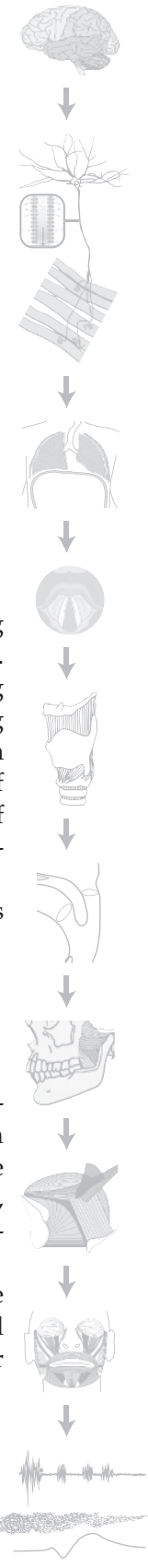
Part I
Getting to Sounds

Chapter 1

The Speech System and Basic Anatomy



Sound is movement. You can see or feel an object even if it – and everything around it – is perfectly still, but you can only hear an object when it moves. When things move, they sometimes create disturbances in the surrounding air that can, in turn, move the eardrum, giving us the sensation of hearing (Keith Johnson's *Acoustic and Auditory Phonetics* discusses this topic in detail). In order to understand the sounds of speech (the central goal of phonetics as a whole), we must first understand how the different parts of the human body move to produce those sounds (the central goal of articulatory phonetics).

This chapter describes the roadmap we follow in this book, as well as some of the background basics you'll need to know.

1.1 The Speech Chain

Traditionally, scientists have described the process of producing and perceiving speech in terms of a mostly feed-forward system, represented by a linear speech chain (Denes and Pinson, 1993). A *feed-forward* system is one in which a plan (in this case a speech plan) is constructed and carried out, without paying attention to the results. If you were to draw a map of a feed-forward system, all the arrows would go in one direction (see Figure 1.1).

Thus, in a feed-forward *speech chain* model, a speaker's thoughts are converted into linguistic representations, which are organized into vocal tract movements – *articulations* – that produce acoustic output. A listener

Articulatory Phonetics, First Edition. Bryan Gick, Ian Wilson, and Donald Derrick.

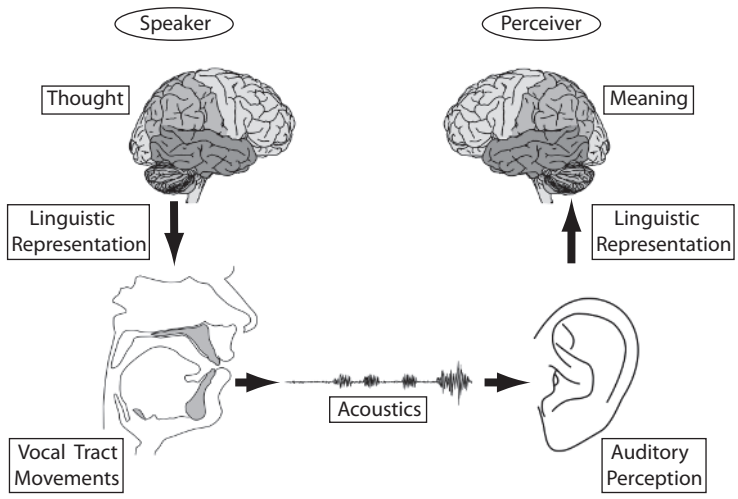


Figure 1.1 Feed-forward, auditory-only speech chain (image by W. Murphey and A. Yeung).

can then pick up this acoustic signal through hearing, or *audition*, after which it is perceived by the brain, converted into abstract linguistic representations and, finally, meaning.

Although the simplicity of a feed-forward model is appealing, we know that producing speech is not strictly linear and unidirectional. Rather, when we speak, we are also constantly monitoring and adjusting what we're doing as we move along the chain. We do this by using our senses to perceive what we are doing. This is called *feedback*. In a feedback system, control is based on observed results, rather than on a predetermined plan. The relationship between feedforward and feedback control in speech is complex. Also, speech perception feedback is *multimodal*. That is, we use not just our sense of hearing when we perceive and produce speech, but all of our sense modalities – even some you may not have heard of before. Thus, while the speech chain as a whole is generally linear, each link in the chain – and each step in the process of speech communication – is a loop (see Figure 1.2). We can think of each link of the chain as a *feedback loop*.

