

PROBLEMS WITH TRADITIONAL ACOUSTIC ANALYSIS OF VOWELS

Data acquisition problems

An experiment was conducted to see how three different field recording methods would influence acoustic analysis. The first two methods are very common in sociophonetic literature: (1) signal acquisition with a Marantz (or comparable) cassette recorder with a built-in microphone; or (2) recording speech with a MiniDisc recorder and an omni-directional lavalier microphone. The third method, still rather uncommon, involves (3) recording speech with a head-set, flat response microphone and a 24-bit digital recorder.

Data acquisition with a standard analog Marantz recorder

Marantz portable cassette recorders have been used in the field for some time. They have a reputation for being sturdy and for producing high quality recordings, particularly for the purposes of news gathering and reporting. The recorder comes with a built-in condenser microphone, capable of capturing broadcast quality sound. Note, however, that audio quality understood in audiophile terms (e.g., “broadcast quality”) does not directly correlate with audio quality in acoustic-phonetic terms, as the former relies on subjective assessment of abstract sound properties, such as “clarity,” “brightness,” or “presence,” (Barlett & Barlett, 1998) while the latter refers strictly to the acoustic fidelity and detail of acquired sounds.

The potential problems with the Marantz-type recorder are three-fold. First, the omni-directional microphone is too close to the recorder’s motor and tape transport mechanism. It therefore “picks up” a lot of low frequency noise. Figure 1 shows an LPC spectrum of the vowel /i/ superimposed on a narrow-band spectrum of the noise produced by the recorder.¹ It is evident that the spectral band of the first formant (labeled “F1”) is similar in level to the low-frequency

¹ Other low-frequency noise generated by a computer fan or an air conditioner would have a very similar spectrum.

components of the noise (labeled “Noise”). This may negatively influence LPC-based formant extraction.

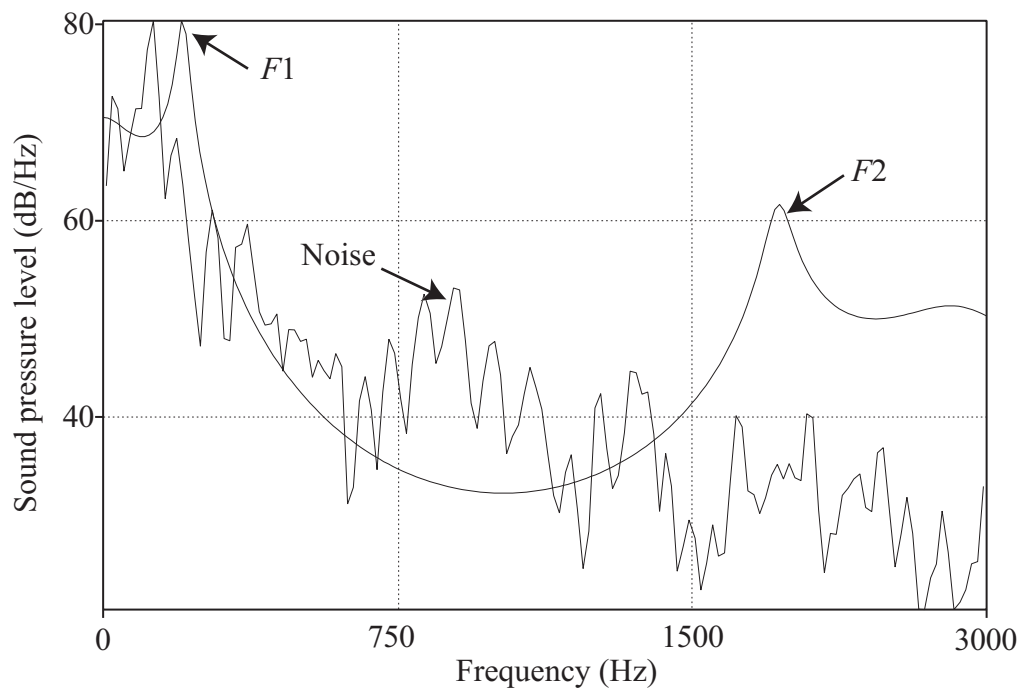


Figure 1. LPC of the vowel /i/ superimposed on the noise spectrum of the Marantz recorder

Second, the Marantz microphone, due to its omni-directional polar pattern, records noise from the environment, as well as the intended speech signal. If the interview is conducted near an air conditioner, open window, refrigerator, or any other source of low frequency energy, the amount of noise recorded on tape can jeopardize the reliability of acoustic analysis. Finally, this kind of microphone cannot be placed close to the talker’s lips, which causes a great deal of attenuation (decrease in amplitude) of the speech signal before it reaches the microphone.

