a scientist.
Euler’s argument is shown in Fig. 1.5. Here, (1) through (3) show a pulsation of compressed air propagating from the bottom of the tube to the opening. Because at the opening of the tube, the density is a constant; the reflected pulsation is made of rarefied air, see (4) through (6). At the bottom of the tube, the particle speed is always zero. The reflected pulsation is again rarefied, see (7) through (9). The pulsation is once more reflected by the open end, to become compressed, see (10) through (12).

The period of a complete cycle $T$ is then four times of travel time along the length $L$:

$$T = \frac{4L}{c},$$

(1.33)

and the frequency is

$$f = \frac{c}{4L}.$$ 

(1.34)

The above result was given by Euler in his monograph on music.

Euler further stated that because of friction and imperfect reflection (for example, due to radiation), the intensity decays with time. Because the loss of energy per period is proportional to the available energy, the decay must be exponential. For the fundamental frequency component, the particle velocity $u(t)$ can be written in a conceptually simple form

$$u(t) = a \sin \frac{\pi ct}{2L} e^{-\kappa t},$$

(1.35)

where $\kappa$ is decay constant, and $a$ is amplitude. If the top of the tube is closed, as shown in Fig. 1.5, the first reflected pulsation is compressed.
Therefore, a complete cycle is $2L/c$. The particle velocity is

$$u(t) = a \sin \frac{\pi ct}{L} e^{-\kappa t}. \quad (1.36)$$

### 1.4 Energy and Power

#### 1.4.1 Power of acoustic wave

The propagation of acoustic waves carries certain energy and power with it. For an acoustic wave traveling to the $+x$ direction, acoustic energy is transferring into the $+x$ direction with a power related to the waveform. On the other hand, an acoustic wave traveling to the $-x$ direction carries acoustic energy into the $-x$ direction. In this section, the relation between the waveform and the power of acoustic wave is studied.

Consider a plane at location $x$, see Fig. 1.6. A perturbation pressure $p(x,t)$ is acting on the plane. The air at that location is moving with a speed $u(x,t)$. First, consider the case of a wave moving towards the $+x$ direction with air velocity $u(x,t) = F(x - ct)$. The instantaneous transfer of energy from the left-hand side to the right-hand side is

$$W = Ap(x,t)u(x,t). \quad (1.37)$$

Because both $p(x,t)$ and $u(x,t)$ are functions of $x - ct$, using Eq. 1.15, we find an expression of $p(x,t)$ in terms of $u(x,t)$ as

$$p(x,t) = \frac{\gamma p_0}{c} u(x,t). \quad (1.38)$$

The instantaneous power is

$$W = \frac{A\gamma p_0}{c} u(x,t)^2. \quad (1.39)$$

Therefore, the energy only transfers into the $+x$ direction.

---

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
On the other hand, for a wave traveling into the $-x$ direction, where $u(x,t) = F(x + ct)$, the instantaneous power is

$$ W = -\frac{A\gamma p_0}{c} u(x,t)^2. \quad (1.40) $$

The energy only transmits into the $-x$ direction, as expected.

Equations 1.36 and 1.36 are useful for estimating the instantaneous power of human voice.

### 1.4.2 Acoustic energy density

Another problem of interest is the acoustic energy density in a tube. The derivation is similar to the derivation of the Poynting vector and energy density in the electromagnetic field. The net instantaneous power influx into a section of unit volume is the rate of increase of energy density,

$$ \frac{dE}{dt} = -\frac{d}{dx} [p(x,t)u(x,t)] $$

$$ = -u(x,t)\frac{dp(x,t)}{dx} - p(x,t)\frac{du(x,t)}{dx}. \quad (1.41) $$

Using Eqs. 1.15 and 1.16,

$$ \frac{dE}{dt} = \rho_0 u(x,t)\frac{du(x,t)}{dt} + \frac{1}{\gamma p_0} p(x,t)\frac{dp(x,t)}{dt} $$

$$ = \frac{d}{dt} \left[ \frac{\rho_0}{2} u(x,t)^2 + \frac{1}{2\gamma p_0} p(x,t)^2 \right]. \quad (1.42) $$

Therefore, the acoustic energy density is

$$ E = \frac{\rho_0}{2} u(x,t)^2 + \frac{1}{2\gamma p_0} p(x,t)^2, \quad (1.43) $$

where the first term is the acoustic kinetic energy density, and the second term is the acoustic potential energy density.

Equation 1.43 points to a significant concept in acoustic energy density. The acoustic potential energy is proportional to the square of the perturbation pressure. For an elementary volume with a positive perturbation pressure, where the pressure is higher than the average atmosphere pressure, there is acoustic potential energy. Similarly, for an elementary volume with a negative perturbation pressure, where the pressure is lower than the average atmosphere pressure, acoustic potential energy also exists. The magnitude of the acoustic energy only depends on the absolute value of the
perturbation pressure, regardless of the sign of the perturbation pressure. This is similar to the case of acoustic kinetic energy. Regardless of the direction of the particle velocity, the acoustic kinetic energy is proportional to the square of local particle velocity. This point is important for the conceptual understanding of the energy conversion process, as in Section 1.5, Section 2.2.6, and Section 4.2.

1.5 Zero-Particle-Velocity Wavefronts in a Tube

In this section, a case of practical importance is studied: the conversion of aerodynamic energy into acoustic energy by blocking a steady air flow in a tube. It is important to remind that the particle velocity, the source of acoustic kinetic energy, is no different from the aerodynamic velocity of air. Both are governed by the Navier-Stokes equations and the ideal-gas state equation. Therefore, a complete conversion is possible.

1.5.1 Heuristic discussions

Figure 1.7 shows the process, step by step. For $t < 0$, shown in Fig. 1.7(A), there is a steady airflow with velocity $u_0$. At $t = 0$, shown in Fig. 1.7(B), a rigid wall is set up at a point $x = 0$. On both sides of the rigid wall, an acoustic process starts to take place. For the right-hand side, a d’Alembert wavefront propagates in the $+x$ direction at velocity $c$. Because at $x = 0$,
the particle velocity is always zero, the rear side of the wavefront always has zero particle velocity. Therefore, the acoustic process leaves behind a region of zero particle velocity, as shown in Fig. 1.7(C) and (D).

Because the velocity of acoustic wave and the initial particle velocity have the same direction, the air in the region left behind is rarefied, and the perturbation pressure is negative. The magnitude can be estimated as follows: As the wavefront moves to $ct$, where $\tau$ is a lapse of time, the continuous motion of the air with velocity $u_0$ makes the actual length to be $(c + u_0)\tau$. Because the conservation of mass, the density becomes

$$\rho = \rho_0 \frac{ct}{(c + u_0)\tau} \approx \rho_0 - \frac{u_0}{c}. \tag{1.44}$$

Therefore, the perturbation density is

$$\rho(x, t) = -\rho_0 \frac{u_0}{c}. \tag{1.45}$$

Using Eq. 1.14, the perturbation pressure is

$$p(x, t) = \gamma \frac{\rho_0 \rho(x, t)}{\rho_0} = -p_0 \frac{u_0}{c}. \tag{1.46}$$

It is negative, which stores acoustic potential energy. On the other hand, for the region $x < 0$, a d’Alembert wavefront propagates in the $-x$ direction at velocity $-c$, also as shown in Fig. 1.7(C) and (D). It also leaves behind a region of zero particle velocity. Because the velocity of acoustic wave and the initial particle velocity have opposite directions, the air in the left-over region is compressed. Similarly, the perturbation pressure is

$$p(x, t) = p_0 \frac{u_0}{c}. \tag{1.47}$$

It is positive, which also stores acoustic potential energy. Therefore, the aerodynamic kinetic energy of the steady airflow is converted into acoustic potential energy, and generates perturbation pressures on both sides.

The case discussed here is related to the mechanism of human voice production in the following way. Assume that the glottis is located at $x = 0$. The negative side corresponds to the trachea. The positive side corresponds to the vocal tract. When the glottis is open, there is a continuous airflow with velocity $u_0$. A glottal closure blocks the airflow at $x = 0$. To observe its acoustic effects, a miniature pressure sensor can be installed below the glottis, referred to as the subglottis pressure sensor; and another above the glottis, referred to as the supraglottis pressure sensor. According to the solutions presented in this section, immediately after a glottal closure,
a positive perturbation pressure should be observed at the subglottis sensor, and a negative perturbation pressure of the same magnitude should be observed at the supraglottis sensor. Indeed those pressure surges were repeatedly observed, and the absolute magnitudes match well with the theoretical estimates. See Section 3.3.

Because of the importance of the process, in the following subsections, a rigorous mathematical treatment is presented.

### 1.5.2 Laplace-transform solution

We first formulate the problem mathematically. The initial condition at \( t \leq 0 \) is a uniform flow of air,

\[
u(x, t) = u_0, \quad -\infty < x < \infty, \quad t \leq 0,
\]

where \( u_0 \) is the velocity of air flow. At \( t = 0 \), a solid wall is set up at the origin, \( x = 0 \). After that, the velocity at \( x = 0 \) is always zero,

\[
u(0, t) = 0, \quad t > 0.
\]

The question is to find the distribution of particle velocity \( u(x, t) \), perturbation pressure \( p(x, t) \) and perturbation density \( \rho(x, t) \) in the entire tube as a function of \( x \) and \( t \). To do that, the left half, \( x < 0 \), and the right half, \( x > 0 \), are treated separately using a Laplace transform in \( x \) for each half.

The Laplace transform of \( u(x, t) \) for \( x > 0 \) is,

\[
U(s, t) \equiv \mathcal{L}\{\nu(x, t)\} = \int_0^{\infty} e^{-sx} \nu(x, t) dx.
\]

The Laplace transform of the second derivative of the particle velocity \( u''(x, t) = \partial^2 u(x, t) / \partial x^2 \) is

\[
\mathcal{L}\{u''(x, t)\} = s^2 \mathcal{L}\{u(x, t)\} - su(+0, t) - u'(+0, t).
\]

Because of the boundary condition Eq. 1.1, both \( u(+0, t) \) and \( u'(+0, t) \) are zero for \( t > 0 \). Therefore, for \( t > 0 \),

\[
\mathcal{L}\{u''(x, t)\} = s^2 \mathcal{L}\{u(x, t)\}.
\]

Using wave equation Eq. 1.3, the differential equation for the Laplace transform of air velocity is

\[
s^2 U(s, t) = \frac{1}{c^2} \frac{d^2 U(s, t)}{dt^2}.
\]
The general solution of Eq. 1.53 is

\[ U(s, t) = C_1 e^{c s t} + C_2 e^{-c s t}. \]  \hspace{1cm} (1.54)

The first term in Eq. 1.54 goes to infinity for large \( t \). Therefore, only the second term is meaningful. The constant \( C_2 \) is determined by the initial condition, Eq. 1.48,

\[ C_2 = U(s, 0) = \int_0^\infty e^{-s x} u_0 dx = \frac{u_0}{s}. \]  \hspace{1cm} (1.55)

The evolution of air velocity \( u(x, t) \) is determined by the inverse Laplace transform of

\[ U(s, t) = \frac{u_0}{s} e^{-c s t} \]  \hspace{1cm} (1.56)

which is

\[ u(x, t) = \begin{cases} 
0 : & 0 < x < c t \\
u_0 : & x \geq c t
\end{cases} \]  \hspace{1cm} (1.57)

Equation 1.57 represents a zero-particle-velocity wave front traveling from \( x = 0 \) to the right at the velocity of sound \( c \). Left over for \( x < c t \) is a portion of air with zero particle velocity.

The perturbation velocity in the left-hand side is similar:

\[ u(x, t) = \begin{cases} 
0 : & 0 > x > -c t \\
u_0 : & x \leq -c t
\end{cases} \]  \hspace{1cm} (1.58)

### 1.5.3 Energy conversion

Before the location \( x = 0 \) is blocked by a solid wall, in the entire tube, there is a constant air velocity, with a uniform density of kinetic energy

\[ \mathcal{E}_k = \frac{1}{2} \rho_0 u_0^2. \]  \hspace{1cm} (1.59)

That kinetic energy is aerodynamic rather than acoustic. After the airflow is blocked at the origin \( x = 0 \), the velocity of the airflow in the tube gradually becomes zero. Where is that aerodynamic kinetic energy going?

As shown in Fig. 1.4(B), at time \( t \), the zero-particle-velocity wavefront moves from \( x = 0 \) to \( x = c t \). Because at that time, the air mass continues to flow forward with velocity \( u_0 \), it accumulates a displacement

\[ \xi(ct, t) = u_0 t. \]  \hspace{1cm} (1.60)

In the interval \( 0 < x < c t \), the displacement has a constant gradient

\[ \frac{\partial \xi(x, t)}{\partial x} = \frac{u_0}{c}. \]  \hspace{1cm} (1.61)
According to Eq. 1.4, in the space interval $0 < x < ct$, the perturbation density is
\[ \rho(x, t) = -\frac{\rho_0 u_0}{c}, \quad 0 < x < ct. \] (1.62)

According to Eq. 1.14, the perturbation pressure is
\[ p(x, t) = -\frac{\gamma \rho_0 u_0}{c}, \quad 0 < x < ct. \] (1.63)

The air package left behind the zero-particle-velocity wavefront is diluted and has a negative perturbation pressure. Nevertheless, according to Eq. 1.43, it gives rise to a positive acoustic potential energy density
\[ \mathcal{E}_p = \frac{1}{2} \frac{\gamma \rho_0}{\rho_0} p(x, t)^2 = \frac{\gamma \rho_0 u_0^2}{2c^2}. \] (1.64)

Notice that the velocity of sound is $c^2 = \gamma \rho_0 / \rho_0$, Eq. 1.64 becomes
\[ \mathcal{E}_p = \frac{1}{2} \rho_0 u_0^2. \] (1.65)

which equals exactly the aerodynamic kinetic energy density, Eq. 1.59. In summary, the propagation of a zero-particle-velocity wavefront converts the aerodynamic kinetic energy into acoustic energy, represented by a negative perturbation pressure.

In the interval to the left of the origin, $-ct < x < 0$, the air velocity also becomes zero. A similar argument leads to the perturbation density
\[ \rho(x, t) = \frac{\rho_0 u_0}{c}, \quad -ct < x < 0. \] (1.66)

According to Eq. 1.14, the perturbation pressure is
\[ p(x, t) = \frac{\gamma \rho_0 u_0}{c}, \quad -ct < x < 0. \] (1.67)

The air left behind the left-hand side zero-particle-velocity wavefront is compressed with a positive perturbation pressure. The acoustic energy density is also identical to the original aerodynamic energy density, Eq. 1.59.

If the tube has a finite length, then the acoustic wave will radiate into open air. In this case, the aerodynamic kinetic energy of a steady airflow is converted into acoustic energy and then radiates into open air.

### 1.6 Fourier Analysis

A fundamental mathematical tool to represent and reproduce voice signals is Fourier analysis, where a voice signal can be expanded into a series of...
sinusoidal wave components. By using Euler’s formula, \( e^{ix} = \cos x + i \sin x \), the sinusoidal waves can be represented by complex exponential functions, a process that is extremely powerful and convenient. The presentation here is designed for voice signals.

### 1.6.1 Amplitude and phase

In 1822, French mathematician and physicist Joseph Fourier discovered that any periodic function \( f(t) \) of period \( T \) satisfying the condition

\[
f(t + T) = f(t)
\]

(1.68)

can be expanded into a fundamental frequency component and a series of overtones,

\[
f(t) = \sum_{n=1}^{\infty} A_n \cos \left( \frac{2n\pi t}{T} - \phi_n \right),
\]

(1.69)

where \( A_1 \) is the amplitude and \( \phi_1 \) is the phase of the fundamental component, whereas \( A_n \) and \( \phi_n \) are the amplitude and phase of the \( n \)-th overtone. In the conventional literature of Fourier analysis, there is also a constant, or DC component. Because of the DC-blocking capacitance in the amplifier circuit, there is no DC component in any voice signal. And the DC component is never audible. Here, the DC term is omitted.

Using a trigonometry identity, Eq. 1.69 can be rewritten into

\[
f(t) = \sum_{n=1}^{\infty} \left[ a_n \cos \left( \frac{2n\pi t}{T} \right) + b_n \sin \left( \frac{2n\pi t}{T} \right) \right],
\]

(1.70)

where

\[
a_n = A_n \cos \phi_n
\]

(1.71)

and

\[
b_n = A_n \sin \phi_n.
\]

(1.72)

The phase \( \phi_n \) can be determined by the coefficients \( a_n \) and \( b_n \). To better define the values of the phase, the two-variable arc tangent function in C programming language is used,

\[
\phi_n = \text{atan2}(b_n, a_n).
\]

(1.73)

The returned value is in the range of \((-\pi, \pi)\). The atan2 function is well defined for every point other than \((0, 0)\), even if \( x = 0 \) and \( y \neq 0 \).