Multimodality and feedback

Speech production uses many different sensory mechanisms for feedback. The most commonly known feedback in speech is auditory feedback, though many senses are important in providing feedback in speech.

(Continued)

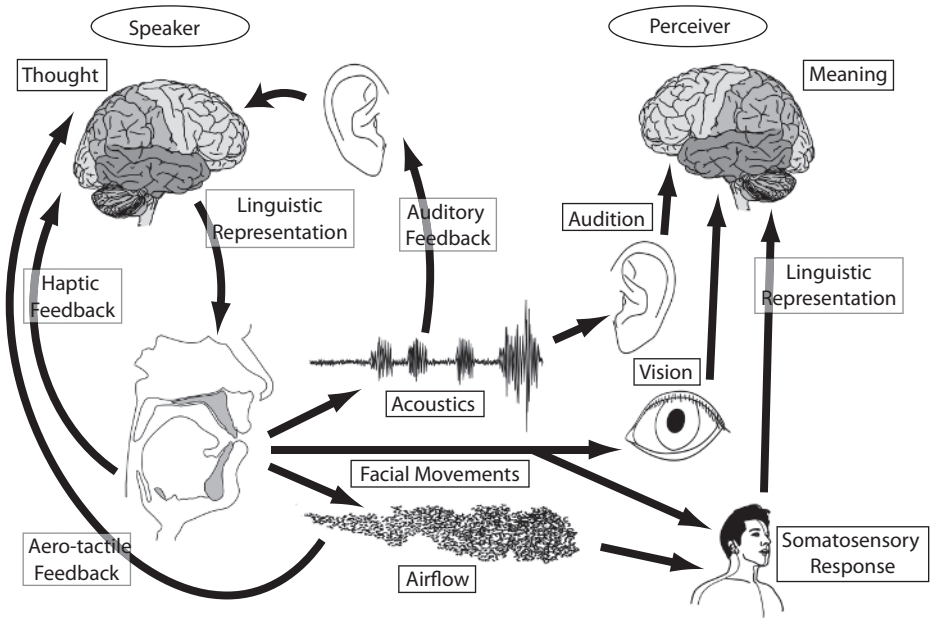


Figure 1.2 Multimodal speech chain with feedback loops (image by W. Murphey and A. Yeung).

Speech is often thought of largely in terms of sound. Sound is indeed an efficient medium for sharing information: it can be disconnected from its source, can travel a long distance through and around objects, and so on. As such, sound is a powerful modality for communication. Likewise, *auditory feedback* from sound provides a speaker with a constant flow of feedback about his or her speech.

Speech can also be perceived *visually*, by watching movements of the face and body. However, because one cannot normally see oneself speaking, vision is of little use for providing speech feedback from one's own articulators.

The *tactile*, or touch, senses can also be used to perceive speech. For example, perceivers are able to pick up *vibrotactile* and *aero-tactile* information from others' vibrations and airflow, respectively. Tactile information from one's own body can also be used as feedback. A related sense is the sense of *proprioception* (also known as *kinesthetic sense*), or the sense of body position and movement. The senses of touch and proprioception are often combined under the single term *haptic* (Greek, "grasp").

