# 9 Sound recordings: acoustic and articulatory data

Robert J. Podesva and Elizabeth Zsiga

## 1 Introduction

Linguists, across the subdisciplines of the field, use sound recordings for a great many purposes – as data, stimuli, and a medium for recording notes. For example, phoneticians often record speech under controlled laboratory conditions to infer information about the production and comprehension of speech in subsequent acoustic and perception studies, respectively. In addition to analyzing acoustic data, phoneticians may employ articulatory methods to observe more directly how speech is produced. By contrast, sociolinguists often record unscripted speech outside of a university environment, such as a speaker's home. Sometimes these recordings themselves constitute the data (e.g., for sociophonetic analysis), while other times they may be transcribed at varying levels of detail (see Chapter 12), with the resultant text serving as the data (e.g., for the analysis of lexical or morphosyntactic variation and discourse analysis). In a similar vein, some language acquisitionists capture naturally occurring conversation in adult–child interactions. The research purposes of these recordings may not be determined until some time after the recordings are made, after a longitudinal corpus for a given child has been collected. It is likewise common for language documentarians to make extensive speech recordings in the field. Some field recordings simply serve as a record of elicitation sessions (e.g., when the researcher is ascertaining phrase structure), while others may be used for acoustic analysis (e.g., if phonetic elements of the language are the object of study). In the latter case, articulatory methods can be employed to more accurately describe phonetic properties of speech, such as a sound's place of articulation or details of the airstream mechanism. As discussed in Chapter 8, sound recordings can also be used as stimuli in perception studies, where listeners may be asked first to listen to a brief audio recording and then to identify whether a particular string of sounds is a real word (Chapter 8); to evaluate how educated the speaker of a brief utterance sounds (Chapter 6); or to rate how accented an L2 speaker sounds (Chapter 7). Linguists may also make use of archival recordings to investigate questions of language change. Proficiency in making sound recordings is thus an increasingly useful skill for linguists of most persuasions.

This chapter provides an overview of how to make sound recordings and collect articulatory data. As the output of speech production and the input to speech comprehension, the acoustic signal occupies the central position in the speech

stream. And since capturing the acoustic signal is important for studies concerned with speech production and comprehension alike, we focus primarily on recording acoustic data in this chapter (Section 2). In Section 3, we describe the most common methods for visualizing, recording, and analyzing the mechanics of speech articulation. We do not cover the design of perception studies here, as the relevant considerations are discussed in Chapters 6, 7, and 8. We conclude in Section 4.

## 2 Acoustic data

When recording audio data, one needs to decide who to record, what to record them saying, how to display recording materials, what equipment to use, and how to instruct speakers to sit and comport themselves in the recording environment. Because making a decision about each of these issues depends largely on the research questions posed, we describe three of the most common scenarios in detail in this section: making recordings in the laboratory, making recordings in the field for sociolinguistics, and making recordings in the field for language documentation. Although we discuss these scenarios separately, and while individual researchers may find one of these scenarios more closely related to the kind of work they do than others, the reader is encouraged to read through all three scenarios. Methods are increasingly borrowed across the subdisciplines, and researchers may find it useful to adopt hybrid methodologies. Before we discuss the particulars of each scenario, we review some considerations that pertain to all recording situations.

### 2.1 General considerations

First, it is important (and perhaps also trivial) to point out that we live in the digital age. Computers cannot represent truly continuous data, so analogue signals are instead encoded as a finite but extremely large number of sequentially ordered discrete bits that, when pieced together, sound continuous (see Ladefoged 1996 and Johnson 2012 for more detailed discussions of *digital signal processing*). While technologies that can capture a sound signal in analogue still exist, few of us still own devices that can play analogue recordings. More importantly, recordings need to be in digital format to do any of the things a linguist might want to do with them – analyze them acoustically (see Chapter 17), manipulate them for use in a perception study (see Chapter 6), upload them to a database, and so on. If recordings ultimately need to be converted to digital form, it is most efficient to record them digitally from the start.

When creating a digital recording, you first need to decide how many times an amplitude value should be recorded over the course of a second. This value is known as the *sampling rate*, which determines the frequency range that can be captured reliably by the digital signal. Only those frequencies up to half of the sampling rate (a value known as the *Nyquist frequency*) are faithfully captured. So

a recording sampled at 44 kHz (CD-quality) can faithfully represent frequencies up to 22 kHz, which represents the upper limit of the frequency range that humans can reliably hear. In practice, this is much higher than is necessary for speech. The highest linguistically meaningful frequencies in the speech signal (e.g., front cavity resonances of fricatives) appear at less than 11 kHz (e.g., Stevens 1998, Ladefoged 2003) – so a sampling rate of 22 kHz is generally sufficient for capturing whatever frequencies a linguist might be interested in. As digital technology progresses, however, recording systems can sample the signal at increasingly higher rates. In fact, some applications do not allow sampling at a rate lower than 44 kHz – which, at present, is the de facto standard sampling rate.

One thing to bear in mind is that the higher the sampling rate, the larger the file size. As disk space is relatively cheap, we recommend against trying to save space by using lower sampling rates. Using a higher sampling rate will also maximize the range of future uses for recordings. For example, data collected for vowel analysis (which only requires a sampling rate of about 10 kHz) can be repurposed for fricative analysis, but only if they were recorded at a sufficiently high sampling rate (22 kHz or more). It is better to sample at a high rate and downsample (or decrease the sample rate by low pass filtering) at a later date, if there is reason to think that a lower sampling rate may improve accuracy (Ladefoged 2003: 26).

A second consideration when creating a digital recording is the *sample size*. The sample size, measured in bits, specifies the number of units the amplitude is divided into. Not all recorders allow you to choose a bit rate, but high-fidelity audio systems typically have a bit rate of at a least 16 bits (which represents $2^{16} = 32{,}000$ gradations in the amplitude domain). Some allow 20- and 24-bit sample sizes (Cieri 2010), though the standard appears to be 16 bits. It is also worth pointing out that not all acoustic analysis software can handle sample sizes larger than 16 bits.

Many recorders allow you to specify the format of the audio data they produce. It is imperative that you choose an *uncompressed* format, what is known as linear pulse code modulated (PCM) format. PCM data can be saved in a number of file formats, such as .wav (waveform audio file format, the main format used on Windows systems) – the most common audio file format used by linguists – and .aiff (audio interchange file format, the main format used on Mac systems). Other formats will be compressed in one way or another, to save disk space. Although most compression algorithms are designed to minimize the *perceptible* distortion of the acoustic signal, they all distort the signal, which calls into question how faithfully the compressed audio signal represents what was actually uttered. Although some research has shown that certain forms of acoustic analysis are still possible with compressed audio, we strongly recommend avoiding compressed formats if at all possible. When using a new recorder, keep an eye out for the default data format – in many cases, it will be MP3! Also bear in mind that much of the data available on the internet is compressed, which limits the kinds of acoustic features that can be reliably analyzed.

An important goal when recording the acoustic signal is to maximize the robustness of the linguistic signal, by achieving as high a *signal-to-noise ratio*

as possible. This can be accomplished in several ways. First, the microphone should be close to the speaker's mouth. According to the *Law of Inverse Squares*, as the source of sound (i.e., a speaker's mouth) moves away from a microphone, the intensity of the sound will decrease at a rate of the square of the distance. Thus, a microphone located 2 feet from a speaker's mouth will be four times less intense than one located only a foot from the speaker's mouth. Second, the recording level should be set as high as possible without clipping (or overloading the signal), through the gain button. The precise level will depend on the recorder being used and how loudly the speaker is talking. Sometimes, the gain is represented as a strip of lights built into the recorder's hardware, arranged as a meter bar (usually green and yellow lights are fine, while red lights indicate clipping), while for other recorders, the gain is represented through the software interface (in the recorder's display window). Either option will suffice, as long as the recorder enables you to adjust the gain as the recording unfolds. As speakers will modulate their volume over the course of a recording, it is important to keep an eye on the recording level, and to adjust the gain as necessary. A final strategy for maximizing the signal-to-noise ratio is to minimize the ambient noise. As the potential sources of noise vary as a function of the recording scenario, I will postpone the discussion of ambient noise until Section 2.2.

Perhaps the most important step in preparing to make a recording is getting well acquainted with the recording equipment. The recording equipment should be tested several times prior to the recording session with the speaker; and even after the speaker has arrived, you should make and listen to a brief test recording to ensure that the data you are about to collect will meet your standards. Once you are sure your recording set-up is functional, and you have obtained whatever permissions are needed (see Chapter 2), begin all recordings with an announcement of the date, time, speaker (or some identifier, if speaker confidentiality is being maintained), the researcher(s) present, and the purpose of the recording. It would be a good idea also to include this information in a text file of metadata that is stored along with the recording, and/or to encode some of this information in the recording's file name, but recording the metadata in the audio record itself ensures that this information will be retained, even if the text file is deleted or the file name changed.

