ARL-TR-9406 ● FEB 2022

Guidelines for Collecting Laboratory Speech Data

by Shan G Lakhmani, Kimberly A Pollard, Daniel E Forster, Andrea S Krausman, Julia L Wright, and Sean M McGhee

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9406 ● FEB 2022

Guidelines for Collecting Laboratory Speech Data

Shan G Lakhmani, Kimberly A Pollard, Daniel E Forster, Andrea S Krausman, and Julia L Wright
DEVCOM Army Research Laboratory

Sean M McGhee
DCS Corporation

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
February 2022	Technical Report	October 2020–February 2022

4. TITLE AND SUBTITLE

Guidelines for Collecting Laboratory Speech Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shan G Lakhmani, Kimberly A Pollard, Daniel E Forster, Andrea S Krausman, Julia L Wright, and Sean M McGhee

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

DEVCOM Army Research Laboratory
ATTN: FCDD-RLH-FD
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TR-9406

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

ORCID IDs: Shan Lakhmani, 0000-0001-6052-439X; Daniel Forster, 0000-0001-8351-2009; Andrea Krausman, 0000-0003-1955-8867; Julia Wright, 0000-0003-3026-1538

14. ABSTRACT

Communication is at the forefront of team research and is useful for providing insights on the team members' cognition and emotions, and on emerging group states such as team cohesion or shared situation awareness. Further, through various analysis methods, speech communication data provide insights into how information is gathered, shared, and used by team members. The goal of this report is to document guidelines and proper procedures for how to collect speech communication data in a laboratory setting. To reach this goal, this report first highlights the considerations in terms of hardware and software that are necessary for recording, storing, transcribing, and analyzing speech data. Then, we outline procedures and best practices for how to ensure data quality and describe various analysis methods that have been used in the human-autonomy teaming studies being conducted at the US Army Combat Capabilities Development Command Army Research Laboratory. A troubleshooting section is provided that describes some common problems and potential solutions. Last, we include a use case for a human-autonomy teaming simulation study and the methods and procedures used to collect, store, and analyze speech data.

15. SUBJECT TERMS

verbal communication, team communication, data collection, microphones, content analysis, prosody

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Shan Lakhmani
a. REPORT	b. ABSTRACT	c. THIS PAGE	UU	51	19b. TELEPHONE NUMBER (Include area code)
Unclassified	Unclassified	Unclassified			(407) 208-3432

Contents

List of Figures

List of Tables

Executive Summary

Speech communication provides a valuable source of data for human research and human-in-the-loop studies. However, collecting, storing, and analyzing speech data can be challenging. The steps are numerous, the equipment and software involved can be complex, and a number of decisions must be made for each study to ensure that usable data results. This report describes the steps of the process, from recording to analysis, highlighting key factors to consider. We provide explanations and recommendations for making the best choices to accommodate the unique needs of each study. We discuss hardware, including microphones, recording devices, storage, and accessories and then present information on software needs and the options available for recording, editing, transcribing, and analyzing the data. We lay out important procedures and analysis methods and end with a troubleshooting Frequently Asked Question section and a use case in which we describe the choices we made for a recent study in our lab. A key conclusion is that one size does not fit all—every study has different goals, needs, and constraints—so no one particular approach to speech data collection is suitable for all research contexts. We hope that discussing the decisions that need to be made and when certain options are most appropriate, will help researchers in their efforts to make the decisions that are best for their study.

1. Introduction

Speech is one of the most important forms of human communication. In daily life, people use speech to communicate ideas or information, express emotions, wants and needs, ask questions, connect socially, and solve problems, to name a few. Verbal communication is one of the most common ways humans convey information in teams (Nonose et al. 2015). Within teams, communication can be defined as a reciprocal process by which team members send and receive information that forms and reforms the larger team's understanding, behavior, and attitudes (Salas et al. 2015). Communication, one of the most widely studied factors in the teams literature, is a hallmark of team performance (Demir et al. 2020; Salas et al. 2015) and is critical for developing and evaluating strategies, coordinating actions, and accomplishing goals (Salas et al. 2005). Further, communication is easily observable and readily collected through unobtrusive methods and provides researchers with a window into team-level behaviors, cognitions, and performance, which allows researchers to make inferences regarding the underlying processes at work. For example, through collection and analysis of speech content during laboratory and field experiments, researchers have been able to identify how speech data are associated with team cohesion (Forster et al. 2020), team trust (Milner et al. 2020; Baker et al. 2020, 2021; Schaefer et al. 2021), workload (Funke et al. 2012), and other team states (Koolagudi and Rao 2012; Scharine 2021). Given the ease with which communication data are obtained and the relationship to processes that are difficult to capture through mere observation, the hope is that the collection and analysis of speech parameters such as content, synchrony, frequency, and prosodics will enable a more comprehensive picture of team dynamics and performance over time, beyond that which surveys or subjective measures can provide.

Although speech data are a rich source of information, there exist some challenges in the process of collecting, transcribing, and accurately analyzing natural language, which are compounded by technological and procedural hurdles. In this report, we provide a guide to collecting speech data, detailing the functions that must be performed and the tools and procedures that can be used to do so. While there are numerous contexts in which the collection of speech data can be useful, this report mainly focuses on communication between team members, in teams of two or more, working on a shared task within a laboratory setting where team members maintain a fixed, seated position and communicate via a simulated radio (e.g., push to talk).

In order to collect speech data, experimenters need hardware that can receive and store audio data from research participants. Experimenters also need software to record, edit, and transcribe the audio data, and to align speech with other relevant experimental inputs and outputs. During experimental setup, experimenters need to enact procedures to ensure clean data, accurate transcription, and a common timeline with the rest of the study data. Once the data are collected, they must be processed so that relevant analyses can be done; different analyses require different kinds of information. This report describes these requirements along with different ways to fulfill them. This report concludes with a list of frequently asked questions, along with answers to those questions, plus a case study wherein the hardware, software, and analysis requirements were addressed for a recent US Army Combat Capabilities Development Command Army Research Laboratory experiment.

2. Hardware Considerations

Collecting, storing, transcribing, and analyzing speech communication data is a multistep process and requires different types of hardware for the different steps—each step requiring considerations as to what is most appropriate for the study at hand. Fundamentally, speech recording requires three basic pieces of hardware: a device to pick up the speech (microphone), a device to record the speech (usually a computer), and a device or media on which to store the recordings (e.g., hard drives). Subsequent transcription and analysis can then occur on different devices (computers). Various types of support hardware or optional hardware accessories also require consideration (e.g., microphone stands, headsets, connectors, batteries, windscreens).

2.1 Recording Devices

The most common speech recording device used in research is simply a computer (laptop, desktop, or tablet). Recording a single audio channel is usually straightforward on a computer and can be accomplished with built-in software. Modern computers generally have more than enough processing power and random access memory (RAM) to handle routine audio recording. However, a machine could become overloaded if too many audio channels are recording at once, if sampling rates are set unusually high, or if too many other processes are running on the same machine at the same time. Pilot testing should reveal if there is an issue, and the hardware or demands can be scaled accordingly. One issue of concern, if using a computer as a recording device, is the processing power of one's computer when running all necessary software simultaneously; if the machines is overloaded, one must upgrade it (e.g., adding RAM), or reduce the number of software packages running simultaneously. Another issue to watch out for is filling up the local storage on the recording computer (see Storage Devices, Section 2.3).

2.2 Microphones

The most challenging hardware decision is likely to arise when selecting a microphone. While some recording devices, including many laptops and tablets, come with built-in microphones, an experimenter has many microphone options. Numerous commercial off-the-shelf (COTS) microphone models are available, differing in mechanism, form factor, wearing location, durability, power requirements, frequency response, directionality, noise cancellation, price, and other factors. A discussion of key factors is provided here to help researchers make the decision that is right for their study.

2.2.1 Microphone Functional Mechanism Types

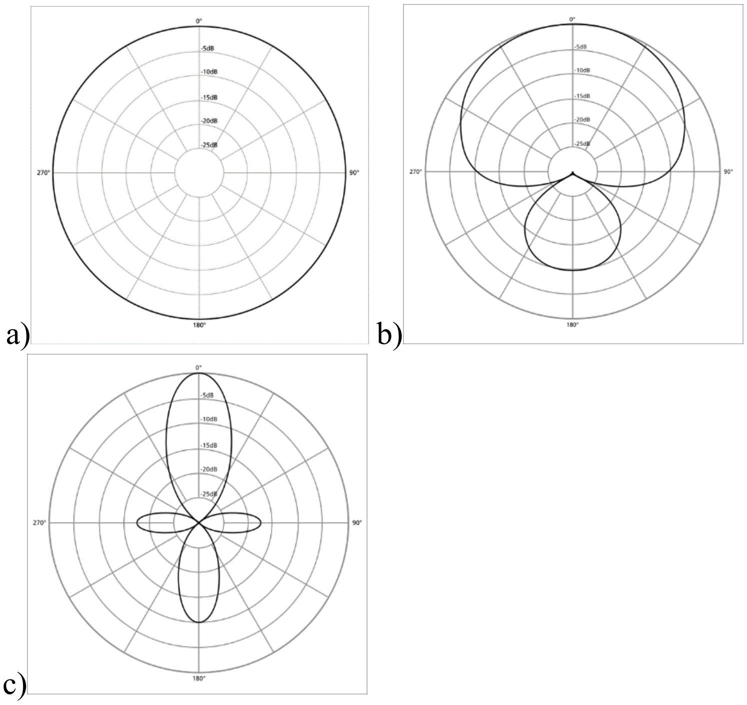
Microphones work by taking acoustic vibrations and translating these into electrical signals which can then be recorded. The mechanical details of microphone function are beyond the scope of this report, but a wide variety of mechanisms are in widespread use. Most COTS microphones used for recording speech during research will fall under the categories of dynamic or condenser microphones and are intended for picking up acoustic vibrations in air. Bone conduction microphones, as a notable exception, are intended for picking up acoustic vibrations from body tissues. Microphone types differ in their sound recording accuracy, robustness to drops and weather conditions, power needs, and cost. Some microphones will incorporate multiple technologies into the same device, including noise cancelation abilities or various features to help mitigate known weaknesses of the particular technologies they use. For these reasons, we will recommend consulting a device’s unique specifications (Sections 2.2.2–2.2.7 each describe a relevant specification to consider) to ensure that the needs of the study are met.

2.2.2 Microphone Frequency Response

Human hearing covers the frequency range of approximately 20 Hz–20 kHz, though many applications work acceptably with narrower frequency ranges. For example, telephones are restricted to 300 Hz–3.4 kHz but still yield sufficient speech intelligibility for communication. For fine detail studies of speech acoustics (e.g., some types of prosody examinations), a microphone with a full or nearly full audible frequency range and a relatively flat response across the range is desirable. This means that the microphone will translate the acoustic waveform into an electrical waveform with good fidelity, for example, without augmenting or diminishing some frequencies. A graph of such a response will have a fairly flat appearance across the desired range of frequencies. Microphone specification documentation should include estimates of the device’s frequency range and a graph of its sensitivity across that range. If speech intelligibility is the key goal (for communication or for automatic speech recognition [ASR]), a flat response is of less concern. We provide a discussion of frequency range and flatness for completeness and to guide the user on where to look for more information, but for general experimental applications this may be an excessive level of detail to consider. For general experimental applications in which speech is recorded, a microphone designed and intended for speech (such as a quality headset microphone) will likely be a suitable choice.

