Audio Engineering Society

Convention Paper 9472

Presented at the 139th Convention
2015 October 29–November 1 New York, USA

Forensic Sound Analyses of Cellular Telephone Recordings

Durand R. Begault[1], Adrian L. Lu[1], and Philip J. Perry[1]

[1] AUDIO FORENSIC CENTER Audio-Video-Digital Media, Charles M. Salter Associates
130 Sutter St., Ste. 500, San Francisco, CA 94104 USA
Durand.Begault@cmsalter.com

ABSTRACT

Recordings involving cellular telephones or personal digital assistants ("PDAs") are increasingly the source evidence in audio forensic examinations, compared to recordings originating with other devices such as hand-held digital recorders. On modern PDA cellular telephones recordings can be made either directly to the telephone or transmitted as voice mail messages. The current investigation focuses on differences in the two types of recordings in terms of dynamic range and linearity of levels. Such information can be important for characterizing the distance of sound sources relative to the microphone and are important for understanding transformation of recorded speech and non-speech sounds.

1. INTRODUCTION

Recordings involving cellular telephones or personal digital assistants ("PDAs") are increasingly the focus of audio forensic examinations, while recordings originating with other devices such as hand-held digital recorders have decreased as these devices are supplanted by PDAs, much as what has happened with consumer-level cameras. On modern PDA cellular telephones, recordings can be made either directly to the telephone, or transmitted as voice mail messages to a particular messaging system. It becomes immediately obvious that the quality of such recordings varies widely between messaging systems and between built-in recording applications; furthermore, the sensitivity of recordings to different sound source levels varies greatly, due to codecs and/or telephonic transmission. An understanding of these transformational properties has important implications for forensic analyses that include speech transcription of "background" voices, gunshots, distance of sound sources, and environmental context analysis (sometimes referred to as "roomprints") [1, 2].

When made directly to the telephone, the recording function is similar to a traditional hand-held digital recorder. For example the native voice memo recording application on the iPhone 5s writes .m4a files that, while technically written using a lossy codec, are both wideband and decent dynamic range (sampling rate 44.1 kHz, bitrate 64 Kbps). A

particular application from a third party may allow alternative means of export (but not necessarily recording). The use of the iPhone microphone as input to a sound level meter application has also been evaluated in the literature [3].

When a recording is saved as a voicemail message, either on another telephone, on a corporate voice mail server, or on an emergency ("911") call log recording system, several additional and varied stages of signal processing can be involved. These stages include algorithms designed to optimize the speech signal against background noise, including compression and voice activity detection algorithms. Most importantly, these stages of signal processing alter the spectrum of speech as a function of level.

The current investigation is necessarily limited in scope, focusing primarily on tests using an iPhone 5s in our laboratory using an AT&T carrier (GSM 4G) in the San Francisco Bay Area. Analyses of actual gunshots from a different older-model cellular telephone are provided in a final section (details have been omitted due to privacy concerns). We have approached understanding of effects from the standpoint of a "black box" analysis, where we provide detailed description of the input and resulting output. The analyses focus on differences in types of recordings in terms of dynamic range and linearity of levels. Such information can be important for characterizing the distance of sound sources relative to the microphone and are important for understanding the impact of recorded speech and non-speech sounds in forensic settings.

An admitted but not fatal limitation of this study is a lack of certainty regarding the technical aspects of the hardware involved, and how the signal was transformed by individual signal processing elements in the communication chain of the cellular telephone system, including those features caused by linear predictive coding, discontinuous transmission, voice activity detection, and the inclusion of so-called "comfort noise" (see e.g. [4] for an informative overview). In other words, we observe what occurs to the signal and its implications for forensic analysis, without analyzing the specific causes.

The frequency response characteristics of the voice mail recordings (AMR file structure) indicate a classic telephony narrow band, fixed rate with a cut-off at around 4k consistent with (but not necessarily indicating) a G711 mu-law codec. There are 3 microphones in the iPhone, one at the bottom, one next to the speaker above the screen, and one between the flash and the lens of the camera on the rear of the phone. Investigation indicates that different applications may access different microphones to optimize the functionality. For example, when using the front facing camera for recording video, the front facing microphone is active. This microphone is deactivated when the camera is switched to rear facing, and the rear facing microphone becomes active. The voice memo appears to use all three microphones to varying degrees, with the main (bottom) mic being the main source. When making voice calls, only the bottom mic is actively transmitting speech, however the iPhone features a noise cancellation control system and it is believed that at least one additional microphone is used for noise cancellation during voice calls when this feature is enabled.