## 2.2 Common recording scenarios

While the issues discussed up to now are relevant to making audio recordings for any purpose, the remaining considerations (e.g., which kinds of recorders and microphones to use, what materials to record, and how to position equipment) depend on specific recording scenarios.

### 2.2.1 Recording in the laboratory

One of the most common sites for capturing audio data is the phonetics laboratory, specifically in a sound-proof recording booth. The most common types

of data collected in this context are recordings intended for subsequent acoustic analysis (see Chapter 17) and recordings intended for subsequent use as stimuli in perception studies (see Chapters 6 and 8).

There are considerable advantages associated with making audio recordings in a laboratory setting. First, the acoustic specifications are as close to ideal as possible, with ambient noise all but eliminated. Second, the equipment set-up in a phonetics lab is more or less stable, so recording a speaker will generally not require extensive reconfiguration of equipment or testing. Finally, laboratory equipment (e.g., recorders, microphones) is typically of very high quality, which further ensures high-quality recordings.

The current standard for digital recording in a lab is to record directly onto a computer's hard drive. In the recent past, labs have used other technologies, such as analogue and DAT (digital analogue tape) recorders, but these technologies have waned as direct-to-computer techniques have become dominant. (Analogue recorders required digitization before recordings could be analyzed acoustically, and DAT recorders required transferring the digital file recorded on the cassette tape to a computer hard drive.) It should be noted that computers are a potential source of noise, as the spinning hard drive and occasional whirring fan can compromise the signal-to-noise ratio, so computers are generally located outside of the recording booth (most booths allow the relevant cables to pass in and out of the booth through a conduit). Another popular technology is the *solid state recorder*, where audio data are stored on flash media instead of a spinning disk. While recording on a solid state recorder will likely produce pristine audio in this environment, when paired with the right microphone, the extra step of transferring audio recordings from the solid state recorder to the computer can be avoided by recording directly to the computer. Data can also be uploaded to a server more easily in the latter case.

Selecting the right microphone is one of the keys to a good audio recording. Most high-quality microphones are *condenser microphones* (i.e., they have their own power supplies). These power supplies can take one of several forms, with the microphone powered by a battery residing in the same unit as the microphone itself; a battery residing in a separate power pack; or phantom power supplied by the recording device or sound mixer.

In addition to the issue of whether a microphone requires a dedicated power source, microphones also differ in terms of directionality. In general, it is preferable to use a *directional microphone* (also known as cardioid or unidirectional), which generally captures the audio coming from a single direction (i.e., the direction the microphone is pointing in). The microphone can therefore be pointed in the direction of the speech signal, which will be picked up more robustly than ambient noise outside of this direct path. In contrast to directional microphones, *omnidirectional microphones* capture noise emanating from all directions (as the name implies); see Section 2.2.3 for an example of how omnidirectional microphones can be useful in the field.

A final consideration relates to how the microphone is held up or *mounted*. Laboratories typically make use of stand-mounted microphones, though other options include head-mounted microphones, lavalier (or tie-clip) microphones, and hand-held microphones. See Figure 9.1 for an example of common microphone mounts. Head-mounted microphones are preferable for obtaining reliable data on intensity, as the distance between the source of speech and the microphone is held constant; on the other hand, speakers are unlikely to move considerably from one moment to the next when seated in front of a table-mounted microphone. See Section 2.2.2 for a discussion of using lavalier microphones in the field. Hand-held microphones are generally not used in linguistics research.

Microphones can attach to recorders in a variety of ways, most often, if not exclusively, through XLR, mini-stereo, and USB jacks, all illustrated in Figure 9.2. While most high-quality recorders and microphones use XLR connections, XLR jacks can be converted to stereo and vice versa via rather inexpensive adapters. Microphones with USB connections are another attractive option, particularly when recordings are made directly to a computer hard drive. At present, the quality of USB microphones is highly variable, though low-noise options are available.

Once the researcher has settled on a recorder and microphone, the speaker needs to be positioned with respect to the equipment. In lab recordings, speakers typically sit in front of a table on which the microphone (usually stand-mounted) is resting. The microphone should never be placed directly in front of the airstream,

Figure 9.1. *Common microphone mounts: stand-mounted (left), head-mounted (middle), and lavalier (right)*

Figure 9.2. *Microphone jacks: XLR (left), mini-stereo (middle), and USB (right)*

but rather at a 45-degree angle from the corner of the speaker's mouth, approximately one open palm's width away. Positioning the microphone in front of the airstream can lead to *clipping* and/or transients in the acoustic signal that correspond not to properties of the airstream in the vocal tract (e.g., stop release bursts, which are of interest to linguists), but rather external properties of the airstream (e.g., the airstream hitting the surface of the microphone, which is of little interest to linguists).

The only remaining consideration at this point is how to display recording materials, which will depend on the nature of the data. Many researchers working on segmental phonetics will ask speakers to read a *word list* that exemplifies the contrasts under analysis (see Chapters 4 and 18). In such cases, all words should be checked in advance with each speaker, to make sure that all the words exist in their lexicon. During the recording, words are typically embedded in a *carrier phrase* like "Please say __ for me" – an utterance that makes sense regardless of what word fills the blank. The target word is usually phrase-medial to avoid the effects of phrase-final lengthening. The researcher should pay special attention to the sounds immediately preceding and following the target word, to facilitate the identification of segment boundaries. If vowel-initial words are under investigation, for example, the word just before the blank should not end in a vowel – since it would then be difficult to isolate the border between two adjacent vowel sounds (see Chapter 17 for more on the acoustic properties of different classes of sounds). Words are most often represented in the language's orthography, though words can also be elicited by having speakers provide translations for English words spoken by the researcher, which may be necessary when working with an illiterate speaker or a language without a standardized orthography. Words can be displayed as a list on a sheet of paper, in which case the paper should be placed on a stand (not held by the speaker, since the rustling of paper will compromise the quality of the recording); individually on cards, though the speaker will need to be instructed not to speak while the cards are being moved; as a list on a computer screen; or individually on a computer screen, perhaps even through a timed PowerPoint presentation (standardizing how long each word is displayed can have the added advantage of standardizing speech rate). In any case, words should be randomized, and it is common for multiple repetitions for each word to be collected. Displaying words individually militates to some extent against speakers producing a list intonation, which can have significant consequences for the phonetic realization of target words.

Researchers interested in connected speech, post-lexical phonological processes, or suprasegmentals may find it useful to record *reading passages*. Passages are sometimes written specifically for fulfilling the needs of a specific study (e.g., when certain words are needed in particular prosodic contexts), but often standard reading passages are used, such as Fairbanks' (1960: 127) Rainbow Passage, which is designed to exemplify a wide range of the sounds of English in a diverse array of phonological contexts. Speakers should be allowed to familiarize themselves with reading passages before beginning the recording.

In spite of all its advantages, one disadvantage of making recordings in a laboratory setting is that it constrains the range of linguistic styles that speakers produce, which tend toward more careful, citation-style speech. For many research questions, this limited range of styles does not pose a significant problem. However, linguists interested in more vernacular speech styles may find it more fruitful to analyze data produced in the field.

### 2.2.2 Recording in the field: sociolinguistics

Sociolinguists most often record unscripted dialogue outside of institutional contexts, usually in the form of *sociolinguistic interviews*, which are generally informal conversations between one or two interviewees and one or two interviewers, intended to elicit unguarded speech (see Chapters 6 and 10 for extended discussions about sociolinguistic interviews and recording social interaction).

While conversational speech is much more likely to exhibit linguistic features of sociolinguistic interest than speech recorded in laboratory contexts, it also makes it more difficult to draw comparisons across speakers (since everyone is saying something different, the features of interest are being produced in different phonological, grammatical, and discourse contexts), though see Chapters 16 and 20 for statistical techniques for dealing with this variability.

Another challenge of recording in the field is reducing the ambient noise captured in the recording. This issue can be addressed in part by choosing the right microphone (see the discussion below), but also by finding the right environment for making recordings. In general, rooms with many hard surfaces should be avoided, as they reflect sound and thus compromise the clarity of the speech signal. Indoor sources of noise include televisions, radios, refrigerators, lighting, air conditioning and heating units, computers, clocks, and phones (Cieri 2010: 27). Noises from outside, such as wind, rain, and traffic, can also disrupt recordings, even when recordings are made indoors. In some environments, like speakers' homes, it is possible to minimize noise by turning off the noisiest of appliances. At the same time, researchers must bear in mind that they are guests and should respect speaker's comfort levels, even if it means that recording quality is compromised. Once, we made a recording in nearly 100 degrees heat and asked the interviewee if we could turn off the air conditioner. She did so willingly, but proceeded to (very audibly) fan herself with a nearby piece of paper (the consent form, incidentally) from time to time. While those segments of the interview were not usable for acoustic analysis, we felt it was more important for her to be comfortable (and safely cool) than for us to have pristine data.