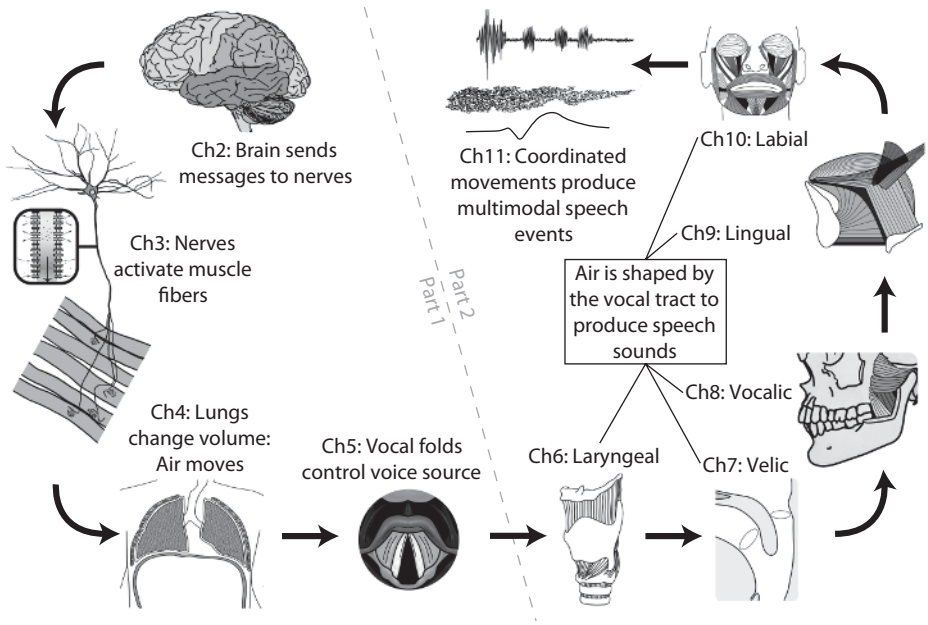


Figure 1.3 Speech production chain; the first half (left) takes you through Part I of the book, and the second half (right) covers Part II (image by D. Derrick and W. Murphey).

1.1.1 The speech production chain

Because this textbook is about articulatory phonetics, we'll focus mainly on the first part of the speech chain, just up to where speech sounds leave the mouth. This part of the chain has been called the *speech production chain* (see Figure 1.3). For simplicity's sake, this book will use a roadmap that follows along this feed-forward model of speech production, starting with the brain and moving in turn through the processes involved in making different speech sounds.

This is often how we think of speech: our brains come up with a speech plan, which is then sent through our bodies as nerve impulses. These nerve impulses reach muscles, causing them to contract. Muscle movements expand and contract our lungs, allowing us to move air. This air moves through our vocal tract, which we can shape with more muscle movements. By changing the shape of our vocal tract, we can block or release airflow, create vibrations or turbulence, change frequencies or resonances, and so on, all of which produce different speech sounds. The sound, air, vibrations and movements we produce through these actions can then be perceived by ourselves (through feedback) or by other people as speech.

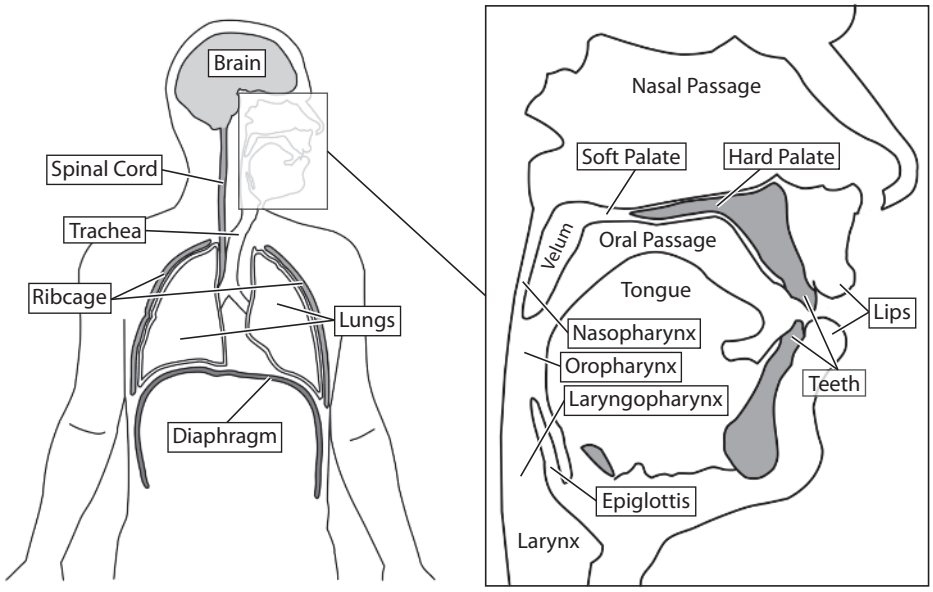


Figure 1.4 Anatomy overview: full body (left), vocal tract (right) (image by D. Derrick).