2. IPHONE TEST SETUP

Tests were conducted in a highly absorptive test room (reverberation time < 0.2 s, background noise level 20 dBA, dimensions ~4.4 m x 5.2 m x 3.2 m), located at Charles M. Salter Associates. A loudspeaker for simulating speech and two different iPhones (model MD644LL/A, A1533, "5s", manufactured ca. 2013) were located off-center and non-parallel with surfaces to minimize modal effects. The loudspeaker had a reasonable approximation of human speech directivity (ADS powered loudspeaker, modified) and was placed at a distance of 1 m from the iPhone. For most of the tests, the iPhone was mounted on a camera stand with its microphone on axis to the loudspeaker, and the body rotated 90 degrees on both its vertical and horizontal axes. In one test ('voiceover") the iPhone was held at a normal mouth position at approximately the same distance. The testing room included a background noise simulation system that was used in some tests to increase the mid-band noise level.

The loudspeaker amplifier level was calibrated by reproducing band-pass (0.2-10 kHz band-pass) pink noise to a level of 75 dB using an ANSI type 1 sound level meter (Brüel & Kjaer 2230). The combined loudspeaker and room frequency response was measured using a 1 s chirp (ref. Figure 1). The "roughness" of the response is due to reflections off equipment and other nearby surfaces but was considered sufficient for the current experiment in the frequency range of speech.

To allow comparison of the phone recordings to a reference, high-quality digital recordings with a reference microphone co-located with the telephone for all measurements were made (ANSI type 1 sound level meters or acoustical measurement microphones

& preamps from Brüel and Kjaer, GRAS; Rion DA-20 digital recorder, 51.2 kHz sample rate, 16 bit).

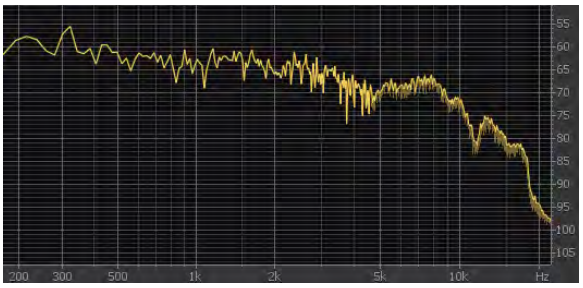

Figure 1. Frequency response, ADS loudspeaker and room response, 1 s sine sweep.

Following calibration, digital audio test files were played back via the analog audio outputs of a laptop computer (MacBook Pro; Sound Forge 11.0, FuzzMeasure 4.0 software) directly to the ADS loudspeaker. The test files used consisted of the following material:

1) "**Stepped level speech**": six similar words ("seat" "seed" "seek" "seem" "seen" "seep") from the Modified Rhyme test of ANSI S3.2 [5]) from 3 different female and 3 different male voices, at three different levels (~76 dBA; ~65 dBA; ~58 dBA), corresponding to a "loud", "raised" and "normal" voice level for a talker at 1 m distance [6]. This test file was used primarily to evaluate the transformation of speech at 1 m to the ATT voicemail system, and emulate a typical forensic scenario where the speech of a background talker is of interest.

Separate tests were conducted with the background noise simulation system off (background noise ~ 20 dBA), and with mid-band noise (frequency range .125-2 kHz; see Figure 2) simulating an **NC 30** and **NC 50** condition. The NC 30 condition is roughly similar to what might be experienced in an indoor environment with a moderately noisy HVAC system operating, and the NC 50 condition could correspond to an outdoor environment.

In a separate "**voiceover test**" using this same test material, an experimenter held the telephone and repeated at regular intervals a series of numbers with intervals of ~1- 3 s silence, so as to modify any level compression of the loudspeaker speech by the nearby talker.

2) "**Noise-speech**": ~5 s of pink noise played at 70 dB(A) followed by ~6 s of female speech at the same level (the same six words from one talker as in the "stepped level speech"). This test was conducted to determine the influence of any voice activity detection software on the phone mail system. Such detectors will cease recording if speech is not detected after a particular interval.

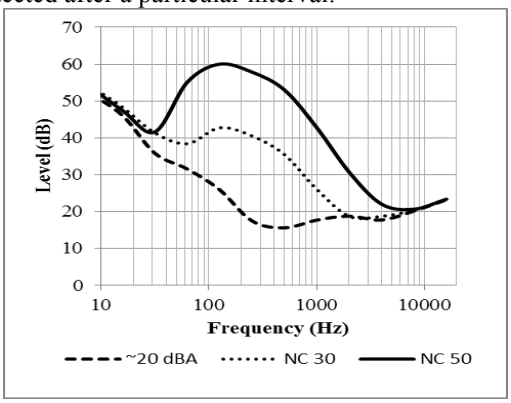

Figure 2. Background noise of test room (20 dBA) and NC 30, NC 50 simulation levels.