60 Hz hum

Sixty Hz hum from power circuits is another common source of noise in field recordings, particularly when unbalanced microphone cables and non-grounded power cables are used.

Figure 2 shows two spectra – one of the 60 Hz hum (solid line) and one of the vowel /a/ in “job” (dashed line). The hum interferes with the speech signal in ways similar to those illustrated in Figure 1. This problem is not unique to Marantz recorders, though the popular PMD 101 model does not have a balanced microphone input and its power supply is, typically, not grounded, which makes this recorder particularly susceptible to 60 Hz hum. Also, often, in respondents’ kitchens and living rooms, grounded power outlets are not available, and the existing wiring used for refrigerators, TVs, air conditioners, and so forth, can potentially cause a great deal of interference. Therefore, the use of professional-grade, balanced, properly grounded, “XLR” (also known as “Canon”) microphone interfaces and cables is necessary to ensure the least amount of electrical interference.

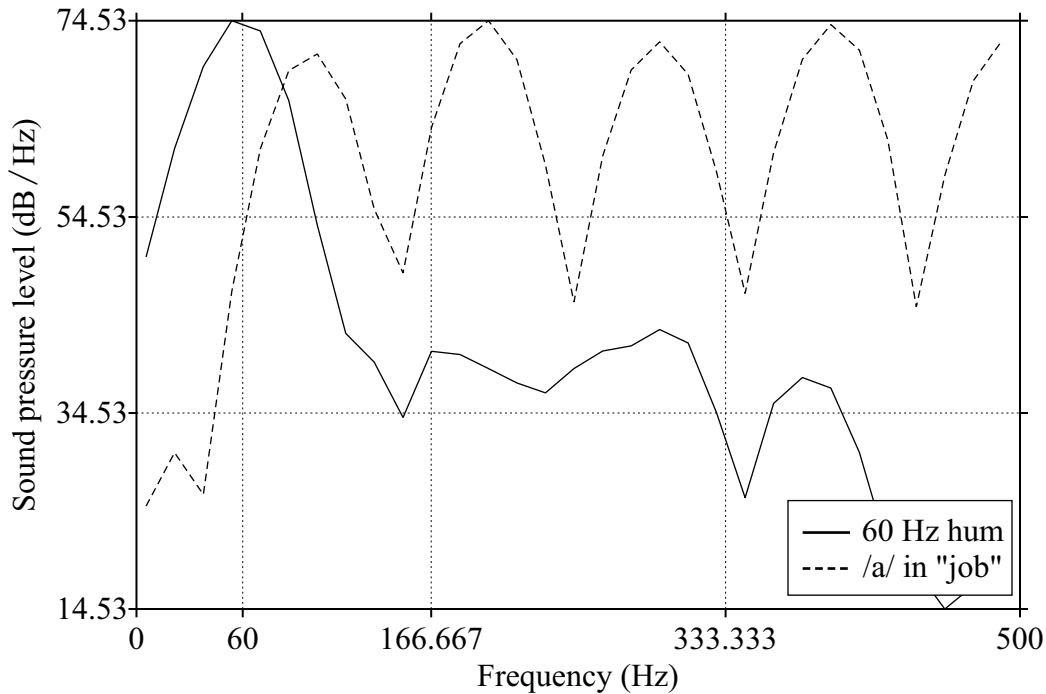


Figure 2. Spectrum of 60 Hz hum and the vowel /a/ in “job”

Digital recording with a MiniDisc player

The Sony MiniDisc form factor first appeared on the market in Japan in 1992. Soon afterwards, due to an effective 1996 advertising campaign entitled “Where the Music Takes You“, the MiniDisc became the most popular portable consumer digital music system (Hunt, 1996). It was also embraced by the fieldworker community, despite the fact that it was never designed to be a quality recording device. Sony designed the MiniDisc recorder as an inexpensive digital option to its Walkman series of cassette players. The MiniDisc has at least three serious problems.

First, the speech signal is altered by the recorder (compressed) in ways over which the researcher has no control. MiniDisc recorders use a lossy psychoacoustic data compression format called ATRAC (Adaptive TRansform Acoustic Coding). Soon after the recorder captures and quantizes an acoustic signal, it converts it to the proprietary ATRAC format at the bit rate of 292 kbps (approximately 1/5 of the uncompressed PCM “CD quality” of 1.41 Mbps). All of the acoustic field data are processed by the algorithm, which is based on psychoacoustic principles whereby the signal is divided into three frequency sub-bands with different data reduction schemes involved in each of them (Pohlmann, 2000). This indicates non-linear digital signal processing whereby the original speech signal is altered in ways determined but by the best compression (compromise between “quality” and bit-rate) options dynamically selected by the algorithm.