1.2 The Building Blocks of Articulatory Phonetics

The field of articulatory phonetics is all about the movements we make when we speak. So, in order to understand articulatory phonetics, you'll need to learn a good deal of anatomy. Figure 1.4 shows an overview of speech production anatomy. The speech production chain begins with the brain and other parts of the nervous system, and continues with the respiratory system, composed of the ribcage, lungs, trachea, and all the supporting muscles. Above the trachea is the larynx, and above that the pharynx, which is divided into the laryngeal, oral, and nasal parts. The upper vocal tract includes the nasal passages, and also the oral passage, which includes structures of the mouth such as the tongue and palate. The oral passage opens to the teeth and lips. The face is also intricately connected to the rest of the vocal tract, and is an important part of the visual and tactile communication of speech.

Scientists use many terms to describe anatomical structures, and anatomy diagrams often represent anatomical information along two-dimensional slices or planes (see Figure 1.5). A *midsagittal* plane divides a body down the middle into two halves: *dextrad* (Latin, "rightward") and *sinistrad* (Latin,

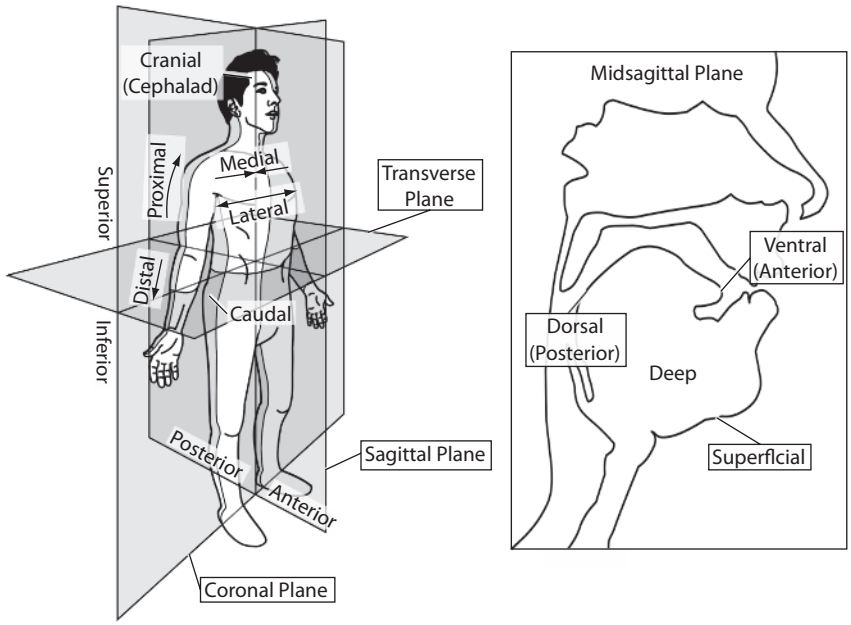


Figure 1.5 Anatomical planes and spatial relationships: full body (left), vocal tract (right) (image by D. Derrick).

"leftward"). The two axes of the *sagittal* plane are (a) vertical and (b) anterior-posterior. Midsagittal slices run down the midline of the body and are the most common cross-sections seen in articulatory phonetics. Structures near this midline are called *medial* or *mesial*, and structures along the edge are called *lateral*.

Coronal slices cut the body into *anterior* (front) and *posterior* (back) parts. The two axes of the *coronal* plane are (a) vertical and (b) side-to-side.

The *transverse* plane is horizontal, and cuts a body into *superior* (top) and *inferior* (bottom) parts.