3) "**stepped pink noise**": one second intervals of pink noise, played at six successively decreasing levels of 10 dB (85, 75, 65, 55, 45 and 35 dB). This test was performed to both evaluate the level sensitivity and compression involved in the various recording systems tested.

4) "**balloon pop**": a balloon was inflated to approximately 0.5 m diameter and then popped using a small knife at the location of the ADS loudspeaker. This test was to emulate a brief high-level (~135 dB peak) impulsive event such as a gunshot and to examine the various recording systems' response. The contribution of reflections in the test room would necessarily follow the impulse response.

Analysis software utilized included MathWorks MATLAB R2015a, PRAAT phonetic analysis software v. 5.43, and iZotope RX4 Professional.

3. RESULTS: VOICE RECORDER

The voice recorder application saves files with a ".m4a" extension using a perceptually based, wideband lossy codec (MPEG-4 AAC LC Advanced Audio Codec, low complexity, 44.1 kHz sample rate, 64.0k bitrate). The frequency response extends to 15.5 kHz in the analyzed files.

For the 50 decibel range of levels tested (~35-85 dBA), the iPhone Voice Memo app showed a nearly linear RMS response to 10 dB attenuation steps of pink noise. This result is not surprising in light of studies that found an accuracy suitable for ANSI Type II criteria (+/- 1 dB) under certain applications [3,7], e.g. for full-bandwidth pink noise at levels

from 65 to 95 dB, A-weighted and unweighted, in 5-dB increments [3].

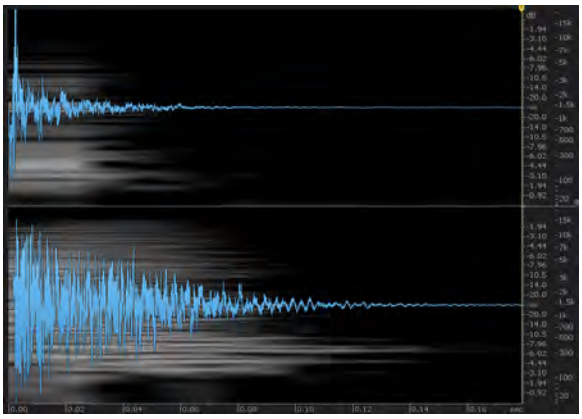

Figure 3. Balloon pop response. Top: measurement mic response; bottom: iPhone recorded to voice memo application. Abscissa: 0-0.18 s; ordinate: 20-20 kHz.

3.1. Balloon pop test

Figure 3, top, shows the acoustical response of a balloon pop following a period of "silence" (20 dBA) from the position of the loudspeaker at 1 m to the microphone. The peak sound pressure level was 134 dB. Following the initial fracture impulse of the balloon are early reflections extending out to ~30 ms. Figure 3, bottom shows the same response as recorded by the iPhone voice memo application. The impulse is captured without evidence of overload, but the reflected energy following it is amplified out to ~140 ms, indicative of a peak compression algorithm. This indicates that the onset time of high-level impulsive sounds such as gunshots are captured but may be more difficult to discern within a waveform display.

4. RESULTS: ATT VOICEMAIL

Files extracted from the ATT "visual voicemail" application have a ".amr" extension using a narrowband lossy codec (AMR-NB Adaptive Multi-Rate Narrowband, 8 kHz sample rate, 12.2 kbps). The frequency response extends to just below 4 kHz in the analyzed files.

4.1. Stepped level speech tests

Figure 4 shows spectrogram and formant frequency estimation plots for the recording playback of 3 male and then 3 female talkers at 76 dBA (loud voice level at 1 m), all saying the word "seed." The spectrograms are synched in time and by talker for two sources: top, from the measurement microphone; bottom, from the voicemail recording. The same result was observed for the 65 dBA playback and the reported results generally apply to the other test words.

What becomes immediately obvious is the transformation of the frequency estimations for the second and third formants, and the instability of the first formant estimation. Less obvious from the graphic but immediately obvious from listening is the intelligibility degradation of the utterances. While the vowel /i/ ("ee" in seed) is more or less maintained, the consonants are varied or transformed (e.g., /d/ becomes /t/ in one instance). The voice mail recording at no point renders an intelligible "seed," and the inter-talker consistency degrades considerably. The rms playback level recorded at the measurement microphone was within ~1 dB for each utterance, but the voice mail recording resulted in much wider variation in levels between talkers (range: ~19 dB). In some cases the initial /s/ consonant was severely attenuated.