Second, the standard MiniDisc recorder does not have a professional microphone interface. As a result, only a small number of amateur quality microphones can be used with it. It is equipped with an electret-condenser interface with so-called “plug-in power.” Some Sony microphones are compatible with it, as are a few of other brands (e.g., Audio-technica AT803b and AT822). While it is theoretically possible to connect a professional-grade microphone to the

MiniDisc player, this is quite difficult, as it requires a special impedance matching in-line transformer.

Finally, an omni-directional, low-quality lavalier microphone that is most typically used with MiniDisc players, produces a noisy recording. Figure 3 shows a waterfall spectrogram of the word “hat” recorded with a MiniDisc recorder and an Audio-technica AT803b microphone. The spectrum appears to contain a great deal of extraneous noise and the formant peaks have unusually low intensity and wide bandwidths when recorded with MiniDisc hardware. MiniDisc recordings are digitized at a dynamically variable bit rate, but they can be converted to a PCM format at 44,100 Hz. The spectrogram in Figure 3, therefore, does contain acoustic information well over 3,000 Hz. However, this information contains very little spectral detail of interest to phoneticians.

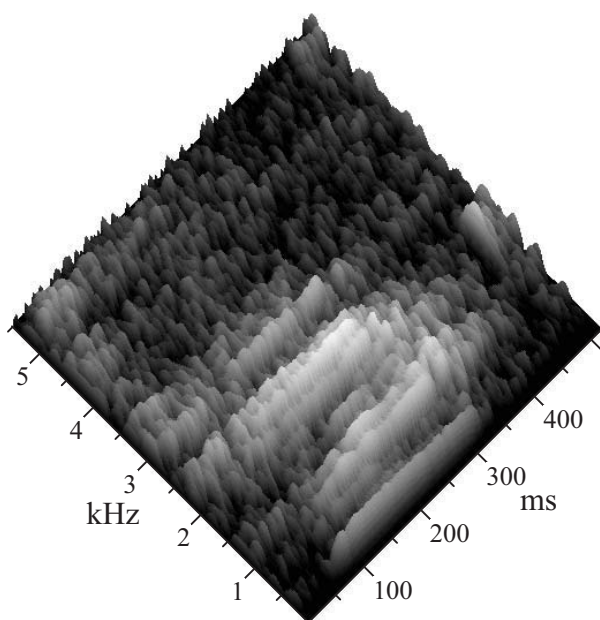


Figure 3 Waterfall plot of the word "hat" by an NCCS-influenced female talker

24-bit digital recording with a close-talking microphone

The third scenario involves a high quality, 24-bit, 48,000 Hz digital hard disc recorder with a flat-response, head-set microphone (Sennheiser HMD25-1). The Sennheiser microphone has been specially designed for recording speech in noisy conditions. It has a directional polar pattern, but despite being a close-talking microphone, it retains a flat response throughout the entire frequency range of up to 16,000 Hz. By being close to the talker's lips, it records speech with a more favorable signal-to-noise (S/N) ratio. The digital recorder, which connects directly to a laptop computer via the USB bus, is equipped with two professional-grade microphone inputs. It reproduces pure 24-bit sound, which contains more acoustic detail and less quantization noise than signals captured with 16-bit digital recorders (such as Tascam DA-P1 portable DAT recorder).

Figure 4 shows a waterfall spectrogram of the word "hat" recorded by the same speaker, at the same time as the sample in Figure 3. The complex spectral features of this vowel are more visibly strong and well defined; there is virtually no unwanted noise. It should come as no surprise that Fast Fourier Transform (FFT) can produce different spectral representations of what are all productions of the same vowel captured with these three different recording methods. Similarly, Linear Predictive Coding (LPC) analysis, which is the standard way of extracting formant values in sociophonetic analysis, can be expected to output significantly different values

