SUGGESTIONS FOR COMPUTER-BASED AUDIO RECORDING OF SPEECH SAMPLES FOR PERCEPTUAL AND ACOUSTIC ANALYSES

Phonology Project Technical Report No. 13

Michael R. Chial
Department of Communicative Disorders
University of Wisconsin–Madison

October 2003

Phonology Project, Waisman Center,
University of Wisconsin–Madison

Preparation of this report was supported by research grant DC00496 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health (Lawrence D. Shriberg, P.I.)

INTRODUCTION

Audio engineers commonly refer to recording and reproduction systems as “chains,” an apt designation because it invites attention to links. Although digital management of audio information offers improvements in signal quality over analog methods, digital systems are not without problems. Aliasing errors, sampling rate jitter, amplitude distortion, intermodulation distortion, spurious output signals, inter-channel cross-talk, inter-channel phase distortion, idle channel noise, and delay distortion can occur and are the subject of technical standards (Audio Engineering Society, 1998). Because digital audio systems by definition accept and produce analog signals, several purely analog issues remain relevant.

Minimally, the digital audio chain includes a microphone, an A-to-D converter, immediate signal storage, recording devices and media, reproduction, a D-to-A converter, some form of audio output, and interconnections among the physical components of the chain. This summary is similar to that offered by Bunta, Ingram, and Ingram (2003) in their review of computerized data collection and analysis of speech and language samples. Rather than duplicate their recommendations, this report emphasizes technical and practical ideas borrowed from professional recording. Possibly unfamiliar terms are italicized and defined in the glossary. A very complete dictionary of professional audio terminology is available online (Rane Corporation, 2003).

The following discussion was motivated by a study comparing analog to digital recording in childhood speech sound disorders (Shriberg et al., 2003). It is hoped that the observations and recommendations noted here will generalize to other applications in communicative sciences and disorders.

DISCUSSION

Signal Sensing

Most modern microphones are based upon magnetic (dynamic) or capacitive (condenser) transduction principles. Condenser devices are inherently high impedance units that require pre-amplification powered either by internal batteries or by *phantom power* supplied by other devices such as recorders, mixers, or microphone amplifiers. Condenser units are preferred in many professional applications because they are more versatile. Dynamic and condenser microphones can be connected to other devices using either unbalanced lines or balanced lines. Unbalanced lines use only two conductors and are subject to electronic noise. *Balanced lines* (highly recommended) use

two conductors for the audio signal and a third as a shield. They are common in professional applications because they reduce electromagnetic noise as well as the stray inductive and capacitive effects of connecting cables (see Fause, 1995; Macatee, 1995; Muncy, 1995; Whitlock, 2002).

Relevant microphone specifications include directional pattern (omni-directional, hemispheric, cardioid, figure-8 or bi-directional, super-cardioid, and hyper-cardioid), sensitivity (expressed in millivolts per pascal), frequency response (in hertz), self-noise (in millivolts), distortion level (in SPL), and dynamic range (in decibels). Directional microphones (cardioid, super- and hyper-cardioid) typically produce increased sensitivity to low-frequency signals (the *proximity effect*).

Specialized devices include coincident and spaced array units for stereo and multi-channel recording; shotgun or parabolic reflector units for long distance recording; boundary-effect, pressure zone microphones (PZM) or wireless microphones for middle distance recording; and close-talk, noise-excluding units for difficult environments. Several of these have distinct advantages in studies of children's speech (e.g., some wireless and some PZM units). See Eargle (2001) and Ballou (2002) for comprehensive reviews of microphone technology and methods of use.

Microphone selection should consider not only device specifications, but also the environment in which the microphones are used, who will use them, how they will be used, and the other equipment with which they will be used. Omni-directional devices offer the least spectral coloration of target signals but capture non-target signals. The same is true for highly sensitive microphones, whether directional or not. Large, obtrusive microphones may cause talkers to behave in uncharacteristic ways. Except in well-engineered laboratory situations, it is unlikely that a single microphone will meet all needs adequately. Professional quality condenser microphones using balanced lines and phantom power are available for less than $300.

Microphone Interface Issues

If balanced-line, phantom-powered microphones are used (as recommended), some means must be found to provide appropriate power and connectors. Preferred solutions offer other advantages as well. Excellent stand-alone, phantom-powered microphone preamplifiers, combined with A-D and D-A converters as well as both analog (line-level) and digital (*USB* or *Firewire)* outputs and inputs, are available for about $250 or less.