The last 10 years have witnessed tremendous advancements in the development of portable digital recorder technology. Many options – solid state recorders of varying sizes, CD recorders, minidisk recorders, cell phones, and laptop computers, to name just a few – have all been successfully employed in sociolinguistic research. We do not recommend using minidisks (an obsolete technology) or CD recorders (due to the inconvenience associated with waiting for the CD to be

burned, and because of the possibility of scratching CDs), but we would like to comment on the other three options.

In a recent study, De Decker and Nycz (2011) report that recordings made on an iPhone (through the Voice Memo app) are of sufficient quality for reliably extracting the first and second vowel formant (though measurements for the third formant were more variable). Subsequently, applications designed specifically for sociolinguistic use in the field have been developed, with some even allowing for files to be automatically uploaded to a cloud. De Decker and Nycz also report that recordings made with a Macbook Pro were sufficient for the analysis of vowel formants. An obvious advantage to recording with laptops and iPhones is that many speakers have grown rather accustomed to the ubiquity of cell phones and computers; they may be less likely to categorize these devices as recording instruments, and accordingly may be more inclined to produce unselfconscious speech. However, given the difficulty associated with faithfully capturing higher frequencies, it may be preferable to use recorders that can better handle frequencies above 3,000 Hz.

We recommend using solid state digital recorders, two examples of which are shown in Figure 9.3, simply to maximize the kinds of analyses that can be conducted. Most solid state recorders can be configured to record uncompressed data onto a flash memory card.

With portable equipment comes the need for portable power; do not rely on the availability of an electrical outlet. Bring batteries, and because you will go through many, it is a good idea to buy rechargeable batteries, which of course necessitates the purchase of a charger. Bring twice as many batteries as you think you might need to each recording session, and get in the habit of charging your batteries every night. Although batteries hold their charge better, over time, if they are completely discharged between chargings, what is gained in battery life is lost in data quality – as it can be extremely disruptive to have to change batteries during the middle of a conversation or story.

As far as microphones are concerned, we recommend using directional *lavalier* (tie-clip, lapel) *microphones* with their own power packs. Using a directional microphone will maximize the likelihood that the speech of the interviewee will


Figure 9.3. *Solid state recorders: Marantz PMD660 (left) and Zoom H2n (right)*

be isolated, and that ambient noise will be minimized (though certainly not eliminated). We recommend lavalier microphones because they are small and can be immobilized by clipping them onto speakers' shirts (ask speakers to clip microphones on for themselves). Higher-end recorders may have two input jacks, for a left and right microphone signal. You may find it beneficial to record the speech of the interviewer with a separate directional microphone or, if there is more than one interviewer, with an omnidirectional microphone. Provided that you feed two separate microphones into the left and right microphone jacks, the recorder will keep the two channels distinct from one another. Separating the left (interviewee) channel from the right (interviewer) channel is trivial with most acoustic analysis software applications. We strongly caution against using built-in microphones, even though most recorders have them, since such microphones are usually unable to isolate the speech signal and over-represent ambient noise.

Compared to laboratory recordings, keeping the signal-to-noise ratio high for field recordings is a significant challenge. It should be noted that the signal-to-noise ratio will be lower for recordings collected in the field than for lab recordings. This is due to the fact that there is more ambient noise outside of controlled laboratory conditions and because people use a much wider dynamic range in conversational speech than they do in the lab, where speakers will produce relatively more consistent loudness levels throughout the recording – so the appropriate gain for one part of the conversation might not be appropriate for other parts.

### 2.2.3 Recording in the field: language documentation

What constitutes "the field" can vary considerably from one project to another. While "the field" will, for some, conjure images of rainforests, deserts, and tundra, a great many more researchers conduct field research work much closer to home, in collaboration with a language consultant, often in the consultant's home or workplace. Whatever the research locale, it is essential – as it was for sociolinguistic recordings – for equipment to be highly portable. In spite of this similarity, recording for the purposes of language documentation differs significantly from recording for sociolinguistic purposes in one main respect: recording equipment need not be made inconspicuous. Language consultants are well aware that their language is under investigation – indeed, they are explicitly asked to reflect on the structure of their language – so seeing a microphone or a recorder in plain sight should have negligible effects on the kind of data collected in this scenario.

For this reason, *head-mounted microphones* are preferred. These directional microphones, located at a constant distance from the speaker's mouth, zero in on the speaker's voice while minimizing other noise. Directional microphones aimed toward speakers' mouths will usually pick up the speech of the researcher as well, though certainly not nearly as robustly. If knowing precisely what the researcher is saying is important, as is typically the case in the elicitation of unfamiliar languages, we recommend capturing the researcher's voice on a separate channel, with a separate microphone. In cases where there is more than one researcher, as in

the case of a field methods class, the researchers can be collectively recorded using an omnidirectional microphone.

It is similarly not very important to make recording devices inconspicuous in this particular recording scenario. So while small portable recorders like those discussed in the previous section are all viable options when making recordings for language documentation purposes, so too are laptop computers.

We have recommended recording strategies that most faithfully and robustly capture the speech signal, even though obtaining high-quality audio recordings is not an important concern for many domains of language description (e.g., research on the structure of relative clauses). Given that disk space is relatively cheap, and because one never knows what research questions might arise in the future, we recommend erring on the side of collecting needlessly clean recordings. This is especially important in the case of an endangered language, where elicitation sessions on clausal syntax may unfortunately come to double as records of the language's sound system. For more issues relating to language description, see Chapter 4.

### 2.2.4 Other recording scenarios

Although we have just presented three rather different scenarios for collecting acoustic data, we do not mean to suggest that the methods that are common in one cannot be imported fruitfully into others. For example, phoneticians may be interested in connected speech processes that are better represented in spontaneous speech than read speech. In these cases, spontaneous speech data can be elicited in the lab, resulting in recordings that, though less controlled in terms of linguistic form, nonetheless still exhibit high signal-to-noise ratios. Similarly, sociolinguists who are primarily interested in conversational interview data, may additionally collect word list data to expand the stylistic range of data collected for each speaker. While recording word lists is a common practice in sociolinguistics, words are often not elicited in the same way as they would be in a phonetic study (e.g., with respect to the issues of randomization and collecting multiple repetitions).

In Figure 9.4, we present a range of alternative techniques for collecting sound recording data. All are worthwhile, but some are better suited to answering particular questions than others. As we move from left to right, we proceed from the least spontaneous speech to the most spontaneous speech. We also go from elicitation tasks that do not closely approximate the speaking situations we most often encounter to tasks that very closely correspond to real-life speaking situations. We also go from methodologies for which it is very easy to compare across speakers, since they are saying the same things in the same linguistic contexts, to methodologies for which it is more difficult to compare across speakers. Finally, the data collection techniques on the left represent approaches that often make use of very visible, and often expensive laboratory equipment, while those on the right represent approaches making use of smaller, yet still relatively expensive equipment. Space constraints prevent us from discussing each technique in detail, though we list the alternatives to provide a sense of what can be done besides word lists and interviews.

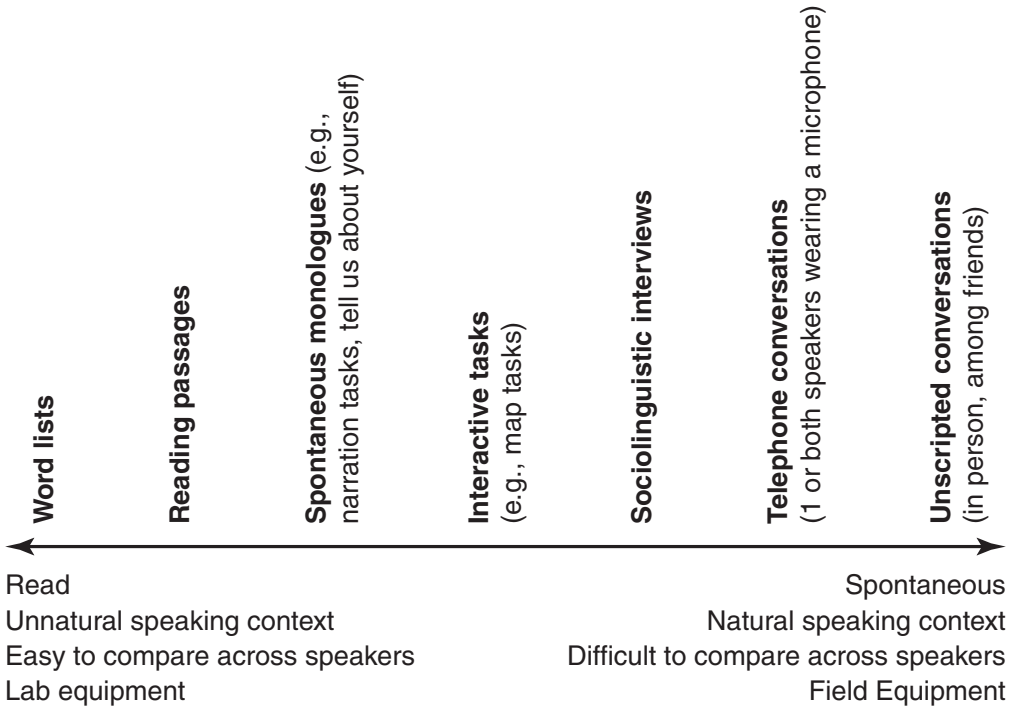


Figure 9.4. *Range of data collection scenarios*

Although Figure 9.4 implies a trade-off between audio quality and the naturalness of speaking situations, this need not be the case. A number of interactional phonetics laboratories (e.g., Tyler Kendall's at the University of Oregon; Norma Mendoza-Denton's at the University of Arizona; Rob Podesva's at Stanford University) are comfortable interactive spaces that have been built with acoustical specifications that approach or equal those of sound-proof and sound-attenuated booths. The goal in such spaces is to collect highly interactive audio (and video) interactions that are also characterized, unlike many field recordings, by a high signal-to-noise ratio.