The Fourier coefficients are

\[
a_m = \frac{2}{T} \int_{-T/2}^{T/2} f(\tau) \cos \left( \frac{2m\pi \tau}{T} \right) d\tau
\]

(1.74)
and
\[ b_m = \frac{2}{T} \int_{-T/2}^{T/2} f(\tau) \sin \left( \frac{2m\pi \tau}{T} \right) d\tau. \] (1.75)

Those expressions of Fourier coefficients can be proved by direct computation. Substituting \( f(\tau) \) by Eq. 1.70, then do the integration in Eq. 1.74. For all the terms where \( n \neq m \), or any of the sine terms, the result is zero. For \( n = m \), the average value of \( \cos^2 x \) is \( 1/2 \). The integral is then \( T/2 \), which cancels the factor \( 2/T \). A similar proof is valid for Eq. 1.75.

### 1.6.2 Complex-variable version

By using the Euler formula
\[ \cos x = \frac{e^{ix} + e^{-ix}}{2}, \] (1.76)

Eq. 1.69 can be written as
\[ f(t) = \sum_{n=-\infty}^{\infty} c_n \exp \left( \frac{2n\pi it}{T} \right), \] (1.77)

where we define
\[ c_n = \frac{1}{2} (a_n - ib_n). \] (1.78)

From Eqs. 1.74 and 1.75, \( a_{-n} = a_n \) and \( b_{-n} = -b_n \). Following Eqs. 1.74 and 1.75, the expression of the complex Fourier coefficients is
\[ c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(\tau) \exp \left( -\frac{2n\pi i\tau}{T} \right) d\tau. \] (1.79)

### 1.6.3 Fourier transform

By increasing the period \( T \) to infinity, Fourier analysis can be applied to any function with finite values in the interval \((-\infty, \infty)\). Here is a proof.

Introducing a variable \( \omega(n) = 2n\pi/T \), Eq. 1.79 becomes
\[ c_n = \frac{1}{T} \int_{-\infty}^{\infty} f(\tau) e^{-i\omega(n)\tau} d\tau. \] (1.80)

Substitute Eq. 1.80 into Eq. 1.77,
\[ f(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \left[ \int_{-T/2}^{T/2} f(\tau) e^{-i\omega(n)\tau} d\tau \right] e^{i\omega(n)t}. \] (1.81)
If \( T \) is large, the sum over \( n \) can be approximated by an integral over \( \omega \). The change of \( \omega(n) \) per incremental change of \( n \) is \( 2\pi/T \). By writing \( \omega \) as a continuous parameter, the sum in Eq. 1.81 becomes an integral,

\[
f(t) \rightarrow \int_{-\infty}^{\infty} \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\tau) e^{-i\omega \tau} d\tau \right] e^{i\omega t} d\omega. \tag{1.82}
\]

The expression in the square bracket of Eq. 1.82 is the Fourier transform of the function \( f(t) \),

\[
F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\tau) e^{-i\omega \tau} d\tau, \tag{1.83}
\]

and Eq. 1.82 becomes

\[
f(t) = \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega. \tag{1.84}
\]

The Fourier transform of function \( f(t) \) can further be represented by a pair of real functions, the amplitude spectrum \( A(\omega) \) and the phase spectrum \( \phi(\omega) \), defined by

\[
A(\omega) = \sqrt{F^*(\omega)F(\omega)} \tag{1.85}
\]

and

\[
\phi(\omega) = \text{atan2}(\text{Im}F(\omega), \text{Re}F(\omega)), \tag{1.86}
\]

where we have

\[
F(\omega) = A(\omega) e^{-i\phi(\omega)}. \tag{1.87}
\]

The amplitude spectrum is a non-negative even function of \( \omega \),

\[
F(-\omega) = F(\omega), \tag{1.88}
\]

and the phase spectrum is an odd function

\[
\phi(-\omega) = -\phi(\omega). \tag{1.89}
\]

A single component in the Fourier series is a sinusoidal wave, described by the frequency or period \( T \), the amplitude \( A \) and the phase \( \phi \):

\[
f(t) = A \sin \left( \frac{2\pi t}{T} - \phi \right). \tag{1.90}
\]
1.7 Numerical Values

1.7.1 Pitch scale

A single sinusoidal acoustic wave has a well-defined pitch, or frequency. In music, the frequency is expressed in a logarithmic scale. The smallest interval in traditional Western music is the semitone, which is defined as a frequency ratio of $\sqrt[12]{2}=1.059463$, see Fig. 1.8.

The numerical values of pitch $p$ are expressed in MIDI, an integer number related to frequency $f$ in hertz by

$$f = 440 \exp \left( \frac{\ln 2}{12} (p - 69) \right) = 440 e^{0.05776 (p - 69)},$$  \hspace{1cm} (1.91)
Fig. 1.9. Ranges of human voice. Usually, human voices are divided into six ranges, three female voices and three male voices. The two middle ranges are sometimes omitted, to become S, A, T, and B.

or

\[ p = 69 + \frac{12}{\ln 2} \ln \frac{f}{440}. \]  \hspace{1cm} (1.92)

Often, frequency is expressed by pitch period \( T \), with a convenient unit of milliseconds. The relation between pitch period and pitch in MIDI is

\[ p = 83.21 - 17.31 \ln T, \]  \hspace{1cm} (1.93)

or alternatively

\[ T = 122.3 e^{-0.05774p}. \]  \hspace{1cm} (1.94)

Usually, human voices are divided into six ranges, see Fig. 1.9. There are three female voices, soprano, mezzo-soprano, and contralto. There are three male voices, tenor, baritone, and bass. In choral music, the two middle ranges are often omitted, and the lower part of female voice is called alto, to become S, A, T, and B. In everyday conversation, the pitch ranges for both genders are often in the lower ranges, A and B.

1.7.2 Intensity scale

In engineering, the intensity of voice is often expressed in a logarithmic scale \[39\]. The standard unit is decibel, or dB, defined by Alexander Graham Bell, as 10 times the logarithm of the sound intensity \( I \) (in W/m\(^2\)) divided by a reference sound intensity \( I_0 \),

\[ I \text{(in dB)} = 10 \log \left( \frac{I}{I_0} \right). \]  \hspace{1cm} (1.95)

The reference sound level is hearing threshold, below which humans cannot perceive. The reference sound intensity is usually defined at 3 kHz as \( 10^{-12} \) W/m\(^2\), to be taken as the zero point of sound level. Therefore, the sound
level in dB is always a positive number. Because the sound intensity is proportional to the square of sound pressure, Eq. 1.95 can be written as

$$I\text{(in dB)} = 20 \log \left( \frac{p}{p_0} \right),$$  \hspace{1cm} (1.96)

where $p$ is the amplitude of sound pressure, and $p_0$ is a reference sound pressure, which is $p_0 = 2 \times 10^{-5}$ Pa. However, the sensitivity of human ear to acoustic power depends on the frequency of the acoustic signal. Through many decades of research, the reference hearing threshold and the loudness scale is defined in an international standard, ISO 226:2003, as shown in Fig. 1.10. Human ears are most sensible to acoustic signals of frequency around 3 kHz. For acoustic signals of 200 Hz or of 10 kHz, the threshold is raised by roughly 20 dB. The dynamic marks on the right-hand side are heuristic rather than quantitative.
Chapter 2
Voice Organs

The science of human voice production involves both physics and physiology. Human voice is produced by a group of human organs. There is a vast amount of literature about the anatomy and physiology of those organs. For a brief overview, see Sataloff’s Scientific American article The Human Voice [74]. More detail can be found in chapter 2 of Sundberg’s Science of the Singing Voice [92], and first four chapters of Titze’s Principles of Voice Production [95]. Even more detail can be found in Clinical Anatomy and Physiology of the Voice [75], a part of Sataloff’s Professional Voice [78]. In this chapter, we make a brief review of the anatomy and physiology of the voice organs as a background for the discussion of human voice production mechanism in the following two chapters.

2.1 Overall Structure

Figure 2.1 is a cross-sectional view of the voice-production organs along the sagittal plane, which divides the human body into left and right halves. The source of voice energy is the kinetic energy of the airflow from the lungs through the trachea, near the bottom of Fig. 2.1. In the process of producing voiced sounds, including vowels such as [a], [i], and [o]; and voiced consonants such as [n], [m], [z] and [ʒ]; the air stream sets the vocal folds to oscillate, causing the air path to close and open frequently. The oscillation of the vocal folds generates an alternation of finite glottal airflow and zero glottal airflow, which triggers the voiced sound. The basic anatomy of vocal folds are presented in the following section.

The “color” of the voiced sound, termed timbre, is determined by the vocal tract, which includes the pharynx, the oral cavity and the nasal cavity. The shape of the oral cavity can be morphed, or controlled by various parts of human organs inside the mouth, including the tongue, teeth and lips. The nasal cavity, although has no way of shaping control, is also an important path for voice generation. In the generation of nasal consonants, the oral path is closed, and the air flow can only go through the nose.
The pharynx is a multiple-function organ: for food ingestion, breathing, and voice production. To prevent food from entering the trachea, the epiglottis functions as a gate. While swallowing food, the path between the pharynx and the vocal fold is closed by the epiglottis, the path of the pharynx and the esophagus is open, and the functions of breathing and producing voice are suspended. Inside the skull, there are many cavities filled with air, which can function as resonance devices for human voice. The largest ones are the frontal sinus and the sphenoidal sinus. The average size of the sphenoidal sinus is 2.7 cm. The average size of the sphenoidal sinus is 2.2 cm. Those cavities are connected with the nasal cavity via apertures. The walls of the sinuses are rather rigid, making them low-loss resonance cavities; and the small dimensions make the resonance frequencies high.

2.2 Vocal Folds

Vocal folds play a pivotal role in voice production. They form a valve to control the airflow from the lungs (through the trachea) to the vocal tract. While closed, vocal folds isolate the vocal tract from the porous lungs to make the vocal tract a high-quality resonance chamber. The article *An Overview of Laryngeal Function for Voice Production* by Ronald J. Baken [4] and the article *Laryngeal Function During Phonation* by Ronald C. Scherer

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
2.2 Vocal Folds

[80] in Sataloff’s *Professional Voice* [78] cover this topic in detail. Here we present a brief review as the background for the next two chapters.

### 2.2.1 Anatomy

Figure 2.2 shows the top view of vocal folds from the pharynx and the cross section views for both open and closed states. The voluntary motion of the vocal folds, *abduction* or moving apart, and *adduction* or moving together, is controlled by the *vocalis muscles*. The oscillation of the vocal folds, however, is involuntary and its frequency is determined by the elastic properties and the masses of the tissue. Figure 2.2 (A) and (B) show the open or *abducted* state. The space between the separated vocal folds is the *glottis*. Through the glottis, the interior of the trachea becomes visible from the pharynx. From the cross sectional view, (B), it becomes clear that *false vocal folds* do not participate in voicing. Rather, they provide some protection to the (true) vocal folds. An important fact to notice is that the vocal folds have a finite thickness, usually greater than 5 mm. During phonation, the upper edge of the vocal folds and the lower edge of the vocal folds do not move synchronously. There is a time lag between the motion.

---

Fig. 2.2. Vocal folds. (A) and (C) show the top view. (B) and (D) show the cross section view along a coronal plane. Abduction of the vocal folds creates a space between them: the glottis. As shown in (B), the vocal folds have a finite thickness. During phonation, the upper edges of the vocal folds and the lower edges of the vocal folds do not move simultaneously. As shown in (D), the lower edges of the vocal folds are in contact, while the upper edges are separated, see Section bernoulli.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
of the upper edge and the motion of the lower edge. We will come back to this fact later on. Figure 2.2 (C) and (D) show the closed or *adducted* state. As shown in Fig 2.2(D), at that moment, the lower edges of the vocal folds are in contact, but the upper edges are separated. There are also moments that the upper edges of the vocal folds are in contact and the lower edges are separated.

### 2.2.2 Strobovideolaryngoscopy

In principle, the motion of the vocal folds can be studied using high-speed video recording technique. However, the vibration frequency of the vocal fold is very high, ranging from less than 100 Hz to more than 800 Hz. To record the motion of the vocal folds within a single pitch period, a speed of several thousands of frames per second is required. Such systems have become available only recently using digital technology, see Chapter 9 of Boehme and Gross [7], but are too expensive for everyday use.

For more than fifty years, the standard technology for imaging the motion of vocal folds has been stroboscopy [7, 45]. A flash light is programmed to produce intensive light pulses periodically, and nearly synchronous with the periodically vibrating vocal folds. Each picture grasps an image of the vocal folds at a different phase. The collection of the images could show a sequence of vocal folds movement. This technique for inspecting the larynx is termed *laryngeal stroboscopy* or *strobovideolaryngoscopy*. Figure 2.3 shows an example, taken with a Kay Elemetnics stroboscope [50].

Laryngeal stroboscopy is a standard instrument for clinic diagnosis of
2.2 Vocal Folds

voice problems. However, it also has disadvantages [7, 45]. It does not provide a continuous series of images. Its time resolution is limited. It can only work with an almost strictly periodic vibration of the vocal folds, which imposes a limitation. It requires a video camera, either a rigid endoscope, or a flexible system using fiber optics inserted through the nose. In either case, it interferes with the speaker’s freedom of phonation.

2.2.3 Laryngeal electromyography

The functions of the vocal folds are controlled by various muscles and nerves in the larynx. The actions of a muscle are accompanied by a change of ion concentrations in the muscles, which is then manifested as a change of electric potential in the muscle. Therefore, by continuously measuring the electric potential of a muscle, the activities of the muscle can be monitored. The technique, electromyography, has been a standard clinic procedure for muscle and nerve systems for several decades. This technique has been used as a clinic examination tool for the larynx, the laryngeal electromyography [77, 79]. In a certain sense, this technique is similar to the application of electrocardiogram for the clinical examination of the heart. There are differences, however. The actions of heart muscles are involuntary, while the actions of the muscles in the larynx is voluntary. The electric pulses of the heart are strong. Therefore, for a cardiogram, the signals can be detected by applying surface electrodes at various places on the surface of the body. Because the laryngeal muscles are small and in close proximity to each other, electrodes must be placed inside the specific muscles to effectively detect the action potentials [77].

Figure 2.4 shows a typical electrode for laryngeal electromyography. It consists of a hollow steel shaft B with a wire A runs through its center, and is insulated for the entire length except a the tip. The outer shaft B is grounded. To perform laryngeal electromyography, the patient is lying on an operating bed, with the neck extended to facilitate the identification of laryngeal muscles. Because local anesthesia could interfere with the muscle actions, and the procedure is not too painful, the needle electrodes are

Fig. 2.4. Needle electrode for laryngeal electromyography. A typical concentric needle electrode contains a central wire, insulated and enclosed in a hollow steel shaft. During examination, a needle electrode is inserted into a laryngeal muscle of interest. The outer shaft is grounded as the reference of the electric potential. After Sataloff et al. [77, 79].
34 Voice Organs

Fig. 2.5. An example of laryngeal electromyograph. When the muscle is at rest, no electric pulses are observed. The frequency of the electrical pulses, each is called a motor unit, increases as the intensity of voluntary contraction of the muscle increases. After Sataloff et al. [77, 79].

usually inserted into the muscles of interest without anesthesia. Because the actions of the laryngeal muscles are voluntary, to verify the accuracy of electrode placement, the patient is asked to take actions, for example, by sniffing or phonating the sound /i/, and the physician watches the changes in the electrical signals as evidence.

An example of the observed electromyograph is shown in Fig. 2.5. As shown, when the muscle is at rest, no electrical pulses are observed. As the intensity of voluntary contraction increases, the frequency of electrical pulses is likewise increased. In combination with other clinic observations, laryngeal electromyography can provide invaluable evidence for an accurate diagnosis of laryngeal conditions [77, 79].

The insertion of needle electrodes to the laryngeal muscles is a minimally invasive medical procedure and must be performed by highly qualified physicians. It is not performed under normal speaking or singing conditions. Therefore, other examination techniques to study the vibrations of vocal folds under normal speaking or singing conditions are needed [45]. In the following, several non-invasive techniques for the study of vocal-folds vibrations are discussed, together with the consequent findings.

2.2.4 Electroglottography

In 1956, French otolaryngologist Philippe Fabre, then a correspondent member of the French National Academy of Medicine, invented an electrical instrument to study the physiology of the vocal folds. The first report, entitled Un procédé électrique percutané d’inscription de l’accélération glot-
2.2 Vocal Folds

tique au cours de la phonetion: glottographie de haute fréquence. Premiers résultats, was published in 1957 on Bulletin de l’Académie nationale de médecine [27]. A schematics of its working principle is shown in Fig. 2.6. Two electrodes are pressed on the skin of the neck near the vocal folds. A weak electrical signal of frequency about 200 kHz is applied on one electrode. Another electrode is the detector of the current passing through the neck. Part of the current passes through the vocal folds. As shown in Fig. 2.6, when the glottis is open, part of the electrical current is blocked; and when the glottis is closed, more electrical current can pass through it. The small difference of conductance due to the opening and closing of the vocal folds is then detected. It is of great advantage to use an electrical signal of a high frequency: First, by using a narrow-range band-pass filter and especially by using the phase-lock technique, the signal-to-noise ratio can be greatly enhanced, and the signal quality can be made very high. Second, the temporal resolution of the signal is high. Using a 200 kHz signal, a temporal resolution of 0.02 msec can be achieved.