Figure 5 shows spectrogram and formant frequency estimation plots for the recording playback of one male saying the word set "seat seed seek seem seen seep" at 76, 65 and 58 dBA (loud, raised and normal levels at 1 m). The spectrograms are synched in time and grouped by decreasing level for two sources: top, from the measurement microphone; bottom, from the voicemail recording. The same characteristic frequency transformation of formant frequencies is evident as in Figure 4. The intra-talker level varied no more than 8 dB rms between "loud" and "normal" levels with respect to the 18 dB variation in input, likely due to a level compression algorithm.

Figure 6 focuses on four examples of the same talker as shown in Figure 5 for the utterance of the word "seat." The example 1 spectrogram is from the reference microphone, 76 dBA at 1 m; examples 2-4 are from the iPhone at 76, 65 and 58 dBA. Listening indicates that the first two iPhone utterances are relatively intelligible, but the final utterance has not successfully reproduced the terminating /t/ consonant.

4.2. Voiceover test

A test was performed to determine whether or not a near talker speaking directly into the telephone in "normal" fashion might affect the compression of background voices during brief silent intervals. In other words, the degradation in signal exhibited for the 58 dB example seen in Figure 6 might be compensated if the voicemail system was "primed" by a nearby voice. The iPhone was removed from the

camera stand and, while the stepped level speech stimulus was played, an experimenter spoke numbers alternating with silence at a regular tempo: "one", (silent "two") "one two" (silent three, four); etc. A comparison of spectrograms and levels from this test to the prior stepped speech level test described in section 4.1 indicated no significant effect of the near talker on background voices.

4.3. Noise-speech test

In the presence of silence, the voice detection algorithms will cause the voice mail to stop recording and ask for push button verification after 3 s ("we did not get your message because you are not speaking…"). The same occurs once the background noise attains a certain level. During various NC 50 conditions, the call would be sometimes terminated before the completion of lower level speech at 58 dBA . The voicemail system recorded 2.3 seconds of pink noise, 3.3 seconds of suppressed noise, then terminated the call. The call did not terminate at the NC 30 and background noise conditions.

The result of playing the noise-speech test stimulus is shown in Figure 7. The upper waveform and spectrogram is the signal as recorded by the reference microphone, and the lower is the same signal recorded by the voicemail system. The formant estimation for the upper spectrogram is noisy but grouped in the area of 1st-3rd formants. For the voicemail recording, several interesting phenomena occur. A time-varying amplitude modulation of the noise level occurs from the start out to ~3 s, with a transient occurring at about 1 s and a 100 ms "gap" at 2 s. At this gap and from 3-5 s, the noise input is effectively squelched into the minimum noise floor level of the system. Over the 3 s period of noise, there is a center frequency variation that can be seen in the formant estimation that gives the noise a "talking" quality. A repeat of the stimulus resulted in the same general phenomena, although the center frequency variation was different.

When the speech began at 5 s the /s/ phoneme of the word "seat" was not recoded by the voice mail system, transforming it into something more similar to "eat". Likely, the noise content of /s/ was interpreted by the codec as a continuation of the preceding noise, as opposed to the start of a speech utterance.

4.4. Stepped pink noise

The voicemail system response to the stepped pink noise stimulus indicated a time-varying level response over the fixed level intervals, as shown in Figure 8. This is likely due to the implementation of a noise suppression algorithm. This suppression occurred suddenly, rather than gradually, during the recorded steps as can be seen in steps one and two. At the third step interval (65 dBA), the voicemail recording terminated after ~1.5 s.

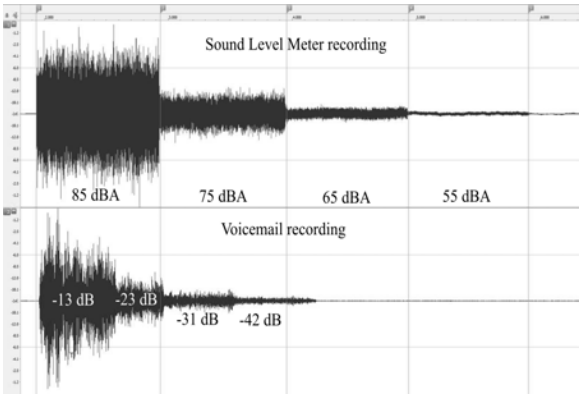

Figure 8. Stepped pink noise response at background noise condition. Top: measurement mic response; bottom: iPhone recorded to voicemail system. Relative levels shown for voicemail recording.

The suppression appears sensitive to background noise level. For example, at NC 50 twenty decibels of suppression was applied to the voicemail-recorded pink noise whereas ten decibels of suppression was applied in the background noise condition. Also, between tests suppression would occur at different times and rates resulting in different recordings of the pink noise.