## 2.3 Managing recordings

After recordings are made, it is imperative that data are backed up immediately. We recommend backing up the data once on a computer hard drive, again on a portable external hard drive (USB drives that do not require a power supply are preferred), and, if an internet connection is available, to a server or cloud. Audio files should be labeled in a systematic way (so develop a file-naming convention that works for your purposes), and metadata should be stored in accompanying text files and ideally also in a database or spreadsheet for your records. As it can be difficult to work with large audio files, you may find it helpful to divide long recordings into more manageable pieces, the size of which will depend on the kind of research being performed. Some researchers may find it useful to take notes on the content of recordings right away, which can be entered into field notes (see Chapter 10), a sound file annotation (see Chapter 17), or transcription software (see Chapter 12).

## 3 Articulatory data

A researcher using acoustic analysis must infer the shape or movement of articulators in the vocal tract by working backwards from the output, using formulas that relate specific acoustic signatures to particular vocal tract states. It is also possible, however, to directly visualize the vocal tract. In this section, we review a number of commonly used devices for directly measuring articulator shape, position, or movement. The discussion is organized around the difficulty and expense of the technique. "Easy" techniques involve equipment you may already have or that is inexpensive to obtain, that requires little or no specialist training, and that can be used anywhere. These include video and static palatography. "Medium" techniques involve equipment that may cost several thousand dollars to obtain, but that any linguist can learn to use and that can be used in a typical departmental linguistics lab or carried into the field. Such equipment includes electropalatography (EPG), sonography, electroglottography (EGG), and masks for aerodynamic measures. To use the "Difficult" techniques, you probably need access to someone with specialized medical training, a medical school, and/or a really large lab budget. While such techniques might be beyond what the readers of this chapter would use themselves, it is likely that they will encounter the results in published research, so it is worth learning how such techniques work. Difficult techniques include endoscopy, magnetic resonance imaging (MRI), the electromagnetic mid-sagittal articulometer (EMMA), and electromyography (EMG).

The set of devices that can be used for articulatory investigations is in principle limited only by ingenuity, and we can cover only the most commonly used methods here. Other more obscure devices (such as the velotrace, plethysmograph, and strain gauges) are described in Horiguchi and Bell-Berti (1987), Ohala (1993), and McGlone and Proffit (1972), respectively. Also, even though X-rays have been important tools for imaging the vocal tract, present-day studies typically avoid the methodology, given the health risks associated with extended exposure. X-ray databases are nonetheless still available (Munhall, Vatikiotis-Bateson, and Tohkura 1995). Finally, we do not discuss methodologies that capture brain function or attention during speech production; for a discussion of these techniques, including eye-tracking, fMRI, PET scans, and ERP, see Chapter 8. For a comprehensive introduction to articulatory phonetics, see Gick, Wilson, and Derrick (2013).

For each technique, we briefly describe the kind of data that can be collected (and why a linguist might care about such data), what is involved in setting up and running an experiment, an example of what data collected with this technique looks like, and a few pros and cons. No matter what technique you decide to use, you should consult someone with experience who can give you more detailed guidance. Here, we aim to give you an idea of what is available, as well as aid you in understanding and interpreting the results of others.

## 3.1 Easy techniques

### 3.1.1 Video

While most of what goes on in the act of speaking happens inside the mouth and thus requires more sophisticated imaging tools, a *video camera* can capture any visible aspects of speech communication. Such aspects might include interpersonal interactions, facial expression, gaze, and gestures with the hands and other parts of the body. This is of course useful for the investigation of signed languages, but studies of the integration of speech with other body movements have turned up interesting data on both interpersonal interaction and general temporal coordination (see Chapter 10 for more on recording interaction). In terms of articulation per se, the linguist might be interested in investigating lip position, to document bilabial vs labiodental place of articulation, for example. In Figure 9.5, two stills extracted from a video clip document two different kinds of bilabial constriction in the Sengwato dialect of Setswana: compression for [ɸ] (left) vs rounding in secondary articulations such as [$s^w$] (right).

For a linguistic video study, you will need only a camera, which should be set up on a tripod for stability. A mirror held at a 45-degree angle to the side of the subject's face can capture a simultaneous side view. Numerous video editing programs are commercially available; one video annotation and editing tool popular with linguists is ELAN, available as a free download from Language Archiving Technology (www.lat-mpi.eu/tools/elan). One thing to be careful of in video studies is subject privacy: you may choose to capture only the lips, as in Figure 9.5, or obscure the eyes, or obtain permission to use the full face image (see Chapter 2 on research ethics).

An obvious drawback to using video to analyze speech production is that video cameras can capture only what can be seen external to the speaker, and only under the proper lighting conditions.

### 3.1.2 Static palatography

*Static palatography* offers a quick and (literally) dirty way to investigate patterns of tongue contact against the palate. It can be used to compare place

Figure 9.5. *Lip position for [ɸ] (left) and [$s^w$] (right) in Sengwato*

Figure 9.6. *Palatogram (left) and linguogram (right) of American English /t/*

of articulation among coronal articulations – for example, to document whether a particular articulation is dental or alveolar, apical or laminal.

Static palatography involves painting the tongue or palate of a subject with a mixture of oil and charcoal. Activated charcoal can be ordered from any pharmacy, without a prescription: its pharmacological use is as a poison antidote, so ingesting a small amount is not harmful. Mix a teaspoon of charcoal with a teaspoon of vegetable oil, and stir until it is the consistency of black paint. A drop of mint extract will make the mixture taste like toothpaste. In addition to your charcoal paint and a small paintbrush, you will also need a small mirror, a few inches square, and a camera to record your results.

To image the pattern of tongue contact on the palate, have your subject stick out his tongue, and paint the tongue with the charcoal and oil mixture, being careful to cover the tip and sides. Go back as far as you can without triggering a gag reflex. You have to work quickly, as the subject cannot close his mouth or swallow. After the tongue is covered, have the subject articulate one consonant – for example, [ata]. The paint will rub off where the tongue touches the palate, leaving the pattern of tongue contact. To get an image of the pattern, hold the mirror at a 45-degree degree angle inside the subject's mouth, and snap a picture of the image in the mirror (as shown in Figure 9.6, left). Afterwards, allow the subject to rinse and spit.

To obtain an image of the part of the tongue that contacts the palate (technically a *linguogram*), paint the palate instead, and have your subject articulate the same consonant. The paint will rub off the palate onto the tongue. Have your subject stick out his tongue, and photograph (Figure 9.6, right).

Static palatography is fun and easy, and gives a good sense of contact in three dimensions, not just the mid-sagittal plane. Drawbacks are that it is messy, and not all subjects are willing to have their tongues painted and photographed. Additionally, the technique only works for coronal consonants in isolation. Because of the gag reflex, and the difficulty of getting a picture, back consonants cannot be investigated. Only a single consonant in isolation can be produced, or the paint will just smear. Finally, the technique does not lend itself to quantification.

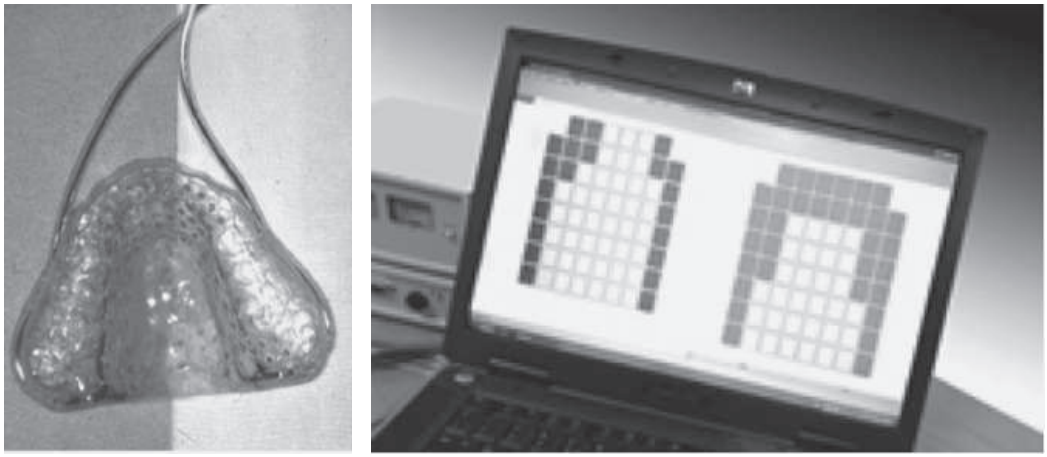

Figure 9.7. *Artificial palate with embedded electrodes (left); sample patterns for /s/ and /t/ (right)*
*http://speech.umaryland.edu/epg.html (left)*
*http://www.rds-sw.nihr.ac.uk/succcess_stories_lucy_ellis.htm (right)*

## 3.2 Medium techniques

### 3.2.1 Electropalatography

*Electropalatography* (EPG) works on the same principle as static palatography, but instead of paint, an artificial palate embedded with electrodes and attached to a computer records the pattern of tongue contact (Figure 9.7, left). When the tongue contacts an electrode, a signal is sent to the computer, which can then compute the pattern (Figure 9.7, right).