A follow-up paper, entitled La Glottographie électrique en haute fréquence, particularités de l’appareillage, published on Comptes Rendus, Société de Biologie in 1959 [28], disclosed a detailed electrical circuit diagram. The core of the circuit is an oscillator based on an LC circuit to produce high-frequency source. The basic circuit has been used in many contemporary products, although the vacuum tubes are replaced by transistors and integrated circuits. The commonly used name of the instrument, electroglottograph, abbreviated EGG, was also established. Fig. 2.7 is a photograph of the electrodes, taken in the author’s recording studio.

![Fig. 2.6. Principle of the electroglottograph. Two electrodes are pressed against the neck, near the vocal folds. A high-frequency electrical signal, typically 100 to 200 kHz, is used to probe the electrical conductance between the two electrodes, thus to probe the status of the vocal folds. (A) While the glottis is open, the conductance between the two electrodes is lower. (B) While the glottis is closed, the conductance is higher, thus the current is higher.](image)

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
In 1992, Baken published a review article about EGG [3]. He said,

EGG is noninvasive, innocuous, inexpensive, and does not interfere with simultaneous measurement of other relevant variables, such as airflow or glottal area. ... (A)nd of paramount importance, EGG provides a view of certain aspects of vocal fold function that is not obtainable by any other means.

In clinical voice laboratories, EGG is used routinely. It allows objective determination of the presence or absence of glottal vibration. It reflects the glottal condition more accurately during the closed phase, and quantitative interpretation of the glottal condition is possible [45].

Since the 1980s, in the speech technology community, a huge amount of high-quality simultaneous voice and EGG signals has been recorded. Designed for research and supported by the government, large-size speech and EGG signal corpus are available publicly. The examples shown in this Section are from the ARCTIC databases, spoken by three speakers, two male and one female, each reads a prepared text of 1132 sentences. The databases are published by Carnegie-Melon University [52]. The recordings also provide a rich resource for the study of general features of EGG signals.

A universal feature of the EGG signals is a very strong asymmetry between glottal closing and glottal opening, as shown in Fig. 2.8. The upper half of the Figure, (A), shows the conductance between the two electrodes. A greater value indicates a closed glottis, and a smaller value indicates an open glottis. The signal is taken at a sampling rate of 32 kHz. The lower half of the Figure, (B), is the time-derivative of the EGG signal, marked as dEGG/dt. It is simply the difference of the value of EGG signal and the value at the previous time point, properly scaled for easy viewing. In Fig. 2.8, the EGG signals in a pitch period are marked by numbers 1 through 8. Point 1 has the lowest conductance, the state of a widely open glottis. At point 2, the vocal folds start to move together, and the conductance is increasing. At point 3, the conductance suddenly increases, and the dEGG/dt signal shows a sharp maximum. At point 4, the rapid increase of conduc-
Fig. 2.8. Typical EGG signals of vowels. (A), the conductance signal. A high conductance indicates a large contact area between the two halves of the vocal folds, therefore a closed glottis. A low conductance indicates a small contact area between the two halves of the vocal folds, therefore an open glottis. (B) the time-derivative of the EGG signal, indicating the speed of conductance variation. Note the sharp peak in dEGG/dt at the closing moments. Source: ARCTIC databases, speaker bdl (US English male speaker), sentence a0002, from 0.86 second to 0.90 second, part of vowel [a].

tance starts to slow down slightly, but continues to rise to a highest point 5. At that point, the conductance reaches a maximum, indicates that the contact area of the two halves of the vocal folds reaches a maximum. At point 6, the conductance is reduced, indicating that the contact area between the two vocal folds has decreased. At point 7, the decrease of conductance accelerates, showing a mild minimum in the dEGG/dt signal, indicating the opening of the glottis. At point 8, the conductance continues to drop, where the glottis opens widely.

The dramatic asymmetry of the huge peak in dEGG/dt at the closing moment (marked by thick arrows) and the weak peak in dEGG/dt at the opening moment (marked by thin vertical lines) is universal in a great majority of simultaneously acquired voice signals and EGG signals. In speech technology community, the sharp maximum of the dEGG/dt signal at the glottal closing moment is taken as the starting point of a pitch period.

The strong asymmetry of glottal opening and glottal closing is critical for the understanding of the voice-production mechanism, and is the basis of pitch-synchronous segmentation of voice signals. In the following Subsection, we will provide an explanation of the strong asymmetry.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
2.2.5 Bernoulli force

In this Subsection, we discuss the effect of Bernoulli force when the glottal width is small, to provide an understanding of the strong and sharp peaks at glottal closing moments in the observed dEGG/dt signals.

Figure 2.9 shows the vocal-fold vibration process according to the body-cover vocal-fold structure proposed by Hirano [46, 87]. Each side of the vocal folds is divided into a body, shaded dark gray, consisting of relatively tight and stiff tissue; and a cover, shaded light gray, consisting of more pliable and flexible tissue. Arrows are added to show velocity distribution, with the arrow length indicating the airflow velocity at that location.

Fig. 2.9. Schematic diagram of vocal-fold vibration. After Hirano [46, 87], with the addition of velocity distributions to show the effect of the Bernoulli force. The length of the arrow indicates the velocity of airflow at that location. The numbers, (1) through (8), are made to match the numbers in Figs. 2.8 and 2.15.
2.2 Vocal Folds

At step (1), the glottis is widely open. Air from the trachea flows through the glottis. At step (2), owing to the elastic force, the vocal folds move together, and the glottal width narrows. Because of the condition of continuity, the speed of airflow through the glottis is increased. At one point, (3), Bernoulli force kicks in. Because the elastic force and the Bernoulli force are in the same direction, and the cover tissue is more pliable and flexible, a vigorous self-reinforcing cycle takes place to rapidly close the glottis, (4). The mucosal wave continues, causing the contact area between the vocal folds to increase, (5). Then, the elastic force pulls the vocal folds apart, first at the lower edges, (6). When the glottis starts to open, the Bernoulli force reappears, (7). However, this time, the Bernoulli force is in the opposite direction of the elastic force. The two forces partially cancel each other. Minor random variations of either of the Bernoulli force or the elastic will show up as the main component of force. Therefore, the opening of glottis is slow and noisy. Eventually, the elastic force overcomes the Bernoulli force. In step (8), the glottis is again widely open.

The Bernoulli force is a critical factor to make the voice production effective. Here we make an analysis to the effect of the Bernoulli force in the oscillation of the vocal folds. Based on conservation of energy, Daniel Bernoulli (1700-1782) discovered that in a continuous flow, the sum of pressure and the kinetic-energy density is a constant,

\[ p_1 + \frac{1}{2} \rho v_1^2 = p_2 + \frac{1}{2} \rho v_2^2, \]  

(2.1)

where \( p_1 \) is the pressure and \( v_1 \) is the velocity at point 1, \( p_2 \) is the pressure and \( v_2 \) is the velocity at point 2. The change of density in this case can be neglected. First, we estimate the Bernoulli pressure in the trachea. The typical rate of air flow is \( Q = 0.5 \) liter/sec while speaking. The typical cross section of the trachea is \( 2.5 \text{ cm}^2 \) [39]. Therefore, the typical velocity is \( 2 \) m/s. The density of air is \( 1.25 \) Kg/m\(^3\). The pressure increases by

\[ \Delta p = \frac{1}{2} 1.25 \times 2^2 \approx 2.5 \text{ Pa}. \]  

(2.2)

It is tiny. However, if the glottis is nearly closed, the pressure could be large. Assuming glottis length is \( L = 10 \) mm [88], when the glottal width is \( 2 \) mm, the velocity is \( 25 \) m/s, and the Bernoulli pressure is about 156 times greater than that in the trachea:

\[ \Delta p = \frac{1}{2} 1.25 \times 25^2 \approx 400 \text{ Pa}. \]  

(2.3)

The force is the product of Bernoulli pressure times the effective area \( A \), which can be estimated from the parameters of the Story-Titze model [88] as

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
follows. The length of the glottis $L = 10$ mm. The total width of the vocal fold is 3 mm. Take one half of it, the effective width is 1.5 mm. Therefore, the effective area is $A = 1.5 \times 10^{-5}$ m$^2$. The Bernoulli force acting on the vocal folds is

$$F_B = 400 \times 1.5 \times 10^{-5} \approx 0.006 \text{ N.}$$

(2.4)

Because the Bernoulli force varies as the square of velocity, it increases rapidly when glottal width is reduced, as shown in Table 2.1. Consequently, when the glottal width is small enough, the Bernoulli force becomes the dominant force to bring the vocal folds together. Once the process is started, the glottal width is further reduced, the air velocity increases further, as the Bernoulli force grows as the square of the air velocity. A self-reinforcing cycle is formed, forcefully and abruptly closing the glottis.

Since the 1960s, numerical models of vocal-fold oscillation were proposed, including the one-mass model of Flanagan and Landgraf [29], the two-mass model of Ishizaka and Flanagan [48], and the three-mass model of Story and Titze [88]. In all cases, Bernoulli force was included as a driving force together with elastic forces [29, 48, 88]. Here we present an analytic treatment to elucidate the effect of Bernoulli force near the glottal closing moment, to explain the observed sharp peaks in the dEGG/dt signals.

Figure 2.10 follows the three-mass vocal fold model of Story and Titze [88]. The three masses are the body mass $m_b$, the lower cover mass $m_l$, and the upper cover mass $m_u$. The values are

$m_b = 0.05 \text{ g, } m_l = 0.01 \text{ g, and } m_u = 0.01 \text{ g;}

and the three corresponding spring constants are

$k_b = 100 \text{ N/m, } k_l = 5.0 \text{ N/m, and } k_u = 3.5 \text{ N/m.}$

The value of $k_b$ is a variable. Here the medium value is taken, which corresponds to a resonance frequency

$$f = \frac{1}{2\pi} \sqrt{\frac{k_b}{m_b}} \approx 190 \text{ Hz,}$$

(2.5)

a pitch frequency shared by male and female speakers and singers.

Table 2.1: Bernoulli force and elastic force

<table>
<thead>
<tr>
<th>Glottal width (mm)</th>
<th>2</th>
<th>1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli force (N)</td>
<td>0.006</td>
<td>0.024</td>
<td>0.096</td>
</tr>
<tr>
<td>Elastic force (N)</td>
<td>0.2</td>
<td>0.1</td>
<td>0.05</td>
</tr>
</tbody>
</table>
2.2 Vocal Folds

Fig. 2.10. Vocal-fold oscillation and Bernoulli force. (A) A three-mass model following Story and Titze [88]. (B) Rapid closing due to Bernoulli force. When the glottis width is small, shown as \( x_1 \), due to Bernoulli force, the speed \( v_1 \) increases rapidly. The time to close a 0.5 mm gap is much less than 0.1 msec.

Assuming the maximum glottal width is \( x_0 = 2 \) mm, the maximum elastic force is

\[
F_k = k_b x_0 \approx 0.2 \text{ N.} \tag{2.6}
\]

Comparing with the values of Bernoulli force in Table 2.1, we see that even the glottal width is at 1 mm, the Bernoulli force is still smaller than the elastic force. When the glottal width becomes 0.5 mm, at the time the elastic force is 0.05 N, the Bernoulli force becomes greater than the elastic force. Furthermore, if the glottal width is smaller than 0.5 mm, the elastic force is insignificant; and the Bernoulli force dominates.

In the following, we make a quantitative analysis of the abrupt glottal closing due to Bernoulli force, see Fig. 2.10. Because the Bernoulli force in the trachea is tiny, only the Bernoulli force of the air flow through the glottis in Eq. 2.1 is effective. Let the flow rate be \( Q \), the anterior to posterior length of the glottis be \( L \), the glottal width be \( x \), the velocity of airflow is \( v = Q/Lx \). When Bernoulli force dominates, shown in step (3) of Fig. 2.9, only the lower cover mass \( m_1 \) is in effect. Let the effective contact area be \( A \), the Newton’s equation for the glottal width \( x \) is

\[
m_1 \frac{d^2 x}{dt^2} = -\frac{1}{2} \rho v^2 A = -\rho Q^2 A \frac{1}{2L^2} \frac{1}{x^2}. \tag{2.7}
\]

Eq. 2.7 can be simplified to

\[
\frac{d^2 x}{dt^2} = -\frac{C}{2x^2}. \tag{2.8}
\]

Taking \( A = 1.5 \times 10^{-5} \text{ m}^2 \), the parameter \( C \) is

\[
C = \frac{\rho Q^2 A}{L^2 m_1} \approx \frac{1.25 \times (5 \times 10^{-4})^2 \times 1.5 \times 10^{-5}}{(10^{-2})^2 \times 0.01 \times 10^{-3}} \approx 6.8 \times 10^{-2} \text{ m}^3/\text{sec}^2. \tag{2.9}
\]
Multiplying both sides by $dx/dt$, Eq. 2.8 can be integrated to

$$
\left( \frac{dx}{dt} \right)^2 = C \left[ \frac{1}{x} - \frac{1}{x_0} \right],
$$

(2.10)

the integration constant $x_0$ is the distance where velocity is zero. At this position, the glottis is widely open and the elastic force dominates. When the glottis width is small, $x \ll x_0$, Eq. 2.10 is simplified to

$$
\frac{dx}{dt} = -\sqrt{\frac{C}{x}},
$$

(2.11)

where the minus sign from the square root indicates an attractive force. Integrating 2.11, the result can be conveniently written as

$$
t = t_0 - \frac{2}{3} \sqrt{\frac{x_0^3}{C}}, \quad \text{with} \quad t_0 = \frac{2}{3} \sqrt{\frac{x_0^3}{C}}.
$$

(2.12)

The meaning of Eq. 2.12 is as follows. If at $t = 0$, the glottal width is $x = x_1$; at time $t_0$, the glottis is completely closed, see Fig. 2.10. For $x_1 = 0.5$ mm, using the estimation of $C$ in Eq. 2.9, one finds $t_0 \approx 0.1$ msec. On the scale of frequency, it corresponds to about 10 kHz, which is above the frequency range of formants. Therefore, on the time scale of formants, the final glottal closing event is instantaneous.

An important issue is the rate of glottal-area declination near the moment of a complete closure, that is, $x \to 0$. Because the glottal area is $\alpha = Lx$, from Eq. 2.11, the declination rate of glottal area is

$$
\frac{d\alpha}{dt} = \frac{d}{dt} Lx = -L \sqrt{\frac{C}{x}}.
$$

(2.13)

Therefore, the rate of glottal-area declination goes to infinity as the glottal width $x$ approaches zero.

Experimental observations of glottal closing process using stroboscopy revealed that the two edges of the glottis are often not parallel, but move as a zipper with a glottic angle [7], see Fig. 2.11. By representing the glottis as a triangle with base $x$ and angle $\phi$, the velocity of airflow is $v = 2Q\phi/x^2$. Following a similar argument leading to Eq. 2.5, the Newton’s equation of the glottal width $x$ is

$$
M \frac{d^2x}{dt^2} = -\rho Q^2 A\phi^2 \frac{2x^4}{2x^4}.
$$

(2.14)

Using a procedure similar to that leading to Eq. 2.10, Eq. 2.14 can be reduced to

$$
\left( \frac{dx}{dt} \right)^2 = 4\rho Q^2 A\phi^2 \frac{3M}{3M} \left[ \frac{1}{x^3} - \frac{1}{x_0^3} \right] = D \left[ \frac{1}{x^3} - \frac{1}{x_0^3} \right].
$$

(2.15)

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
Assuming $\phi = 0.1$, the constant $D$ in Eq. 2.15 is

$$D \approx \frac{4 \times 1.25 \times (5 \times 10^{-4})^2 \times 1.5 \times 10^{-5} \times 0.1^2}{3 \times 0.01 \times 10^{-3}} \approx 6.25 \times 10^{-9} \text{ m}^5 \text{sec}^{-2}. \quad (2.16)$$

When $x \ll x_0$, the equation of motion becomes

$$\frac{dx}{dt} = -\sqrt{\frac{D}{x}}.$$  \quad (2.17)

where the minus sign from the square root indicates an attractive force. Integrating 2.17, a result similar to Eq. 2.12 is obtained,

$$t = t_0 - \frac{2}{5} \sqrt{\frac{x_0^5}{D}}, \quad \text{with} \quad t_0 = \frac{2}{5} \sqrt{\frac{x_1^5}{D}}. \quad (2.18)$$

Using the value of $D$ in Eq. 2.16, for $x_1 = 0.5$ mm, we have

$$t_0 \approx \frac{2}{5} \sqrt{\frac{(5 \times 10^{-4})^5}{6.25 \times 10^{-9}}} \approx 0.028 \text{ msec}, \quad (2.19)$$

with is even shorter than that for a parallel glottis, as expected.

Some basic conclusions are as follows:

(1) Whenever there is a complete glottal closing, the Bernoulli force will come into effect at some moment, to make the final closing explosive.

(2) Once the Bernoulli force comes into effect, the time of closing, from a finite airflow to zero, is instantaneous on the time scale of formants.