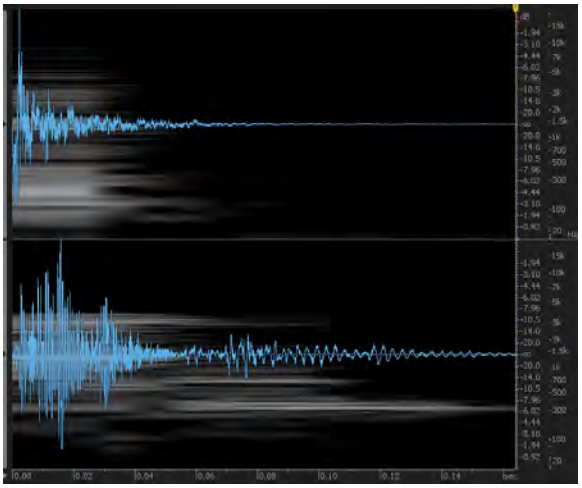

Figure 9. Balloon pop response. Top: measurement mic response; bottom: iPhone recorded to voicemail system. Abscissa: 0-0.16 s; ordinate: 20-20 kHz.

4.5. Balloon pop test

Figure 9, top, shows the acoustical response of a balloon pop following a period of "silence" (20 dBA) from the position of the loudspeaker at 1 m to the microphone, as recorded by the voice mail system. As in Figure 7, the peak sound pressure level was 134 dB. Following the initial fracture impulse of the balloon are early reflections extending out to ~30 ms. Figure 9, bottom shows the same response as recorded by the voicemail system. The impulse is captured without evidence of overload, but the amplitude peak occurs 18 ms from the time of the onset, and a signal processing "echo" can be discerned from 70–160 ms. The echo appears as a low-pass version of the initial burst.

This peak delay and echo has implications for gunshot analysis of timings. A separate test conducted with a different telephone (Motorola RAZR) compared the response of an indoors gunshot recorded through an entire 911 emergency recording system at different locations through a residence, to a recording from a reference microphone located 1 m away. The reference recording showed a characteristic brief impulse of a gunshot in an acoustically damped room (~150 ms). The 911 recording had a different pattern of signal processing echoes that lengthened the response time to about 1 s.

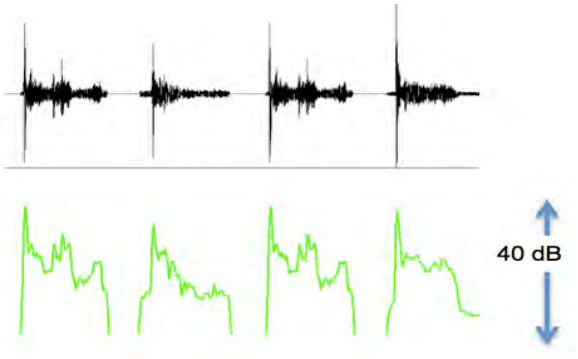

Figure 10. Gunshot responses under controlled conditions, transmitted by a Motorola RAZR telephone indoors and recorded by a 911 dispatch center. Top: waveform; bottom: amplitude envelope. Duration of each burst is ~1 s.

5. OVERALL CONCLUSIONS

We have offered several examples of transformations of an acoustic signal by a cellular telephone, focusing on a specific voice mail system. The non-linearity of responses under different test conditions are largely explained by a system that is optimized to transmit intelligible speech from a nearby talker in an efficient, economical manner, under a variety of noise conditions and other factors. This optimization causes a number of either unpredictable or "irregular" results for signals the device was not intended for, but that become of interest to the forensic audio investigator, such as distant speech or gunshots. Linear predictive coding, noise cancellation and speech detection-compression signal processing are likely causes. Notably, the voice codec used here affects speech formants quite differently than other codecs tested (G.723 or MSV LPEC-SP) [8].

The implications for timing analysis of impulsive events, voice comparison, and transcription of speech are amongst the results described here that should be considered in forensic analysis of audio originating from cellular telephones.

6. ACKNOWLEDGEMENTS

We would like to acknowledge the support of Anthony Nash and our colleagues at Charles M. Salter Associates.

7. REFERENCES

[1] E. B. Brixen, "Acoustics of the crime scene as transmitted by mobile phones," *Audio Eng. Soc. 126th Convention,* Munich, DE, May 2009. Preprint 7699.

[2] A. H. Moore, M. Brookes, and P. A. Naylor, "Room identification using roomprints", *Proc. of the 54th Audio Eng. Soc. Int. Conference, "Audio Forensics: Techniques, Technologies and Practice,"* London, UK, pp. 47-53, June 2014.

[3] C. A. Kardous, P. B. Shaw, "Evaluation of smartphone sound measurement applications," *J. Acoust. Soc. Am.*, vol. 135, EL186-EL192, 2014.