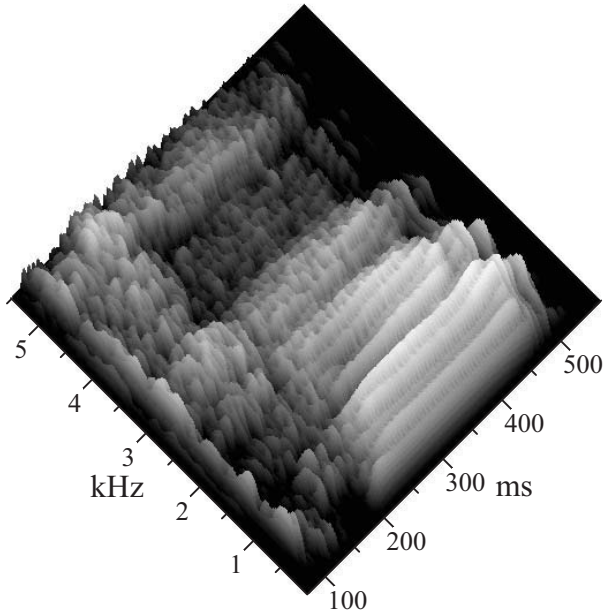


Figure 4 Waterfall plot of the word "hat" by an NCCS-influenced female talker

The role of the microphone

The role of the microphone in acquiring reliable speech signals is at least as important as the role of the recording device and the recording medium. The ideal microphone must have a wide and flat frequency response, and must not cause an increase in low-frequency amplitudes when placed close to the talker's lips (so-called "proximity effect"). Proximity effect is a natural consequence of placing a dynamic, cardioid microphone close to the sound source. Interestingly, most manufacturers are not interested in eliminating proximity effect as the extra low-frequency boost is often desired by stage performers and newscasters (Huber & Williams, 1998). Figure 5 shows a spectrum of the vowel /a/ in "job" recorded with the AKG C-420 microphone placed approximately 6 cm from the talker's lips. This sample shows a significant increase in amplitude in low frequencies around 100 Hz due to proximity effect. While the measurements of $F1$ and $F2$ in this particular case might not be significantly affected by proximity effect, the measurements of $F0$ and other parameters measured in the vicinity of the first 2 or 3 harmonics will be considerably biased.

