Undergraduate Research Symposium
Madison, Wisconsin
13 April 2017
http://waisman.wisc.edu/vocal/

# Effects of Sampling Rate of Speech Waveform Acoustic Measurements

Madison Wagner, Paul Milenkovic, Ray D. Kent, and Houri K. Vorperian

Vocal Tract Development Lab, Waisman Center, University of Wisconsin - Madison

## I. BACKGROUND / INTRODUCTION:

- In digital signal analysis, a fundamental issue is the selection of sampling rate, the rate at which samples of the digitized waveforms are obtained. For studio-quality recordings, the usual practice is to cover the frequency range of human audition (20 to 20,000 Hz). A sampling rate of 48 kHz achieves this goal.
- LPC (Linear Predictive Coding) is an analysis method for speech. Although not required in TF32 (2), speech signals are typically *downsampled* for LPC (by processing the speech signal to reduce its sampling rate). This is motivated, in part, by the acoustic model on which analysis is based. This model strictly applies to frequency below 4 kHz and only loosely applies beyond 4 kHz.

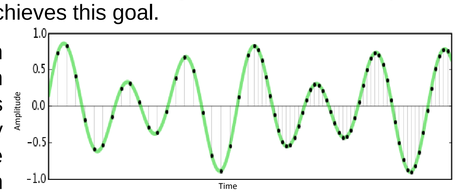


**Figure 1:** Black dots on a waveform indicating sampling points* on the waveform. Note the variation in sampling rate over the course of time, with lower sampling rate on the left and higher sampling rate (downsampling) on the right.
* http://manual.audacityteam.org/man/digital_audio.html

- One of the most common goals of speech analysis is the estimation of formant frequencies, which are the resonant frequencies of the particular speech sound. For most purposes, vowel sounds can be characterized with the frequencies of the lowest formants: F1, F2, and F3. The addition of F4 is useful to account for speaker variations in vocal tract anatomy.
  - Formants are the resonances of the vocal tract and are often identified by observing peaks in a Fourier spectrum analyses expressing variation in signal intensity across its frequency range. As the shape of the vocal tract changes during speaking, the formant frequencies change.
  - LPC gives the variation in signal intensity across its frequency range in a simplified form allowing automatic determination of formant peaks. This form derives from a physical approximation of the vocal tract where the acoustic waves locally satisfy plane-wave propagation (wave fronts perpendicular to the vocal tract wall), a condition best satisfied for frequencies below 4 kHz. Output of automatic formant analysis from LPC may be displayed as "tracks" overlaying a time-frequency Fourier spectrogram.
- Formant frequencies broadly vary in inverse relation with the length of the vocal tract, and therefore with the age and sex of the speaker. Men have low values of formant frequencies, women have somewhat higher values, and young children have even higher values. Acoustic analyses must accommodate these speaker-dependent variations because analyses suited to the speech of adult males are not always optimal for adult females or children.
- It is not clear if downsampling, typically done in speech analysis systems either improves or degrades the accuracy of formant frequency estimation for vowels produced by different groups of speakers.

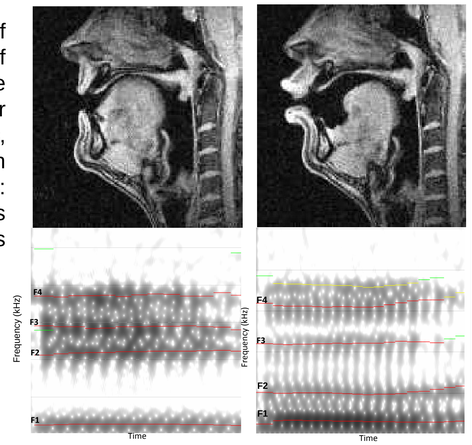


**Figure 2:** Vocal tract shapes and corresponding acoustic analyses (spectrograms). Top panel shows the MRI midsagittal plane of and adult male producing the vowels /i/ (left) and /u/ (right)*. Bottom panel shows the spectrograms of the respective vowels. Time is on the horizontal axis, frequency is on the vertical axis, and signal intensity is represented by greyscale. Vocal tract resonant frequencies (formants) are marked F1, F2, F3, and F4. The faint horizontal lines are at 1 kHz frequency spacing. The red lines are the automatic formant tracks based on LPC.
* https://www.quora.com/Do-you-think-a-200-letter-alphabet-would-be-better-for-English-and-French