[4] J. Kotus, A. Ciarkowski, A. Czyzewski, "Auto adaptation of mobile device characteristics to various acoustic conditions," *Audio Eng. Soc. 136th Convention*, Berlin, Germany, April 2014.

[5] ANSI/ASA, *Method For Measuring The Intelligibility Of Speech Over Communication Systems*, ANSI Standard S3.2-2009 (R2014).

[6] H. Levitt, J. C. Webster, "Effects of Noise and Reverberation on Speech," in Handbook of Acoustical Measurements and Noise Control, 3rd ed. McGraw-Hill, 1991, ch. 16.

[7] F. Tavera, "Acoustical measurement software housed on Mobile operating systems test," *Audio Eng. Soc. 135th Convention*, NY, USA, Oct. 2013.

[8] E. B. Brixen, and D. R. Begault, "Validity of Bit Compressed Digital Voice Recordings for Spectrographic Analyses: Input for a Database, a Preliminary Test, " *Proc. of the 26h Audio Eng. Soc. Int. Conference, "Audio Forensics in the Digital Age" Denver, CO,* pp. 47-53, July 2005.

.

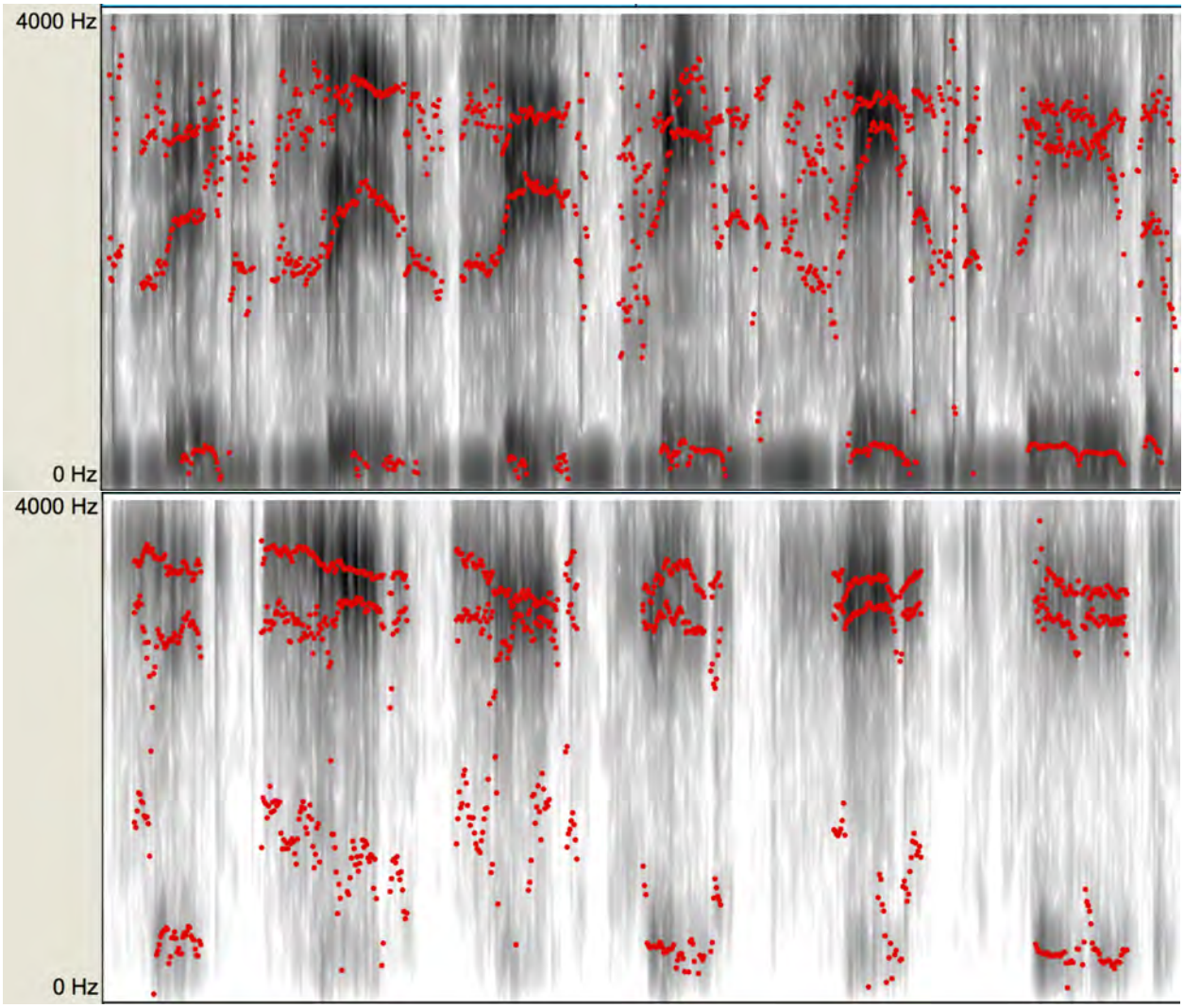

Figure 4. Spectrogram of 3 male followed by 3 female talkers, uttering the word "seed." Top: recording from reference microphone. Bottom: same, from voicemail recording.