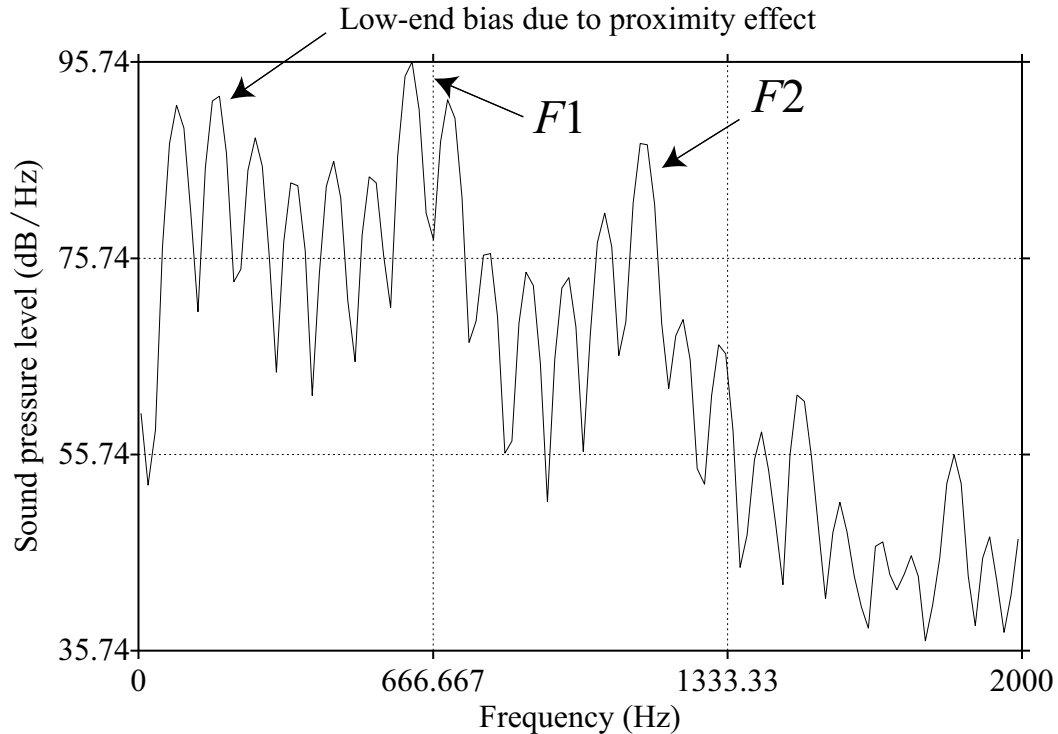


Figure 5. Low-end bias due to proximity effect of the AKG C-420 microphone

There are very few, specially designed microphones that retain broad and flat frequency response from 20 to 20,000 Hz. Such microphones, particularly when coupled with 24-bit digital recorders, are capable of capturing high quality, unbiased speech signals. Figure 6 compares frequency response of Shure Beta 87a² and Earthworks M30 microphones, according to the manufacturers. The frequency response of the Shure Beta 87a microphone has two significant peaks – one around 5,000 Hz (solid line), and one around 100 Hz (dashed line). The increase in low frequency amplitudes (dashed line) occurs when the microphone is placed close to the sound source (below 6 cm). Conversely, the low-frequency amplitude decreases when the microphone is moved away from the sound source (above 60 cm). The Earthworks M30 microphone, on the other hand, retains a relatively flat frequency response throughout the entire human hearing

² Shure SM48, a dynamic microphone similar to Shure Beta 87a, was once distributed with the Kay Elemetrics Computerized Speech Lab (KayElemetrics, 1998).

range (dotted line), regardless of its distance to the sound source – subject to the limitations of the hardware used in the remainder of the recording circuit.

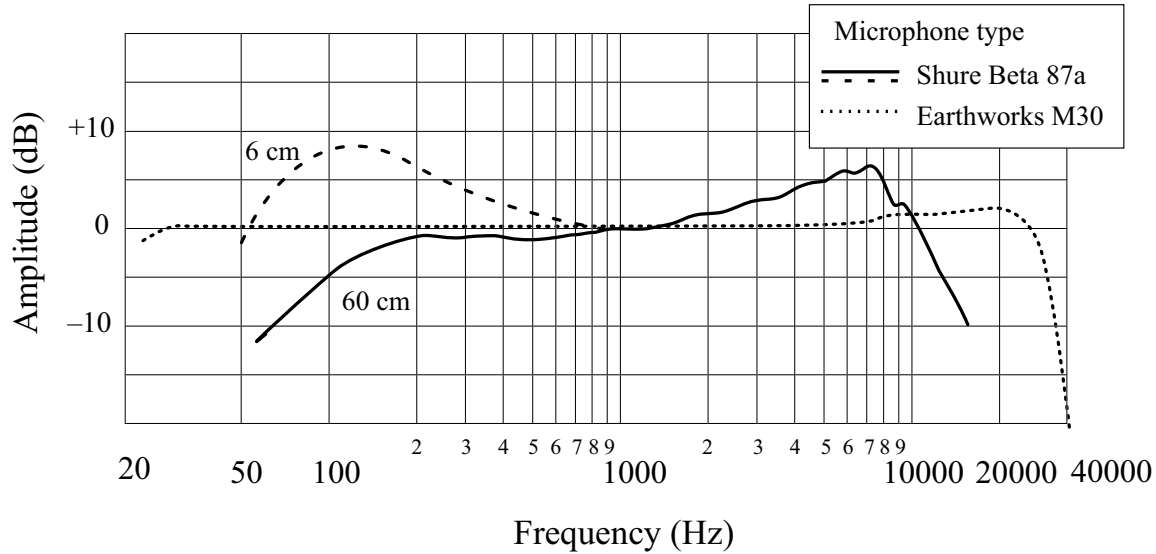


Figure 6. Frequency response of Shure Beta 87a and Earthworks M30

Different signal acquisition methods return different formant values

Given the information in the previous section, one might suspect that LPC analysis would output significantly different formant values across the three field recording methods. An experiment was designed to test this hypothesis. A female talker with an NCCS-influenced vowel system was recorded reading a wordlist containing words with four American English vowels (/a/, /æ/, /ε/, and /aɪ/). The first three vowels are participants in NCCS, and the fourth one, /aɪ/, was selected because of its diphthongal quality. The talker was recorded with a Marantz recorder, an Audio-technica AT803b lavalier microphone connected to a Sony MiniDisc recorder, and a Sennheiser head-set microphone plugged into a Sound Devices USB Pre unit simultaneously.

A total of 99 vowel tokens were recorded and transferred to a computer workstation. The first two sets were transferred via the digital S/PDIF³ interface, and the analog recording was digitized at 24-bit and 48,000 Hz. The recordings were then downsampled to 16,000 Hz and analyzed acoustically. Acoustic analysis was performed by means of *Akustyk* (Plichta, 2004).

How to avoid researcher bias in formant analysis?