(3) From Eq. 2.11, towards the closing moment when $x \to 0$, $dx/dt$ approaches infinity. The glottal-area declination rate goes to infinity.
(4) In the opening process of the glottis, at one point, the elastic force and the Bernoulli force will be equal in magnitude and opposite in sign. The two forces will cancel each other at some point. Minor variations of the two competing forces could dominate the event, and random noise could be clearly observed. Therefore, comparing with glottal closing, glottal opening is slow and noisy.

(5) In the interpretation of EGG signals, we should note that the instant of maximum conductance is not the instant of glottal closure. It is the moment of maximum contact area between two halves of vocal folds. Instead, the very fast variation of the EGG signal between point 3 and point 4 is the time of closing, and point 7 is the time of the glottis opening event. In other words, the peaks in the \(d\text{EGG}/dt\) charts are correlated to the closing and opening of the glottis.

2.2.6 Water-hammer analogy

The rapid closing of the glottis due to Bernoulli force points to a mechanism of human voice production proposed by Ronald Baken as an analogy to the water-hammer effect in hydrodynamics [4]:

The sharp cutoff of flow is particularly crucial, because it is this relatively sudden stoppage of the air flow that is truly the raw material of voice. To understand why, think of an experience that you may have had with a poorly designed plumbing system. The faucet is wide open, and the water is running at full force. The tap is then quickly turned off. Water flow stops abruptly and there is a sudden THUMP! from the pipes inside the walls. (Plumbers call this “water hammer.”) This happens because, in the simplest terms, the sudden cessation causes moving molecules of water to collide with those ahead of them (like the chain-reaction collision caused when a car suddenly stops on a highway). This generates a kind of “shock wave”. When the pipe is jolted by this shock, it moves, creating the vibrations in the air that we hear as a thump. The relatively sudden cutoff of flow that characterizes the glottal wave creates very much the same effect in the vocal tract. An impulse-like shock wave is produced that “excites” the vibration of the air molecules in the vocal tract. That excitation is the voice in its unrefined form.

Figure 2.12 shows water-hammer effect in a water supply system with a upstream pipe, a valve, and a downstream pipe. When the valve is open, water flows continuously from the upstream pipe to the downstream pipe.
The flowing water carries a kinetic energy $E_k$

$$E_k = \frac{1}{2} \rho ALv^2,$$

where $\rho$ is the density of water, $A$ is the cross section of the pipe, $L$ is the length of the pipe, and $v$ is the velocity of water flow.

After a sudden closing of the valve, due to inertia, water continues to flow. In the upstream pipe, the flowing water collides with the valve. To make a numerical estimate, we use an example given by the article *water hammer* on Wikipedia as follows. In a 14 km water tunnel of 7.7 meter diameter, full of water traveling at 3.75 m/s, the kinetic energy of the flowing water is 8000 megajoules. The closing of a valve means that amount of kinetic energy must be rested. Because each kilogram of TNT corresponds to 4.7 megajoules of energy, the kinetic energy of the flowing water is equivalent to a monstrous bomb of 1.7 ton TNT. That energy can blow up a well-built water tunnel. On the downstream side, the inertia of flowing water will create a vacuum. The pipe near the valve could collapse due to the suction force created by the kinetic energy of the flowing water. To prevent pipe destruction, relief devices are often required.

Human voice production is similar to the process in the plumbing system, Fig. 2.12. The trachea is equivalent to the upstream pipe. The glottis is the valve. The vocal tract is equivalent to the downstream pipe. When the glottis is open, there is a steady airflow. When the glottis suddenly...
closes, a process similar to the water hammer effect in plumbing systems takes place. This time, it is air instead of water. First, air is 800 times lighter than water. Second, air is compressible. There is kinetic energy on both sides of the glottis that needs to be released, but the effect is much milder. The voice organs would not be destroyed. A zero-particle-velocity d’Alembert wavefront is created on each side, see Section 1.5. On the upstream side, the trachea, pressure builds up near the closed glottis. Such a pressure buildup is actually observed experimentally using a miniature pressure sensor placed inside the trachea, see Section 3.3. Because the trachea is closed, there is no audible sound. On the downstream side, the vocal tract, a zero-particle-velocity d’Alembert wavefront creates a column of rarefied air with a lower pressure. The pressure drop in the pharynx was also experimentally observed using a miniature pressure sensor placed directly above the glottis, see Section 3.3. Notice that a column of rarefied air contains the same acoustic potential energy as being compressed with the same percentage, see Section 1.4.2, it contains enough acoustic energy to become audible voice. Because the length of vocal tract is finite, the air disturbance triggered by a sudden glottal closing resonates in the vocal tract to create a decaying acoustic wave, and radiates. The waveform of that decaying acoustic wave is determined by the geometry of the vocal tract. This is the essence of human voice production, see Section 4.2.

2.2.7 Incomplete closures

The ACRTIC database has 3396 sentences [52]. A very high percentage of the EGG signals of vowels look similar to that in Fig. 2.8. However, exceptions are observed, for example the incomplete closures, shown in Figs. 2.13 and 2.14. The percentage of incomplete closures is not high, and depends on the manner of the speaker. In the ARCTIC databases, speaker bdl has 1.2%, slt has 5.35 %, and jmk has 6.13 %. Nevertheless, it is important in the understanding of the vocal-fold oscillation process.

Incomplete closures often take place near the beginning and the end of a voiced segment. At the beginning of a voiced segment, the vocal folds start to oscillate, first with a small amplitude not able to cause a full closure; then gradually increasing to make full closures. Near the end of a voiced segment, the amplitude of vocal-fold oscillation gradually decreases, the time of closing decreases, eventually becomes zero. However, even without full closure, the vocal folds continue to oscillate, as shown by the periodic variation of the EGG signals. In all cases, it is found that the vibration frequency of the vocal folds, with or without closure, is continuous. There is no abrupt change of pitch period. This experimental fact seems to indicate that the elastic force is the primary factor to determine the vibrational

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
2.2 Vocal Folds

Fig. 2.13. Incomplete closures (1). (A), the EGG conductance signal. (B) the time-derivative of the EGG signal. 1. having closures, but the time of closed phase 4, decreases gradually. 2. the vocal folds continue to vibrate without closings. 3. closure resumes with an increasing time of the closed phase 4. Source: ARCTIC databases, speaker jmk, sentence b0353, from 0.86 second to 1.01 second [52].

Displayed here are several interesting cases where the closures cease for a time period then resume, often with very minor change of the vibrational frequency over the entire course. Figure 2.13 is from a male speaker of rather deep tone and strong EGG signals. At the beginning, 1, there are clear closings. Then the duration of closed phase 4 is shrinking gradually,

Fig. 2.14. Incomplete closures (2). Same as Fig. 2.13. Source: ARCTIC databases, speaker slt, sentence a0042, from 1.12 second to 1.22 second.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
Table 2.2: Incomplete glottal closures

<table>
<thead>
<tr>
<th>Speaker</th>
<th>bdl</th>
<th>jnk</th>
<th>slt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete closures</td>
<td>261,503</td>
<td>184,447</td>
<td>368,864</td>
</tr>
<tr>
<td>Incomplete closures</td>
<td>3,142</td>
<td>11,318</td>
<td>19,767</td>
</tr>
<tr>
<td>Percentage of incomplete closures</td>
<td>1.20%</td>
<td>6.13%</td>
<td>5.35%</td>
</tr>
<tr>
<td>Intra-voice incomplete closures</td>
<td>7</td>
<td>292</td>
<td>590</td>
</tr>
</tbody>
</table>

to enter a section with no closures, 2. Then the closure resumes with an increasing closed-phase duration, 3. Figure 2.14 is from a female speaker of bright tone. The vocal fold oscillation continues without a closure for 12 periods of steady amplitude. During the entire time of 100 msec, the change of pitch period is less than 6%. The accompanying voice signal does not change significantly with or without closures. We will return to this point in Chapter 5.

Table 2.2 shows that number and percentage of such incomplete glottal closures in the ARCTIC databases. As shown, the percentage of incomplete

Fig. 2.15. Isolated closures. Three isolated glottal closures recorded by EGG. Source: ARCTIC databases, speaker slt (US English female speaker), sentence a0312, from 1.51 second to 1.58 second. Because the average pitch frequency is 240 Hz, the distance between adjacent closures is more than 3 times the average pitch period.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
2.2 Vocal Folds

glottal closures depends on speaker, and ranges from 1% to about 6%. The number of intra-voice incomplete closures, where a section of incomplete closures is bounded by two sections with full closures, is very small.

### 2.2.8 Isolated closures and glottal stops

In the ARCTIC databases and the SpeechOcean databases, a substantial portion of glottal closures do not belong to a quasiperiodic series; but look like isolated events. Those isolated glottal closures are either events in a vocal-fry section, or *glottal stops*, which constitute an indispensable element of speech signals. Showing in Fig. 2.15 are three individual closures found in the recordings of a US English female speaker in the ARCTIC databases. Because the average pitch frequency is 240 Hz, the distance between adjacent closures is more than 3 times the average pitch period. The related phenomena will be discussed in Chapters 4 and 5.

### 2.2.9 Videokymography

In section 2.2.2, strobovideolaryngoscopy is briefly presented as a powerful method to study vocal fold oscillations. However, the time resolution of that technology is not enough to observe the fast events during phonation. The sequence of strobovideolaryngoscopic images is not a true sequence in time, but a series of sporadically chosen samples from a large number of pitch periods. Videokymography provides an alternative solution [7, 93].

The working principle of videokymography is shown in Fig. 2.16. Similar to stroboscopy, a CCD video camera is inserted in the larynx. Instead of taking full video pictures, the vertical scan is disabled, and only the horizontal scan is functioning. The line scan is fixed to a single anterior-

![Fig. 2.16. Working principle of videokymography.](image)

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
The videokymography image of a healthy person (A), with explanations in (B). The numbers, (1) through (8), are made to match the numbers in Fig. 2.8. Source: References [7, 93].
2.2.10 Closed quotient and voice intensity

An important experimental observation using EGG and videokymography is the relation between the ratio of the duration of the closed phase and the pitch period with respect to voice intensity [7, 62]. The closed quotient, abbreviated CQ, the ratio of the duration of closed phase and the pitch period, is a significant parameter to characterize the quality voice. Sometimes, the ratio of the duration of the open phase and the pitch period, the open quotient, abbreviated OQ, is also used. Obviously,

\[ CQ + OQ = 1. \]

(2.21)

It is a general observation that high CQ is correlated to loud phonation, and low CQ is correlated to soft phonation [7, 62]. The typical experimental finding of CQ is as follows: For female voices, 0.39 for soft phonation, and 0.51 for loud phonation. For male voices, 0.35 for soft phonation, and 0.53 for loud phonation, see page 109 of Reference [7]. In the study of singing, a frequent observation is that high-intensity singing voice is correlated with high closed quotients, see pages 40–41 of Reference [62]. We will provide a quantitative explanation of the correlation between voice intensity and closed quotient in Chapter 4.

2.3 Vocal Tract

An overall diagram of the vocal tract is in Fig. 2.1. It consists of the larynx, the oral cavity, and the nasal cavity. The oral cavity can undergo extensive modifications by the motions of the soft palate, the tongue, teeth and lips. The overall dimensions are shown in Table 2.3, after Stevens [86].

Vowels are produced by combined actions of vocal folds and vocal tract. Although for vocal music, vowels are the central element, in speech, consonants sometimes hold a more important role than vowels. Evidence is

<table>
<thead>
<tr>
<th>Table 2.3: Dimensions of the vocal tract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average length in mm</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Vocal tract length</td>
</tr>
<tr>
<td>Pharynx length</td>
</tr>
<tr>
<td>Oral cavity length</td>
</tr>
</tbody>
</table>

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
that in many writing systems in the world, only consonants are explicitly written, such as Arabic.

Fricatives are also universal. Unvoiced fricatives are generated at various places of the vocal tract. Vocal folds are widely open. No acoustic role is played by vocal folds in the production of unvoiced fricatives. As an example, fricative [s] is presented.

Since the first systematic study of human voice production by Robert Willis in 1829 [99], it is universally recognized that the timbre of vowels is determined by the shape of the vocal tract. Willis showed that for vowels, the effect of the vocal tract can be understood in terms of Euler’s transient resonator. We will analyze the vowels [a] and [u] as examples. More details will be presented in Chapter 3 and Chapter 4.

2.3.1 Plosives

According to Peter Ladefoged and Ian Maddieson [55], stop consonants are the most universal sound of the world’s languages. Unvoiced stop consonants are generated in the front part of the oral cavity. Vocal folds are in an abducted state, functioning only as a through hole, same as at breathing. As an example, the case of plosive [k] is presented.

Compared with other sounds, the physics and physiology of the production of plosives are relatively simple. As an example, unvoiced stop [k] is discussed here. As we will show below, the actual sound depends on the vowel that follows. Do a five-second experiment by yourself and you will be convinced. Pronounce a devoiced “cut” and a devoiced “cool”. Even before you make a plosive, the positions of the lips and tongue are different. The perceived vibration frequencies of two versions of [k] sound quite different. In Fig. 2.18, the case of [ka], such as in “cut”, is shown.

The production of a velar stop [k] has two steps. In the first step, the tongue is pressing against the hard palate, and the soft palate is pressing against the back side of the pharynx, as shown in (1) of Fig. 2.18(A). The entire air path is completely blocked. Air pressure is building up in the pharynx, as shown in Fig. 2.18(A). The stop consonant starts with the voluntary release of the tongue from the hard palate, as shown by (2) of Fig. 2.18(B). A burst of air flows into the front oral cavity, as shown by (3) of Fig. 2.18(B). Because the following vowel is [A], the teeth are close together, and the lips are not far from the teeth. The space between the soft palate (2) and teeth (4) forms an Euler transient resonator. An example of the observed waveform is shown by the dotted curve in Fig. 2.19. The thin solid curve is an analytic approximation of the observed data using a decaying sinusoidal wave with frequency 1.40 kHz,
2.3 Vocal Tract

Fig. 2.18. Production of plosive [k\#]. (A) As a preparation, the tongue presses against the hard palate to block the airflow. (B) After a voluntary release of the tongue from the hard palate, (2), air bursts into the oral cavity (3), which acts as an Euler transient resonator. A decaying acoustic wave is formed then radiates into open air, (5).

\[ u(t) = 1500 \sin(2\pi \times 1.40 t)e^{-0.485 \tau} , \quad (2.22) \]

where \( t \) is in milliseconds. The length of the Euler transient resonator can be estimated by Eq. 1.34. Using millimeters and milliseconds as units, one obtains \( L = 352/(4 \times 1.40) \approx 63 \) mm. It is consistent with the typical length from the back of the hard palate to the teeth for a male speaker.

Fig. 2.19. Waveform of plosive [k\#]. Dotted line: experimental waveform, source: SpeechOcean Database for Mandarin speech synthesis, sentence 041429, 2.557 sec to 2.567 sec. Thin solid curve is a decaying sinusoidal function, Eq. 2.22.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
Fig. 2.20. Production of plosive [ku]. (A) As a preparation, the tongue presses against the hard palate (1) to block the airflow. (B) After a voluntary release of the tongue from the hard palate, (2), air bursts into oral cavity (3), which acts as an Euler transient resonator. A decaying acoustic wave is formed and then radiates, (5).

Another example of plosive [ku], as in “cool”, is shown in Fig. 2.20. Even at the step of preparing the plosive, the contact point of the tongue against the palate is located further back, and the lips are protruded further out, making the length of the Euler transient resonator much greater than the case of [k\textalpha]. A recorded waveform is shown as the dotted line in Fig. 2.21. A decaying sinusoidal wave, represented by the thin solid curve, is

\[ u(t) = 2400 \sin(2\pi \times 0.753 t) e^{-0.309 t}. \]  

(2.23)
2.3 Vocal Tract

2.3.2 Fricatives

Fricatives are also universal in all languages of the world. As an example, the positions of the oral cavity during the enunciation of fricatives [s] in “see” and “soon” is shown.

The source of the sizzling sound can be identified by doing a five-second experiment by yourself, as shown in Fig. 2.23. First, pronounce a voiceless fricative with devoiced [i] in “see”, as in Fig. 2.23(A). The sound comes from the space between the tip of tongue and the teeth. Next, pronounce a voiceless fricative with devoiced [u] in “soon”, as in Fig. 2.23(B). The sound comes from the space between the tip of tongue and the protruding lips. The difference of the vibrational frequencies can be perceived. During the enunciation of either sizzling consonants, open your mouth but do not change anything else. The consonant immediately stops. There is no sound coming from the throat.

The experimentally observed amplitude spectrum of the signal further clarifies the source. Fig. 2.24(A) shows the spectrum of [si]. It is generated by cutting a 5 msec portion of the signal, take the Fourier transform, then smooth it over a 0.1 kHz interval. Four such examples are displayed. The peak of the spectrum is at 5.5 kHz. Fig. 2.24(A) shows the spectrum of [si].
As shown, the peak shifts to 4.4 kHz. Additional features around 10 kHz are also observed.