2.2.3 Microphone Directionality

Microphones also differ in their directionality and areas of sensitivity. This is important to consider to ensure that the desired sounds are recorded well and recording of undesired sounds is minimized. The sensitivity areas of microphones are typically drawn as polar graphs, such as those shown in Fig. 1. The graph indicates how well the microphone picks up sounds coming from different directions. A microphone with a spherical or omnidirectional polar pattern will pick up sound from all directions. An omnidirectional microphone could be an appropriate choice if the experiment is being performed in a quiet room with a single participant speaking at a time. Microphones with cardioid, supercardioid, or similar sensitivity maps are more directional. That is, they will tend to pick up sound coming from one direction better than from other directions. When the area of greatest sensitivity is pointed toward the intended speaker, that person's speech sounds will be recorded more strongly than sounds coming from other directions, all else being equal. A supercardioid or other directional microphone could be an appropriate choice (one for each participant) if the experiment is being performed in a room with multiple participants speaking simultaneously from different locations. Extreme directional microphones also exist, such as long-distance or shotgun microphones. The sensitivity graphs of these microphones are so tightly focused that sounds can be recorded well even from distant targets while diminishing the recording of sounds coming from the sides. Long-distance or shotgun mics may be desirable in the case of field studies where participants are moving extensively or at a distance. Common microphones, including headset microphones, lapel or lavalier microphones, freestanding microphones, and boom microphones generally have non-extreme polar patterns (more directional than an omnidirectional microphone but less directional than a shotgun microphone). They differ in terms of intended placement. Depending on the positioning of the speaker and the noisiness of the environment, each of these microphones may be appropriate. Microphone location is discussed below (see Section 2.2.4).



Reprinted from Galak76 via Wikimedia Commons under GNU Free Documentation License: (a) https://commons.wikimedia.org/wiki/File:Polar_pattern_omnidirectional.png [25 April 2007], (b) https://commons.wikimedia.org/wiki/File:Polar_pattern_supercardioid.png [25 April 2007], and (c) https://commons.wikimedia.org/wiki/File:Polar_pattern_directional.png [1 May 2007].

Fig. 1 Polar graphs depicting examples of different microphone directionalities: a) omnidirectional, b) supercardioid, c) highly directional lobar/shotgun. The thick black line indicates how sensitive the microphone is to sounds coming from the directions indicated by the degrees around the circle, with 0° representing the direction that the microphone is pointed and 180° representing the direction opposite where the microphone is pointed, with the microphone located at the center of the graph. The grey rings of the graph represent levels of microphone sensitivity, with the outermost ring indicating highest sensitivity and each inner ring showing a 5 decibel (dB) drop in sensitivity. Polar graphs are a 2-D representation of what is in reality a 3-D distribution of sensitivity to sound waves from different angles.

2.2.4 Microphone Wearing Locations

Of great practical importance is how the microphone is designed to be worn or where it is intended to be placed. Popular types of microphones include those attached to headsets which typically sit about an inch from the lips on a small boom, those intended to be clipped to the user's clothing on their collar or chest area ("lapel" or "lavalier" mics), and those intended to be held in stands near the participant, approximately a foot or a few feet away. Bone conduction microphones are intended to be worn in direct contact with the user's head. Other microphones may be designed to be held above participants' heads on a boom, hanging from the ceiling, held by stands at a distance, or held by personnel who will be actively following and recording the participant (e.g., shotgun mics). Some helmets, virtual reality systems, or other head-mounted displays have built-in microphones.

Microphones are also built into laptops and cell phones and can be used for some applications, although usually with some loss of quality compared to dedicated separate microphones. It is important to choose a microphone that is suitable for the experimental task setup. For example, a lavalier may be inappropriate if participants have to move around a lot, as a clothing-attached mic can pick up swishing and scratching sounds. Similarly, a head-mounted mic may be inconvenient if an electroencephalogram (EEG) cap or eye tracker will also be worn. Headset mics and lapel mics are good choices for targeting the intended speaker, as they are always close to the speaker. Freestanding mics perform well if there are no other speakers in close proximity. Highly directional long-distance or shotgun mics may be appropriate if the participant is at range, for example outdoors, especially if they are moving and active. Headset mics can also be an excellent choice in this situation if the fit is good. As always, pre-piloting is key to ensure that the mic chosen works well in the particular setup.

2.2.5 Bone Conduction Microphones

Bone conduction microphones are a special case in terms of microphone type and wearing location. Their frequency response also differs from that of most air microphones and is influenced by how the bone conduction microphones are used. Here we provide background on bone conduction and discussion of relevant considerations to help experimenters make decisions on what is best for their study. In bone conduction recording, speech sounds travel through the bones and soft tissues of the user's head and are picked up by a contact microphone worn against the head. The device must be held in direct contact with the user's skin, ideally with a static force of 200–300 gf (~1.9–2.94 N; Toll et al. 2011), which provides good contact while still being comfortable for the user. Higher force levels are generally uncomfortable. Due to impedance mismatch between body tissues and air, bone conduction mics—which are designed to pick up sound from body tissues—are relatively insensitive to background noise in the air. This makes bone conduction microphones an excellent choice for use in high noise environments. Bone conduction mics are also a good solution for picking up speech when air-conducted sound transmission is disrupted, such as when participants are wearing face masks, respirators, or other facial personal protective equipment (Levin et al. 2021; Pollard et al. 2014; Round and Isherwood 2021).

Some bone conduction microphones are designed for wear at a particular skull location, such as on the cheek in front of the ear (typically on the mandibular condyle), on the forehead or temple, or on the bone structure behind the ear (typically the mastoid process). Other bone conduction microphones are designed to fit inside the ear canal and can pick up speech signals from that location.

Different skull locations have different effects on the speech recordings. The tissues and bones that the sound must pass through, their varying impedances (Dobrev et al. 2019), and the different angles and reflections of sound as it travels from the throat, mouth, and nasal cavity to the microphone location mean that different recording locations may emphasize or degrade sound in different frequency bands (Tran et al. 2013). As a consequence, different locations yield different levels of speech intelligibility (Tran et al. 2008; McBride et al. 2011). Individual differences in vocal characteristics (such as voice fundamental frequency), demographics (such as gender), and skull morphology (such as head breadth) also influence the sound transmission and resulting speech intelligibility (McBride et al. 2008; Pollard et al. 2015; Pollard et al. 2017). In general, the forehead and mandibular condyle often yield acceptable speech intelligibility for communications applications. In-the-ear bone mics have also demonstrated good speech intelligibility (Pollard et al. 2014). Bone conduction microphones worn on the throat can yield poorer speech intelligibility because the sound is picked up largely before it has been shaped by articulators higher in the vocal tract (e.g., mouth, tongue, lips), making some phonemes hard to distinguish (Acker-Mills et al. 2006). If a bone conduction mic is selected for a particular application and yields unacceptable speech intelligibility, moving it to a different skull location, perhaps with different locations for different users, can be a helpful workaround. However, there are many factors that affect speech intelligibility, so it is important to pre-pilot to find a setup that works for the particular study conditions.

Because bone conduction microphones often yield recordings with different frequency content than air recordings and may sound muffled, they are generally not recommended if the goal of the study is to conduct detailed analyses of acoustic aspects of speech communication (e.g., prosody analyses). Air mics would be preferred in this case. There is also concern that speech intelligibility of bone conduction recordings might be low for ASRs. Major ASR algorithms are typically trained on large sample corpora of *air-conducted* speech recordings. A recognizer algorithm will likely need significantly more point-of-use training to recognize bone conducted speech at an acceptable accuracy level. Alternatively, or in conjunction, the bone conducted speech signals can be pre-processed using appropriate filtering techniques to make it more suitable for processing with ASR algorithms that were built using air-conducted speech models and air-conducted speech training sets. However, we are aware of no standard or well-accepted filters to achieve this.

2.2.6 Microphone Accessories

Many microphones (air or bone) require a power source or powered pre-amplifier to work properly. This power source may be conveniently supplied by the same cable as is used for data transmission from the microphone to the recording device. Some microphones may have a separate power cord or require a battery. It is important to check the chosen microphone's power needs and ensure the required power-related accessories are available. If recording outdoors or in an area with lots of fans or air movement, a windscreen (sometimes called a windshield or windsock) can be used to help reduce wind sounds interfering with the recordings. A variety of form factors are available. If recording indoors with participants close to a microphone, a pop filter or foam covering can be used to reduce interference from plosive puffs of air and to protect microphones from saliva.

2.2.7 Microphone and Device Connections

If participants are stationary, the preferred connection is a hard-wired one as opposed to a Bluetooth mic. While Bluetooth mics provide quality sound and freedom of movement, they are subject to latency issues not found in hard-wired connections. When there are two or more nearby devices using Bluetooth, they can interfere with each other since they use a limited number of frequency bands. For hardwired connections, a significant challenge arises when matching up different connectors and cords to get each piece of hardware to communicate as needed with each other piece. Our recommendation is to pay close attention to what types of connector ports are available on the recording device and on the microphone (and on any amplifiers, mixing boards, or other intermediaries). Few things are more frustrating than setting up for a pilot run and finding that no cables are available to connect an XLR to a USB to a Firewire. We recommend diagramming the connections if necessary to ensure all the required types of connectors and cables are acquired. Converter connectors are helpful, but it is a good idea to use as few as necessary. Extra converters and cables can pick up hum and noise that diminishes the quality of recordings, and they add more points of possible failure to the system.

2.3 Storage Devices

Audio recordings, particularly in the uncompressed or lossless file formats needed for prosodic analysis high-quality transcription, can be quite large, so it is critical to have a storage solution that can handle large amounts of data. A larger internal drive on the recording computer translates to a less frequent need to migrate the files off the working machine to free up space. External drives or network drives can be used as the main long-term storage for the recordings. However, we recommend against using external drives or network drives *during* recording, as this introduces additional potential points of failure that could ruin an experimental run (e.g., if the network goes down or a cord jiggles loose). Another critical factor to keep in mind is the nature of the data. Voice recordings are considered personally identifiable information (PII) and must be safeguarded as part of human research subject's protection procedures. Drives or external media containing participant voices are best kept password protected and/or in locked containers, and access should be limited to those permitted according to the study's Institutional Review Board (IRB) protocol.

3. Software Considerations

Although hardware varies in its ability to detect the appropriate signals, the signal processing from data storage to automated transcription is done through software. Much like hardware, software varies in its ability to effectively and efficiently process signals. Further, there will likely be several software requirements—some for recording data in a consistent format, some for filtering and boosting different frequencies, and others for translating audio data to text. To understand the appropriate software requirements and how they may affect speech data collection, we discuss each step in the signal processing chain, each of its unique challenges, the solutions offered by extant software, the shortcomings of current software solutions, and what the future may hold for automating some of the more difficult processes. Next, we will address each of these points for sound recording, postprocessing, data synchronization, and automated speech recognition software.