The direction of the head is *cranial* or *cephalad*, and the direction of the tail is *caudal*. Also, *ventral* refers to the belly, and *dorsal* refers to the back. So, for creatures like humans that stand in an upright position, ventral is equivalent to anterior, and dorsal is equivalent to posterior. There are also terms that refer to locations relative to a center point rather than planes. Areas closer to the trunk are called *proximal*, while areas away from the trunk, like hands and feet, are *distal*.

Finally, structures in the body can also be described in terms of depth, with *superficial* structures being nearer the skin surface, and *deep* structures being closer to the center of the body.

1.2.1 Materials in the body

Anatomical structures are made up of several materials. Nerves make up the nervous system, and will be discussed in Chapters 2 and 3. As we are mostly interested in movement in this book, though, we'll mainly be learning about bones and muscles.

The "hard parts" of the body are made up of bones and cartilages. Bony, or *osseous* material is the hardest. The skull, ribs, and vertebrae are all composed of *bone*. These bones form the support structure of the vocal tract. *Cartilaginous* or *chondral* (Greek, "cartilage, grain") material is composed of semi-flexible material called *cartilage*. Cartilage is what makes up the stiff but flexible parts you can feel in your ears and nose. The larynx and ribcage contain several important cartilages for speech. Bones and cartilages are also the "hard parts" in the sense that you need to memorize their names, whereas most muscles are just named according to which hard parts they are attached to. For these reasons, we usually learn about the hard parts first, and then proceed to the muscles.

Muscles are made up of long strings of cells that have the specialized ability to contract (we'll look in more detail at how muscles work in Chapter 3). The word "muscle" comes from the Latin *musculus*, meaning "little mouse" (when you look at your biceps, it's not hard to imagine it's a little mouse moving under your skin!). In this textbook we'll study only striated, or skeletal, muscles. *Striated muscles* are often named by combining their *origin*, which is the larger, unmoving structure (usually a bone) to which they attach, and their *insertion*, which is usually the part that moves most when a muscle contracts. Many muscle names take the form "origin-insertion." For example, the muscle that originates at the *palate* and inserts in the *pharynx* is called the "palatopharyngeus" muscle (this is a real muscle you'll learn about later)! As you can see, if you know the origin and insertion of a muscle, then in many cases, you can guess its name.

Muscles seldom act alone. Most of the time, they interact in agonist-antagonist pairs. The *agonist* produces the main movement of an articulator, while the *antagonist* pulls in the opposite direction, lending control to the primary movement. Other muscles may also act as *synergists*. A synergist does not create movement, but lends stability to the system by preventing other unwanted motion.

Depending on whether a muscle is attached closer to or farther from a joint, it can have a higher or lower *mechanical advantage*. A muscle attached farther from a joint has a higher mechanical advantage, giving the muscle greater strength, but less speed and a smaller range of motion. A muscle attached closer to a joint has a lower mechanical advantage, reducing power but increasing speed and range of motion.