The observed spectrum can be understood in terms of Euler tube resonators, see Fig. 2.23. For [s] in “see”, the constriction at point 1 of Fig. 2.23(A) is the source of turbulence, which can be considered as a random emission of pulses. Each pulse excites a resonance wave in the Euler transient resonator 2. The peak frequency 5.5 kHz is associated with a length 16 mm, which is about right as the distance between the constriction 1 and the teeth 3. For [s] in “soon”, Fig. 2.23(B), the resonator is longer. The peak frequency, 4.4 kHz, is associated with a length 20 mm, which is about right as the sum of cavities 5 and 6. The features around 10 kHz are associated with shorter Euler transient resonators, probably 6 in Fig. 2.23(B).

A significant difference of the fricative from vowels is the phase spectrum, as shown in Fig. 2.25. It is a universal observation that the phase spectrum
of fricatives is completely random, reflecting the turbulent nature of the excitation arisen at the constriction of the tube resonator.

2.3.3 Vowels

While plosives and fricatives are produced solely in the oral cavity, vowels and voiced consonants are produced jointly by the vocal folds and vocal tract. In 1829, Robert Willis published the first detailed study of the mechanism of vowel production [99]. As a capable mechanical engineer, Willis built a number of mechanical models to demonstrate the production of vowel sounds, a few of them are illustrated in Fig. 1 in the Preface. At the incoming end of the demonstration device, he used a reed to mimic the vocal folds to generate periodic air puffs at a given pitch frequency. A tube connected to the reed mimics the vocal tract. Willis applied Euler’s theory of transient resonators for explanation: each time a pulsation is emitted by the reed, it will resonate back and forth in the tube to form a decaying wave. The wavelength of the decaying wave is four times the length of the tube. Willis showed that by changing the length of the tube, sound of different colors can be produced to mimic different vowels. With a fixed tube length, while the pitch frequency is changed by the structure of the reed, the timbre of the vowel is unchanged. In other words, he showed that the character of the vowel only depends on the resonance tube, not the reed. The single resonance frequency model of Willis is definitely an oversimplification for most vowels. However, for a few vowels, there is a dominant format, and the frequency of that dominant formant can be explained by simplifying the vocal tract as a tube with uniform cross section, such as [u] and [a]. Note that this intuitively inspiring model is for conceptual understanding only. In reality, the structure of the vocal tract is much more complicated, and a rich array of formants can be produced. See Chapter 4 for details.
First, we study the production of vowel [u], as shown in Fig. 2.26. During the enunciation of vowel [u], the lips are rounded to form a small orifice of diameter about 10 mm, and the tongue is lowered. The entire length of vocal tract, from the vocal folds to the lips, forms an Euler transient resonator, see Fig. 2.26. For male speakers, the average length of the vocal tract is 169 mm. With protruded lips, the total length is about 190 mm. According Euler’s formula, Eq. 1.34, the resonance frequency is

\[ f = \frac{c}{4L} = \frac{352000}{4 \times 190} \approx 460 \text{ Hz.} \]  

(2.24)

Figure 2.27(A) shows the waveform of a vowel [u] pronounced by a male speaker with pitch frequency about 100 Hz, which is typical. As shown, each pitch period starts with a strong peak, then gradually decays. The waveform in each pitch period oscillates four to five times. Figure 2.27(B) is the amplitude spectrum of a single pitch period. As shown, except for the peak at about 100 Hz which is the fundamental-frequency component, the main feature of the spectrum is a strong peak at about 460 Hz.
The oral cavity is widely open. The narrow tube between the glottis and the soft palate becomes an Euler resonator. The typical length of that section is about 110 mm, corresponding to a resonance frequency of about 800 Hz.

Next, we study the production of vowel [a], see Fig. 2.28. While enunciating vowel [a], the mouth is widely open, and the back end of the tongue is close to the soft palate. The mouth is effectively a part of the open air. The pharynx is essentially the resonance tube. The typical length is 110 mm. The expected resonance frequency is 800 Hz. Figure 2.29(A) shows an observed waveform. In each pitch period, there are about 8 decaying sinusoidal cycles. Figure 2.29(B) shows that the vocal-tract resonance frequency is nearly 800 Hz, in line with the expectation. The actual geometry of the vocal tract is more complicated, which gives rise to the second formant of the vowel [a] near 1400 Hz and other features.

Fig. 2.28. Oral cavity while enunciating [a]. The oral cavity is widely open. The narrow tube between the glottis and the soft palate becomes an Euler resonator. The typical length of that section is about 110 mm, corresponding to a resonance frequency of about 800 Hz.

Fig. 2.29. Waveform and amplitude spectrum of [a]. (A) Waveform of vowel [a]. (B) Amplitude spectrum of [a], showing a main formant peak at 800 Hz.
Chapter 3

Experimental Facts

In this chapter, we present basic experimental facts of human voice, collected by various instruments. Based on those facts, in Chapter 4, the physics of human voice production is expounded, which becomes the foundation of the mathematical representations of human voice in Part II.

3.1 Microphones and Voice Signals

The most abundant experimental data of human voice are collected by microphones. In this section, we present the working principles of microphones and observations from voice signals collected by microphones.

3.1.1 Types and working principles of microphones

To understand the nature of microphone signals and to evaluate its quality, the modern microphones and their working principles are presented. Early types of microphone, such as the carbon microphone, are not presented here because of the signal quality is low. The microphones used in modern human-voice recordings are of two categories, the condenser microphone and the dynamic microphone. There are two types of condenser microphones. The membrane type, requiring a phantom power, is used in high-quality recordings. The electret type is used in portable electronics.

The condenser microphone, invented in the early 20th century, is currently the standard microphone for high-quality voice recordings [36, 51, 100]. Photographs of a typical condenser microphone is shown in Fig. 3.1. Figure 3.2 shows its working principle and frequency response curve. Figure 3.1(A) is an internal view after the cover is removed. The sensor of the microphone is a thin membrane plated with gold, typically 5 μm thick, extended on a frame with a tension of around 2000 nN/m against a perforated backplate with an air-gap of typically 25 μm, see Fig. 3.2(A). A capacitor of about 50 pF is formed. A DC voltage source, called phantom power and typically 48 V, is applied to polarize the capacitor. Utilizing the available DC power, a built-in preamplifier is mounted inside the cover, as shown in Fig. 3.1(A) and Fig. 3.2(A). The output impedance is converted
Fig. 3.1. Condenser microphone. Photographs of a condenser microphone Røde NT1-A, taken in the author’s recording studio. (A). Internal view without the cover. The diameter of the membrane is 25 mm. A silicone shock mount and a preamplifier are inside the cover. (B). A complete setup with popshield and external shock mount.

from as high as 10 MΩ to as low as 100 Ω, to make it immune to external electrical noise. The microphone is protected against mechanical vibration by an internal shock mount made of silicone, as well as an external shock mount, as shown in Fig. 3.1. The sensitivity of condenser microphones is typically 25 to 50 mV/Pa. The frequency response, as shown in Fig. 3.2(B), is within a few dB from 20 Hz to 20 kHz.

The condenser microphone is extremely sensitive to any motion of air.

Fig. 3.2. Working principle and frequency response of condenser microphone. (A). Working principle. A 5 μm thick membrane is extended on a frame against a perforated backplate with a 25 μm air gap to form a capacitor. The membrane responds to acoustic pressure to generate a voltage, which is then amplified by the built-in preamplifier. (B). Frequency response. Source: www.rodemic.com.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
3.1 Microphones and Voice Signals

While pronouncing stop consonants such as [p], [t], and [k], the air jet can generate a wild gyration in the output signal. For human voice recording, a popshield is installed in front of the microphone, see Fig. 3.1(B).

Another important type of microphone is dynamic microphone. It is more rugged than the condenser microphones and does not require a phantom power. Therefore, it is used more often than the condenser microphone. Figure 3.3 shows photographs of a typical dynamic microphone, and Fig. 3.4 shows its working principle and frequency response. Figure 3.3(A) shows an internal view after the cover is removed. Although it is much less sensitive to the motion of air than the condenser microphone, an internal foam-plastic pop shield is built in, see Fig. 3.4(A). The dynamic microphone often includes a radio-frequency emitter to make it wireless, and it can then be held by hand during recording. As shown in Fig. 3.4(A), the basic structure of the dynamic microphone is identical to a dynamic loudspeaker. When the coil moves in a magnetic field, an electromotive force is generated. As the

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
condenser microphone responds to air particle displacement, the dynamic microphone responds to air particle velocity. As shown in Fig. 3.4(B), at frequencies lower than 100 Hz and higher than 15 kHz, the response of dynamic microphone becomes lower.

For portable electronic devices, including laptop computers and smartphones, the most common type of microphone is electret microphone. A photograph of a MCE-100 electret microphone is shown in Fig. 3.5(A). The working principle is identical to the studio condenser microphone. Instead of using a phantom power to polarize the capacitor, a thin membrane of permanently polarized material, the electret (a word coined by combining electric and magnet), to form an electric field inside the air gap between the membrane and the baseplate, see Fig. 3.5(B).

### 3.1.2 Source of voice signals

The voice samples used in this book are listed in Table 3.1, recorded with high-quality microphones under well-controlled conditions by professional speakers, with simultaneous electroglottograph signals.

<table>
<thead>
<tr>
<th>Source of database</th>
<th>Speaker</th>
<th>Sentences</th>
<th>Sample rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCTIC databases, CMU</td>
<td>bdl (male)</td>
<td>1132</td>
<td>32 kHz</td>
</tr>
<tr>
<td>ARCTIC databases, CMU</td>
<td>jmk (male)</td>
<td>1132</td>
<td>32 kHz</td>
</tr>
<tr>
<td>ARCTIC databases, CMU</td>
<td>slt (female)</td>
<td>1132</td>
<td>32 kHz</td>
</tr>
<tr>
<td>King-TTS-003, Speechocean</td>
<td>female</td>
<td>19509</td>
<td>44.1 kHz</td>
</tr>
<tr>
<td>King-TTS-012, Speechocean</td>
<td>male</td>
<td>15000</td>
<td>44.1 kHz</td>
</tr>
</tbody>
</table>

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
3.1.3 Vowels

An essential feature of human voice is that the pitch frequency, defined as the inverse of the pitch period, is constantly varying. The production process of vowels can be understood by analyzing experimental facts of vowels under varying pitch frequencies. The examples shown here are sections of vowels near the end of a breathing unit. In each section, the pitch value starts in the high or middle part of the speaker’s pitch range, then gradually lowers. There are thousands of such events in the databases of Table 3.1.

The exposition of information in each of the figures is as follows: In part (A), the top chart is the raw waveform in PCM, the middle chart is the time derivative of PCM, and the lower chart is short-time power averaged over every two msec displayed in dB scale. The last four periods in the speech signal are segmented and ends-matched; then a Fourier analysis is executed on each pitch period, shown in (B). The peaks in the amplitude spectra, that is, the formant frequencies, are marked on each peak.

Vowel [a]. Figure 3.6 shows the experimental data of vowel [a]. At the beginning of the time interval, the pitch frequency is 184 Hz, in the middle-high range of the male voice. The pitch frequency gradually decreases to 118 Hz. Inside each pitch period, the short-time averaged power gradually decreases as well. The dB chart clearly shows an exponential decay within each pitch period with a fairly consistent rate of $-5$ dB per msec. The amplitude spectra of the last four pitch periods, Fig. 3.6(B), are almost identical despite huge pitch period differences. The waveforms of the earlier pitch periods strikingly resemble the starting part of the waveform of later periods. If an early waveform is allowed to evolve freely, each one will display a full decaying wave, resembling the last one.

![Figure 3.6](image)

Fig. 3.6. Vowel [a]. Source: Sentence 050007 of King-TTS-012, 2.23 sec to 2.28 sec. (A). Waveform and short-time power. (B). Amplitude spectra of four pitch periods.
Vowel [i]. Figure 3.7(A) shows the experimental data of vowel [i]. At the beginning, the pitch is 263 Hz, in the high end of the speaker’s pitch range. The decay of the short-time power, in dB scale, has two stages. At the beginning of a pitch period, the short-time power decays $-7 \, \text{dB per msec}$, which is the decay rate of the 2-3 kHz formants. Later, the decay rate becomes $-2 \, \text{dB per msec}$, that of the 330 Hz formant.

Vowel [u]. As shown in Fig. 3.8, the experimental data of vowel [u] displays features similar to vowels [a] and [i]. The starting pitch, 253 Hz, is in the high end of the speaker’s pitch range. The dominant features in the amplitude spectrum are two formants in 400 Hz and 790 Hz. The short-time power decays at a constant rate of $-2 \, \text{dB/msec}$. 

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
Vowel [e]. As shown in Fig. 3.9, the experimental data of vowel [e] displays features similar to vowel [i]. There are three high-frequency formants which decay at $-7$ dB/ms within each pitch period, and the 540-Hz formant decays at $-2$ dB/ms within each pitch period. The corresponding amplitude spectrum shows a sharp formant peak at 540 Hz.

Vowel [o]. As shown in Fig. 3.10, the experimentally observed features are similar to those of vowel [u]. The starting pitch, 161 Hz, is in the middle-high portion of the speaker’s pitch range. The amplitude spectrum shows a sharp formant at 620 Hz. The short-time power shows a consistent $-2$ dB/ms decay rate, apparently related to that formant.
3.1.4 Superposition principle

In this subsection, a summary of the experimental observations described in the previous section is presented. The experimental facts can be understood in the light of the superposition principle, which was formulated by Robert W. Scripture in 1930 [84].

By looking through Figs. 3.6 through 3.10, it is apparent that although the pitch period changes by a factor of two, the underlying elementary waveforms representing the color of a vowel are essentially identical. For example, in Fig. 3.7(A), the number of first-formant cycles in a pitch period starts with 2, increases to 3 and then 4. Nevertheless, the two feature groups in the first pitch period look identical to the first two feature groups in later pitch periods, and so on. The amplitude spectra of the same vowel are essentially identical while the pitch varies by more than an octave.

In Figs. 3.6 through 3.10, it is clear that for each pitch period, the signal starts with a strong impulse, then decays. The high-frequency components between 2 kHz to 4 kHz decay faster, at a rate of $-7 \text{ dB/ms}$. The medium frequency components, around 1 kHz, decay more slowly, at $-5 \text{ dB/ms}$. The low-frequency components, between 0.3 kHz and 0.5 kHz, decays even more slowly, at $-2 \text{ dB/ms}$. Again, if those formant oscillations were allowed to evolve freely, each one would continue until it disappears.

To summarize, for each pitch period, the signal starts with a strong impulse, then decays. For each vowel, the underlying elementary waveforms are essentially identical. The entire signal can be constructed by the superposition principle, which is a property of linear systems.

Because the wave equation is linear, the superposition principle should be valid. It states that the response caused by two or more stimuli is the sum of the responses caused by each stimulus individually. Assuming at $t = t_1$, a stimulus produces a response $F_1(t - t_1)$, and at $t = t_2$, another stimulus produces a response $F_2(t - t_2)$, and so on. The observed signal should be the sum of all responses,

$$S(t) = F_1(t - t_1) + F_2(t - t_2) + F_3(t - t_3) \ldots$$ (3.1)

If the configuration of the vocal tract stays unchanged, the response functions $F_n(t)$ should only differ by a constant factor, determined by the strength of each stimulus. The vowel signal thus produced could be represented by Eq. 3.1 with similar response functions. By looking through Figs. 3.6 through 3.10, it is obvious that the above statement is true.

To synthesize human voice of a vowel with an arbitrary pitch contour and arbitrary intensity profile, only one elementary waveform of that vowel, allowed to decay to the end, is needed. Figure 3.11 shows a simple case with identical intensity and equal time interval between the starting moments.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
The elementary wave, here for a vowel [e], is extracted from a recorded voice, as displayed in Fig. 3.9. (A) through (C), three elementary waves of vowel [e], each separated by a pitch period $T$. (D) through (F) show synthesized voice using the superposition principle with pitch values varying over one octave. The synthesized voice is smooth, but sounds robotic because of the constant pitch. Using Eq. 3.1, with varying pitch periods and intensities, more human-like voice can be synthesized, see Chapter 8.

The superposition principle presented here is phenomenological: it is solely based on observations on recorded voice signals. For a better understanding of the underlying physics, especially the nature of the stimulus that starts the elementary wave of a vowel, other experimental facts need to be studied. These are presented in the following sections.

### 3.2 Electroglottograph and Voice Data

In section 2.2.4, the working principle and observations of the electroglottograph (EGG) are presented. As shown, the peaks in the derivative of the EGG signal point to the closing moment and opening moment of the glottis. As a result of the Bernoulli force, the $d$(EGG)/$dt$ signal at glottal closings...
is much sharper than that at glottal openings. The glottal closing is the defining moment of a pitch period. In this Section, the temporal correlation of the d(EGG)/dt signal and the voice signal is investigated.