Local Environment Control

Goals for the local control of the recording environment include both freedom from noise and short reverberation times. Ambient sound levels can be monitored with inexpensive sound level meters. Levels above 40 dB(A) are probably excessive. A simple test of reverberation is a handclap: if the sound of the clap audibly persists, there is too much reverberation.

The Problem of Digital Decisions

Some problems associated with digital audio are less a matter of technology than of user choice. Others involve device and data compatibility. Apart from technology issues (though still influenced by them) is the matter of standards by which audio data can be stored and shared while increasing confidence that instrumentation does not become an impediment. Related challenges involve organization and control of data archives.

User decisions include sampling rate and data word width (or length). Sample rates (the frequency at which analog signals are converted to discrete digital amplitude values) range from 5 kHz to 192 kHz. Generally, the highest frequency component of a digitized signal will be one-half the sampling rate (the Nyquist limit). Low-pass filtering of frequency components above the sample rate is required to limit *aliasing* errors. One currently available audio workstation program offers the following rates (all in kHz): 5, 7, 11, 11.1, 22.05, 22.25, 24, 32, 44.1, 48, 96, and 192 kHz. While slower rates increase storage capacity (Bunta et al., 2003), sample rates below 22 kHz will exclude some components of speech and should be avoided (Olsen, 1988; see also Kent & Read, 2002). Of the rates just listed, only three are considered standard by the recording industry and, hence, anticipated by a wide range of recording and reproduction devices: 44.1 kHz for CD-audio, 48 kHz for digital audio tape (DAT), and 44.1 and 96 kHz for digital versatile disk (DVD) recordings. Given the current costs of recorders and media, the CD-audio sample rate of 44.1 kHz appears most appropriate.

Data word width (the number of bits reserved for each sampled amplitude) can range from 8 to 64 bits. Word width defines the maximum dynamic range in decibels of captured information: 8 bits give a range of 48 dB, 16 bits give 96 dB, 20 bits give 120 dB, and 24 bits give 144 dB. These are

best-case values. Equipment and environmental noise typically produce poorer performance. Recognized standards are 16 bits (for CD-audio) and 24 bits (for DVD recording). Given these factors and the nature of speech signals, a word width of 16 bits appears most appropriate. Compared to CD-audio standards (with 22 kHz bandwidth and 96 dB dynamic range), analog audio recording and reproduction (using Phillips cassette tapes) typically produce bandwidths of 6–10 kHz and dynamic ranges of 30–40 dB.

Signal Recording and File Formats

Until recently, A-D conversion of audio signals for computer processing was performed exclusively with plug-in sound cards or other conversion circuits built into computers. While these will persist for entertainment purposes, professional audio work increasingly uses A-D converters that are physically separated from computers to limit electromagnetic noise. They are also recommended because such devices are inherently cross-platform (see ***Microphone Interface Issues***, above).

Once digitized, audio signals sent to a computer are either stored in random access memory (RAM) or sent directly to a hard drive. RAM is fast but also volatile and limited in capacity. Modern hard drives can store hours of sound, but they must be fast enough (at least 7200 rpm) to do so without data loss.

Storing sound files on portable media requires attention to data formats—the ways in which data are structured into files. Two formats, WAV files for Windows and AIFF files for Macintosh computers, were originally proprietary but can be easily read and converted by most audio editing software (e.g., Sonic Foundry Sound Forge and Cakewalk Home Studio for Windows; MOTU Digital Performer, Emagic Logic, Bias Peak, and Macromedia Sound Edit for Macintosh), including newer professional software available for both computer platforms (Digidesign ProTools LE and Steinberg Cubase). These programs do much more than waveform editing; many can display respectable speech spectrograms and waterfall spectra. Other sound file formats include U-Law, MP3 (or MPEG Layer 3), and Sound Designer. Any of these formats can be used to store sound on common removable media (Imation Zip disks, CD-ROMs, or various trademarked card media such as compact flash, multimedia [MMC] cards, memory stick, secure digital, etc.). However, formatted sound can be reproduced only by software capable of recognizing and decoding the same format.

Recording audio on CDs that can be reproduced by stand-alone CD-audio players requires attention to older standards designated by the recording industry for such devices. The prevailing standard is the "Red Book," the first of several *Compact Disc "Color Books"* that define data and control protocols for various applications. CD-audio does not use WAV or AIFF formats, but instead employs an earlier pulse code modulation (PCM) method independent of computers. Red Book compliant CD-audio media can be reproduced by the greatest range of devices. For this reason, they are recommended for off-line storage of audio material.