3.1 Sound Recording

Sound recording software helps ensure that the signals are recorded consistently at the highest quality. Understanding how this software will interface with the hardware is a critical step and will likely affect, or determine, how data are formatted and aligned with other data streams. Several commercial and open-source audio recording software packages are available to record speech data (e.g., Adobe Audition, Audacity). The packages are designed for sound editing (e.g., filtering frequencies, mixing multiple channels, normalizing volume), which can be useful for any post hoc audio processing after the data are collected, but may be more feature laden than most researchers would prefer. Furthermore, researchers interested in these capabilities can import and edit any audio file regardless of whether it was recorded natively on the software. Though recording software has major advantages in postprocessing, dealing with a user interface can become burdensome for researchers, especially when managing multiple components of data collection. Recording audio with independent software also uses more of a system's resources than using a computer script that runs in the background, which could add unnecessary demand on the system when collecting several data streams (e.g., physiological sensors, real-time behavior, creating an immersive virtual environment).

Additionally, Voice over IP (VoIP) software applications (e.g., Mumble, Ventrilo, TeamSpeak), can also be used to record audio data. These systems are typically designed for use by gamers and other individuals collaborating over software or virtual spaces, so this type of software package may be useful if the study needs to record multiple participants speaking to one another while working in a shared virtual environment. One of the primary benefits of VoIP software is it often allows one to create custom networks, enabling researchers to create unique communication systems that help address specific research problems (e.g., how information is transferred from Person A to Person C when their only way to relay information is through Person B). This type of software is also great for aligning multiple speech data streams, making it easier to understand who is speaking to whom and when. However, if the participants are not directly interacting with one another (e.g., multiple human-computer dyads, all working toward a shared goal), then the use of a VoIP may not be worth the added complexity of this software.

In addition to dedicated audio recording software, researchers should also explore the possibility of writing a simple program to record audio directly to the soundcard, without interacting with a specific user interface. Getting a computer to record audio is trivial in most programming languages and doing so gives researchers greater flexibility in exactly when and how their audio is recorded. This could be especially advantageous if a study is already using a programming language to manage other aspects of data collection. For example, the Python library `sounddevice` has a function `rec` that simply requires the duration of recording, the sample rate, and number of channels:

```
import sounddevice
from scipy.io.wavfile import write

newsound = sounddevice.rec(duration, samplerate, channels)
sounddevice.wait() # wait for recording to end
write(‘output.wav’, samplerate, newsound)
```

By recording audio using a simple script in a commonly used language, researchers should have greater flexibility in how data are recorded, stored, and aligned with other data streams, especially if those other data streams are managed using the same language. This particular script saves the audio data as a .wav file. While file compression can be used to reduce the size of the output, one must take care not to reduce the quality, as doing so removes subtle differences in prosodic information and can reduce the quality of transcriptions. To avoid this reduction in quality, we recommend keeping files in uncompressed formats such as .wav, if possible. If file size becomes a serious problem, using a lossless compression format could help, but keep in mind that not all audio software can handle every compression format.

3.2 Postprocessing

After audio is recorded, the files can be processed by applying filters for eliminating frequencies, reducing noise, and enhancing wave forms commonly associated with speech. As mentioned previously, audio files can be imported into any sound editing software (e.g., Audacity, Adobe Audition) to apply these filters. To process several audio files, one may find it easier to use a scripting language, such as Python, to apply the same set of filters to multiple audio files without interacting directly with an interface. To get the best of both worlds, one could determine the exact specification of the desired audio postprocess by using an audio editing software interface, then apply the same process to multiple audio files using a scripting language.

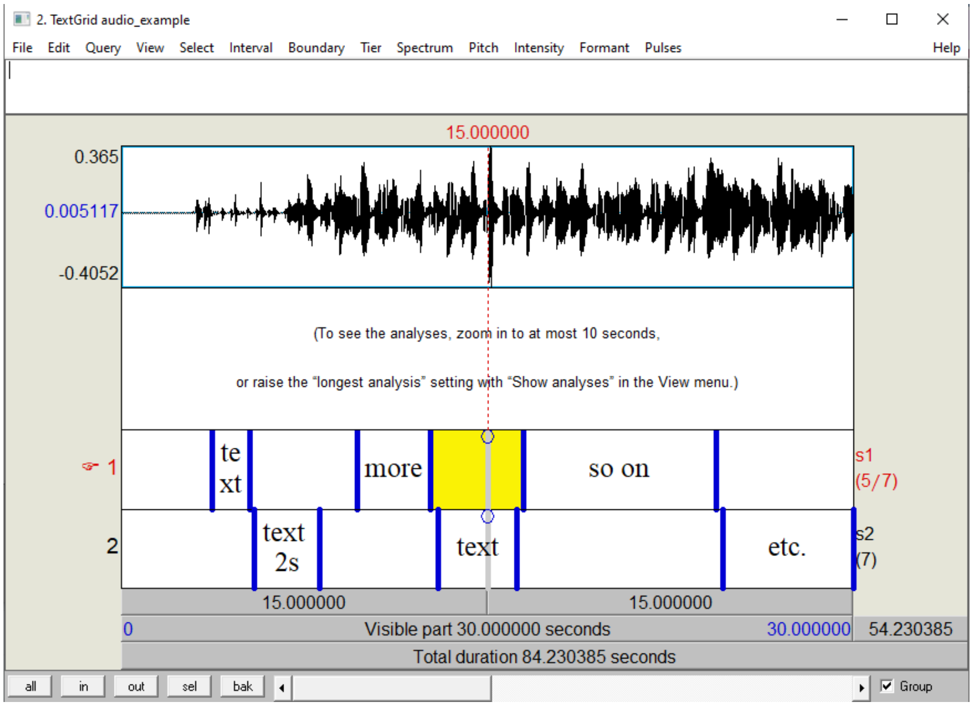
3.3 Transcription

Transcription is the process of converting speech into the corresponding written language. The decision to transcribe speech data is based on the research questions and may not always be needed. Transcription is typically done if the researcher is interested in analyzing speech content, but there are a number of other analysis methods, covered later, that do not require transcription. If transcription is needed, there are two options for speech transcription, manual and automated.

3.3.1 Manual Transcription

Manual transcription requires humans to listen to previously recorded audio files and type words as they hear them. While manual transcription is attractive because of the accuracy it provides, a major drawback is transcription time. Estimates suggest it can take approximately 4 h to manually transcribe each hour of recorded audio for one individual speaker (Britten 1995).

Because communication researchers are often interested in who is speaking to whom and when each speech act occurs, transcribers should be capable of logging these features while also accurately transcribing the content. Some software packages, such as Praat (Boersma 2001; Bonial et al. 2019; Boersma and Weenink 2021), make time stamping transcriptions a relatively trivial task by allowing users to view an audio file's wave form and add annotations at any time window within that file (see Fig. 2 for an example). Additionally, Praat allows users to create multiple channels, which can help with creating separate transcriptions for multiple speakers on a single audio file, and to add supplementary information such as paralinguistic vocalizations (e.g., sighs) and other event markers that can be detected in the audio (e.g., environmental sounds that people may respond to). To ease the manual transcription process, Praat has some basic functionality for adjusting pitch and applying filters, which may help reduce noise and add clarity to speech sounds. Once transcribed in the Praat environment, users can output a data file with transcriptions for each row, along with corresponding time stamps and channel information (e.g., speaker 1, speaker 2), which can easily be manipulated and analyzed using the various methods we discuss later. Though software platforms like Praat make the transcription process easier than it has historically been, manual transcribers will quickly learn that the work is painstaking and tedious, requiring lots of patience and dedication to ensure that each word is transcribed accurately. For a less burdensome process, researchers could also pay for a professional service that employs humans to manually transcribe audio files. For a lower cost (time and money) approach, researchers may instead turn toward automated transcription approaches.



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

Fig. 2 **Screen shot of Praat's annotation window, which allows users to select time segments corresponding to an audio file and add transcriptions or other information. Praat can add multiple "tiers" (labeled "s1" and "s2" in this example) that can be used to distinguish different types of audio information (e.g., transcripts for separate speakers).**

3.3.2 Automated Transcription

Automated transcription works by using speech recognition algorithms that are trained to perform the same task as humans—that is, identifying phonemes, grouping them into words, and grouping words into sentences. In recent years, these algorithms have seen tremendous advancements due to developments in machine learning (especially deep learning) and the increasing availability of massive databases for use as training corpora. Automated speech transcription performs the same tasks as human transcribers by analyzing audio wave forms and ascribing written symbols to represent the sounds, while assigning time stamps to each utterance. As with any automated process, the main benefit of automated transcription is that it can work with very little human intervention, thus freeing an immense amount of time and money. This is especially true when considering that automated transcription can be accomplished in real time. However, automated transcription comes with various drawbacks, which can sometimes differ dramatically between software packages. Here, we will focus on the benefits and challenges by discussing software packages offered by two companies: Nuance (Dragon NaturallySpeaking) and Google[*] (Speech-to-Text On-Prem[†]).

When selecting transcription software, it is important for researchers to consider their transcription goals pertaining to the following questions: Does transcription need to be real time? Does speech transcription need to work in tandem with other programs or work with specific programming languages? Should software work with an a priori understanding of what speech might be used (i.e., customizable dictionaries)? Should software have voice recognition capabilities, enabling it to understand an individual's speech idiosyncrasies? Do data need to be processed on local hardware or can data be processed on the cloud (whether constrained by equipment, policy, or IRB determination)? With one exception, both Dragon and Google address criteria concerning all of the aforementioned questions—that is, both software packages can transcribe in real time, can be programmed using various languages (e.g., Python), can take specialized dictionaries to help it adjust probability weights for sound-word associations, and can recognize individual voice profiles. Cost may be another factor that impacts researchers' decisions: whereas Dragon software can be a one-time purchase for anywhere from $200 to $500, Google charges nothing for the first 60 min of transcription and $0.006 for

[*] Google has released a second transcription software package (Recorder). This transcription software offers real-time, offline transcription, but it is only available on the Google Pixel phone series. While this software can still be used despite its limitations (and is used in the *Use Case for Speech Data Collection* as a backup for Dragon NaturallySpeaking), we will only be discussing the features of Google Speech-to-text in this report.

[†] Google's Speech-to-Text On-Prem can be run on localized hardware, unlike its cloud-based counterpart Google Speech-to-Text

every additional 15 s (for $200, one could pay Google to transcribe over 8000 h of speech). However, regardless of the software being used, the transcription should, to some extent, be reviewed by a human to assess the general level of accuracy and to note any consistent and relevant errors (such as repeatedly mistranscribed homonyms or consistently incorrect words or phrases). If a frequent phrase is repeatedly mistranscribed, it is important to know this so that the experimenters can judge whether that phrase is important enough to be corrected in the transcriptions. No current transcription platform is 100% accurate. The extent to which transcripts are reviewed is up to the capabilities and time available to the researchers.