EPG is an improvement over the paint-and-charcoal technique, in that it can image tongue contact in running speech (the electrodes do not smear). The researcher can see patterns of contact changing as the constriction is formed and then released, not just maximum constriction (frame rates are typically 100–200 Hz). EPG also allows quantification, as the number of electrodes and specific pattern activated can be compared across different articulations.

The drawbacks of EPG include cost: the system itself will cost several thousand dollars, and the artificial palates must be custom-made from a dental cast, at significant additional cost for each subject. For subjects, the palates may take some getting used to (they feel like an orthodontic retainer), so speech may not be entirely natural. Finally, the technique can measure where the contact is made on the palate, but not which part of the tongue is making it.

### 3.2.2 Sonography

*Sonography* is in some ways the opposite of electropalatography: with this technique, you can see tongue position, but not palate contact (at least not directly). Like EPG, sonography involves an initial expense to acquire the equipment, in the order of $25,000 at the time of writing. Once acquired, however, it costs very little more to use. Portable sonographs, not much bigger than a laptop, are available for use in fieldwork.

In linguistic sonography, a transducer is held under the subject's chin (Figure 9.8, left). Gel spread on the skin facilitates unbroken contact. The transducer emits a series of sound waves that travel up from the transducer through the skin and tongue muscle, and then bounce back when they reach the border

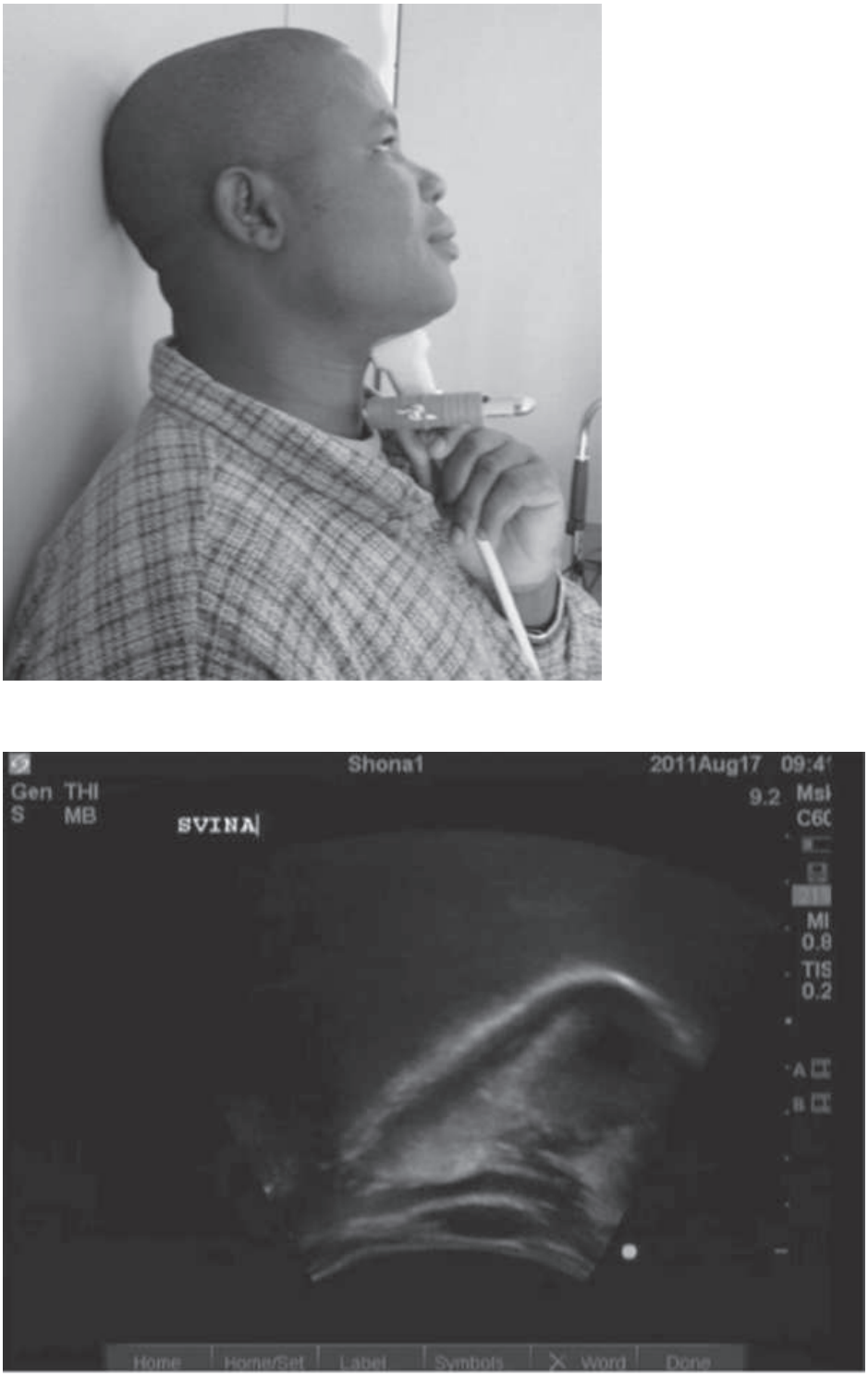


Figure 9.8. *Subject holding a sonograph transducer (top); sonograph image for the vowel /i/ (bottom)*
*Gick 2002*

between the tongue surface and the air inside the vocal tract. The equipment measures the time delay between transmission and reception, and converts that measure to a distance between transducer and tongue surface. Repeated measurements produce an outline of the surface of the tongue, as shown in Figure 9.8 (right).

The pros of sonography are that it is direct and non-invasive, and can record changes in tongue shape over time, in real time (although the acquisition rate, typically 40 ms per frame, may not be fast enough to capture fast-moving articulations such as taps). Subjects enjoy watching the moving images of their own tongues (though they should be allowed to do this before and after the experiment, not during, so that they do not get distracted.) Programs for tracing and quantitatively comparing different tongue shapes are widely available.

Because of its non-invasiveness, sonography has grown in popularity, not only in the field of phonetics, but also sociolinguistics, to investigate questions of language variation and change. Recent studies have demonstrated that articulatory variation can surface in the absence of significant variation in the acoustical signal (Lawson, Stuart-Smith, and Scobbie 2008; Mielke, Baker, and Archangeli 2010; De Decker and Nycz 2012). For example, De Decker and Nycz (2012) draw on ultrasound data to show that some speakers achieve tense variants of /æ/ with a raising/fronting tongue gesture, while others exhibit no evidence of such a gesture.

One disadvantage of sonography is that is not always possible to image the tongue tip, if there is not a direct line, through muscle only, from transducer to tongue tip. An air space under the tongue tip, or interference from the hyoid bone, may prevent the sound waves from reaching the very front of the tongue. Additionally, the tongue and palate cannot be imaged at the same time, so that patterns of tongue-to-palate contact or constriction cannot be measured directly. In order for an image of the palate to be obtained, the subject can be asked to hold a swallow of liquid in the mouth, eliminating the air border at the top of the tongue, so sound waves travel through the tongue and through the liquid, bouncing back when they hit the palate, allowing an outline of palate shape to be imaged. Then, in order to discover tongue position in relation to the palate, as would be necessary to investigate place of articulation, the two separate images of tongue and palate must be overlaid. In order for this overlay to work, it is crucial that neither the subject's head nor the transducer move at all during the imaging session, so as not to change the alignment. Finding an effective head-stabilization technique that does not compromise the comfort of the subject is probably the most challenging aspect of using sonography. Some approaches involve immobilizing the subject's head and the transducer (see Davidson and De Decker 2005 for an inexpensive and portable method); other approaches allow head movement, but measure the movement and compensate for it (see Whalen et al. 2005 for a description of HOCUS, the Haskins optically corrected ultrasound system).