3.2.1 Temporal correlation

Since the advance of unit-selection speech synthesis research started in the 1980s, large speech databases of simultaneous voice and EGG data have been collected. Some of them are shown in Section 3.1.2. Figure 3.12 shows an example. As shown in Section 2.2.4, the opening moment and the closing moment divide each pitch period into a closed phase and an open phase. Some general observations are as follows:

- The voice signal starts after the glottal closing moment with a constant delay $\tau$, typically 1 msec.
- The voice signal in the closed phase is much stronger than the voice signal in the open phase.
- The effect of opening moment on the voice signal is minor and variable. Sometimes it causes mild noise signals.

The observed facts can be explained by looking at Fig. 3.12. The linear distance from the glottis to the microphone, marked on Fig. 3.12 as a gray curve $L$, is about 350 mm. The time for an acoustic pulse to travel from Fig. 3.12. Microphone signals and electroglottograph signals. At a glottal closing moment, an acoustic pulse is generated. The time for an acoustic pulse to travel from the glottis to the microphone through the curve $L$ is about 1 msec. The voice signal in the closed phase is strong, and the voice signal in the open phase is weak. Source: ARCTIC databases, speaker bdl, sentence b0274, time 1.454 sec to 1.476 sec.
the glottis to the microphone is about 1 msec. Therefore, the *glottal closure* is the origin of the stimulus. In other words, at the time the glottis closes, an impulse emerges from the glottis and starts the elementary wave of the vowel of each pitch period. That impulse takes about 1 msec to travel along the gray curve to reach the microphone. Figure 3.13 shows the correlation between voice signal and EGG signal for six vowels. The above facts are universally observed.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
The above experimental facts are counterintuitive at first glance. Naturally, one would think that the process of vowel production is similar to the production of plosive [k], presented in Section 2.3.1: During the closed phase, the pressure in the trachea builds up, as marked by a hollow arrow in Fig. 3.12. It forces the vocal folds to open in an explosive manner, emitting a puff of air to excite the voice signal. If such a picture were true, then the voice signal would peak at about 1 msec after a glottal opening. Nevertheless, the experimental fact is, very little action is observed within a millisecond or so after the opening of the glottis. On the contrary, the opening of glottis often weakens the voice signal. It is a general observation that the voice signal in the open phase is weaker than the voice signal in the closed phase. This puzzle will be solved in Chapter 4.

3.2.2 Glottal stops

In Section 2.2.4, we have shown isolated EGG closure signals not connected with a quasi-period train. Each of such closures triggers a piece of voice signal. These voice signals are known as the glottal stop in phonetics, with a IPA symbol [ʔ]. An example of such voice signals, by a female US English speaker slt in the ARCTIC database, is shown in Fig. 3.14.

Fig. 3.14. Glottal stops. Source: ARCTIC databases, speaker slt, sentence a0312, time 1.521 sec to 1.557 sec. The sections of waveform (A) and (B) are Fourier analyzed to generate amplitude spectra (C) and (D). The formant structure is essentially identical to that of the typical formant structure of vowel [a].

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
In Fig. 3.14, two consecutive glottal stops, with a time elapse of 17 msec, are shown. From a perception point of view, the two events are perceived as a single glottal stop, similar to the multiple velar plosive shown in Section 2.3.1. The correlation of the glottal closure detected by EGG signal and the starting of the glottal stop signal is identical to the production of vowels, see Section 3.1.3: a delay of about 1 msec originated from the propagation time of the line \( L \) in Fig. 3.12. In the closed phase, clean decaying signals are observed. In the opening moment, certain noise is generated, also similar to the phenomenon observed in vowels.

The segments of waveforms extracted from the signal, marked as (A) and (B), are Fourier-analyzed, and the amplitude spectra are shown as (C) and (D), respectively. As shown, the formant peaks match well with those of vowel \([a]\). Clearly, although those events are characterized as stop consonants, the spectral nature is identical to the vowel \([a]\), see Fig. 3.6. If those events are allowed to repeat with a time interval of for example 5 msec, a sustained vowel \([a]\) is produced.

### 3.3 In-Vivo Pressure Measurements

In the 1960s, semiconductor miniature pressure transducers for medical research were developed. Those pressure transducers, mounted on a flexible catheter of 2 mm in diameter, originally designed for inserting into the blood vessels, with a frequency response up to 20 kHz, are ideal to probe into the larynx to study the mechanisms of human voice production. A number of reports of subglottal pressure and supraglottal pressure measurements during phonation were published\(^1\) showing informative temporal correlations with simultaneously acquired EGG signals [21, 61, 82].

Figure 3.15 shows a typical configuration of experimental setup. The pressure sensors are inserted into the larynx through the nasal cavity. One pressure sensor is placed beneath the vocal folds, and another one is placed above the vocal folds. Therefore, the variation of subglottal pressure and the supraglottal pressure can be measured simultaneously with a time resolution of better than 0.1 msec while vocalizing. The voice signals through a microphone and the EGG signals can be acquired simultaneously. To ensure that the measurements are synchronized, a multiple-channel data recording device is utilized.

The most salient and persistent feature of the observed subglottal pressure and supraglottal pressure is the temporal coincidence of maximum sub-

---

\(^1\)In an article by Yasuo Kioke on page 181 of *Vocal Fold Physiology* [87], several reports on sub- and supraglottal pressure variation measurements were cited. Nevertheless, no simultaneous EGG signals were acquired and compared.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
Experimental Facts

glottal pressure and minimum supraglottal pressure with the glottal closing moment identified by EGG signals. Typical results are shown in Fig. 3.16. As shown by lines marked A, A', and A", the maximum subglottal pressure and the minimum supraglottal pressure occur simultaneously at a glottal closure instant. Furthermore, the subglottal pressure at the glottal closure moment is the absolute maximum pressure in the entire pitch period, and the supraglottal pressure at the glottal closing moment is the absolute minimum pressure in the entire pitch period.

On the other hand, around each glottal opening moment as detected by EGG signals, very weak variations are noticed in both subglottal pressure and supraglottal pressure. As shown in Fig. 3.16, before the glottal opening moments, as marked by B and B', the subglottal pressure is at a relatively low value, that is, in the lower range of the pressure scale. After the glottal opening, immediately to the right side of the lines marked as B and B', the subglottal pressure does not change noticeably.

Another remarkable observation is that at each glottal closing moment, the absolute value of the subglottal pressure increase at the glottal closing moment roughly equals the absolute value of the supraglottal pressure decrease. And the rate of pressure change during the glottal closing moment

Fig. 3.15. In-vivo pressure measurement apparatus. Two miniature semiconductor pressure sensors, one supraglottal and one subglottal, are placed using a flexible catheter through the nasal cavity. Pressure reading with time resolution down to 0.1 msec can be acquired simultaneously with voice signal and EGG signal. Using a multiple-channel data recording device, those measurements are well synchronized. After Schutte and Miller [82].

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
3.3 In-Vivo Pressure Measurements

Fig. 3.16. Time-sequence of glottal pressures. The maximum of subglottal pressure and the minimum of supraglottal pressure occur at the glottal closing moment, A, A' and A". Near the moments of glottal opening, marked by B and B', the subglottal pressure is in the lower range of the pressure scale, and does not change much near a glottal opening. The absolute magnitude of the rise of subglottal pressure, about 1 kPa, is approximately equal to that of the drop of supraglottal pressure. After Miller and Schutte [61].

is the greatest within the entire pitch period.

The above observations are confirmed by the measurements of another research group, shown in Fig. 3.17. In that Figure, the glottal closing moments are marked by an asterisk “*”, and the glottal opening moments are marked by a circle “o”. As shown, at the glottal closing moments, the subglottal pressure reaches to an absolute maximum in each pitch period, and the supraglottal pressure reaches to an absolute minimum. At the glottal opening moments, the pressure remains in a moderate value with very little variation around the moment of glottal opening.

The results of a follow-up experiment on five different vowels are shown in Fig. 3.18. For all vowels, the maximum subglottal pressure occurs at glottal closure instants, and the waveforms of subglottal pressure are similar for all vowels. Nevertheless, the waveforms of the supraglottal pressure show noticeable difference for different vowels. That fact can be understood from the nature of the pressure measurement: After the first valley of supraglottal pressure at the glottal closing moment, the timing of the next valley of the supraglottal pressure should be the moment when the acoustic wavefront
Fig. 3.17. Pressure measurements for vowel [a]. The glottal closing moments are marked by an asterisk *, and the glottal opening moments are marked by a circle o. Here, (A) is the EGG signal, (B) is the supraglottal pressure, (C) is the subglottal pressure, and (D) is the speech of airflow immediately above the glottis. As shown, at the glottal closing moments, the subglottal pressure reaches to an absolute maximum in each pitch period, and the supraglottal pressure reaches to an absolute minimum. At the glottal opening moments, the pressure remains in a moderate value with very little variation. After Cranen and Bovis [21].

reflecting from the vocal tract arrives. For vowel [a], the time is shorter, and for vowel [u], the time is longer, see Section 2.3.3. For the waveforms of the subglottal pressure, the time of pressure peak following the first pressure peak at the glottal closing moment is determined by the time the acoustic wave reflecting from the lower end of the trachea arrives. Because the length of the trachea does not change with the identity of the vowel, the waveforms of the subglottal pressure do not depend on the vowel, as expected.

For more than one century, there has been a popular opinion that in each pitch period, a vowel sound is excited by a puff of air at the opening moment of the glottis [23]. When the glottis is closed, the pressure from the lungs and trachea builds up. At a certain point of time, the subglottal air
Fig. 3.18. Pressure variation of several vowels. After the initial surge, the subsequent evolution of subglottal pressure for all vowels are similar. However, after the initial drop, the supraglottal pressure for different vowels diverge significantly, indicating an acoustic effect depending on the instantaneous geometry of the vocal tract. After Schutte and Miller [82].

pressure becomes so high that it forces the vocal folds to separate rapidly, releasing an explosive puff of air. And that puff of air at the glottal opening moment is the source of vowel signals. According to such point of view, immediately before a glottal opening, the subglottal pressure should be the highest. Immediately after a glottal opening, the subglottal pressure should drop quickly to a minimum. None of those phenomena are observed experimentally. This is because that popular opinion is based on an intuition in fluid mechanics. In reality, the processes in the trachea and vocal tract near the vocal folds are governed by the law of acoustics, or more precisely, by the wave equation, rather than by fluid mechanics. In Chapter 4, by resolving the wave equation with proper boundary conditions and proper initial conditions in the vocal tract and in the trachea, a quantitative explanation of the observed pressure variations is provided.
Chapter 4
The Physics of Voice Production

In Chapter 3, the basic experimental facts of human voice are presented. In this Chapter, a theory is formulated to explain the experimental observations as completely as possible. The focus is vowels and voiced consonants, especially to explain voice waveforms with simultaneous EGG signals, videokymography information, and \textit{in vivo} pressure measurements. The production of plosives and fricatives is relatively straightforward, which has been outlined in Section 2.3 in Chapter 2.

4.1 A Brief Summary of Experimental Facts

Following is a brief summary of the experimental facts of vowels:

1. The vowel signals satisfy the superposition principle. A segment of vowel signal can be expressed as a linear superposition of individual elementary waves $F_n$ started at time $t_n$,

   \[ S(t) = F_1(t - t_1) + F_2(t - t_2) + F_3(t - t_3) + \ldots \quad (4.1) \]

2. Each elementary wave starts at a glottal closing moment.
3. The glottal closing is abrupt and forceful.
4. The voice signal is strong in the closed phase.
5. The glottal opening is slow, weak, and sometimes noisy.
6. The voice signal is weak in the open phase.
7. The higher-frequency components of the voice signal (1 kHz to 4 kHz) concentrate at the beginning of the closed phase and decay faster.
8. The lower-frequency components of the voice signal ($<$1 kHz) decay more slowly and dominate the later portion of the elementary wave.
9. High closed quotient corresponds to high voice intensity.
10. Low closed quotient corresponds to low voice intensity.

11. The subglottal pressure rises to a maximum at glottal closings.

12. The supraglottal pressure drops to a minimum at glottal closings.

Obviously, the key to understand the production of vowels is the acoustic process in the vocal tract after a glottal closing. We present a conceptual picture first, then present a rigorous mathematical treatment, followed by explanations of various experimental facts.

4.2 The Concept of Timbrons

The experimental facts can be explained by the following theory:

1. Immediately before a glottal closing, there is a steady and uniform airflow in the vocal tract with a typical velocity of 0.5 m/s to 5 m/s. The airflow varies slowly, the resulting acoustic excitation is weak.

2. A glottal closing abruptly terminates the airflow from the trachea into the vocal tract. Because of inertia, air in the vocal tract continues to flow with the steady velocity.

3. A zero-particle-velocity wavefront is created at the glottis, propagating at the speed of sound into the vocal tract, converting the aerodynamic kinetic energy of the airflow into acoustic energy.

4. The above elementary acoustic wave resonates in the vocal tract and radiates. Its waveform is determined by the geometrical configuration of the vocal tract, thus representing the instantaneous timbre.

5. The glottal closing does not supply energy to the voice. It triggers the conversion of the kinetic energy of the steady airflow immediately before a glottal closing in the vocal tract into acoustic energy.

6. A glottal opening connects the lungs to the resonance cavity of the vocal tract, often accelerating the decay of the acoustic wave.

7. If a glottal closing takes place before the previous elementary acoustic waves disappear, the acoustic wave triggered by the new glottal closing superposes on the tails of the previous elementary acoustic waves.

The waveform of the elementary acoustic wave triggered by a glottal closing is determined by the instantaneous geometrical configuration of the vocal tract.
Fig. 4.1. Acoustic waves triggered by a glottal closure. (0), the glottis is open, there is a uniform and steady airflow in the trachea and the vocal tract. (1) through (3), a glottal closing generates a zero-particle-velocity d’Alembert wavefront propagating towards the lips. (4), the d’Alembert wavefront reflects at the lips with a reversed air-particle velocity. (4) through (6), the reversed d’Alembert wavefront propagates towards the glottis. (7), the d’Alembert wavefront reflects at the glottis to generate a d’Alembert wavefront of compressed air. (7) through (9), the d’Alembert wavefront of compressed air propagates towards the lips. (10), the d’Alembert wavefront reflects at the lips to generate a d’Alembert wavefront with particle velocity pointing to the open air. (10) through (12), the d’Alembert wavefront propagates towards the glottis to recover the initial state similar to (1).

The vocal tract thus it represents the timbre at that moment. The above elementary wave is reasonably to be termed as a *timbron*. Timbrons are elementary building blocks of voice, in a similar sense as atoms are elementary building blocks of matter. For convenience, the theory presented here is termed *timbron theory*, in a similar sense to that of the atomic theory of matter.

To illustrate the above concept step by step, a single-tube model of the vocal tract shown in Fig. 4.1 is analyzed. The medium gray background indicates the unperturbed air density $\rho_0$. The light gray area represents rarefied air particles. The dark gray area represents compressed air particles. The small arrows indicate the particle velocity, initially $u_0$. The large arrow indicates the velocity of sound. Steps (0) through (12) are marked.

Let us trace the process step by step. Before a glottal closing, step (0), a steady airflow with velocity $u_0$ runs from the trachea through the glottis and the vocal tract into open air. The density of air over the entire length is uniform. Immediately after a glottal closing, (1), the velocity of air at the glottis is abruptly forced to zero. Because the air particles keep moving up with velocity $u_0$, a zero-particle-velocity wavefront is created and propagates into the vocal tract. At time $t$, the wavefront is at $x = ct$. Due
to particle velocity $u_0$, the particle initially at $x = ct$ is moved to

$$x' = (c + u_0)t.$$  (4.2)

The air column left behind by the wavefront is expanded by a factor of $1 + u_0/c$. Because no additional air mass could come through the glottis, conservation of mass requires that the air column is rarefied,

$$\rho_0 \rightarrow \frac{\rho_0 x}{x'} = \frac{ct \rho_0}{(c + u_0)t} \approx \rho_0 - \frac{u_0}{c}.$$  (4.3)

The perturbation density of the air column is negative,

$$\rho = -\rho_0 \frac{u_0}{c}.$$  (4.4)

The relation between perturbation pressure and perturbation density, Eq. 1.14, predicts a negative perturbation pressure,

$$p = -\gamma \rho_0 \frac{u_0}{c}.$$  (4.5)

A typical particle velocity is $u_0 = 1.5$ m/s. The perturbation pressure is

$$p = -1.4 \times 100 \text{ kPa} \times \frac{1.5}{352} \approx -0.6 \text{ kPa},$$  (4.6)

which explains the drop of supraglottal pressure at a glottal closure.