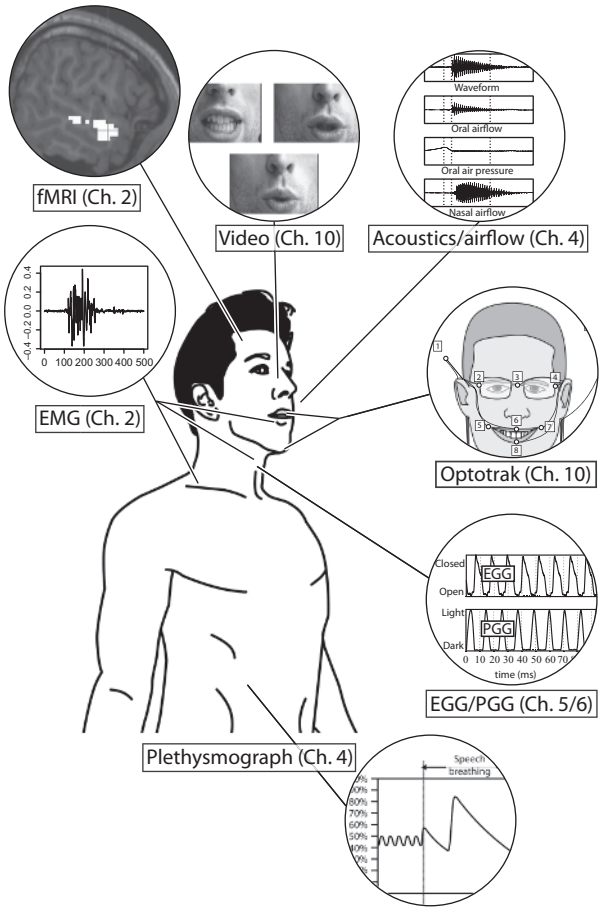


Figure 1.6a Measurement Tools for Articulatory Phonetics (image by D. Derrick).

1.3 The Tools of Articulatory Phonetics

Scientists interested in articulatory phonetics use a wide array of tools to track and measure the movements of articulators. Some of these tools are shown in Figures 1.6a and 1.6b, including examples of data obtained using each tool. Each tool has important advantages and disadvantages, including time and space resolution, subject comfort, availability and setup time, data storage and analysis, and expense. The issues related to each tool are discussed in depth in their respective chapters.

Decades ago, when graphical computer games first came out, the objects on the screen made choppy movements and looked blocky. Their movements were choppy because it took a long time for early computers to redraw images on the screen, indicating poor temporal resolution. *Temporal resolution* is a term for how often an event happens, such as how often a

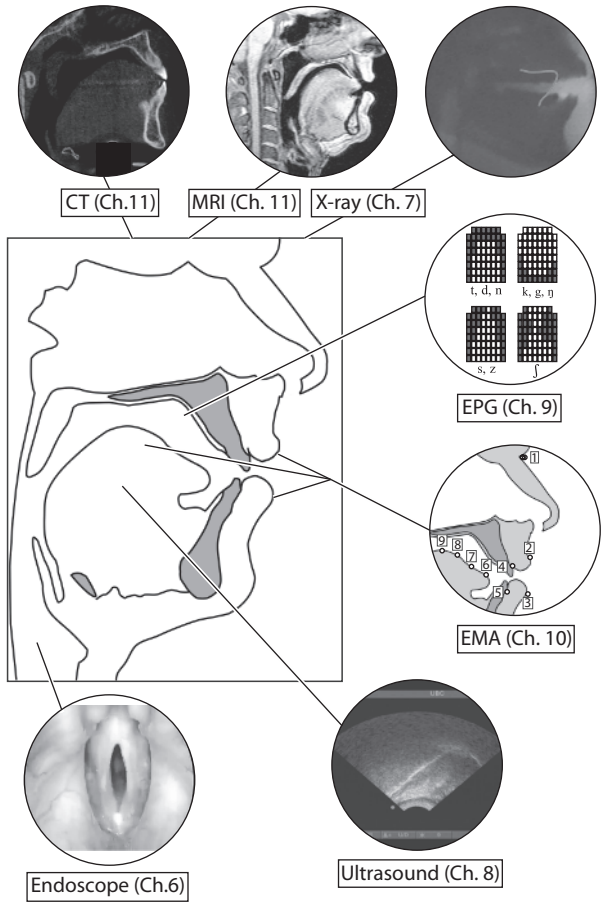


Figure 1.6b Measurement Tools for Articulatory Phonetics (image by D. Derrick).

recording, or sample, is taken. The term "temporal resolution" is often interchangeable with "sampling rate," and is often measured in samples per second, or *Hertz* (Hz). The "blocky" look was because only a few square pixels were used to represent an object, indicating low spatial resolution. From a measurement point of view, *spatial resolution* can be thought of as a term for how accurately you can identify or represent a specific location in space. Temporal and spatial resolution both draw on a computer's memory, as one involves recording more detail in time and the other requires recording more detail in space. Because of this, temporal and spatial resolution normally trade off, such that when one increases, the other will often decrease.