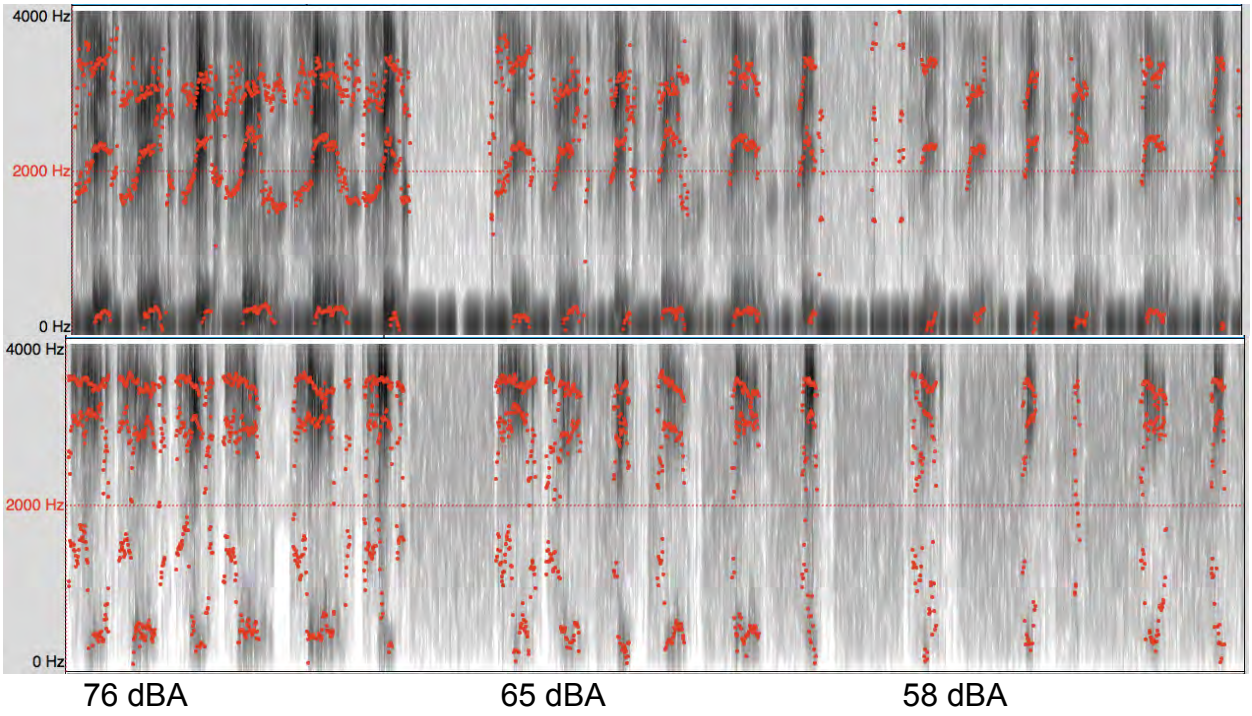

Figure 5. Spectrogram, one male talker uttering the word "seat seed seek seem seen seep." Top: recording from reference microphone. Bottom: same, from voicemail recording. Grouped by level indicated below.

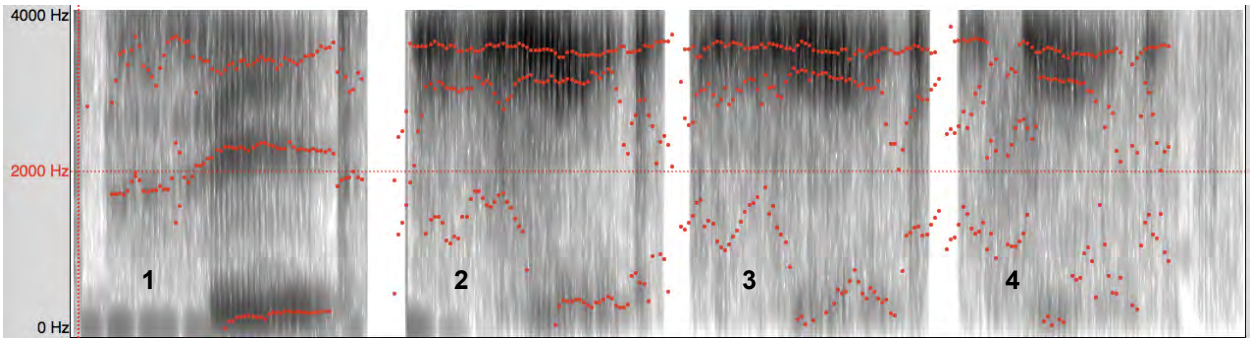

Figure 6. Spectrogram, one male talker uttering the word "seat," loudspeaker at 1 m. Key- 1: reference microphone, 76 dBA at 1 m; 2: from iPhone at 76 dBA; 3: from iPhone, 65 dBA; 4: from iPhone, 58 dBA.