It is difficult to run a test comparing three different sets of recordings without researcher bias. This bias is caused by the subjectivity in identifying the steady state, an area of relative stability in the frequency domain, where formants are to be measured (see, for example Hillenbrand, Getty et al. (1995) for a more detailed discussion on steady state identification). Therefore, instead of identifying steady state and measuring formants statically, a dynamic approach was designed. The vowel nucleus was divided into 0.025s Gaussian analysis windows. Formants and bandwidths were measured and recorded for each such window. Based on these data, the software calculated mean formant and bandwidth values for the entire duration of the vowel, as well as their average cumulative variation based on the algorithm given below, where v is the average variation over time, x_n is the frequency at point n and d is vowel duration:

$$v = \frac{\sum_{n=1}^T (x_{n+1} - x_n)}{d} \quad (1)$$

Statistical analysis of formants, bandwidths, and cumulative variation

Multivariate analysis of variance (MANOVA) was performed to test the hypothesis that the three recording modes return different formant values. The dependent variables included formant values in Hertz ($F1$ through $F3$), formant bandwidths in Hertz ($B1$ through $B3$), as well

³ S/PDIF (Sony/Philips Digital Interface) is a widely used digital audio interface capable of carrying stereo data quantized at 24-bit and 96,000 Hz.

as total cumulative variation the first three formants in Hertz/s ($F1v$ through $F3v$), while the three different recording modes constituted the independent variables. MANOVA was significant for recording type, $F(18,188)=8.318$, $p<.001$, Wilks' $\Lambda=0.312$. Because the overall MANOVA was significant, a series of post-hoc tests was performed to discover which acoustic parameters were significantly affected by recording type, and it was found that recording type was significant for all of $F1$ parameters. The post hoc Bonferroni multiple comparisons test (at the significance level of $p<0.05$) showed that all $F1$ -related parameters were significantly different from one another across the three different recordings types. In addition, the tests found that all vowels individually, exhibited a similar kind of variability. Table 1 summarizes the results, where $F1$ is the mean first formant frequency, $B1$ is the mean first formant bandwidth, and $F1v$ is the mean cumulative $F1$ variation.

Tests of between-subjects effects			
	$F1$	$B1$	$F1v$
Overall	$F(2,102)=17.741$, $p<.001$	$F(2,102)=17.741$, $p<.001$	$F(2,102)=3.781$, $p<.025$
/ ϵ /	$F(2,21)=1.882$, $p=.177$ (ns)	$F(2,21)=8.042$, $p<.001$	$F(2,21)=.900$, $p=.421$ (ns)
/ a /	$F(2,21)=7.94$, $p<.025$	$F(2,21)=4.544$, $p<.05^4$	$F(2,21)=1.076$, $p=.359$ (ns)
/ \ae /	$F(2,36)=6.12$, $p<.025$	$F(2,36)=3.773$, $p<.05$	$F(2,36)=4.15$, $p<.025$
/ ai /	$F(2,15)=8.737$, $p<.025$	$F(2,15)=31.106$, $p<.001$	$F(2,15)=2.972$, $p=0.82$ (ns)

Table 1. Summary of statistical results of the data acquisition test

⁴ With this type of between-subject follow-up test, it might be argued that the ANOVA has to be significant at the .025 level.

The box plot in Figure 7 is based on the so-called “quartiles,” which divide the distribution into four equally filled intervals. The upper and lower edges on the boxes are located at the first and third quartiles of the data, respectively (Jacoby, 1998). The vertical lines above and below the boxes (so-called “whiskers”) extend to the upper and lower adjacent values, or values that correspond to the maximum and minimum values in the data set. It can be seen in Figure 7, for example, that the mean $F1$ computed from the recordings obtained with the lavalier microphone has the shortest whiskers, indicating a relatively small distance between the highest and lowest $F1$ frequency in the data set. The horizontal line inside each box is the median line. If the line is off-center, the plot indicates an asymmetrical density of data points. Figure 7, for instance, shows that the distribution of $F1$ values obtained with the head-set microphone is the most symmetrical, while the $F1$ distribution obtained with the built-in microphone is spread over the broadest range of values. At the same time, none of the methods show any values beyond the adjacent values, which indicates that no “unusual data” or outliers have been found.