At step (3), with time $t = L/c$, the wavefront reaches the opening of the vocal tract. Because of the negative perturbation pressure, according to Eq. 1.43, the acoustic potential energy density in the vocal tract is

$$\mathcal{E} = \frac{1}{2\gamma \rho_0} p^2 = \frac{\gamma \rho_0 u_0^2}{2 c^2}.$$  (4.7)

On the other hand, according to Eq. 1.20, the velocity of sound is

$$c^2 = \frac{\gamma \rho_0}{\rho_0},$$  (4.8)

and then Eq. 4.7 can be converted into

$$\mathcal{E} = \frac{1}{2} \rho_0 u_0^2,$$  (4.9)

which equals the kinetic energy density of airflow originally in the vocal tract. Therefore, after the first phase of wavefront propagation, the aerodynamic kinetic energy of the airflow in the vocal tract immediately before a
4.2 The Concept of Timbrous

Glottal closing is converted into acoustic potential energy. After the waveform moves over the vocal tract into open air, at step (3), the particle velocity becomes zero. At that time, the air pressure in the open air is $p_0$, but the air pressure in the vocal tract is lower by an amount $p$, as shown in Eq. 4.6. Driven by that pressure difference, air particles flow from the open air into the vocal tract, see step (4). A waveform towards the glottis is created, propagating with the velocity of sound, with a particle velocity pointing to the glottis, see step (5). After a time $t = L/c$, the waveform reaches the glottis, see step (6). Because the glottis is now closed, the particle velocity must be zero, the acoustic waveform reflects back, see step (7), then propagates into the $+x$ direction, see step (8). This time, the particle velocity and the velocity of sound wave have the opposite direction. The air particle originally at position $x = ct$ moves to

$$x' = (c - u_0)t.$$  (4.10)

The volume of air mass left behind by the waveform is reduced by a factor of $1 - u_0/c$. Therefore, the air density is changed to

$$\rho_0 \rightarrow \frac{\rho_0 x}{x'} = \frac{ct \rho_0}{(c - u_0)t} \approx \rho_0 + \rho_0 \frac{u_0}{c}.$$  (4.11)

The perturbation density of the air column is positive,

$$\rho = \rho_0 \frac{u_0}{c}.$$  (4.12)

The air column is compressed, and the perturbation pressure is positive,

$$p = \gamma p_0 \frac{u_0}{c}.$$  (4.13)

Nevertheless, neglecting radiation loss, the acoustic potential energy density is identical to the case of rarefied air particle, Eq. 4.7,

$$\mathcal{E} = \frac{\gamma p_0 u_0^2}{2 c^2}.$$  (4.14)

After the waveform reaches the opening of the vocal tract, see step (9), the particle velocity becomes zero. Now, the air pressure inside the vocal tract is higher than the pressure of open air by an amount $p$. Air particles flow from the vocal tract into open air to relieve the excess pressure. A waveform propagating towards the glottis is created, see steps (10) and (11). After a time $t = L/c$, the waveform reaches the glottis, see step (12). A condition similar to step (1) is recreated. The cycle then repeats.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
The entire cycle takes four phases, each one has a time duration of $L/c$. The entire cycle has a period of

$$T = \frac{4L}{c}. \quad (4.15)$$

When the wavefront reaches the open end of the vocal tract, part of the acoustic energy radiates into open air. Therefore, the acoustic wave decays with time. It creates a situation similar to the Euler transient resonator presented in Section 1.3. The waveform of the elementary wave triggered by a glottal closing is determined by the configuration of the vocal tract, representing the timbre at that moment, thus termed a timbron. Timbrons are elementary building blocks of human voice. Furthermore, an isolated glottal closing would trigger an isolated timbron, see Section 3.2.2. Triggered by a more or less equally timed array of glottal closings, an array of timbrons makes a vowel in the modal voice register, see Section 3.1.3. Triggered by an array of slow and somewhat irregularly timed glottal closings, an array of timbrons makes a vowel in the vocal-fry register.

### 4.3 Acoustic Waves in the Trachea

A glottal closing triggers a timbron in the vocal tract. It also triggers an acoustic wave in the trachea. Although the acoustic wave in the trachea is hardly perceptible by humans, it is detectable by miniature pressure sensors placed inside the trachea.

Figure 4.2 shows the acoustic process in the trachea after a glottal closing. As shown, step (0) is the state before a glottal closing, where a steady airflow runs from the trachea through the glottis into the vocal tract. After a glottal closing, see step (1), the air particle velocity at the glottis becomes zero. A zero-velocity wavefront is created and propagates downwards into the trachea. At time $t$, the wavefront is at $x = -ct$. Due to the particle velocity $u_0$, the particle initially at $x = -ct$ is moved to

$$x' = -ct + u_0 t. \quad (4.16)$$

The volume of air column left behind by the wavefront is reduced by a factor of $1 - u_0/c$. The air density is changed to

$$\rho_0 \rightarrow \frac{\rho_0 x}{x'} = \frac{-ct \rho_0}{-(c - u_0) t} \approx \rho_0 + \rho_0 \frac{u_0}{c}. \quad (4.17)$$

The perturbation density of air column left by the wavefront is positive,

$$\rho = \rho_0 \frac{u_0}{c}. \quad (4.18)$$
4.4 An Analytic Solution of the Wave Equation

In the previous sections, using intuitive arguments, we studied the acoustic waves in the vocal tract and the trachea after a glottal closing. In this section, we study the evolution of the acoustic wave after a glottal closing using the wave equation,
\[ \frac{\partial^2 u(x,t)}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u(x,t)}{\partial t^2} \]  

The perturbation pressure is also positive, with a magnitude equal to the drop of supraglottal pressure, Eq. 4.5,
\[ p = \gamma p_0 \frac{u_0}{c}, \]  
which explains the surge of subglottal pressure at a glottal closing.

The bronchi and lungs have more open space than the trachea, which create a situation similar to the open end of the vocal tract. The rest of the evolution process is almost identical to the process in the vocal tract. We leave it to the readers to complete the thinking.

4.4 An Analytic Solution of the Wave Equation

Fig. 4.2. Acoustic waves in the trachea. (0), the glottis is open, there is a uniform and steady airflow in the trachea and the vocal tract. (1) through (3), a glottal closing generates a zero-velocity d’Alembert wave front of compressed air propagating towards the bronchi. (4), the d’Alembert wavefront reflects by the bronchi, and the perturbation velocity of the air particles is reversed. (4) through (6), the reversed d’Alembert wavefront propagates towards the glottis. (7), the d’Alembert wavefront reflects at the glottis to generate a d’Alembert wavefront of rarefied air. (7) through (9), the d’Alembert wavefront of rarefied air propagates towards the bronchi. (10), the d’Alembert wavefront reflects at the bronchi to generate a d’Alembert wavefront with perturbation velocity pointing downwards. (10) through (12), the d’Alembert wavefront propagated towards the glottis to recover the initial state similar to (1).
where \( u(x, t) \) is the particle velocity, and \( c \) is the velocity of sound, see Fig. 4.3. Closed-form analytic solutions are obtained. That solution is derived with mathematical rigor; thus it is more convincing.

4.4.1 Initial conditions and boundary conditions

Naturally, we set the origin of time \( t = 0 \) at a glottal closing moment. Before the glottal closing, there is a steady airflow in the vocal tract with velocity \( u_0 \). Therefore, at \( t = 0 \), the initial conditions are

\[
\begin{align*}
  u &= u_0 \quad \text{at} \quad 0 < x < L, \quad t = 0; \\
\end{align*}
\]

(4.21)

and because the velocity is a constant,

\[
\frac{\partial u}{\partial t} = 0 \quad \text{at} \quad 0 < x < L, \quad t = 0.
\]

(4.22)

At \( t > 0 \), the glottis is closed. The particle speed at \( x = 0 \) is zero,

\[
  u = 0 \quad \text{at} \quad x = 0.
\]

(4.23)

Beyond the lips, \( x > L \), there is open air. The density is a constant,

\[
  \rho = 0 \quad \text{at} \quad x > L.
\]

(4.24)

Using the continuity equation 1.6, the boundary condition at \( x = L \) is,

\[
\frac{\partial u}{\partial x} = 0 \quad \text{at} \quad x = L.
\]

(4.25)
4.4 An Analytic Solution of the Wave Equation

The meaning of the boundary condition Eq. 4.25 is as follows: the particle density is stable in the open air; thus the particle speed is also stable.

4.4.2 Acoustic waves in the vocal tract

In this subsection, we solve the wave equation in the vocal tract after a glottal closing using Fourier series. Using the boundary conditions, Eqs. 4.24 and 4.25, we can write down the most general mathematical form of the solution using Fourier series,

$$u = \sum_{n=1}^{\infty} g_n(t) \sin \left( \frac{(2n-1)\pi x}{2L} \right).$$

(4.26)

The correctness of the expression can be verified by direct inspection: Because each term of the sine function satisfies the boundary conditions, Eqs. 4.23 and 4.25, the sum should also satisfy such conditions. And the theory of Fourier series guarantees that the solution is the most general one. The functions of time, $g_n(t)$, are to be determined by the wave equation, Eq. 4.20, and the initial conditions, Eqs. 4.21 and 4.22.

Inserting Eq. 4.26 into the wave equation, Eq. 4.20, we find the differential equation for the functions $g_n(t)$ in Eq. 4.26,

$$\frac{d^2 g_n(t)}{dt^2} = - \left[ \frac{(2n-1)\pi c}{2L} \right]^2 g_n(t).$$

(4.27)

The general solution is

$$g_n(t) = a_n \cos \left( \frac{(2n-1)\pi ct}{2L} \right) + b_n \sin \left( \frac{(2n-1)\pi ct}{2L} \right),$$

(4.28)

where the constants $a_n$ and $b_n$ are to be determined by the initial conditions, Eqs. 4.21 and 4.22. Because of initial condition Eq. 4.22, $b_n = 0$. At $t = 0$, from Eqs. 4.21, 4.26 and 4.28,

$$\sum_{n=1}^{\infty} a_n \sin \left( \frac{(2n-1)\pi x}{2L} \right) = u_0, \quad 0 < x < L.$$  

(4.29)

The Fourier coefficients $a_n$ can be easily evaluated:

$$a_n = \frac{4u_0}{(2n-1)\pi}.$$  

(4.30)

Put the above results together, the solution is

$$u = \frac{4u_0}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \cos \frac{(2n-1)\pi ct}{2L} \sin \frac{(2n-1)\pi x}{2L}.$$  

(4.31)
Using trigonometric identity
\[ \sin a \cos b = \frac{1}{2} [\sin(a + b) + \sin(a - b)] \] (4.32)
and the identity [2]
\[ \sum_{n=1}^{\infty} \frac{1}{2^{n-1}} \sin \left( \frac{(2n-1)\pi x}{2L} \right) = \frac{\pi}{4} \text{sgn} \sin \frac{\pi x}{2L}, \] (4.33)
a closed-form solution is found,
\[ u = u_0 \left[ \text{sgn} \sin \frac{\pi(x + ct)}{2L} + \text{sgn} \sin \frac{\pi(x - ct)}{2L} \right]. \] (4.34)

The sign function \( \text{sgn}(x) \) in Eqs. 4.33 and 4.34 is defined as: for \( x > 0 \), \( \text{sgn}(x) = 1 \); and \( x < 0 \), \( \text{sgn}(x) = -1 \). Therefore, the combined function \( \text{sgn} \sin(x) \) is \( +1 \) for \( (2n-1)\pi < x < 2n\pi \), and \( -1 \) for \( 2n\pi < x < (2n+1)\pi \). Equation 4.34 represents a piecewise d’Alembert solution with reflections at the glottis and lips.

### 4.4.3 Analysis of the solution

Although the solution exists for all real values of \( x \) and \( t \), only the values of \( u \) inside the vocal tract, \( 0 < x < L \), are meaningful. We will use period \( T \) in Eq. 4.15 as a scale of time. During time interval \( 0 < t < T/4 \), the first term in the square bracket of Eq. 4.34 is always 1, and the second term is 1 for \( x > ct \), and \( -1 \) for \( x < ct \). In other words, \( u = u_0 \) for \( x > ct \), and \( u = 0 \) for \( x < ct \). Therefore, Eq. 4.34 describes a d’Alembert wavefront of zero particle velocity moving from the glottis towards the open end of the vocal tract at the velocity of sound \( c \). Because of the equation of continuity, Eq. 1.6, behind the zero-particle-velocity d’Alembert wavefront, air is stagnant but rarefied. In fact, by integrating the equation of continuity across the wavefront, the left-hand side is
\[ \int_{x = ct - \epsilon}^{x = ct + \epsilon} \frac{\partial p}{\partial t} dx = \int_{x = ct - \epsilon}^{x = ct + \epsilon} \frac{\partial p}{\partial t} \frac{dx}{dt} dt = c \rho, \] (4.35)
here \( \epsilon \) is a small positive constant, and the right-hand side is
\[ -\rho_0 \int_{x = ct - \epsilon}^{x = ct + \epsilon} \frac{\partial u}{\partial x} dx = -\rho_0 u_0. \] (4.36)
Therefore, the perturbation density is
\[ \rho = \frac{u_0}{c} \rho_0. \] (4.37)
The air behind the wavefront is rarefied by a factor \( u_0/c \). During the time interval \( T/4 < t < T/2 \), the first term in the square bracket of Eq. 4.34 is \( -1 \) for \( x > c(t - T/4) \), and \( +1 \) for \( x < c(t - T/4) \), and the second term is always \( +1 \). Therefore, Eq. 4.34 describes a zero-particle-velocity d’Alembert wavefront of \( u = u_0 \) moving from the open end of the vocal tract toward the glottis at the velocity of sound \( c \).

During the time interval \( T/2 < t < 3T/4 \), the first term in the square bracket of Eq. 4.34 is always \( -1 \), and the second term is \( +1 \) for \( x > c(t - T/2) \), and \( 1 \) for \( x < c(t - T/2) \). Therefore, Eq. 4.34 describes a zero-particle-velocity d’Alembert wavefront of \( u = -u_0 \) moving from the glottis toward the open end of the vocal tract at the velocity of sound \( c \). Using a similar argument in Eqs. 4.35 through 4.37, the air behind the wavefront is compressed by a factor of \( u_0/c \).

During the time interval \( 3T/4 < t < T \), the first term in the square bracket of Eq. 4.34 is always \( +1 \), and the second term is \( -1 \) for \( x > c(t - T/4) \), and \( +1 \) for \( x < c(t - T/4) \). Therefore, Eq. 4.34 describes a zero-particle-velocity d’Alembert wavefront of \( u = -u_0 \) moving from the open end of the vocal tract toward the glottis at the velocity of sound \( c \).

A complete cycle takes time \( T = 4L/c \), and the inverse

\[
F_1 = \frac{1}{T} = \frac{c}{4L}
\]

(4.38)

is the first formant frequency. From Eq. 4.31, there is a series of formant frequencies \( F_2 = 3F_1, F_3 = 5F_1, \ldots \), with intensities proportional to \( F_n^{-2} \), falling off at a rate of \( -6 \text{ dB per octave} \). This fits well to the experimental observation of the vowel shwa [ə].

### 4.4.4 Numerical solutions for various vowels

The single-tube model of the vocal tract enables an analytic solution to make the process transparent. It approximates the production of the vowel shwa [ə]. It is apparent that by using a variable-cross-section model of the vocal tract, a numerical solution of the time-dependent wave equation should imitate the production of various vowels.

### 4.5 Explanations of Experimental Facts

In this Section, we will show how the experimental facts are explained by the timbronic theory of voice production.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
4.5.1 Superposition behavior of vowel signals

According to the timbros theory, the vowel signal is a superposition of a series of elementary waves, each starting with a stimulus, and decaying during each single pitch period. For vowels, the principle stimulus to start an elementary wave is the closing event of the glottis. After a glottal closing, the bottom of the vocal tract is closed, while the mouth and/or nose are open. The kinetic energy of the air in the vocal tract is converted into acoustic energy with a well-defined resonance characteristics. The subsequent opening of the glottis has a much smaller effect on the acoustic wave than the glottal closing event. It can bring a weak random noise to the acoustic signal, or accelerate damping by opening the vocal tract to the lungs.

4.5.2 Efficiency of voice production

For a long time, the efficiency of voice production was an unresolved puzzle. The power of breathing can be estimated as follows. From actual measurements, for example Figs. 3.16 and 3.17, the cross-glottal pressure is typically 2 kPa. On the other hand, the measurements of airflow during speaking is well documented, see Baken et al. [5]. For male speakers, the average rate of airflow is 150 cm$^3$/s.\(^1\) The average power is

$$P = 2 \times 10^3 \times 150 \times 10^{-6} \approx 0.3 \text{ W.}$$

(4.39)

However, the average power of human voice is only 100 $\mu$W. The efficiency is about 3$\times$10$^{-5}$. The efficiency gap is tremendous. The timbros theory provides insights into the problem. According to the timbros theory, only the kinetic energy of airflow inside the vocal tract immediately before a glottal closure can convert into acoustic energy. The air flow velocity in the vocal tract immediately before a glottal closing is about 1 m/s. The cross section of the pharynx is about 5 cm$^2$, and the average length of the pharynx is 89 mm, see Stevens [86]. Immediately before a glottal closure, the kinetic energy of airflow in the vocal tract is

$$E = \frac{1}{2} \times 1.2 \times 1^2 \times 5 \times 8.9 \times 10^{-6} \approx 26 \mu\text{J.}$$

(4.40)

At a pitch frequency of 100 Hz, the power of the kinetic energy inside the vocal tract is about 2.6 mW, and the efficiency of voice production from that power is about 4%. It is a reasonable number.