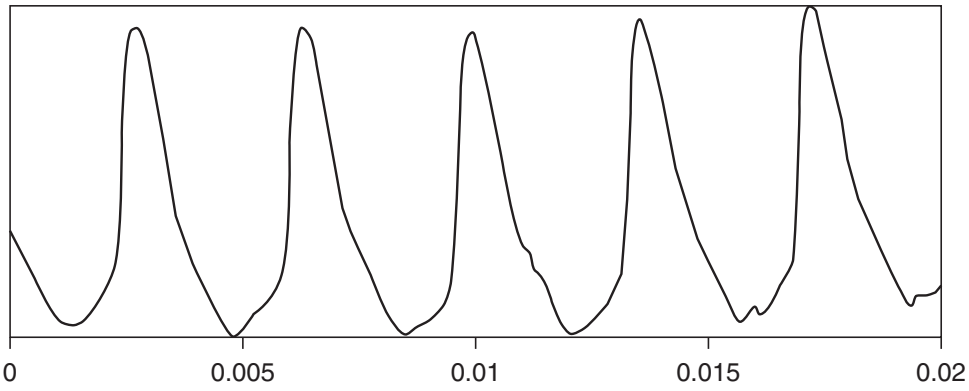


Figure 9.9. *Example of an EGG waveform during modal voicing*

### 3.2.3 Electroglottography

*Electroglottography* (EGG) uses electrical impedance to measure opening and closing of the glottis. It is often used for studies of voice quality. In EGG, electrodes are held against the skin of the neck, on either side of the larynx. Typically, a Velcro strap holds them in place. Then, a very weak current is passed between the electrodes – the current is so weak it cannot be felt at all by the subject, but the strength of the current can be detected by the technology. Electrical impedance between the two electrodes is greater when the vocal folds are open than when they are closed, so that a graph of the measured impedance shows the relative opening and closing of the glottis (Figure 9.9).

EGG is non-invasive, and involves no discomfort other than a snug Velcro collar. It allows direct measurement of glottal state, bypassing the vocal tract filter. Initial cost is again several thousand dollars, but there is no additional cost per use. Placing the electrodes properly, directly on either side of the vocal folds, can be tricky, depending on the subject's body type. Because of differences in laryngeal anatomy, EGG may work better on male subjects, where the location of the larynx is often more readily apparent, than on female subjects.

### 3.2.4 Aerodynamic measures

*Aerodynamic measures* record oral and/or nasal airflow. For certain sounds, it matters a lot how much air is flowing where. A linguist might want to measure the degree of vowel nasalization, for example, or the pressure differential in front of and behind the constriction in a fricative.

The technique involves a mask, similar to an oxygen mask, that is held over the face while the subject is speaking (Figure 9.10, left). The mask may be split, to have separate chambers for the nose and mouth. Screens in the mask allow air to move in and out, so that the subject can continue to breath and speak, while transducers in the mask measure air flow and air pressure. To measure pressure behind a constriction, the end of a small plastic tube can be placed just behind the lips (or, with slightly more care, just behind the tongue front). Figure 9.10 (right) shows pressure build-up behind the lips during a bilabial fricative.

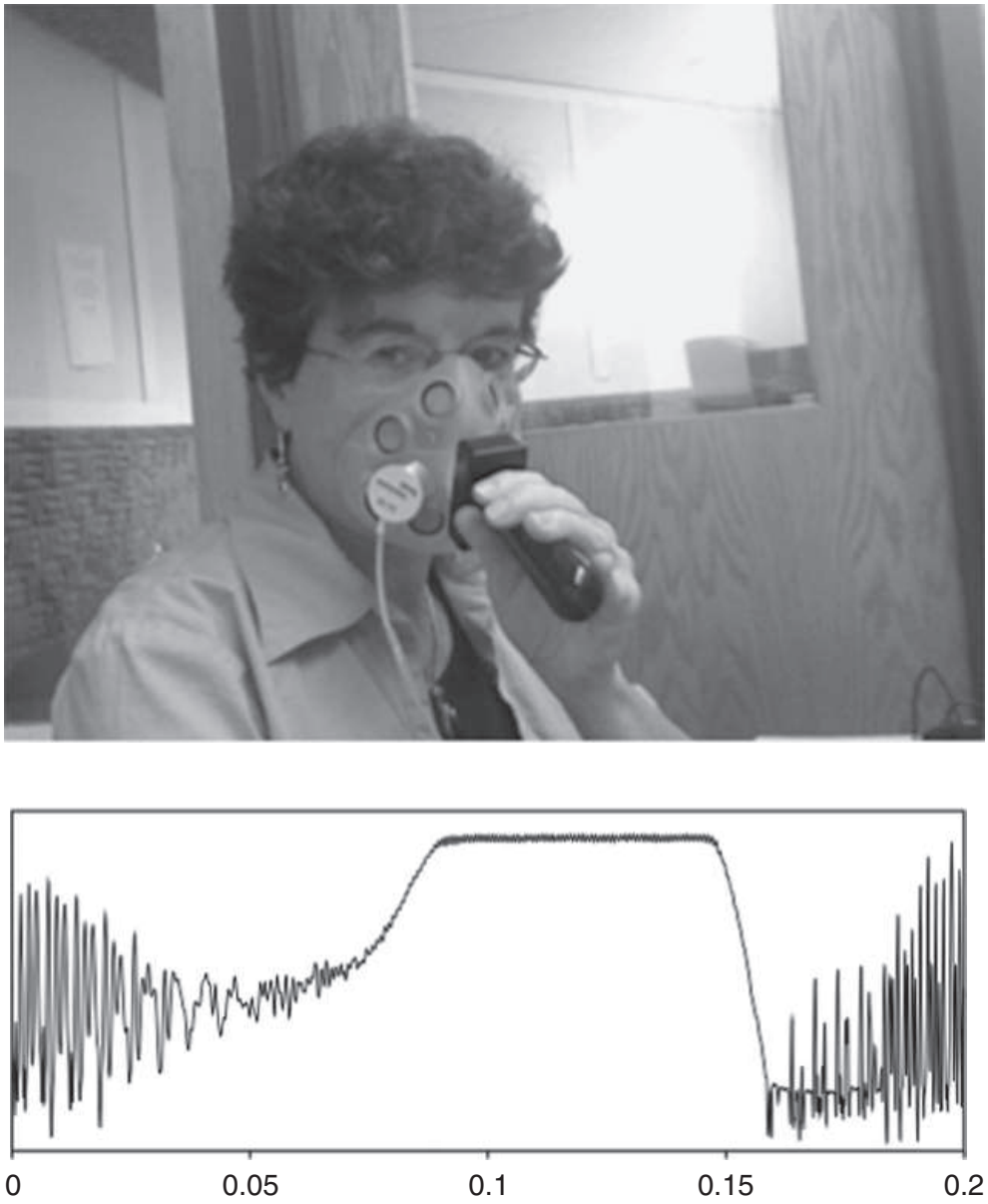


Figure 9.10. *Using a pressure/airflow mask (top); trace of pressure at the lips during [aφa] (bottom)*

A direct measure of airflow can be very useful, because airflow and intraoral pressure are very hard to infer from the acoustic record, if it can be done at all. A drawback is that airflow measures from the transducers are hard to calibrate. Further, while muffled speech can be heard through the mask, one cannot collect a clear acoustic record while the mask is being used.

## 3.3 Difficult techniques

### 3.3.1 Endoscopy

For linguistic research, an *endoscope* is used to take video or still images of the larynx, and thus can be used to investigate states of the vocal folds

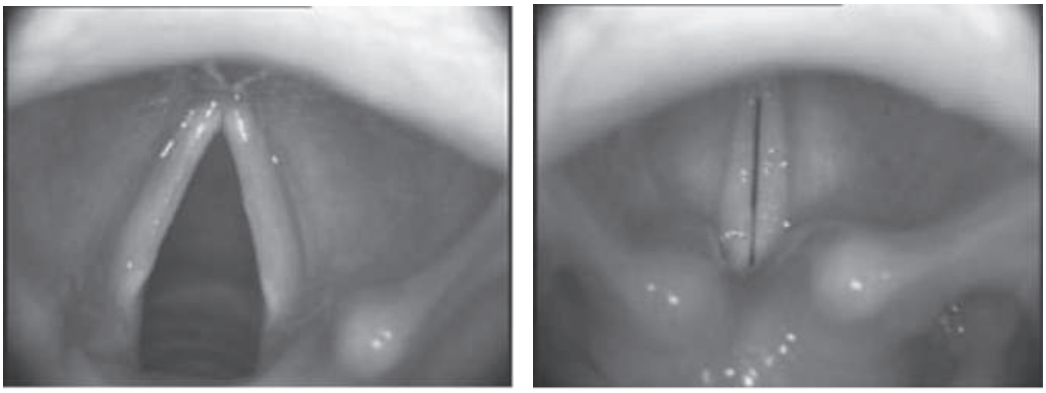

Figure 9.11. *Pictures of abducted (left) and adducted (right) vocal folds, taken via flexible endoscope*
*http://www.voicedoctor.net/media/normal-vocal-cord*

during different types of phonation or articulation (Figure 9.11). The technique involves positioning a camera in the vocal tract. With a rigid endoscope, the camera is at the end of a rigid tube that is held toward the back of the mouth, with the camera pointing downward to image the larynx. With a flexible endoscope, the tube is inserted through the nasal passages, until it passes through the velar port and hangs down in the back of the throat. The flexible endoscope thus allows direct visualization of the larynx without interfering with articulation: the subject can speak normally while images are being captured. If a numbing agent is sprayed into the nose prior to insertion, any discomfort is more psychological than physical.