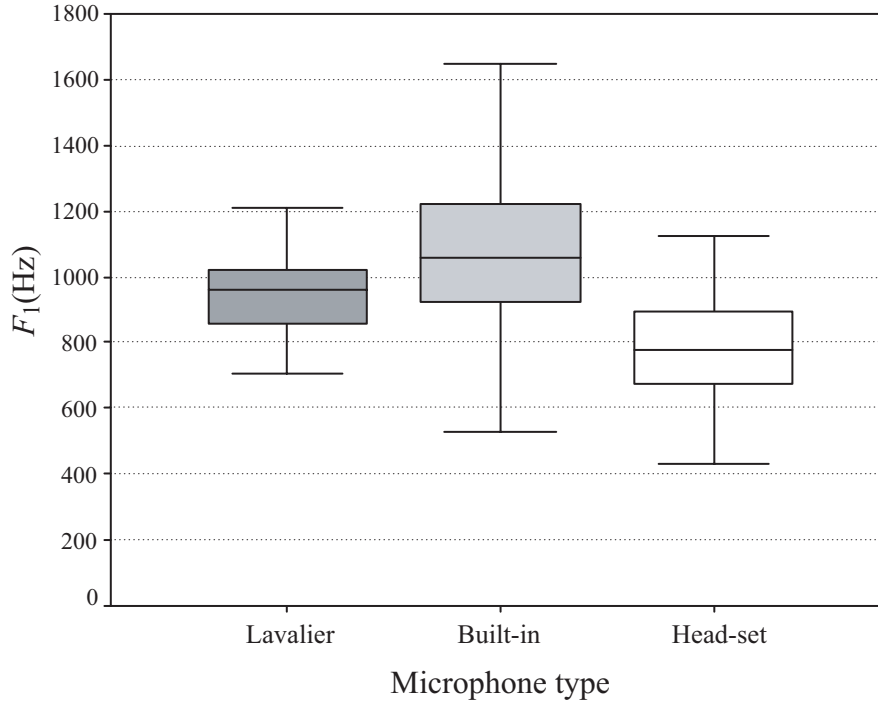


Figure 7. Box plot of overall between-subject comparisons of the data acquisition test

Summary of the signal acquisition experiment

The choice of data acquisition technology and methodology is an important consideration. Depending on the signal acquisition method used, significantly different results may be obtained across the F_1 domain. Researchers should be encouraged to follow best practices in the area of speech recording, processing, and analysis, and should use modern technology to their advantage (Plichta, 2002). There should also be no doubt that signals acquired with the head-set microphone and a 24-bit digital recorder are the most accurate, and, therefore, the most reliable. Such signals contain the greatest amount of spectral detail and the least amount of unwanted noise.

The intent of this dissertation is not to advocate the abandonment of traditional methods, nor is it to claim that all traditional sociophonetic research is unreliable. On the contrary, there is an extremely rich sociophonetic tradition of acoustic analysis that has furthered our

understanding of language variation and change probably more than any other branch of modern sociolinguistics. Still, NCCS-influenced vowels, due to their complex spectral features, are difficult to analyze, and caution must be exercised in this type of analysis.

To illustrate this point, Figure 8 shows three LPC spectra of the vowel /a/ in the word “lot,” obtained from the same corpus. LPC was applied at the same point in time and with the same parameters (analysis width of 0.025s, the sample rate of 16,000 Hz, LPC filter order of 13, and a mid-range pre-emphasis filter starting at 50 Hz). The vowel /a/, one of the major participants in NCCS, is problematic for an LPC algorithm because $F1$ and $F2$ are quite close to each other in frequency.

In poor quality, highly attenuated recordings (such as that in Figure 3), the distinction between the two peaks is blurred and LPC can return an incorrect reading. Figure 8 shows three LPC filter response graphs superimposed on top on one another. The box below the LPC graph contains the windowed (0.025s) portion of the waveform. The dashed line represents the sample obtained with a built-in microphone, and, as can be seen, it captures only one peak in the vicinity of $F1$. The same is true of the sample recorded with the lavalier microphone (dotted line). It is only the recording obtained with the head-set microphone that allows the LPC filter to return two formant values and a very realistic-looking spectral envelope. There are many examples of this kind, and the data acquisition test carried out earlier in this chapter supports this point as well.