\(^1\)Averaged from the measured values of mean airflow for all male speakers between age 19 to 65, Table 9-10, page 363, Baken and Orlikoff [5].
4.5.3 Role of the closed quotient

It is observed experimentally that high closed quotient corresponds to high voice intensity, and high open quotient corresponds to low voice intensity [7, 62]. The timbron theory provides a satisfactory explanation. When the glottis is closed, the vocal tract becomes a tight resonance cavity. The vocal folds are squeezed together. Because the finite thickness, moving air can hardly affect the shape of the vocal folds. Nevertheless, once the glottis opens during a resonance process, the lungs are connected to the resonance cavity. Because the lungs are porous, power decay is accelerated. The longer the open phase, the more the voice signal decays.

On the other hand, the voice intensity is not determined by the average flow rate over the entire time of a period, but the instantaneous velocity of airflow immediately before a glottal closure. If the average airflow rate is a constant, then the shorter the open phase, the higher the instantaneous air velocity immediately before a glottal closing. Because the air velocity rises slowly near the glottis opening moment, then reaches a high value before a glottal closure [4], one can describe the air velocity by a simple analytic expression such as

$$u = u_0 \sin^2 \frac{\pi (t - T_C)}{T_O},$$

where the origin of time is the glottal closing moment, $T_C$ is the duration of the closed phase, and $T_O$ is the duration of the open phase. The sum is

![Image](image_url)

**Fig. 4.4. Open quotient of an opera singer.** Using videokymography, the opening and closing of an opera singer during singing are recorded. The data is more certain than using electroglottography. From the image, the closed quotient is much greater than 0.5. It is associated with a intensive and bright voice. Courtesy of Donald Miller.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
the pitch period $T$,
\[ T_C + T_O = T. \] (4.42)
the relation between average glottal airflow $V$ is
\[ u_0 = \frac{2T}{T_C} \frac{V}{S} = \frac{2}{OQ} \frac{V}{S} \] (4.43)
where $V$ is the average flow rate, and $S$ is the cross section of the pharynx. For the male speaker discussed in Section 2.2.10, for soft phonation, $OQ=0.65$. Take the average value of flow rate, 150 cm$^3$/s, and $S = 5$ cm$^2$,
\[ u_0 = \frac{2}{0.65} \frac{150}{5} \approx 90 \text{ cm/s} = 0.9 \text{ m/s}. \] (4.44)
For loud phonation, $OQ = 0.47$. The instantaneous air velocity could become 1.3 m/s. The voice power, which is proportional to the square of the air-glow velocity, is doubled. For opera singers, the open quotient can be even smaller, and the voice is even louder, see Fig. 4.4.

4.5.4 Supraglottal pressure and subglottal pressure
The timbron theory provides a clear and quantitative explanation of the observed drop of supraglottal pressure and the surge of subglottal pressure at the glottal closing moment as well as the pressure variation during the entire pitch period. At a glottal closing moment, step (1) of Fig. 4.2, the supraglottal pressure drops. If the air velocity immediately before a glottal closure is 1.5 m/s, according to Eq. 4.6, the pressure changes by
\[ p = -1.4 \times 100 \text{ kPa} \times \frac{1.5}{352} \approx -0.6 \text{ kPa}, \] (4.45)
which explains the experimentally observed pressure drop.

After one half of the period of the lowest formant, see step (7) in Fig. 4.2, the supraglottal pressure should become positive,
\[ p = 1.4 \times 100 \text{ kPa} \times \frac{1.5}{352} \approx 0.6 \text{ kPa}. \] (4.46)
The duration of the half formant period depends on the vowel. For vowels [i] and [u], the first formant is about 300 Hz, similar to the pitch period. The maximum of supraglottal pressure takes place in the middle of a pitch period, as expected. For vowels [e] and [o], the surge of supraglottal pressure comes earlier. For vowel [a], the surge comes even earlier, as expected.

The subglottal pressure undergoes a similar cycle. As shown in Fig. 4.3, at the glottal closing moment, the subglottal pressure should surge with a identical magnitude as in Eq. 4.46. Nevertheless, the length of the trachea does not change with vowels, thus the pattern of subglottal pressure variation does not change with vowels.
4.5.5 Radiation and decay of formants

According to the detailed theoretical treatment of the radiation of acoustic waves by Phillip Morse [63], for a tube with an opening of radius \( a \), the radiation power is proportional to the square of the ratio of the radius of the opening and the wavelength \( \lambda \) of the acoustic wave,

\[
\frac{z_p}{z} \sim \frac{1}{2} \left( \frac{2\pi a}{\lambda} \right)^2.
\] (4.47)

Therefore, if at the beginning, the resonance wave in the vocal tract has a number of components with different frequencies, the high-frequency components will radiate faster than the low-frequency components. At the beginning of a timbron, the high-frequency components are strong but they decay faster. The low-frequency components last longer.

The different decay rates of formants at different frequencies provide an adequate explanation of the observed waveforms of vowels. For vowel [i] in Fig. 3.7 and vowel [e] in Fig. 3.9, there are two groups of formants. The formant groups with frequency around 3 kHz decay at a rate of -7 dB/msec. It is also apparent in the waveforms that the high-frequency signals are strongest immediately after the glottal closing, and almost disappear after a few milliseconds. The low-frequency formants, 330 Hz for [i] and 540 Hz for vowel [e], decays much more slowly, at about -2 dB per msec. This is clear in both the power curve in dB and the waveforms. The vowels [u] and [o], in Figs 3.8 and 3.10, are each dominated by one low-frequency formant, which are 400 Hz for [u], and 620 Hz for [o], decaying at -2 dB per msec. The formants of vowel [a] in Fig. 3.6 have frequencies around 1 kHz. The decay rate is predictably in between, -5 dB/msec.

4.6 The Harp Analogy

It is instructive to compare vowel production and singing with a string instrument. Based on the similarity of sound, violin seems to be an obvious choice. Intuitively, the vocal folds, previously called vocal cords to hint at a similarity to violin strings, are the source of sound. The vocal tract is analogous to the top plate of the violin, which selectively amplifies the sound generated by the vocal folds. The mouth resembles the f-holes. Nevertheless, there is a paradox of size. The fundamental frequency of the violin sound is determined by the length of the string. Even the fundamental frequency of a female voice is often lower than G3, the lowest note of a violin. For a male voice, a cello or a double bass is needed to generate the fundamental frequency, requiring a string length comparable to the entire
human body. However, the size of vocal folds is only about one centimeter, which is too small to function as the source of voice.

According to the timbronic theory presented in Section 4.2, vowel production resembles a harp, rather than a violin. When a harpist pulls a string with a fingertip, no sound is produced. After the string is abruptly released by the fingertip, the elastic potential energy loaded on the string is converted into acoustic energy. This is similar to how an elementary acoustic wave of vowel sound is produced: During the open phase of the glottis, air flows into the vocal tract, but no sound is produced. An abrupt glottal closing starts the conversion of the kinetic energy of airflow in the vocal tract into acoustic energy. The vocal folds are not the source of sound. They constitute a valve to control the airflow. The kinetic energy of flowing air loaded in the vocal tract is the source of acoustic energy, similar to the elastic potential energy loaded by the harpist’s fingertip on a string. Incidentally, the size of fingertips is approximately the size of vocal folds.

Individual notes are often played on a harp. Humans also make sounds with individual glottal closures, such as events in the vocal-fry register or glottal stops. More often, humans make repeated glottal closures with approximately equal time intervals to produce sustained vowels. The analogy in a harp is tremolo, where the repetition frequency of plucking is equivalent to the fundamental frequency of the vowel. There is a subtle difference between human voice organs and the harp. On a harp, tremolo cannot be played by simply plucking a single string repeatedly. A fingertip placed on a vibrating string damps it immediately and completely. To play a convincing tremolo, a harpist must first set the pitches of two adjacent strings to the same target pitch using pedals. By plucking the two adjacent strings alternatively, same as playing a trill, a note on one string starts before a ringing note on another string disappears. A continuous tremolo is therefore produced. In human voice organs, a glottal opening during the vibration of air in the vocal tract only causes some additional damping, but the resonance wave continues into the next pitch period.

Another subtle difference of the harp and human voice is that each string on a harp can only produce a sound of one fixed frequency. In human voice production, each configuration of the vocal tract can produce a number of frequencies, or formants. This difference is similar to the difference of the original Willis mechanism and the refined Willis mechanism proposed by Peter Ladefoged [54], as mentioned in the Preface.
Bibliography


synchronous spectral parameters. *U.S. Patent 8,942,977*, Jan. 27,
2015.

Statistical acoustic processing method and apparatus for speech recog-
nition using a toned phoneme system. *U.S. Patent 5,751,905*, May
12, 1998.

methods in continuous mandarin speech recognition. *Eurospeech’97*,

for recognizing tone languages using pitch information. *U.S. Patent

ing pitch information on the main vowel of each syllable. *ICASSP’01*,

continuous text using statistical approaches. *U.S. Patent 5,806,021*,
Sep. 8, 1998.


[21] B. Cranen and L. Boves. Pressure measurements during speech pro-
duction using semiconductor miniature pressure transducers: Impact
1985.


[23] R. Daniloff, G. Schuckers, and L. Feth. *The Physiology of Speech and

sentations for monosyllabic word recognition in continuously spoken

2003.

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.


Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.


Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.


Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.


Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.


Index

acoustic energy density, 15–17
amplitude spectrum, 24, 100–102, 117–120, 145, 147, 153, 154
consonant [l], 135
consonant [n], 135
fricative [θ], 134
fricative [ʃ], 134
fricative [f], 134
fricative [s], 56, 134
plosive [k], 124
vowel [a], 59, 65, 73, 120, 150
vowel [ɔ], 133, 150
vowel [e], 67, 133, 150
vowel [i], 65, 122, 133, 150
vowel [o], 67
vowel [u], 58, 65, 122, 150
ARCTIC databases, xv, 36–49, 64, 70, 106, 110, 115, 132, 146, 149, 154, 155, 164, 175, 189
Bernoulli force, 37–44, 55, 70
causality condition, 98, 99, 118, 124, 139, 199
consonant, 4, 29, 52, 56, 57, 63, 73, 79, 112, 114, 116, 117, 120, 124, 133, 137, 156, 165, 177, 180, 185, 188
voiced, 79, 112, 117, 136, 156, 177
d’Alembert solution, 11–12
d’Alembert wavefront, 17–21, 80–89, 147–148, 150
dispersion relations, see Kramers-Kronig relations
efficiency of voice production, 90

EGG, see electroglottography
electroglottography, xii, xv, 3, 34–37, 46, 48, 69, 70, 72, 73
Euler, vii, viii, 4, 5, 13, 57
formula, 21, 23
transient resonator, 12–15, 52, 53, 55, 56, 58
fast Fourier transform, 117, 126, 161, 172, 173, 175, 184
feature vector, 131, 139, 157, 160, 161, 165, 181–183, 185
FFT, see fast Fourier transform
formant, ix, x, 5, 43, 58, 65, 67, 72, 92, 93, 120, 124, 129, 132, 139
synthesizer, 181, 182
Fourier transform, 23–24, 56, 100, 101, 139, 200
frequency, x, 5, 13, 24, 35, 47, 49, 56, 58, 62, 112, 122, 141, 146, 153, 200
and decay rate, 93, 120
formant, ix, 67, 122, 182
fundamental, 22, 94, 117
human perception, 128
pitch, x, xii, 65, 90, 119, 197
positive and negative, 142
ranges, 26, 142, 156, 161, 173
resolution, 101, 125, 175
resonance, 30, 55, 57, 58
response, 63, 73
sampling, 118
scale, 124, 127, 128, 174
vibration, viii, 52
fricative, 52, 55–57, 79, 124, 132, 135, 150, 153, 154, 157, 172, 176, 177, 181
Gegenbauer polynomials, 191, 192
glottal closing, xii, xiii, 3, 6, 19, 48, 51, 70, 71, 75, 79, 86, 90, 92, 98, 105, 108, 151
glottal opening, xi, 44, 70, 72, 73, 77, 80
glottal stops, 72
glottis, xi, 4, 19, 32, 35, 36, 39–41, 44, 50, 58, 70, 72, 77, 80–82, 85, 87–90, 94, 105, 147
human voice, vii, xi, 3, 7, 16, 26, 83
intensity scale, 27
power, 90
ranges, 26
incomplete glottal closure, 46–49, 106–107
intensity, xi, 5, 14, 26, 27, 51, 69, 80, 91, 103, 122, 131, 139, 157, 164, 168, 172–174, 181, 185, 188, 197
and closed quotient, 51
distribution within a pitch period, 103
scale of human voice, 27
Kramers-Kronig relations, 99, 148, 151, 152, 161, 173, 184, 199–200
Laguerre function, 97, 128–130, 132, 133, 164, 172, 179, 183, 201–202
laryngeal electromyography, 33–34
linear predictive coding, xiv, 97, 107, 125, 127, 132, 159, 166, 168, 170
LPC, see linear predictive coding
mel-frequency cepstral coefficients, xiv, 97, 126–128, 132, 136, 159
MFCC, see mel-frequency cepstral coefficients
microphone, xii, 3, 61–197
MIDI, see musical instrument digital interface
music scale, 25
musical instrument digital interface, 25, 164, 197
PCM, see pulse-code modulation
phase spectrum, 24, 56, 98, 100, 118, 128, 139, 141, 144, 145, 147, 149, 153, 154, 157, 161, 173, 180, 184
fricative s, 57
from voice signals, 145
recovery, 139–144
phonograph, vii, ix, 5
pitch, viii–x, xii, 5, 13
pitch scale, 25–27
pitch-synchronous, 105, 108, 118, 124
pitch-synchronous overlap add, xv, 102, 159, 165, 182
plosive, 52–55, 57, 72, 79, 99, 114, 124, 165, 177, 180
Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.
Index

power, 15–17, 65, 66, 90, 92, 112, 156, 174, 189
vowel [o], 133
vowel [e], 133
vowel [i], 133
vowel [a], 150
vowel [o], 150
vowel [e], 150
vowel [i], 150
vowel [u], 150
timbre, viii, xi, 6, 30, 52, 57, 81, 84, 104, 120, 122, 125
timbre vector, 98, 125–137, 182, 184, 187, 188
consonant [l], 135
consonant [n], 135
consonant [θ], 134
consonant [s], 134
fricative [θ], 134
fricative [s], 134
fricative [f], 134
fricative [s], 134

pressure
perturbation, 8–9, 12, 18, 21
subglottis, 19, 73–77, 80, 82
supraglottis, 19, 73–77, 80, 82
profile function, 107–108, 115, 161, 171, 179
PSOLA, see pitch-synchronous overlap add
timbre, 4, 6, 80–89, 93, 118, 151, 161, 173, 182, 199
timbre extraction, 99–124
timbre recovery, 149–156
timbre theory, 80–89, 92–93

radiation, 83, 90
and decay of formants, 93
Rayleigh, Lord, 5, 7

velocit
air-flow, 91
of sound, 10, 11, 13, 20, 81, 83, 88
particle, 8–9, 11, 15, 17, 19, 64, 81, 83, 88

velocity

velocity
acoustic, 7, 29

vocal folds, viii, ix, xi, 3, 30–31, 35, 37, 44, 47, 50, 52, 91, 94, 112
vocal tract, viii, 3, 19, 29, 51–59, 77, 80–84, 86–89, 93, 134, 147, 182

voicelessness index, 131, 139, 161, 183
fractional, 132, 150, 153–157

vowel, viii, 29, 36, 52, 57–59, 65–69, 75, 77, 80, 84, 89, 92, 94, 101, 116, 120, 125, 132, 133, 136, 142, 149, 153, 155, 177, 180, 185, 187, 189

water-hammer analogy, 44–46

velocity

velocity
acoustic, 7, 29

vowel, viii, 29, 36, 52, 57–59, 65–69, 75, 77, 80, 84, 89, 92, 94, 101, 116, 120, 125, 132, 133, 136, 142, 149, 153, 155, 177, 180, 185, 187, 189

water-hammer analogy, 44–46

velocity

velocity
acoustic, 7, 29

vowel, viii, 29, 36, 52, 57–59, 65–69, 75, 77, 80, 84, 89, 92, 94, 101, 116, 120, 125, 132, 133, 136, 142, 149, 153, 155, 177, 180, 185, 187, 189

water-hammer analogy, 44–46

velocity

velocity
acoustic, 7, 29
elementry, 71, 79, 80, 89
mucosal, 50
power, 15
resonance, 55
sinusoidal, 24, 52
wave equation, xiii, 3, 7–10, 20, 68, 77, 85, 87–89
waveform, 52, 58, 59, 65, 67, 68, 73, 80, 89, 93, 98, 103, 129, 132, 148, 151, 169, 173, 185
and EGG signals, 71
glottal stop, 55, 72
plosive [k], 52
plosive [ku], 53
stop, 52, 53, 55, 114
vowel [a], 58, 65
vowel [e], 67
vowel [i], 65
vowel [o], 67
vowel [u], 58
waveform recovery, 138
wavefront, 17, 21, 80, 82–84
Willis, 4, 5, 52, 57, 94