It is very likely that DVD storage of audio information will become common in the future, in part because DVD media have seven times greater storage capacity than CD media. Thus, DVD-ROM is a clear alternative to CD-ROM for storing WAV, AIFF, or other audio formats designed for computer use. Although standards exist for DVD-audio, disagreements within the recording industry about security issues have resulted in very few commercially available DVD players capable of accommodating such standards. DVD-video recording of audio is feasible, but existing software developed for video plus audio work (in which a major goal is multi-channel surround sound) is not particularly pertinent to managing recorded speech for perceptual or acoustic analyses.

Reproduction

Outboard D-A converters have the same advantages as outboard A-D converters: greater noise immunity and cross-platform compatibility. Many such devices also include line-level headphone drivers with level controls (see ***Microphone Interface Issues***, above).

The listening environment may be less critical than the recording environment, but it is still pertinent. Headphones are preferable for several reasons, as are listeners with confirmed normal hearing. Circumaural headsets have the advantage of greater attenuation of ambient sound. While speech has a narrower bandwidth than music, headphone frequency response remains important. Fully adequate headphones are available for about $100.

Although most speech signals captured for phonetic transcription, prosody-voice coding, or language analyses are recorded on a single channel, listeners may perform more accurately (or with greater confidence) if signals are presented diotically. Doing so may require external headphone amplifiers to drive both right and left earphones from a single audio channel. Such devices are readily available.

The levels at which listeners attend to reproduced signals are seldom specified beyond a "comfortable listening level." Because the subject matter of phonetic transcriptions can vary by nearly 50 dB under controlled conditions (and perhaps more for conversational speech), it may be desirable to standardize methods for establishing listening levels to ensure audibility of all components of speech, while also avoiding signal distortion during reproduction.

The following observations suggest strategic considerations and several things to avoid.

Strategic Considerations

1. Consider audio A-D and D-A systems external to computers. Most of these employ the USB or IEEE-1394 bus and provide superior microphone preamplifiers, optional phantom power, balanced-line connections, and headphone outputs. Consumer-grade plug-in sound cards are not designed for use with balanced-line microphones, cannot supply phantom power, and are subject to electronic noise generated within computers.

2. If sample rate, data word width, or numbers of channels can be changed by hardware, software (or both), verify settings before each recording session.

3. Prefer industry standard sampling rates and word widths (e.g., the CD-audio standard of 44.1 kHz at 16 bits or the DAT standard of 48 kHz at 16 bits). The cost of storage media is probably less than the cost of information lost or unshared due to uncommon standards or inconsistent choices of rates and word width.

4. To maximize signal-to-noise ratio, place the microphone between 10 and 12 inches (about 0.3 meters) from the talker's mouth. Use a foam pop filter to limit plosive, sibilant, and air flow noise. Orientation of the microphone diaphragm to the talker depends upon directional characteristics and should be explicitly specified. Microphones should be supported by stands, booms, lavaliere clips, or wall mounts (for PZM units), not hand-held or placed on tables subject to mechanical noise. Consistent with environmental and instrumental noise sources (e.g., computer fans), use microphone cables no longer than necessary.

5. Monitor (listen to) the signal as recorded to ensure acceptability prior to each recording session.

6. Prefer standardized, cross-platform interface systems (USB 1.1, USB 2.0, IEEE-1394, or IEEE 1394b) and data storage formats (*ISO-9660* for CD-ROM, and Red Book for CD-Audio).

7. Prefer cross-platform software for recording CDs, or at least software sources with a record of supplying and updating similar programs for major operating systems. Roxio, for example, markets Toast for Macintosh systems as well as Easy CD and DVD Creator for Windows systems. These and companion products allow Red Book compliant recording.

8. Consider the benefits of Red Book CD-audio over other recording schemes (raw Microsoft WAV or Macintosh AIFF files recorded to CD-ROM). Advantages include the abilities to define up to 99 tracks (each analogous to a song on a commercial music CD) and up to 49 index points per track. If Red Book standards are followed, recorded material can be reproduced on consumer CD players and controlled with commonly supplied remote controls. If reproduced on a CD player internal to a computer, tracks and index points can be used to locate sound segments with CD player control software routines supplied with operating systems. If the research community were to adopt this as a standard (or some other data and control information protocol), it would be easier to share recordings for the purpose of assessing inter-laboratory reliability of analysis and coding outcomes.