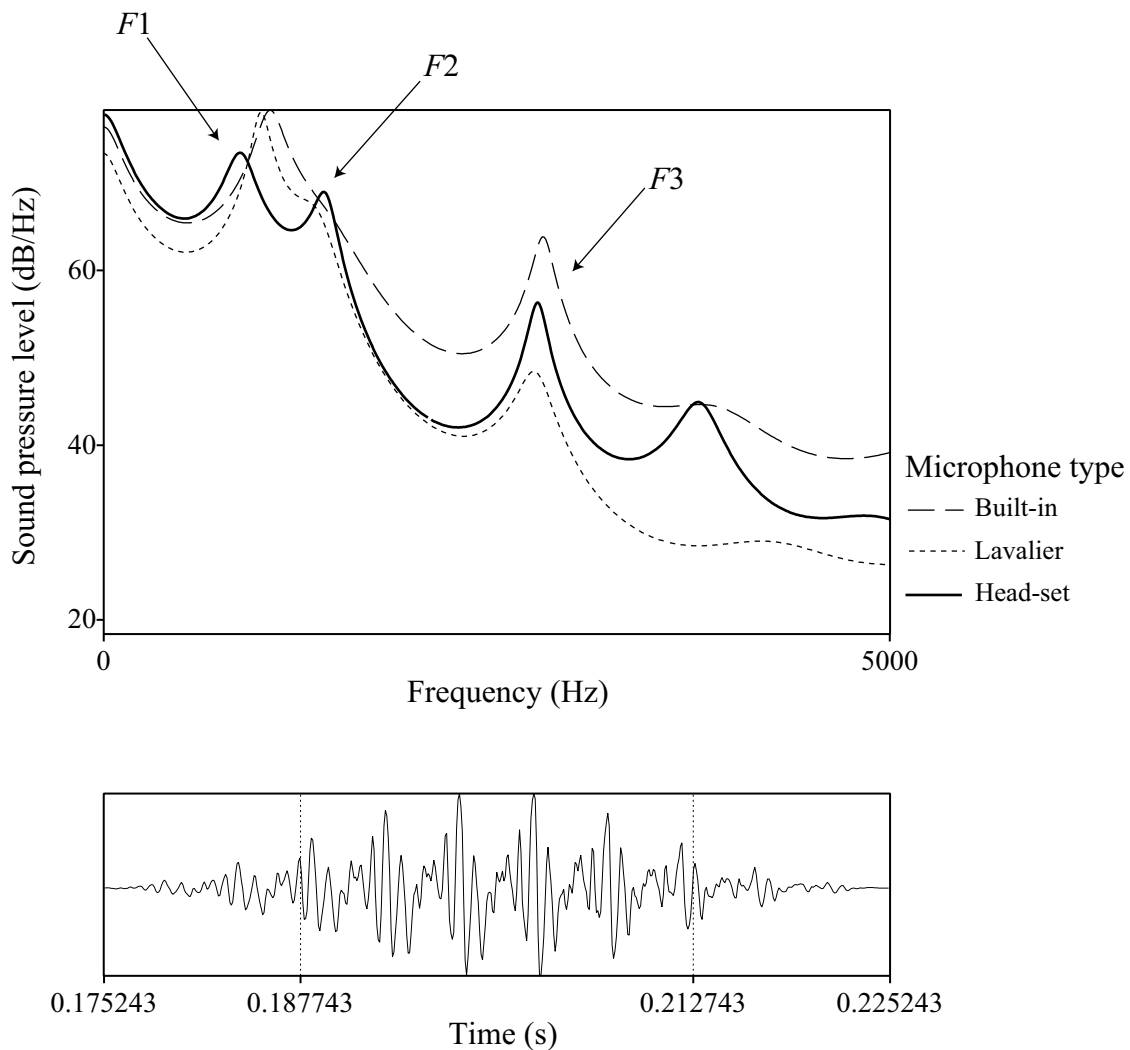


Figure 8. LPC spectra of the /a/ vowel in "job" by a female talker with an NCCS-influenced vowel system acquired with by different methods

REFERENCES

- Barlett, B., & Barlett, J. (1998). *Practical recording techniques* (Second ed.). Boston: Focal Press.
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*(97), 3099–3111.
- Huber, D. M., & Williams, P. (1998). *Professional microphone techniques*. Emeryville, CA: Mix Books.
- Hunt, K. (1996, Aug 27). Sony revives MiniDisc in package deal. *Los Angeles Times*, p. 5:1.
- Jacoby, W. G. (1998). *Statistical graphics for visualizing multivariate data*. Thousand Oaks: Sage Publications.
- KayElemetrics. (1998). *Computerized Speech Lab*. Lincoln Park: Kay Elemetrics Corp.
- Plichta, B. (2002). Best practices in the acquisition, processing, and analysis of acoustic speech signals. *U. Penn Working Papers in Linguistics*, 8.3.
- Plichta, B. (2004). *Akustyk for Praat (Version 1.7.2)*. East Lansing: Michigan State University.
- Pohlmann, K. C. (2000). *Principles of Digital Audio* (Fourth ed.). New York: McGraw-Hill.