This technique is probably between medium and difficult. The technology is not any more expensive than other "medium" techniques, it is pretty easily portable, and technically one does not need medical training to insert a tube up a subject's nose. It is, however, a lot more invasive than holding a transducer under a subject's chin, and is not a technique that every subject or every linguist would be comfortable with.

### 3.3.2 Magnetic resonance imaging

*Magnetic resonance imaging* (MRI) can provide the linguist with beautiful, clear pictures of the whole vocal tract. In MRI imaging, the subject is placed in a magnetic field – a large plastic tube surrounded by a huge magnet. When the magnet is turned on, all hydrogen atoms in the subject's body align to the field. A radio pulse sent to a specific depth and location is used to disrupt the field and knock the atoms out of alignment. After the pulse passes, the atoms return to alignment, but in doing so they give off energy, which is detected by the technology. The amount of energy is correlated with the amount of hydrogen, which is correlated with type of tissue and tissue density, so boundaries between different types of tissue show up crisply.

The ability to image the whole vocal tract simultaneously is especially useful. MRI can be used to create a series of images over time, although acquisition rate is

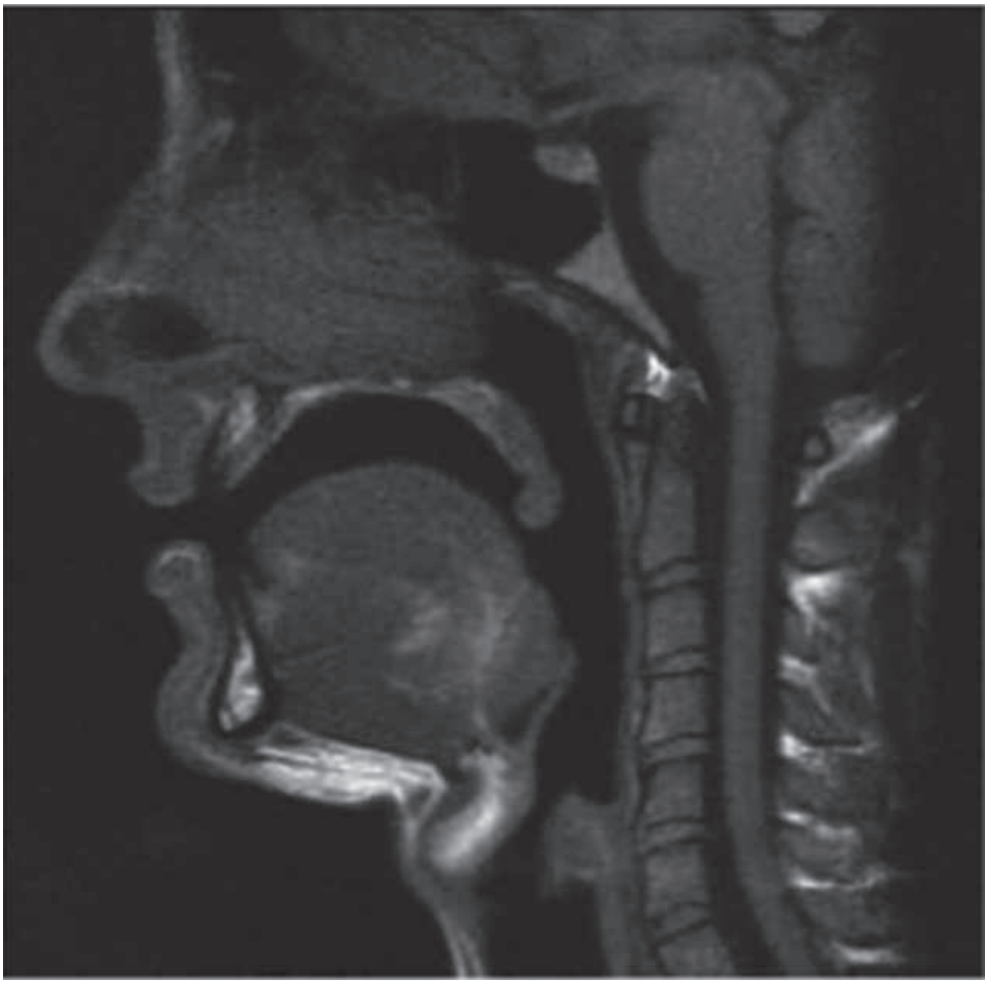

Figure 9.12. *MRI image of Portuguese [ã]*
*Martins et al. 2008*

somewhat slow as of this writing: while MRI movies are possible, the technology is mostly used to capture steady-state images. The technology also allows the linguist to visualize a slice in any dimension, showing, for example, grooving of the tongue during fricatives, or a cross-section of pharyngeal width.

The main drawback of MRI imaging is that it is very expensive – a machine is more expensive than a linguistics department could afford (even if it had the space). Linguists generally work in collaboration with a hospital or medical school, which will negotiate charges by the hour. The equipment is definitely not portable – you must bring your subjects to the lab, and not everyone is comfortable in the small tube. Also, the magnets make a lot of noise, so you cannot get good acoustics at the same time as the image.

### 3.3.3 Electromagnetic mid-sagittal articulometry

A substitute for MRI can be *electromagnetic mid-sagittal articulometry* (EMMA). This technique shows how articulator position changes over time, and can be used to determine velocity as well. In EMMA, small pellets are affixed (with non-toxic adhesive) to surfaces in the vocal tract: along the surface of the

tongue to measure tongue movement, on the lips to measure their movement, on the lower teeth to track the jaw, and on the upper teeth as a landmark. Due to the gag reflex, the back of the tongue and velum cannot be imaged.

The articulometer consists of a plastic frame that the subject sits inside (Figure 9.13, left). The frame holds three magnetic coils. As the subject speaks, the pellets move through the magnetic fields created by these coils, and pellet movement, in either two or three dimensions, can be tracked. An example movement track is shown in Figure 9.13 (right).

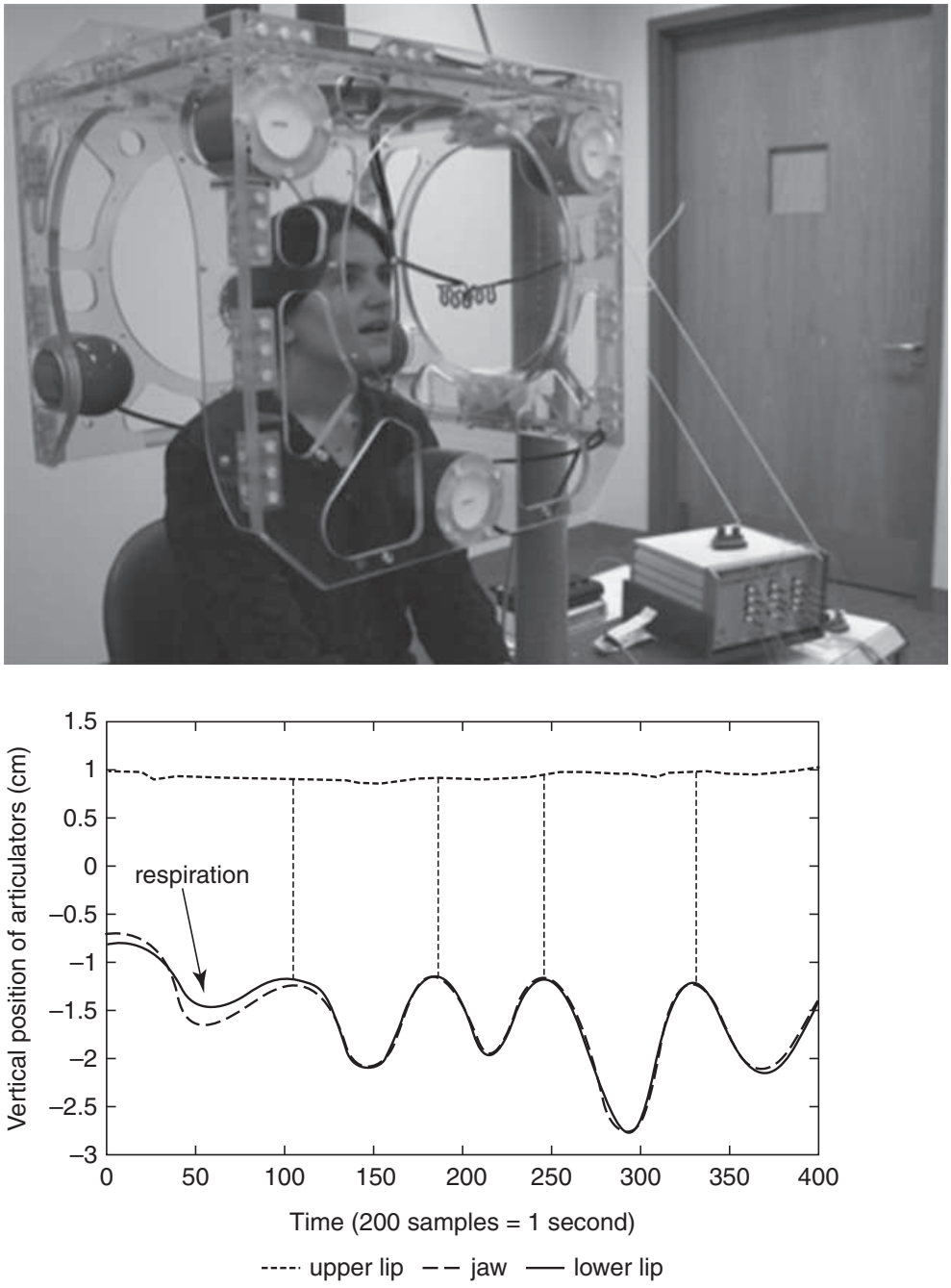


Figure 9.13. *EMMA apparatus (top); ample movement trace (bottom) http://beckman.illinois.edu/news/2007/10/100307 (top); Fagel and Clemens 2004 (bottom)*