3.4 Data Alignment

Aligning speech data with other data streams is essential if a goal is to understand how speech dynamics fluctuate with other study factors. Certainly, many research questions can be addressed by looking solely at speech data, such as how speech works differently across people or over time. However, if we are to understand who is speaking to whom and at what times, how speech changes following specific events, or how speech covaries over time with other measured constructs (e.g., physiology), then it is necessary to use a mechanism for aligning data. We cover a few methods for handling data alignment, from simple to sophisticated, which may depend on the complexity of data.

3.4.1 Aligning Speech from Multiple Speakers

Recording speech and understanding a sequence of events within those recordings is trivial—speech occurs over the course of time. But when recording speech from three people, for example, how can we determine when Person A is talking with Person B, or when Person C is addressing both Person A and Person B? One method for doing so is to manually align all three speech files and use human coders to determine these dynamics based on context clues from the conversation itself, or on other information that may help resolve ambiguity (e.g., whether it is unclear if Person A is talking to Person B, Person C, both, or neither). However, this may be prone to errors when multiple speakers have similar voices or when the intended recipient is unclear. To be absolutely sure about who is talking to whom and at what times, researchers should consider creating communication networks that enable participants to decide who they want to speak to and when, such as by pressing one button to speak to Person A, another to speak to Person B, and another to address to Person A and Person B (see information on VoIP, Section 3.1). However, if the hardware and programming infrastructure is not available to support communication networks, researchers could also encourage the use of call signs so that, prior to each utterance, the speaker calls to a specific recipient. When using call signs, the speaker’s identity could also be stated at the outset, providing identification of the speaker and intended recipient. The speaker could also be easily identified if each speaker is equipped with their own microphone, thus making speaker identification possible by determining which microphone had the strongest signal for a given utterance. Finally, researchers could also rely on speaker diarization algorithms, which use machine learning to separate a single audio source into multiple channels based on voice profiles.

3.4.2 Aligning Speech Data with Other Data

While the problem of aligning speech streams with other speech streams has multiple solutions, how could one align speech streams with other data streams, such as physiology or study events? This can be especially challenging because different data streams can be sampled at different rates, meaning that audio may provide 44,100 samples per second, whereas heart rate may provide only 64 samples per second. Of course, this would not be a problem if both data streams are using the same clock to provide time stamps, but researchers should not expect different sources to reference a common time. Indeed, any computational system that is tracking time can drift significantly from other systems (Marouani and Dagenais 2008), which can be difficult to retrace once data have already been collected. Therefore, if sequencing speech with several streams at short timescales is important to the researcher, it is best to address this issue in advance. Lab Streaming Layer (LSL; https://labstreaminglayer.readthedocs.io/info/intro.html), an open-source system for the unified collection of measurement time series, was developed to address the issue of aligning different data sources collected at different frequencies and being processed by different computational systems. If a study uses multiple data streams—and thus requires a common timeline and time stamps to analyze the data—a software like LSL should be used to ensure data alignment. Accounting for drift between multiple data sources is an active area of research and development (e.g., Hauweele and Quoitin 2020), though detailing every solution is beyond the scope of this report. Regardless of whether manual or automated alignment methods are used, it is always good practice to use a test data set and pilot testing to ensure all data are being recorded and are aligned properly (i.e., using the same time clock). These issues and other procedures will be covered next.

4. Procedures

As with all research, the creation of procedures for equipment setup, data collection, data coding, and analysis is critical. Established procedures offer consistency between data collectors, between participants, and certainty that the equipment is set up and operating properly each time. Following well-constructed procedures results in less data loss and reduced variability (particularly unattributed variance), which together result in improved data quality. Considering that speech data is often a mix of qualitative and quantitative data, consistency in methodology could be the difference between useful, informative data, and poor quality or lost data. This section briefly discusses procedures for three stages of data collection, configuration, collection, and storage. The details of these procedures can be selected and implemented depending on the needs of the study.

4.1 Configuration

4.1.1 Hardware Configuration

Hardware for speech data collection can include headsets, microphones, and even loudspeakers depending on how multiple individuals might interact. A hardware configuration for one participant would be very different than that for multiple participants, so care should be taken to ensure the setup is suitable.

Data collection stations should be set up so that any stand or directional microphones are placed within reach of a power source, whether an AC outlet, computer, or other power source, and in close enough proximity to pick up all voices of interest. If external microphones and speakers are used, care must be taken in their placement so that feedback is not an issue and there is no overlap in what each participant hears (each should hear only what is meant for them, no one else). Ensure that appropriate cables, power supplies, and equipment accessories are available. This is an easy step to overlook during experimental setup. If multiple people will be speaking, adjust equipment settings (e.g., microphone gain) so that the target participant's voice can be adequately picked up while avoiding picking up anyone else's voices. Make note of this configuration and all equipment settings, and note them on the checklist for data collectors to verify prior to each session. When gauges or dials have no numbers or detents, it is good practice to include a photo showing the proper position or setting (e.g., the indicator line is straight up). Pilot testing should help to identify the settings that result in the quality of data needed for the study.

All equipment positions and locations should be clearly marked in the lab space. If a piece of equipment is inadvertently moved it should be able to be placed in the exact location/position it had been previously. If the data collection is to encompass an extended period of time, best practice indicates securing stationary equipment so that it cannot be moved.

It is important to confirm that adequate data storage is available for the entirety of the study. Audio recording files can be very large, so use sample recordings to estimate the files' sizes. If the initial storage area can store at least one full session of data but not the entirety of the study's audio data, then we recommend creating a procedure for transferring the data to a secondary location at the end of each session. We also recommend backing up all sessions to an external drive as well, if possible.

4.1.2 Software Considerations

First and foremost, we recommend double-checking that all necessary software used in recording and/or transcribing speech data is installed on the devices being used. Ensure that the software version is appropriate for the hardware being used and is the same version on all devices.

It is advisable to create multiple communication channels for data collection. Different configurations allow for the analysis of different communication patterns (patterns of interaction) between various participants. Thoroughly test each channel to ensure the correct people are speaking and being heard on the correct channels and that the required data is being transmitted along the correct channels.

Information architecture should be customized so that it matches the study's needs. This includes properly labeling files and including relevant details like the source of the communication channel or station. Properly and consistently labeling the files allows easy identification when analyzing data. If making individualized voice profiles, it may be useful to set up the system or procedure to save these prints as separate files from the data.

While some transcription software can convert audio directly to text without the user providing a voice profile (e.g., Google Speech-to-Text), some transcription software (e.g., Dragon Naturally Speaking) requires the use of a voice profile. There are two approaches that can be used, either create a general voice profile or create individualized voice profiles for each participant. Individualized voice profiles have to be set up during the data collection process, and thus will be discussed in Section 4.2. A general voice profile can be set up at initial configuration using an experimenter's voice. The process of setting up a general voice profile includes writing a script (or adapting an existing script) and recording it using the existing audio system. If communication will include specialized jargon, consider using that jargon within the script. The voice profile serves as training data to improve the accuracy of the transcription software algorithm's output.

4.2 Collection

To begin collection, turn on the hardware and boot up the software. This includes turning on any special features in the hardware, such as noise reduction.

Before each experimental day or session, it is advisable to ensure that microphone preamps are on (if needed) and that microphone phantom power is being supplied (if needed). If these or any other devices are battery-powered, it is important to check that the batteries have not depleted and of course to keep fresh batteries on hand!

Researchers should confirm that all audio channels are working properly, and all team members or participants can hear and respond to one another. Depending on the system and supporting software, it may be necessary to ensure the headset and/or microphone is connected properly, and that any speech-related software is running properly.

Configurations should be checked to make sure they have not changed. Related to this, we recommend not updating any software once data collection has begun, as that might change the configurations.

A signal should be included to demarcate the beginning and end of a session. This demarcation can be done using software—having a program send a signal when an experimental testbed starts up, or can be done manually—marking time through a stopwatch or including an audio signal.

A unique and clear signal occurring simultaneously, across multiple data streams and channels, facilitates time alignment. This could include a button press triggering a message, a tone that can be picked up through the audio capture, or even a vocal phrase.

If using transcription software, creating individual voice profiles for transcription may increase transcription accuracy. If pursuing this procedure, make sure that participants read aloud from the same script and they speak steadily and clearly. This process should be done before you transcribe the audio. We recommend that you do so at the beginning of a data collection session; it only takes a few minutes to complete. Finally, confirm that these voice profiles are recorded and placed in the desired storage space.

4.3 Storage

We recommend checking each day (or before each session) to make sure that the files are recording and being saved to the correct location and that they are of the expected size.

The data should be stored and backed up in the locations (physical or virtual) prepared before the study began. It is important to remember that speech data are considered PII, so care must be taken to make sure these data are properly stored and protected.

If any of the hardware needs to be recharged (e.g., wireless headsets), then plan to recharge them between sessions.

5. Analysis

The analyses you want to run determines the type of data you need, which, in turn, affects the hardware and software best suited for collecting those data. For example, if a researcher wishes to analyze the change in participants' vocal pitch, then one needs to collect voice recordings with all the prerequisites that that entails. This section discusses a series of different analyses that can be performed with speech data and what in general needs to be done to perform these analyses.

5.1 Frequency Analyses

There are a number of different methods to assess team communication; the frequency with which team members interact with one another is a common one (Marlow et al. 2018). Here, frequency refers to the volume (amount) of communication between team members. Examples include individual message length (e.g., number of words per message), total communication volume (e.g., total number of words), and patterns of interaction (e.g., frequency of particular interactions between sender and receiver; Marlow et al. 2018; Tiferes and Bisantz 2018; Khaleghzadegan et al. 2020). To run these kinds of analyses, one needs to transcribe the speech that occurs during the study so that the aforementioned variables can be analyzed. While one can learn something from comparing communication volume between different conditions, examining patterns of interactions allows one to answer a wider scope of questions. Measuring patterns of interactions requires one to indicate who originated the speech being counted and who is the sender and receiver of the message. In addition to assessing speech frequency, one can also measure speech quality. Speech quality refers to the effectiveness and clarity of communication between team members (Marlow et al. 2018). To measure speech quality, one must segment communication into meaningful sequences and categorize them (Nonose et al. 2015). Examples include anticipation ratio (ratio of information "pushes" and "pulls") and category frequency (e.g., categorizing and counting queries and task relevant communication; MacMillan et al. 2004; Tiferes and Bisantz 2018; Khaleghzadegan et al. 2020). This work requires coding the speech that is observed so the measurement requirements include recording speech—either audio recordings or transcriptions—and a coding rubric to allow for consistent categorization.

5.2 Content Analyses

Analyzing communication content requires transcription (whether manual or automatic, see Section 3.3 for more information). Considering how flexible human language is, it can be difficult to tie a specific research question to a single linguistic analysis. Overall, there are two main approaches to content analysis: top-down and bottom-up. A top-down approach takes an a priori stance on what a word means and how it relates to a construct of interest. For example, a researcher may first propose a dictionary that contains all the known words with positive and negative sentiment, and can use this dictionary to compute sentiment scores in observed speech (e.g., total positive, total negative, total difference). A bottom-up approach, by contrast, infers words meanings based on the observations. The logic of the bottom-up approach is that words obtain their meaning based on the words with which they co-occur, so words that appear close together are likely to be relevant to the same topics and different words that occur in the same context are likely to be synonyms, or have similar meanings. Here, we provide a brief overview of some of the software and models available to analyze communication content using these two broad approaches.