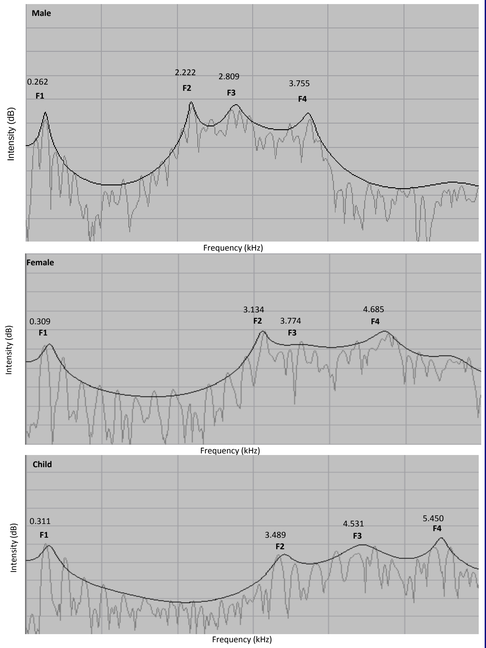


**Figure 3:** Time spectral slice showing intensity as a function of frequency for vowel /i/ as in "eat" of an adult male (top panel), adult female (middle panel) and child (bottom panel) speakers. The grey plot depicts the Fourier spectrum of an interval of the speech signal; the darker plot is the LPC spectrum giving a simplified intensity variation locating the peaks designating formant frequencies F1-F4.

## II. OBJECTIVE:

This study investigates how downsampling the speech signal affects measurement of formant frequencies by the LPC method in the speech of men, women and children.

## III. METHODS:

Participants: Speech recordings of 12 participants without speech or language impairments were taken from an acoustics database. Participants included eight adults (4 males, 4 females) ages 20-24 years, and four children (2 males, 2 females) age 5 years.

Recording Procedures: Participants' speech had been recorded using a studio-quality microphone and digital audio recorder. Participants were instructed to repeat one syllable words containing the vowels /i/ as in "eat", /u/ as in "hoot", /ae/ as in "hat", and /ɑ/ as in "hot". Words were recorded at a sampling rate of 48 kHz.

Analysis Protocol:

- The most stable vowel interval between 25 and 50 ms in length was selected from each word to capture the vowel steady state.
- The acoustic speech signal on each selected interval was downsampled to 16 kHz and to 10 kHz using "foldover" filter, "gap" filter, and no filter.
  - The 'resample' function in MATLAB applied to the 48 kHz signal allows specifying an anti-alias filter to reject spectrum components outside of the reduced spectrum frequency ranges for the sampling rates of 16 and 10 kHz -- the respective frequency ranges for the three sampling rates are 24, 8, and 5 kHz.
  - The "foldover" filter distorts the spectrum shape less by allowing some foldover of higher frequency spectrum components into the frequency range after downsampling; the "gap" filter eliminates foldover by introducing a gap in the spectrum within the frequency range after downsampling; "no filter" distorts the spectrum least but allows maximal foldover.
- TF32 (2) was selected for acoustic analysis in this study because it does not require downsampling.
- The fundamental (F0) and first through fourth formant frequency measurements (F1-F4) were made for each sampling rate and filter setting at the midpoint of the steady state vowel interval of each word. Manual intervention allowed overriding an assignment of an LPC frequency to a particular formant number (F1, F2, F3, F4) when considered an error while retaining the automatically determined frequency value.
- The number of LPC coefficients was adjusted for speaker age and sex because the smaller vocal tract lengths cause their formants to be more widely spaced than those in adult males.
- The 'smooth' post processing operation in TF32 was applied to the automatic LPC formant tracking.
- When the signal was not strong enough for the automatic LPC tracking, the manual selection of the peaks from the LPC spectral slice was used to determine the formant frequencies.

## REFERENCES:

(1) Kent, R. D., & Read, C. (2002). *Acoustic Analysis of Speech: Second Edition.* Albany, NY: Singular Thomson Learning.