Like MRI, EMMA can provide data from more than one articulator at a time. Unlike MRI or sonography, EMMA tracks movement of a set of specific points, rather than overall articulator shape (which can be either a plus or a minus). As with sonography, EMMA cannot directly measure contact, although contact can be inferred from changes in velocity. The technique is also somewhat invasive, similar to EPG, in that sensors must be placed inside the mouth. Additionally, the pellets sometimes fall off, and data are lost.

### 3.3.4 Electromyography

The final technique to be covered is *electromyography* (EMG). This technique directly measures electrical activity in a muscle. EMG involves inserting tiny wire probes ("hooked wire electrodes") into the muscle under examination. When the muscle contracts, the electrode picks up the electrical signal given off by the firing muscle cells, and sends the signal to a connected computer. By coordinating the EMG signal with the speech signal, a researcher can determine which muscles are contracting for which speech sounds.

This technique has been used to study laryngeal muscles and tongue muscles, and can be the only way to get information on their specific activity. What laryngeal muscles are activated during glottal opening, or pitch lowering? What tongue muscles are active during fronting? Figure 9.14 shows some sample EMG data from Thai.

Unfortunately, EMG is not pain-free for the subject. Generally, linguists only use EMG on themselves or willing colleagues, and it must be performed by a medical doctor. Even so, it can be difficult to get accurate readings. Laryngeal

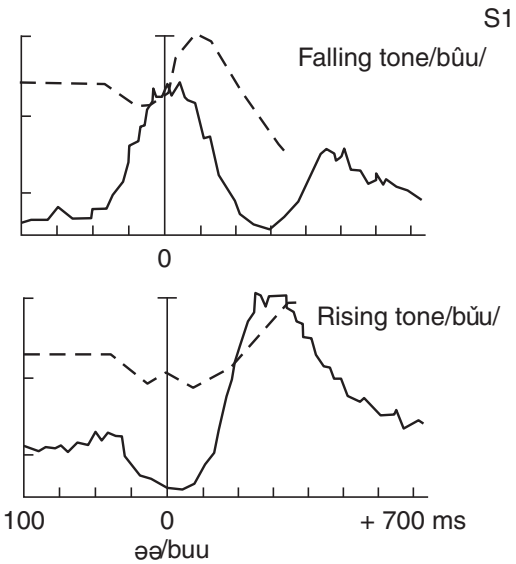


Figure 9.14. *EMG trace (solid line) shows a burst of activity in the cricothyroid muscle during pitch raising (dotted line) in Thai falling and rising tone Erickson 1976*

muscles are small and relatively inaccessible, tongue muscles are intertwined, so it can be hard to be sure that the electrode is in the right place.

Some of these more difficult techniques notwithstanding, articulatory measurements are not beyond the reach of the typical linguist or linguistics lab. And all linguists can benefit, if only by reading articulatory studies, from the information that such studies provide. For more detail about the techniques discussed here, see Gick (2002), Ladefoged (2003), and Stone (2010).

## 4 Concluding remarks

Whenever a researcher makes a sound (acoustic or articulatory) recording for the purposes of linguistic research, there are many considerations to bear in mind, and many things can go wrong. We cannot emphasize strongly enough the importance of extensive practice with the recording procedure. When recording acoustic data, the researcher should always think about how to reduce ambient noise and ensure that the microphone is sufficiently close to the speaker's mouth. Similarly, when recording articulatory data, the researcher should make sure that speakers are properly positioned with respect to the equipment. In both cases, pay special attention to ensure that speakers are comfortable (see Chapter 2).

Although sound recordings fall squarely under the purview of research methods in phonetics, their utility across the subdisciplines of our field is becoming increasingly evident. While all kinds of linguists can likely identify some useful purpose for acoustic recordings, we would like to encourage further thinking about how recording speech articulation might shed light on issues outside of phonetics proper. The fact that multiple articulatory configurations can result in similar acoustic outputs (e.g., Mielke, Baker, and Archangeli 2010; De Decker and Nycz 2012) raises questions about the nature of contrast (phonology), how children acquire such patterns (language acquisition), the role that articulatory variation might play in language change (historical linguistics), and whether such variation is socially meaningful (sociolinguistics). As the field of linguistics becomes more interdisciplinary, we hope that the methods we have discussed here will be used to address an ever expanding set of questions.

## References

Cieri, C. 2010. Making a field recording. In M. Di Paolo and M. Yaeger-Dror, eds. *Sociophonetics: A Student's Guide*. London: Routledge, 24–35.

Davidson, L. and P. De Decker. 2005. Stabilization techniques for ultrasound imaging of speech articulations. *Journal of the Acoustical Society of America* 117: 2544.

De Decker, P. and J. Nycz. 2011. For the record: which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics* 17: 51–9.

2012. Are tense [æ]s really tense? The mapping between articulation and acoustics. *Lingua* 122: 810–21.

Erickson, D. M. 1976. A physiological analysis of the tones of Thai. Unpublished Ph.D. dissertation, University of Connecticut.

Fagel, S. and C. Clemens. 2004. An articulation model for audiovisual speech synthesis: determination, adjustment, evaluation. *Speech Communication* 44: 141–54.

Fairbanks, G. 1960. *Voice and Articulation Drill Book*, 2nd edn. New York: Harper and Row.

Gick, B. 2002. The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association* 32: 113–21.

Gick, B., I. Wilson, and D. Derrick. 2013. *Articulatory Phonetics*. Malden, MA: Wiley-Blackwell.

Horiguchi, S. and F. Bell-Berti. 1987. The Velotrace: a device for monitoring velar position. *Cleft-Palate Journal* 24: 104–11.

Johnson, K. 2012. *Acoustic and Auditory Phonetics*, 3rd edn. Malden, MA: Wiley-Blackwell.

Ladefoged, P. 1996. *Elements of Acoustic Phonetics*, 2nd edn. University of Chicago Press.

2003. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Malden, MA: Blackwell.

Lawson, E., J. Stuart-Smith, and J. Scobbie. 2008. Articulatory insights into language variation and change: preliminary findings from an ultrasound study of derhoticization in Scottish English. *University of Pennsylvania Working Papers in Linguistics* 14: 102–10.

Martins, P., I. Carbone, A. Pinto, and A. Teixeira. 2008. European Portuguese MRI based speech production studies. *Speech Communication* 50: 925–52.

McGlone, R. E. and W. R. Proffit. 1972. Correlation between functional lingual pressure and oral cavity size. *Cleft Palate Journal* 9: 229–35.

Mielke, J., A. Baker, and D. Archangeli. 2010. Variability and homogeneity in American English /ɹ/ allophony and /s/ retraction. In C. Fougeron, B. Kuehnert, M. Imperio, and N. Vallee, eds. *Laboratory Phonology*, 10 vols. Berlin: Mouton de Gruyter, Volume X, 699–719.

Munhall, K. G., E. Vatikiotis-Bateson, and Y. Tohkura. 1995. X-ray film database for speech research. *Journal of the Acoustical Society of America* 98: 1222–4.

Ohala, J. J. 1993. The whole body plethysmograph in speech research. *Journal of the Acoustical Society of America* 93: 2416.

Stevens, K. N. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stone, M. 2010. Laboratory techniques for investigating speech articulation. In W. Hardcastle, J. Laver, and F. E. Gibbon, eds. *Handbook of the Phonetic Sciences*, 2nd edn. Malden, MA: Wiley-Blackwell, 9–38.

Whalen, D., K. Iskarous, M. K. Tiede, D. J. Ostry, H. Lehnert-Lehouiller, E. Bateson-Vatikiotis, and D. S. Hailey. 2005. *Journal of Speech, Language, and Hearing Research* 48: 543–53.