5.2.1 Top-down Content Analysis

Perhaps the most widely used software for top-down content analysis is Linguistic Inquiry and Word Count (LIWC; Pennebaker et al. 2015). This software boasts hundreds of validated dictionaries and scoring metrics that measure how often words from a dictionary appear in the observed transcript. Some of the constructs covered by these dictionaries include sentiment, function words (e.g., prepositions, conjunctions), positive and negative affect, planning, and several others. The computations underlying LIWC typically rely on classifying a word into a category and outputting the percentage of times words in any given category appear in the content, making it easy for users to upload custom dictionaries to capture specific constructs. The ease of matching words in a document to words in a dictionary also means that this functionality is accessible in several software packages, though very few software packages directly compete with the size of proprietary dictionaries available in LIWC. Generally, top-down approaches are best for understanding what proportion of speech pertains to some pre-specified construct, using either validated or exploratory dictionaries. This can be beneficial for researchers who wish to generate testable hypotheses regarding speech content, as well as for researchers looking for straightforward interpretations of exploratory results.

5.2.2 Bottom-up Content Analysis

In addition to top-down content analyses, researchers may also consider bottom-up approaches, which take the content itself to extract underlying patterns of word relations. There are several bottom-up approaches, including Latent Semantic Analysis (LSA; Landauer et al. 1998), Latent Dirichlet Allocation (LDA; Blei et al. 2003), and word2vec (Mikolov et al. 2013). Generally, these methods use a framework that processes words (or, more accurately, n-grams, which can be one word, two words) in their associated documents (e.g., a speaker, an utterance, speech-relevant events). LDA and LSA, in particular, use algorithms to split word groupings across multiple dimensions, yielding different topics that represent word groups. Though word2vec also examines word associations, it relies on the use of context words to infer word meanings, such that words closer together in their numeric representation are also closer together in their meaning. To further contrast word2vec from LDA and LSA, the words underlying a similar topic from LDA, for instance, are not considered to have similar meanings, only to be mutually relevant to a similar construct. The details of how to work with these models is beyond the scope of this report. Ultimately, if choosing to analyze speech content using bottom-up approaches, researchers should be aware that these models work best with large amounts of data. However, researchers with smaller data sets could still benefit from these approaches by training their models on existing large data sets, such as online forums, Wikipedia articles, or any number of openly available natural speech databases (e.g., Litman et al. 2016). Then, researchers could apply these trained models to the speech content they obtained from their studies, which may provide more valid estimates of when people changed discussion topics or whether people used words with similar meanings. Bottom-up approaches could also be used to quantify semantic similarity (e.g., the extent to which people discussed similar topics; Babcock et al. 2014) and communication density (i.e., rate of semantic information per word; Gorman et al. 2003).

5.3 Prosody

Another important aspect of speech for researchers to consider is prosody. Prosody refers to elements of speech such as pitch changes (intonation), timing, loudness, vocal quality, and linguistic stress that convey meaning beyond the words themselves (Lausen and Hammerschmidt 2020). One way to think of it is to think of prosody as providing additional information beyond what can be extracted from a text transcript alone. In spoken English, prosody helps to distinguish a statement from a question or a sincere comment from an ironic one. It is how we signal what part of a phrase is the most important or provides new information. Prosody can convey emotion, opinion, demographics, health, conversational roles, and social identity (Banse and Scherer 1996; Kreiman et al. 2005; Cheang and Pell 2008; Xu et al. 2013' Podesva and Callier 2015; Scherer et al. 2016; Lausen and Hammerschmidt 2020). Measurable prosodic elements are numerous (e.g., see lists in Banse and Scherer 1996; Kreiman et al. 2005; Wright et al. 2019), and understanding the complex interplay of acoustic elements with one another and with situational, lexical, gestural, and listener-side information is an active field of research. We present here just a few basic examples.

5.3.1 Loudness

Speech can vary in its loudness, both across utterances and within a single utterance, and can vary within a single syllable or phoneme. The shapes and degrees of these variations can be measured and compared. For example, in Fig. 3 the relative loudness changes during the utterance ("the truck is here"). In Fig. 3a, it is loudest on the vowel in "truck" and changes according to the intensity contour shown in yellow. In this case, the speaker is stressing the word "truck," emphasizing that the truck is the thing that is here. The relative intensity (which we generally perceive as loudness) of the sound is one of the prosodic changes indicating this linguistic stress. In the second utterance (Fig. 3b), "here" shows the highest relative intensity and is from a case where the speaker is stressing the word "here," emphasizing the location of the truck.

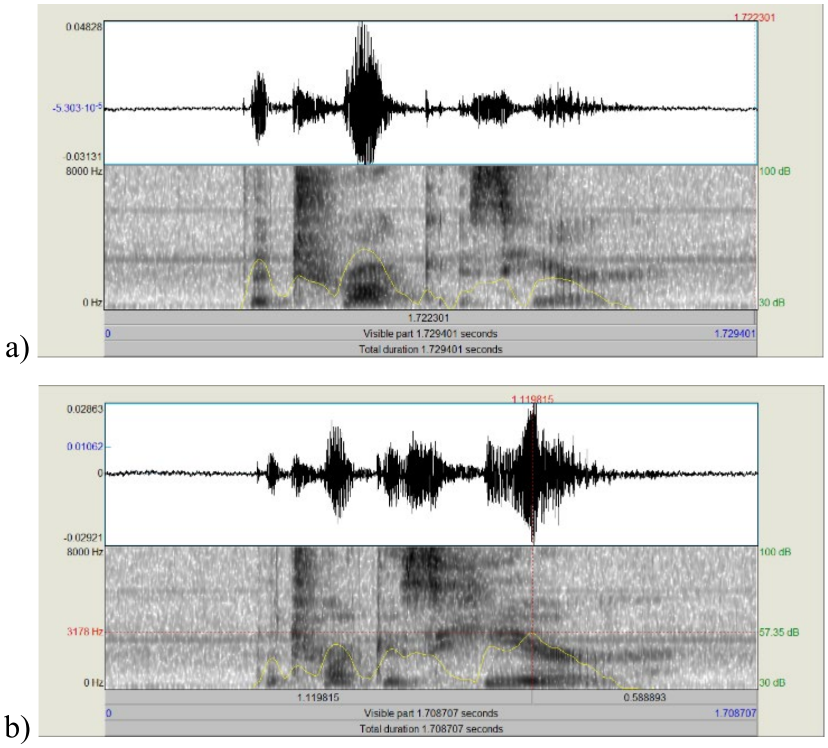


Fig. 3 Waveform and spectrogram with curves of relative intensity (which we generally perceive as loudness) marked in yellow. The top image in each is the waveform, which is a representation of signal amplitude (y-axis) vs. time (x-axis). The bottom image in each is the spectrogram, which is a representation of intensity (darker colors indicate higher intensity) vs. time (x-axis) vs. audio frequency (which we generally perceive as pitch, y-axis). Intensity is shown in decibels (dB), time is shown in seconds (s), and frequency is shown in hertz (Hz, which is cycles per second). The dB scale for the yellow intensity curve is shown on the right side of the spectrogram. The two utterances are a) "the *truck* is here" and b) "the truck is *here*". Red lines on the image illustrate a measurement taken at the point of highest relative intensity, showing it to be 57 dB occurring at 1.12 s from the start of the recording.

5.3.2 Pitch

Similarly, pitch can also vary within and across speech segments, and the shapes and degrees of these variations can be compared. In this example, the spoken utterance "the truck is here" is spoken first as a declarative statement (Fig. 4a) and second as a question (Fig. 4b). The pitch contour is shown in blue. In spoken English, a rising pitch at the end of an utterance may suggest the speaker is asking a question.

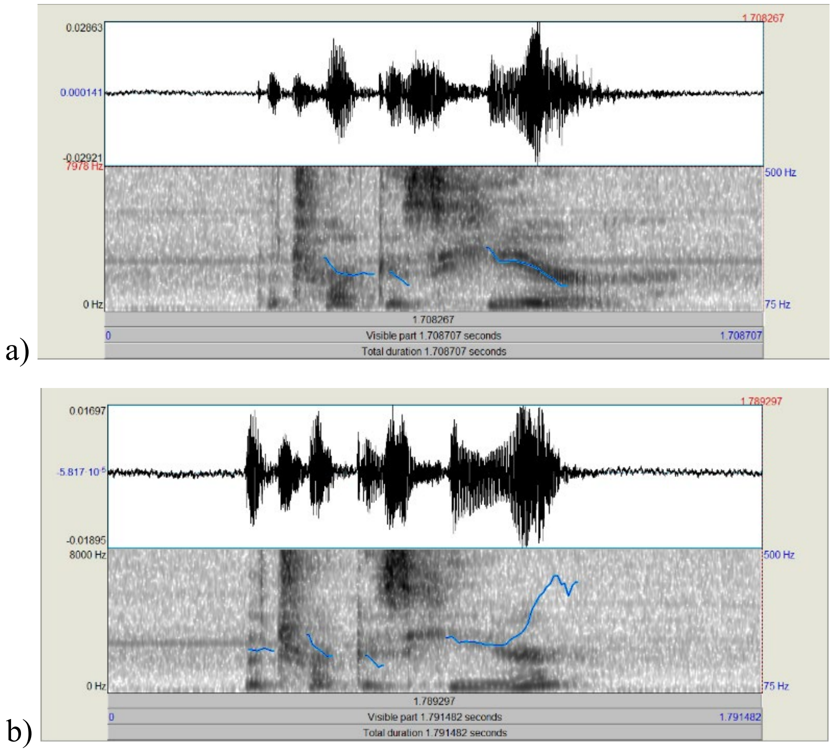


Fig. 4 Waveform and spectrogram with curves of pitch highlighted in blue. The Hz scale for the blue pitch curve is shown on the right side of the spectrogram. Note that this is a different scale than the Hz scale on the left, which is for the spectrogram as a whole. The two utterances are a) "The truck is here." and b) "The truck is here?"

5.3.3 Other Prosodic Elements

Timing (i.e., rhythm or tempo) elements in speech also vary. For example, speaking rate can indicate emotional state (Banse and Scherer 1996; Scharine 2021) or can convey sarcasm in some contexts (Rockwell 2000; Cheang and Pell 2008). The placement and length of pauses between words, sentences, or phonemes can also convey prosodic meaning, as can the duration of words or syllables (e.g., Lausen and Hammerschmidt 2020). Multiple aspects of vocal quality also can vary (Kreiman et al. 2005). Some examples include use of breathy voice or creaky voice (Fig. 5), harmonics-to-noise ratio, or fine-scale pitch variations (jitter) and fine-scale loudness variations (shimmer), which may be related to emotion, health, demographics, and social identity (Podesva and Callier 2015; Wright et al. 2019; Lausen and Hammerschmidt 2020).