(2) Milenkovic, P. (2016). Time-frequency analysis software program for 32-bit Windows (Alpha Version 3) [Computer software]. A version of the TF32 software is available at http://userpages.chorus.net/cspeech/

## IV. RESULTS:

As shown in Figure 4, the differences of formant frequencies for adult males and females from 48 kHz sampling rate were greater for 10 kHz than 16 kHz. The accuracy of the higher formant frequencies in children was compromised by the lowest sampling rate, 10 kHz, irrespective of filter used. The higher formant frequencies either could not be measured, or were measured incorrectly. Fundamental frequency (F0) was not affected by sampling rate. Discrepancies between filters were minimal for all frequencies and therefore only the filter information for the "foldover" filter was graphed in Figure 4.

- **Adult Males:** mild effect of SR for higher formants; no effect on filter condition
- **Adult Females:** mild effect of SR on higher formants; no effect on filter condition
- **Children:** extreme effect of SR on higher formants; no effect on filter condition

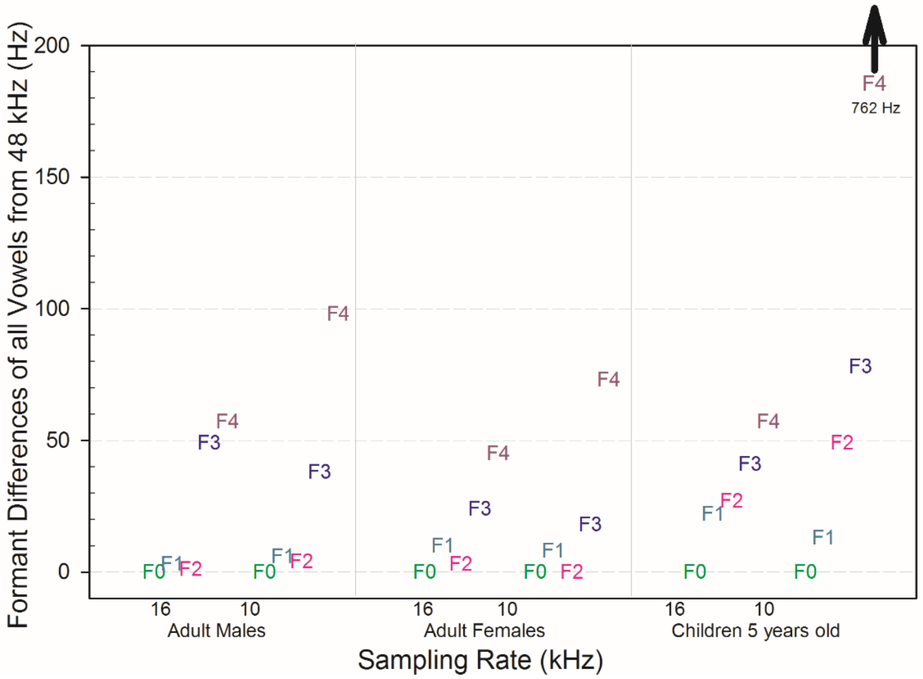


Figure 4: Formant Frequency differences between 48 kHz sampling rate (SR) and both 16 kHz SR and 10kHz SR. These values have been averaged over all vowels for adult males, adult females, and children 5 years old.

## V. CONCLUSION:

- The acoustic signal of speech encodes information of human communication, and therefore it is important to develop accurate measures of formant frequencies to better understand this process (1).
- The measurements of formant frequencies throughout the human lifespan are important because they correlate with changes to the vocal tract that occur with development and aging. This study is pertinent to ongoing work in the Vocal Tract Development Lab to help establish acoustic and anatomic correlates using data from individuals with and without speech or language impairments. Such information can in-part help us understand how persons with developmental disabilities, such as Down syndrome, differ in their speech.
- As the highest formant frequency in children's speech can exceed 5 kHz, accurate measurement requires a sampling rate higher than 10 kHz. This study shows that it is important to select a number of LPC coefficients appropriate to the subject population.
- Formant values for men and women showed larger differences between 48 kHz and 10 kHz than between 48 kHz and 16 kHz, but the filter condition at a given sampling rate made a much smaller difference.
- For accurate formant measurements, higher sampling rates such as 16 kHz or 48 kHz are recommended for all populations but particularly for children.

**ACKNOWLEDGEMENTS:**
This work was supported by NIH research grant R01 DC006282 from the National Institute on Deafness and other Communication Disorders (NIDCD), and core grants P30 HD03352 and U54 HD090256 from the National Institute of Child Health and Human Development (NICHD).