9. Manage cables and interconnections to minimize noise and interference: (a) keep cables short, (b) use balanced-line connections for microphones, (c) keep audio cables away from power cables, (d) use digital interconnections whenever possible (e.g., *S/PDIF*), and (e) use optical cables whenever possible (e.g., *TOS,* an optical derivative of S/PDIF).

10. Keep the recording-reproduction chain as short as possible by minimizing the number of components.

Things to Avoid

1. Avoid technologies likely to become orphaned due to device or media cost, limited recording capacity, or storage media longevity problems (e.g., DAT). Similarly, avoid proprietary technologies likely to be abandoned by manufacturers (e.g., Sony Mini Disc).

2. Avoid inexpensive peripheral components (microphones, cables, connectors, and earphones). They are more likely to perform poorly, or to fail.

3. Never connect (or remove) a phantom-powered microphone to (or from) the next link in the recording chain when the next link is powered. Connect first, and then activate power; or deactivate, and then disconnect.

4. Never test a microphone by taping it. Just speak.

5. Avoid re-recording, including sequential A-D and D-A conversions (e.g., by using a DAT recorder to capture signals, then routing the DAT analog output to the analog input of a computer-based audio workstation). Each time a signal is converted, some quality is lost and opportunities for errors increase (e.g., by inadvertently confusing sample rates).

6. Avoid down-sampling (converting from a faster sample rate to a slower one) and up-sampling (converting from a slower sample rate to a faster one). Both invite errors.

7. Avoid plug-in or external A-D and D-A systems for which the major selling point is surround sound (e.g., Dolby Digital 5.1 or DTS) if your recorded signals are mono or stereo.

8. Avoid signal compression (e.g., MPEG3 or Real Audio) unless your purpose is to transmit audio over the Internet, the major motivation for these systems. Compression technology has not been evaluated for scientific or clinical applications.

CLOSING COMMENT

None of the systems or practices discussed above were developed specifically for use by clinical phoneticians. This state of affairs is, of course, simply a contemporary re-telling of earlier times and technologies. One thing is abundantly clear: over time, systems will become faster, less expensive, and more portable. Increased manufacturing yields of integrated circuits and competition will reduce the cost of RAM, hard disks, removable media, and record/play devices. Competition also will increase the variety of devices, software, and technical standards. Some technologies will disappear. These changes will be driven by initiatives independent of our concerns. But as in the past, systems originally developed for use by the entertainment industry are ripe for adaptation to our purposes.

GLOSSARY

ADAT Lightpipe: A professional optical interconnection system that caries eight digital channels, developed by Alesis Corporation (Maguire & Louderback, 2002).

***AES/EBU*:** A professional stereo digital interconnection system using three-conductor XLR connectors, codified by the Audio Engineering Society and the European Broadcasting Union (see Finger, 1992).

***Aliasing*:** Spurious information created when digitally sampling a signal of a frequency higher than one-half the sampling rate. Corrected by low-pass filtering to attenuate frequency components higher than one-half the sampling frequency (see Pohlmann, 1988).

***Balanced line* (circuit):** A two-conductor, analog circuit for which all conductors and everything with which they connect have the same electrical impedance relative to ground and to all other conductors. This requires a third conductor (shield) that serves as a ground (see Whitlock, 2002). Connectors used with balanced lines include XLR or Canon plugs, and stereo 1/4-inch diameter phone plugs (called "tip-ring-sleeve" or TRS).

Compact Disc "Color Books": Shorthand for internationally codified standards for physical dimensions, data coding, and control coding for each of several applications of compact discs. The Red Book pertains to CD-audio, the Orange Book to CD-R, the Yellow Book to CD-ROM, the Green Book to CD-interactive, the White Book to CD-video, and the Blue Book to CD-enhanced multi-session (Pohlmann, 1988; Pohlmann, 2002).

***Firewire or iLink**:* A general-purpose interface bus allowing connection between a computer and peripheral devices, based upon one of two upward compatible standards: IEEE 1394 (400 Mbps) or IEEE 1394b (800 Mbps). Both are adequate for audio.

***ISO-9660*:** An international standard coding/decoding format for CD-ROMs, allowing transparent use of data recorded on CDs on Windows, UNIX, Linux, and Macintosh operating systems.