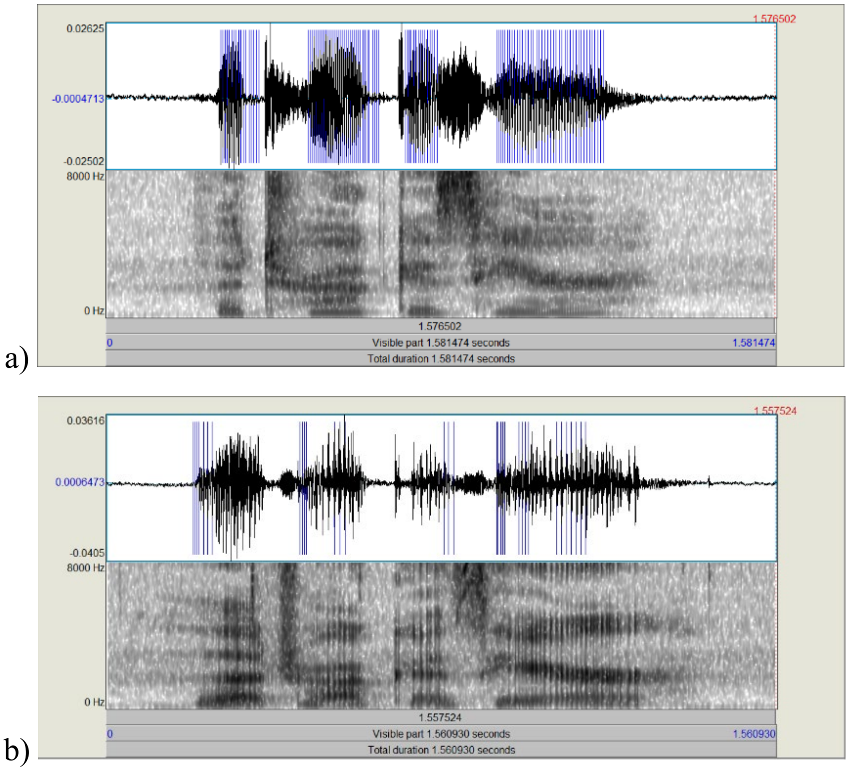


Fig. 5 Example of a) breathy voice and b) creaky voice in the phrase "the truck is here." Breathy voice and creaky voice differ in several acoustic parameters, one of which is pulse spacing, highlighted by dark blue lines in the waveform.

5.3.4 Recording for Prosody

For prosody analysis, the prime data requirement is to have quality audio recordings of speech. Achieving this requires a similar approach to obtaining quality recordings for ASR: recording the target speaker clearly and fully while minimizing the recording of cross talk from other speakers and background noise. Thus, microphones and the remaining setup should be chosen accordingly (i.e., flat response air microphones with a full audible frequency range, see Section 2.2, Microphones, for more detail) Lossless or uncompressed file formats (such as .wav) are preferred to preserve as much acoustic detail as possible. For examination of audio frequency-based prosodic parameters, the best option would be to ensure a sampling rate of at least approximately 40 kHz and to avoid use of filters during recording, if feasible given the conditions, in order to capture the broadest spectral range used in human vocal communication. Sampling rates of 44.1 and 48 kHz are standard and typically the default in recording software. Acoustic editing software, such as Audacity, Praat, or Audition can be used to cut longer recordings into manageable pieces or to extract focal words or utterances for prosody analysis. Praat or similar software tools can then be used to measure the prosodic elements of interest.

5.3.5 Measuring Prosody

Audio recordings or visualizations of the recordings (e.g., waveforms, spectra, spectrograms) are necessary to analyze prosodic elements. Perhaps the most popular software tool for examinations of prosody is Praat (Boersma 2001; Boersma and Weenink 2021). Praat is a free program and can be run on various operating systems. Praat can generate waveforms, spectra, and spectrograms for measurements (Figs. 3–5), and it has built-in features to automatically measure key variables of interest. It also allows scripting and plugins to add additional functionality, and a variety of prosody analysis scripts and collections have been developed (e.g., ProsodyPro, Xu 2013; and GSU Praat Tools, Owren 2008). Other acoustic visualization and measurement software packages are also suitable for some types of prosodic measurements (see Urbani 2011). Examples include Audacity, WinPitch, and SIL Speech Analyzer, as well as larger commercial software packages like Adobe Audition. Custom machine-learning algorithms have also been developed for prosodic comparisons (e.g., Scherer et al. 2016), and code repositories exist for acoustic speech processing (e.g., COVAREP, Degottex et al. 2014). As previously discussed, in order to measure sound frequency elements of prosody with the best accuracy, the audio data one collects should fall under the full or nearly full audible frequency range and be recorded with a relatively flat response across the range.

5.3.6 Transcribing Prosody

Many, perhaps most, prosody studies measure acoustic features directly and do not perform formal annotation. However, annotation can be useful for some analyses or to prepare a language corpus for use by other researchers. Prosodic elements can be annotated/transcribed using standardized protocols. The ToBI system (Tone and Break Indices, Silverman et al. 1992; Pitrelli et al. 1994; Beckman and Elam 1997) is a popular annotation scheme for prosody developed collaboratively by researchers across multiple fields (Wightman 2002). The intonation transcription in ToBI uses letters and symbols and is intended to be machine-readable, making annotated corpora analyzable by computerized methods. Manual annotation with ToBI can be time consuming (Syrdal et al. 2001), but there have been efforts to speed this up with automated systems (Syrdal et al. 2001; Rosenberg 2010). Other prosody annotation schemes are also in use (e.g., Breen et al. 2012).

6. Troubleshooting

This initial set of considerations can serve as a useful starting point for the development of a research study collecting speech data. Once these initial considerations have been addressed, one must transition from preparation to testing.

The process of implementing these decisions into a working research setup can result in unexpected challenges. Several common questions that arise during the implementation process and some potential solutions one can use to address them are discussed next.

Question: When using headsets for communication, what if a participant can’t hear their teammates?

Possible Solutions:

- Make sure the headset (and microphone, if separate) is properly connected to the appropriate port on the computer—this can be verified by tracing the cable from the headset to the port.
- Make sure the headset is set as the audio output device (and input device, if relevant) in the computer’s audio settings.
- Make sure any secondary software applications (e.g., Mumble, Team Speak) are running on the computer and they are set to the correct channel.
- Make sure the microphone is NOT muted—on some headsets it is easy to inadvertently press the mute button when putting the headset on. There should be an indicator on the headset which indicates the mic is muted.
- Make sure headphone volume is at the correct and appropriate (safe) level.

Question: What if speech is not being recorded?

Possible Solutions:

- Make sure headsets are properly connected and mics are not muted.
- Make sure speech recording software is running; if it is running, try exiting and restarting the application.
- Make sure any secondary software applications (e.g., Mumble, Team Speak) are running on the computer and that they are configured correctly (e.g., set to the appropriate channel).
- Make sure all hardware is connected properly and hardware and software are compatible—may need to update software and/or software drivers.

Question: What is Crosstalk? What should I do if it becomes an issue?

Crosstalk is when voices of non-target individuals are recorded onto the focal participant’s audio track, and, if using transcription software, may also be transcribed by ASR on the focal participant’s transcripts. Crosstalk is an issue because it leads to transcript data that are contaminated by data from other participants, making participant-based comparisons difficult. If cross talk becomes an issue, there are a few solutions to try—some on the recording side, and some on the transcription or analysis side. Depending on the setup it may not be possible to eliminate cross talk completely. It may have to be addressed post hoc depending on how often it appears in the data files.

NOTE: It is best to test the potential for cross talk prior to an experiment so any adjustments or modifications to the experimental setup can be made.

Possible Solutions (for Recording):

- If running multiple participants, separate them spatially or put them in different rooms if possible, depending on the needs of the study design. If participants need to be in proximity, one could also insert a divider between them to reduce the volume of external voices being picked up through the microphones.
- Use headset mics, directional mics, noise-canceling mics, and/or bone conduction mics to further focus the recording on just the intended individual's voice. Look for mics specifically designed to be used in noisy environments (e.g., those designed for helicopter cockpits), as they are built to reduce ambient noise pollution and maximize accuracy. **NOTE:** Bone conduction mics require additional considerations. Please refer to Section 2.2.5, Bone Conduction Microphones, for discussion.
- Adjust the microphone sensitivity (gain) to minimize the likelihood that other voices in the room are picked up
- Use a push-to-talk setup for recording, rather than a continuous hot mic, especially in situations where discrete communication occurs. This will ensure that the participants' communication channels are being recorded only when they press a button to speak, which will reduce instances of other voices being recorded.

Possible Solutions (during Transcription or Analysis):

- Even with taking measures to reduce cross talk, there may be instances when it still occurs, especially when speakers must remain in close physical proximity to one another. When recording or transcribing speech from multiple speakers in the same audio file, speech separation algorithms also known as diarization algorithms may be helpful. These algorithms segment the speech into clusters categorized by speaker. These algorithms employ various processing techniques such as speech enhancement and target speaker extraction (Park et al. 2021), which would help isolate the main speaker in instances where cross talk occurs.

Question: What if the Automatic Speech Recognition (ASR)/Transcription Software performs poorly?

Poor ASR can occur when the ASR software incorrectly identifies the words spoken, leading to transcripts that do not accurately represent what was said. Very poor ASR transcript utterances may read as nonsense or babbling, making analysis difficult or even impossible.

NOTE: We should not assume that any ASR technology will afford 100% accuracy; however, accuracy will likely improve as artificial intelligence (AI) algorithms for voice recognition advance. Common errors that occur with ASR include incorrectly transcribing homonyms (e.g., piece, peace) or words that sound very similar (e.g., quarter, corner). Further, acoustically similar phonemes, such as /b/ and /d/ or /p/ and /t/ can lead to transcription errors.

Possible Solutions:

- The simplest solutions include ensuring the proper microphone placement and instructing speakers to speak clearly, enunciate their words, and to maintain an even pace. However, in many cases this is not feasible or desirable if the goal is to capture natural language in a particular context.
- Excessive background noise can also degrade speech signals. Consider mitigation if consistent with study design. Ensuring the microphone is close enough to the speaker's mouth should improve recognition accuracy. Another option is to use a directional microphone, which may help mitigate the effects of background noise (see microphone options in Section 2, Hardware Considerations).
- If the transcription software being used requires a voice profile, consider using an individualized voice profile. While ASR systems can be robust even if using a general voice profile, using individualized voice profiles for each speaker should help increase accuracy, especially if using the script uses the sorts of language and phrases that are expected for the experimental task. This can also be particularly useful if the person is speaking English in a regional accent different from that used in the ASR's training data and general profile (e.g., southern United States, Irish, Scottish). Many ASRs have modules built for different regional accents.
- A similar approach to using individual voice profiles is to train the software on individual voices prior to the experiment (also called calibration).

Periodic accuracy checks are also a good idea, regardless of what training method is used. After the first session or day of a study, read through an ASR transcript generated during the study and determine whether most of the text seems reasonable. Note: ASR will not be perfect, but you should see most utterances as something that reasonably could have been said in the study context.

- Many of the cross talk solutions mentioned previously can also help with poor ASR. If a microphone is picking up bits of speech from non-focal participants, this degraded audio might be particularly poorly transcribed. It also can muddle the participant's speech signal if they speak at the same time, leading to instances of poor ASR.