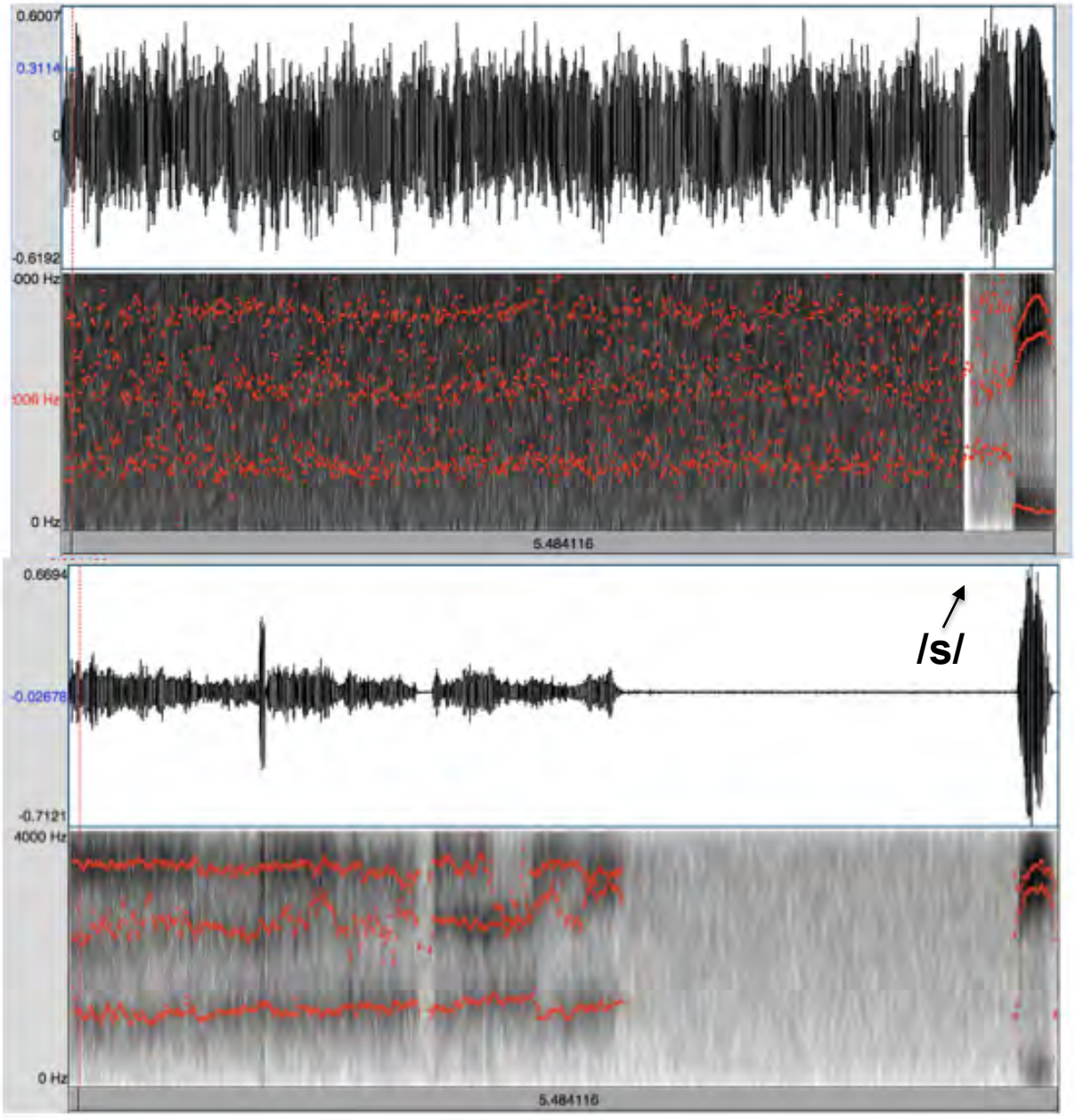

Figure 7. Spectrogram, pink noise followed by female talker uttering the word "seat," 70 dBA, loudspeaker at 1 m. Top: reference microphone, bottom: iPhone voice mail.