***Phantom Power*:** For professional condenser microphones, a way to provide power (typically a dc bias voltage from 9V to 52V) to polarize the capacitor capsule and to match microphone output impedance to preamplifier input impedance. Invariably implemented with three-conductor XLR connectors. Power is provided by the preamplifier stage of a recording device, mixer, or stand-alone microphone preamplifier (Ballou, 2002).

***Proximity Effect*:** For directional microphones, an increase in output level when the microphone is operated close to a sound source. More specifically, distance and frequency interact to increase level: as frequency or distance from the diaphragm decreases, output level increases. At 500 Hz, this effect can produce a 3 dB boost at 11 cm and a 10 dB boost at 5.5 cm; at 100 Hz, the increase at 11 cm can be 14 dB, and 20 dB at 5.5 cm (Eargle, 2001).

***S/PDIF*:** A consumer-grade stereo interconnection system developed by Sony and Phillips using coaxial copper wire fitted with RCA connectors (Maguire & Louderback, 2002).

***TOS, TOS Link**:* An optical version of S/PDIF developed by Toshiba (Maguire & Louderback, 2002).

Universal Serial Bus (USB): A general-purpose interface bus allowing connection between a computer and peripheral devices. USB 1.1 offers up to 12 Mbps; USB 2.0 offers up to 480 Mbps. Both are adequate for audio.

REFERENCES

Audio Engineering Society. (1992). *AES 3-1992. AES recommended practice for digital audio engineering: Serial transmission format to two-channel linearly represented digital audio data.* New York, NY: Audio Engineering Society.

Audio Engineering Society. (1998). *AES 17-1999. AES standard method for digital audio engineering: Measurement of digital audio equipment.* New York, NY: Audio Engineering Society.

Ballou, G. (2002). Microphones. In G. Ballou (Ed.), *Handbook for sound engineers* (3rd ed.). Boston: Focal Press.

Bunta, F., Ingram, K., & Ingram, D. (2003). Bridging the digital divide: Aspects of computerized data collection and analysis for language professionals. *Clinical Linguistics and Phonetics, 17*(3), 217–240.

Eargle, J. (2001). *The microphone book.* Boston: Focal Press.

Fause, K. (1995). Fundamentals of grounding, shielding and interconnection. *Journal of the Audio Engineering Society, 43*(6), 498–516.

Finger, R. (1992). AES3-1992: The revised two-channel digital audio interface. *Journal of the Audio Engineering Society, 40*(3), 107–116.

Kent, R., & Read, C. (2002). *Acoustic analysis of speech* (2nd ed.). San Diego: Singular/Thompson Learning.

Macatee, S. (1995). Considerations in grounding and shielding audio devices. *Journal of the Audio Engineering Society, 43*(6), 472–483.

Maguire, J., & Louderback, J. (2002). *TechTV's secrets of the digital studio: Insider's guide to desktop recording.* Indianapolis: Pearson Professional Education.

Muncy, N. (1995). Noise susceptibility in analog and digital signal processing systems. *Journal of the Audio Engineering Society, 43*(6), 435–453.

Olsen, W. (1988). Average speech levels and spectra in various speaking/listening conditions: A summary of the Pearson, Bennett and Fidell (1977) report. *American Journal of Audiology, 7,* 21–25.

Pohlmann, K. (1988). The compact disc: Formats, technology and application. *Journal of the Audio Engineering Society, 36*(4), 250–287.

Pohlmann, K. (2002). Compact discs, SACD and DVD. In G. Ballou (Ed.), *Handbook for sound engineers* (3rd ed.). Boston: Focal Press.

Rane Corporation. (2003). *Pro audio dictionary.* Retrieved September 16, 2003, from Rane Corporation Web site: http://www.rane.com/digi-dic.html

Shriberg, L. D., McSweeny, J. L., Anderson, B. E., Campbell, T. F., Chial, M. R., Green, J. R., Hauner, K. K., Moore, C. A., Rusiewicz, H. L., & Wilson, D. L. (2003). *Transitioning from analog to digital audio recording in childhood speech sound disorders.* Manuscript submitted for publication.

Whitlock, B. (2002). Grounding and interfacing. In G. Ballou (Ed.), *Handbook for sound engineers* (3rd ed.). Boston: Focal Press.

COMMERCIAL SOURCES

The following firms supply professional-quality recording and reproducing components and systems. Selling prices are normally less than list prices, particularly for educational institutions.

B & H Photo-Video: http://www.bhphotovideo.com/

Full Compass: http://www.fullcompass.com/

Roxio: http://www.roxio.com/

Sweetwater Sound: http://www.sweetwater.com/