7. Use Case for Speech Data Collection

As part of its modernization efforts, the US Army is investing in advanced, disruptive combat vehicle capabilities that deliver improved crew and equipment performance through optimized application of autonomous systems and novel crew compositions. The Next Generation Combat Vehicle (NGCV) Army modernization priority seeks to improve vehicle and crew performance, crew awareness, rapid decision-making, and to reduce crew workload through teaming of Soldiers and autonomous systems. The DEVCOM Army Research Laboratory (ARL) created the Human Autonomy Teaming Essential Research Program (HAT ERP) to address the challenges associated with integrating humans and autonomous systems into teams that work cooperatively and achieve their mission. As part of the HAT ERP, DEVCOM ARL researchers have been investigating novel methods for assessing crew state in order to understand crew interactions over time and what these interactions tell us about the larger team dynamic and the team's overall performance. One promising method for understanding the team dynamic and interactions within multi-human human-autonomy teams is the analysis of crew communication during a mission. The following use case documents the specific choices we made while collecting crew communication data for a team-based simulation experiment aimed at effectively teaming humans and autonomous systems.

The use case describes a team-based study with seven crew members seated closely in the same room, communicating with one another throughout multiple scenarios (see Fig. 6). Since we were interested in transcription of communication data, we needed to minimize ambient noise in the environment as well as the potential for cross talk. Therefore, we used a COTS gaming headset with active noise reduction (ANR) and a directional boom-type microphone. Further, these headsets were hard-wired and connected directly to the computer via a USB port, so they did not require batteries or a separate charging device. Using a hard-wired headset also helped reduce the likelihood of speech latency, which can be characteristic of Bluetooth headsets, as well as interference with other Bluetooth sensors or devices being used to record other data (e.g., EEG, heart rate).

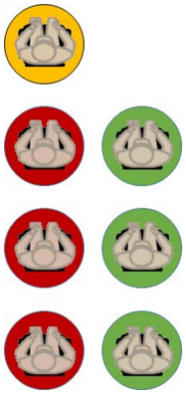


Fig. 6 Visualization of participant layout in the use case, where seven participants were seated near one another. Participant on the top left (yellow) was the leader, each participant on the left (red) was in a gunner role, and each participant on the right (green) was in a driver role.

To collect audio, a number of different software solutions were explored, including a server running VoIP software (e.g., Team Speak, Mumble) and audio recording through a Python script. The study developers settled on using a Python script to record audio to reduce the number of different software suites running simultaneously during data collection. Lossless or uncompressed audio file formats (such as .wav) are preferred if possible, and this is what we used. For best results with ASR and to allow the possibility of prosodic examination of frequency parameters, we used a 44.1-kHz sampling rate. No additional audio cleanup or postprocessing was performed.

Given the large amount of speech data to be processed, different transcription software options were considered, but ultimately Dragon Naturally Speaking was selected for real-time transcription. This decision was based on previous work indicating that Dragon provides reasonable accuracy rates, Dragon is readily available, and it a COTS product that is able to transcribe in real time. Using Dragon thus allowed time savings compared to manual transcription or post hoc automatic transcription (Krausman et al. 2019).

Additionally, Dragon was able to provide more accurate time stamps (through writing custom software to leverage the log files it generated), it ran on PC without need of an emulator, and it could run without an Internet connection. Dragon was used for real-time audio transcription so that transcripts would be immediately available for analysis. However, the recorded .wav files were also processed through Google Recorder* post hoc to check against Dragon's accuracy. Because Google Recorder is only available on the Pixel phone series, this process had to be done acoustically (microphone next to speaker) and in real time, which made it a lengthy process, but yielded more accurate transcriptions than Dragon.

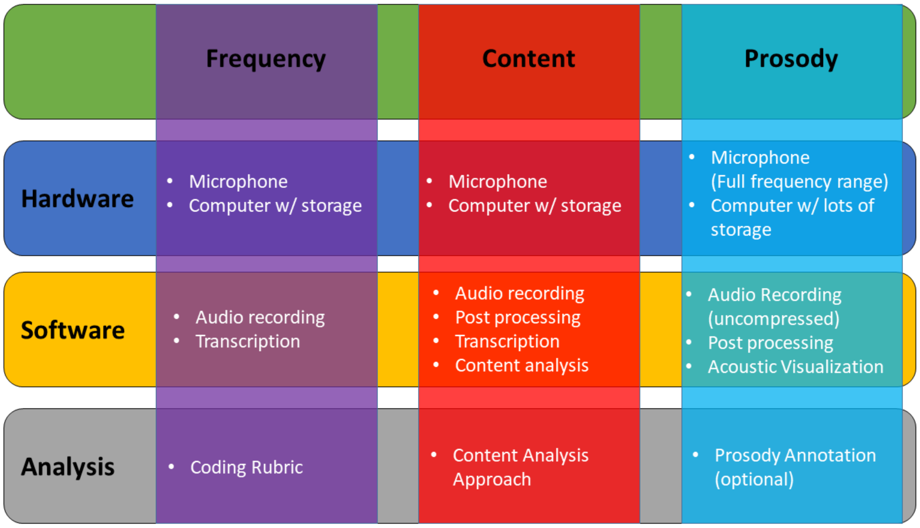
Prior to the start of experimental sessions, a communications check was done to confirm that all channels were working and all team members could hear one another only on the appropriate channels. For the communication check, each crew member pressed the "push to talk" button and uttered a brief phrase such as "this is Red One on section net" to the other team members who would raise their hand to verify they heard the speaker. This procedure was followed for all crew positions for each of the different communication channels.

All crew members were provided with a PowerPoint illustration of the study's input interface (a yoke/steering wheel) that labeled buttons used for the push-to-talk functionality and identified which buttons corresponded to the different communication channels.

Once the data were collected, individual transcripts were processed so that analyses could be done. Communication volume measures (e.g., number of interactions, timing for information requests) were tallied to determine who was communicating and how often/much. These transcripts were also used to run dictionary-based analyses through LIWC, looking at the kind of language used (e.g., tone, descriptiveness). We were interested in speech frequency and content, so we used the aforementioned hardware and software solutions; see Fig. 7 for the decisions that would have been made if we had been interested in speech quality or prosody.

Transcripts, acoustic recordings, and multiple other data streams (including performance measures and questionnaires) are available for future communications analyses for this study.

* At the time, Google's Speech-to-Text software required the use of their cloud platform. Using software that requires use of a cloud platform was not feasible due to privacy and clearance issues. Future researchers are advised to use Google's Speech-to-Text On-Prem instead of Google Recorder.



	Frequency	Content	Prosody
Hardware	- Microphone - Computer w/ storage	- Microphone - Computer w/ storage	- Microphone (Full frequency range) - Computer w/ lots of storage
Software	- Audio recording - Transcription	- Audio recording - Post processing - Transcription - Content analysis	- Audio Recording (uncompressed) - Post processing - Acoustic Visualization
Analysis	- Coding Rubric	- Content Analysis Approach	- Prosody Annotation (optional)

Fig. 7 Suggested prerequisites for collecting speech data, organized by data category. Columns detail the category of data collected, rows describe the category of hardware or software needed, and cells list suggested items to collect the desired communication data.

8. Conclusion

Speech communication is a core part of human interaction and is critical for many team processes. As such, laboratory-collected speech communication data can provide rich insights into the ways people interact and work together. These data can reveal otherwise hidden states such as emotion, attitudes, cognition, situation awareness, and team cohesion. Acquiring speech communication data is a complex process, however. It involves multiple steps, each step laden with details to consider and numerous decisions to make. Our aim in this report is to help make this process a bit easier. In this report, we laid out the steps of the speech data collection pipeline from recording to analysis, and discussed key considerations for each step. We provided a review of the issues and requirements, along with recommended guidelines, best practices, and rationale gained from our combined experience, as well as insights gained from others who have done similar work; for further detail, see the recommended reading list in Table 1. We organized the considerations into four major categories: hardware, software, procedure, and analysis. We included a troubleshooting section and a case study of a recent research project for which we collected and analyzed speech communication data.

Every study is unique, so each study must handle speech data collection in its own way. We hope that reviewing the considerations here, along with the Frequently Asked Questions and the use case, will facilitate the decision-making process as researchers plan exciting new speech communication studies.

Table 1 Recommended reading for topics related to collecting communication data

Topic	Citation
Transcription: Dragon	Krausman et al. 2019
Transcription: Praat	Boersma 2001; Bonial et al. 2019
Speech Frequency and Speech Quality	Marlow et al. 2018
Content Analysis: LIWC	Pennebaker et al. 2015
Content Analysis: LSA	Landauer et al. 1998
Content Analysis: LDA	Blei et al. 2003
Content Analysis: word2vec	Mikolov et al. 2013
Prosody	Kreiman et al. 2005

9. References

Acker-Mills BE, Houtsma AJM, Ahroon WA. Speech intelligibility in noise using throat and acoustic microphones. Aviat Space Environ Med. 2006;77(1):26–31.

Babcock MJ, Ta VP, Ickes W. Latent semantic similarity and language style matching in initial dyadic interactions. J Language Soc Psychol. 2014;33:78–88.

Baker AL, Fitzhugh SM, Forster DE, Brewer RW, Krausman A, Scharine A, Schaefer KE. Team trust in human-autonomy teams: analysis of crew communication during manned-unmanned gunnery operations. CCDC Army Research Laboratory; 2020 May. Report No.: ARL-TR-8969.

Baker AL, Fitzhugh SM, Huang L, Forster DE, Scharine A, Neubauer C, Lematta G, Bhatti S, Johnson C, Krausman A, et al. Approaches for assessing communication in human-autonomy teams. Human-Intelligent Syst Integration 2021;3:99–128. https://doi.org/10.1007/s42454-021-00026-2.

Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Personality Soc Psychol. 1996;70(3):614–636.

Beckman ME, Elam GA. Guidelines for ToBI labeling, version 3. Ohio State University; 1997.

Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Machine Learn Res. 2003;3:993–1022.

Boersma P. Praat, a system for doing phonetics by computer. Glot International. 2001;5(9/10):341–345.

Boersma P, Weenink D. Praat: doing phonetics by computer. Version 6.1.51. 2021 [accessed 2021 Aug 20]. http://www.praat.org/.

Bonial C, Henry C, Artstein R, Marge M. Transcription guidelines for Army Research Laboratory (ARL) human-robot dialogue corpus. CCDC Army Research Laboratory; 2019 Oct. Report No.: ARL-TR-8832.

Breen M, Dilley LC, Kraemer J, Gibson E. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). Corpus Linguistics and Linguistic Theory. 2012;8(2):277–312. https://doi.org/10.1515/cllt-2012-0011.

Britten N. Qualitative interviews in medical research. British Med J. 1995;311:251–253

Cheang HS, Pell MD. The sound of sarcasm. Speech Comm. 2008;50(5):366–381. https://doi.org/10.1016/j.specom.2007.11.003.

Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP - A collaborative voice analysis repository for speech technologies. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014. p. 960–964. https://doi.org/10.1109/ICASSP.2014.6853739.