Many of the data-recording tools we'll look at in this book are ones that researchers use to measure anatomy relatively directly: imaging devices and point-tracking devices. *Imaging* devices take "pictures" of body

structures, so that they tend to show more spatial detail, but often run more slowly than other kinds of devices. For example, using current ultrasound or x-ray technology, a two-dimensional view of the tongue can be captured no faster than about 100 times per second, but a more common rate is about 30 times per second, as in standard movie-quality video (relatively low temporal resolution). These tools allow the measurement of slower articulator motion but show a picture of an entire section of the vocal tract (high spatial resolution). Imaging tools we'll look at in this book include: functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and electroencephalography (EEG) (Chapter 2); endoscopy (Chapter 6); x-ray film (Chapter 7); ultrasound (Chapter 8); electropalatography (EPG) (Chapter 9); video (Chapter 10); and computed tomography (CT) and magnetic resonance imaging (MRI) (Chapter 11). While most of these produce 2D images, some of them, such as CT and MRI, provide many slices of 2D information that can be combined to create 3D shapes. Electropalatography (EPG), which shows tongue-palate contact, has the highest temporal resolution of the imaging tools, but records only a few dozen points, giving it the lowest spatial resolution.

Unlike imaging devices, *point-tracking* systems track the motion of a small number of fixed points or markers very accurately, giving them high spatial resolution for only those points, and they can often capture this information at 1000 or more times per second, giving them high temporal resolution as well. Optotrak, Vicon MX, x-ray microbeam, and electromagnetic articulometers (EMA) (all discussed in Chapter 10) are all point-tracking systems.

We'll also look at other measurement devices in this book. These typically have higher temporal resolution, but may not directly convey any spatial information at all. For instance, a CD-quality sound file contains audio information captured 44,100 times per second. Electromyographs (EMG) (Chapter 3), airflow meters, plethysmographs (Chapter 4), and electroglottographs (EGG) (Chapter 5) are other measurement tools that can capture information at many thousands of times per second.

Now that we have some basic background, and an idea of where we're headed, the next chapter will start with the central nervous system, looking at the first steps of how the brain converts ideas into movement commands.

Exercises

Sufficient jargon

Define the following terms: feed-forward, speech chain, articulations, audition, feedback, multimodal, feedback loop, auditory feedback, vibrotactile

feedback, aero-tactile feedback, proprioception (kinesthetic sense), haptic, speech production chain, midsagittal, dextrad, sinistrad, sagittal plane, medial (mesial), lateral, coronal plane, anterior, posterior, transverse plane, superior, inferior, cranial (cephalad), caudal, ventral, dorsal, proximal, distal, superficial, deep, bone (osseous), cartilage (chondral), striated muscle, origin, insertion, agonist, antagonist, synergist, mechanical advantage, temporal resolution, spatial resolution, imaging devices, point-tracking systems.

Short-answer questions

1 Which anatomical plane(s) can show parts of both eyes in the same plane?
2 When perceiving speech, one can use vision as well as audition. List ten English consonants that normally provide clearly visible cues (use IPA where possible; where the IPA character is not available on a standard keyboard, standard English digraphs and/or broad phonetic descriptions are acceptable; for example, some English vowels could be listed as follows: a, u, lax u, open o, schwa, caret).
3 The *sternohyoid* muscle connects the *sternum* (a bone in the chest) and the *hyoid* (a bone in the neck). Which of these bones is the origin? Which is the insertion? What do you think is the name of the muscle that originates at the sternum and inserts into the *thyroid* (a cartilage in the neck)? What muscle originates at the thyroid cartilage and inserts into the hyoid bone? If the *palatoglossus* muscle runs between the palate and the tongue, what muscle inserts into the tongue and originates at the hyoid bone?
4 According to the above definition of spatial resolution, rank the following tools in terms of their spatial resolution (from lowest to highest): electropalatograph, Optotrak, ultrasound.

References

Denes, P. B. and Pinson, E. N. (1993). *The Speech Chain: The Physics and Biology of Spoken Language* (2nd edition). New York: W. H. Freeman.

Johnson, K. (2011). *Acoustic and Auditory Phonetics* (3rd edition). Chichester, UK: Wiley-Blackwell.