1 • Introduction

The singing voice is the oldest musical instrument, but its versatility and emotional power are unmatched. Through the combination of music, lyrics, and expression, the voice is able to affect us in ways that no other instrument can. The fact that vocal music is prevalent in almost all cultures is indicative of its innate appeal to the human aesthetic. Singing also permeates most genres of music, attesting to the wide range of sounds the human voice is capable of producing. As listeners we are naturally drawn to the sound of the human voice, and, when present, it immediately becomes the focus of our attention.

To meet the demands of vocal communication, the vocal apparatus is extremely flexible, yet also remarkably self-consistent. An endless variety of sounds can be produced by a fairly simple physical system of vibration and resonance. Describing the distinctive character of a voice, however, is difficult without resorting to vague and subjective terms (e.g., “squeaky”) that have no objective correlates. These qualities are believed to be a combination of innate physical characteristics and learned patterns of expression, but quantifying, extracting, and modeling these features have proven to be a difficult task. Likewise, the singing voice has also proven difficult to simulate.
convincingly because of its greater physical variation compared to other instruments. In order to pronounce different words, a singer must move his or her jaw, tongue, teeth, etc., changing the shape and thus the acoustic properties of the vocal mechanism. This range of acoustic variation is difficult to capture in a low-dimensional model, and analysis and synthesis techniques which have been successful for other musical instruments often do not apply well to speech or singing.

Numerous models for analysis and synthesis of singing have been designed for a variety of applications. Some applications, such as singing detection and singer identification, focus only on analysis, while others, such as singing voice coding, require both analysis and synthesis components. This chapter explores the prevalent models used in modern singing voice analysis and synthesis and highlights the features and targeted applications of each.

Section 2 discusses background principles of singing voice modeling, such as vocal physiology and the source–filter formulation. Sections 3 and 4 examine multiple models used for the glottal source and vocal tract filter, respectively. This is followed by a description of various approaches to overall singing synthesis in Section 5.

2 • Principles of Singing Voice Modeling
2.1 The Anatomy of the Singing Voice

The anatomy of the voice consists of three primary collections of organs: the lungs, the larynx, and the oropharynx. Exhalation from the lungs results in the air pressure changes necessary for singing. The larynx consists of a skeleton of cartilage enclosing and supporting two structures of muscle and ligaments covered by mucous membranes known as the vocal folds. The oropharynx is more commonly known as the vocal tract. The key characteristic of the vocal tract is its ability to assume a wide range of different shapes, which are easily altered by articulating the positions of the jaw, tongue, and lips.

The process of singing begins with breath pressure produced by the lungs. In the case of voiced sounds (vowels, e.g., [a], [i], and semivowels, e.g., [m], [r]), muscles initially adduct (close) the vocal folds, but the breath pressure forces them to open. The airflow through the opening is uneven, with the air adjacent to the folds traveling a greater distance than the unimpeded air flowing through the opening. The result is a pressure differential, which causes the vocal folds to be sucked back together by the Bernoulli force. Rapid repetition of this process is called phonation, and the frequency of this repetition correlates to our