Demir M, McNeese NJ, Cooke NJ. Understanding human-robot teams in light of all-human teams: aspects of team interaction and shared cognition. Int J Hum Comput Stud. 2020;140:102436. doi:https://doi.org/10.1016/j.ijhcs.2020.102436.

Dobrev I, Sim JH, Pfiffner F, Huber AM, Röösli C. Experimental investigation of promontory motion and intracranial pressure: stimulation site and coupling type dependence. Hearing Res. 2019;378:108–125.

Forster DE, McGhee SM, Perelman BS, Neubauer C, Schaefer KE, Krausman A. Human-autonomy teaming: using latent semantic analysis for assessing team cohesion from communication. CCDC Army Research Laboratory; 2020 Sep. Report No.: ARL-TR-9067.

Funke GJ, Knott BA, Salas E, Pavlas D, Strang AJ. Conceptualization and measurement of team workload: a critical need. Hum Factors. 2012;54:36–51.

Gorman JC, Foltz PW, Kiekel PA, Martin MJ, Cooke NJ. Evaluation of latent semantic analysis-based measures of team communications content. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2003;47(3):424–428.

Hauweele D, Quoitin B. Toward accurate clock drift modeling in wireless sensor networks simulation. Comput Comm. 2020;163:1–11.

Khaleghzadegan S, Kazi S, Rosen MA. Unobtrusive measurement of team cognition: A review and event-based approach to measurement design. Contemporary Res. 2020;95–113.

Koolagudi SG, Rao KS. Emotion recognition from speech: a review. Int J Speech Tech. 2012;15:99–117. https://doi.org/10.1007/s10772-011-9125-1.

Krausman A, Kelley T, McGhee S, Schaefer KE, Fitzhugh S. Using Dragon for speech-to-text transcription in support of human-autonomy teaming research. CCDC Army Research Laboratory; 2019 Nov. Report No.: ARL-TN-0978.

Kreiman J, Vanlancker-Sidtis D, Gerratt BR. Perception of voice quality. In: Pisoni DB, Remez RE, editors. The handbook of speech perception. Blackwell Publishing Ltd; 2005. p. 338–362. https://doi.org/10.1002/9780470757024.ch14

Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes. 1998;25(2–3):259–284.

Lausen A, Hammerschmidt K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. Humanities Soc Sci Comm. 2020;7(1):2. https://doi.org/10.1057/s41599-020-0499-z.

Levin M, Zhou K, Sommer E, McHugh T, Sommer D. Ambient noise levels and wireless headsets for communication in aerosolizing otolaryngology surgery during COVID-19. Otolaryngology Head Neck Surg. 2021;1–4.

Litman D, Paletz S, Rahimi Z, Allegretti S, Rice C. The teams corpus and entrainment in multi-party spoken dialogues. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov. p. 1421–1431.

MacMillan J, Entin EE, Serfaty D. Communication overhead: the hidden cost of team cognition. In: Salas E, Fiore SM, editors. Team cognition: understanding the factors that drive process and performance. American Psychological Association; 2004. p. 61–82. https://doi.org/10.1037/10690-004.

Marlow SL, Lacerenza CN, Paoletti J, Burke CS, Salas E. Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. Org Behav Hum Decision Process. 2018;144:145–170.

Marouani H, Dagenais MR. Internal clock drift estimation in computer clusters. J Comput Syst Network Comm. 2008;2008:583162.

McBride M, Hodges M, French J. Speech intelligibility differences of male and female vocal signals transmitted through bone conduction in background noise: implications for voice communication headset design. Int J Indust Ergon. 2008;38(11–12):1038–1044.

McBride M, Tran P, Letowski T, Patrick R. The effect of bone conduction microphone locations on speech intelligibility and sound quality. Appl Ergon. 2011;42:495–502.

Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv preprint arXiv:1301.3781.

Milner A, Han Seong D, Brewer R, Baker AL, Krausman A, Chhan D, Thomson R, Rovira E, Schaefer KE. Identifying new team trust and team cohesion metrics that support future human-autonomy teams. In: Cassenti D, Scataglini S, Rajulu S, Wright J, editors. Advances in Simulation and Digital Human Modeling. AHFE 2020. Advances in Intelligent Systems and Computing; 2020. Vol. 1206. Springer, Cham. https://doi.org/10.1007/978-3-030-51064-0_12.

Nonose K, Kanno T, Furuta K. An evaluation method of team communication based on a task flow analysis. Cogn Tech Work. 2015;17:607–618. https://doi.org/10.1007/s10111-015-0340-4.

Owren MJ. GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software. Behav Res Meth. 2008;40(3):822–829. https://doi.org/10.3758/BRM.40.3.822.

Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S, Narayanan S. A review of speaker diarization: recent advances with deep learning. 2021. arXiv preprint arXiv:2101.09624.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. University of Texas at Austin; 2015.

Pitrelli JF, Beckman ME, Hirschberg J. Evaluation of prosodic transcription labeling reliability in the ToBI framework. 1994; 123–126.

Podesva RJ, Callier P. Voice quality and identity. Ann Rev Appl Linguistics. 2015;35:173–194. https://doi.org/10.1017/S0267190514000270.

Pollard KA, Garrett L, Tran P. Bone conduction systems for full-face respirators: speech intelligibility analysis. Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6883.

Pollard KA, Tran PK, Letowski T. The effect of vocal and demographic traits on speech intelligibility over bone conduction. J Acoustical Soc Am. 2015. 2015;137(4):2060–2069. https://doi.org/10.1121/1.4916689.

Pollard KA, Tran PK, Letowski T. Morphological differences affect speech transmission over bone conduction. J Acoustical Soc Am. 2017;141(2):936–944. https://doi.org/10.1121/1.4976001.

Rockwell P. Lower, Slower, louder: vocal cues of sarcasm. Journal of Psycholinguistic Res. 2000;29(5):483–495.

Rosenberg A. AuToBI – A tool for automatic ToBI annotation. Proceedings of INTERSPEECH. 2010:146–149.

Round M, Isherwood P. Speech intelligibility in respiratory protective equipment - implications for verbal communication in critical care. Trends Anaesthesia Critical Care. 2021;36:23–-29.

Salas E, Shuffler ML, Thayer AL, Bedwell WL, Lazzara EH. Understanding and improving teamwork in organizations: a scientifically based practical guide. Hum Resource Manage. 2015;54(4):599–622.

Salas E, Sims DE, Burke CS. Is there a "big five" in teamwork? Small Group Res. 2005;36:555–599. doi:10.1177/1046496405277134.

Schaefer KE, Brewer RW, Baker AL, Krausman A, Neubauer C, Fitzhugh S, Forster D, Chhan D, Milner A, Seong DH, et al. Wingman joint capabilities technology demonstration: trust metrics for manned-unmanned lethality teams. DEVCOM Army Research Laboratory; 2021 Apr. Report No.: ARL-TR-9182.

Scharine A. Development of a neural network algorithm to detect Soldier load from environmental speech. In: Wright JL, Barber D, Scataglini S, Rajulu SL, editors. In: Advances in Simulation and Digital Human Modeling. AHFE 2021. Lecture Notes in Networks and Systems, vol 264. Springer, Cham; 2021. https://doi.org/10.1007/978-3-030-79763-8_7.

Scherer S, Gratch J, Rizzo A, Morency L-P. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. IEEE Trans Affective Comput. 2016;7(1):59–73. https://doi.org/10.1109/TAFFC.2015.2440264.

Silverman K, Beckman M, Pitrelli J, Ostendor M, Wightman C, Price P, Pierrehumber FJ, Hirschberg J. ToBI: a standard for labeling English prosody. 1992;867–870.

Syrdal AK, Hirschberg J, McGory J, Beckman M. Automatic ToBI prediction and alignment to speed manual labeling of prosody. Speech Comm. 2001;17.

Tiferes J, Bisantz AM. The impact of team characteristics and context on team communication: an integrative literature review. Appl Ergon. 2018;68:146–159.

Toll LE, Emanuel DC, Letowski T. Effect of static force on bone conduction hearing thresholds and comfort. Int J Audiol. 2011;50:632–635.

Tran PK, Letowski TR, McBride ME. The effect of bone conduction microphone placement on intensity and spectrum of transmitted speech items. J Acoustical Soc Am. 2013;133(6):3900–3908.

Tran P, Letowski T, McBride M. Bone conduction microphone: head sensitivity mapping for speech intelligibility and sound quality. Proceedings of International Conference on Audio, Language and Image Processing (ICALIP 2008); 2008. p. 107–111.

Urbani M. Instruments and methods for the analysis of prosody. Linguistics Appl. 2011;4:104–115.

Wightman C. ToBI or not ToBI? Proceedings of Speech Prosody. 2002;1:25–29.

Wright R, Mansfield C, Panfili L. Voice quality types and uses in North American English. Anglophonia. 2019;27. https://doi.org/10.4000/anglophonia.1952.

Xu Y. ProsodyPro—A tool for large-scale systematic prosody analysis. Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013). 2013;7–10.

Xu Y, Lee A, Wu W-L, Liu X, Birkholz P. Human vocal attractiveness as signaled by body size projection. PLoS ONE. 2013;8(4):e62397. https://doi.org/10.1371/journal.pone.0062397.

List of Symbols, Abbreviations, and Acronyms

2-D	two-dimensional
3-D	three-dimensional
AC	alternating current
AI	artificial intelligence
ANR	active noise-reduction
ARL	Army Research Laboratory
ASR	automatic speech recognition
COTS	commercial off-the-shelf
DEVCOM	US Army Combat Capabilities Development Command
EEG	electroencephalogram
HAT ERP	Human Autonomy Teaming Essential Research Program
IRB	Institutional Review Board
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LSA	Latent Semantic Analysis
LSL	Lab Streaming Layer
NGCV	Next Generation Combat Vehicle
PII	personally identifiable information
RAM	random access memory
ToBI	Tone and Break Indices
USB	Universal Serial Bus
VoIP	Voice over IP

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLD DCI
TECH LIB

1 DEVCOM ARL
(PDF) FCDD RLH B
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL
35898-7290

1 DEVCOM ARL
(PDF) FCDD HSI
J THOMAS
6662 GUNNER CIRCLE
ABERDEEN PROVING
GROUND MD 21005-5201

1 USN ONR
(PDF) ONR CODE 341 J TANGNEY
875 N RANDOLPH STREET
BLDG 87
ARLINGTON VA 22203-1986

1 USA NSRDEC
(PDF) RDNS D D TAMILIO
10 GENERAL GREENE AVE
NATICK MA 01760-2642

1 OSD OUSD ATL
(PDF) HPT&B B PETRO
4800 MARK CENTER DRIVE
SUITE 17E08
ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

19 DEVCOM ARL
(PDF) FCDD RLH
J LANE
Y CHEN
P FRANASZCZUK
K MCDOWELL
FCDD RLH F
K OIE
J GASTON
FCDD RLH FA
G BOYKIN
A W EVANS
D FORSTER
FCDD RLH FB
J GARCIA (A)
H ROY
FCDD RLH FC
S MCGHEE
K POLLARD
T ROHALY
J TOURYAN (A)
FCDD RLH FD
A MARATHE
S LAKHMANI
A KRAUSMAN
J WRIGHT