Effects of sampling rate and type of anti-aliasing filter on linear-predictive estimates of formant frequencies in men, women, and children

Paul H. Milenkovic, Madison Wagner, Raymond D. Kent, Brad H. Story, and Houri K. Vorperian

Citation: The Journal of the Acoustical Society of America **147**, EL221 (2020); doi: 10.1121/10.0000824
View online: https://doi.org/10.1121/10.0000824
View Table of Contents: https://asa.scitation.org/toc/jas/147/3
Published by the Acoustical Society of America

JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0000824

Effects of sampling rate and type of anti-aliasing filter on linear-predictive estimates of formant frequencies in men, women, and children

Paul H. Milenkovic,[1,a)] Madison Wagner,[2] Raymond D. Kent,[2] Brad H. Story,[3] and Houri K. Vorperian[2]

[1]*Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, Wisconsin 53706, USA*

[2]*Vocal Tract Development Laboratory, Waisman Center, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, Wisconsin 53706, USA*

[3]*Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85718, USA*

phmilenk@wisc.edu, madisonashawagner@uwalumni.com, kent@waisman.wisc.edu, bstory@email.arizona.edu, vorperian@waisman.wisc.edu

Abstract: The purpose of this study was to assess the effect of downsampling the acoustic signal on the accuracy of linear-predictive (LPC) formant estimation. Based on speech produced by men, women, and children, the first four formant frequencies were estimated at sampling rates of 48, 16, and 10 kHz using different anti-alias filtering. With proper selection of number of LPC coefficients, anti-alias filter and between-frame averaging, results suggest that accuracy is not improved by rates substantially below 48 kHz. Any downsampling should not go below 16 kHz with a filter cut-off centered at 8 kHz.

[Editor: Charles C. Church] Pages: EL221–EL227

Received: 8 November 2019 **Accepted:** 10 February 2020 **Published Online:** 4 March 2020

1. Introduction

Linear predictive coding (LPC) calculates the coefficients of a digital filter from the acoustic speech signal (Atal and Hanauer, 1971). The frequency response of this filter approximates the acoustic spectrum with a small number of these coefficients. Although this all-pole model has known limitations (Shadle *et al.*, 2016; Vallabha and Tuller, 2002), it is not subject to the time-frequency tradeoff of the Fourier spectrum, giving it the potential to identify narrow and/or closely spaced formant peaks. LPC is widely used to derive estimates of formant frequencies, but the accuracy of such estimates may vary with sampling rate and anti-alias filter as well as with the characteristics of the speech samples that are under analysis.

Telephone signals sampled at the 8 kHz standard are an important early use of LPC. This low rate diminishes speech intelligibility, especially of fricatives. By allowing examination of energy at higher frequencies, a higher sampling rate benefits research along with clinical studies of speech. Although low-cost recording and data-storage devices sample at 44.1 or 48 kHz to meet requirements of the studio recording industry, recent speech studies often downsample data stored at a higher rate prior to LPC analysis (Alku *et al.*, 2013; Plichta, 2002; Shadle *et al.*, 2016). Many widely disseminated software packages impose such downsampling (Burris *et al.*, 2014).

Reasons for downsampling include (1) higher rates require proportionally more LPC coefficients, resulting in a high computation load, (2) the spectral valleys at higher frequencies could make the LPC calculation numerically unstable (Makhoul, 1975), and (3) the simplified plane-wave acoustic-tube underlying the all-pole model (Markel and Gray, 1976) breaks down at higher frequencies, on account of higher-order modes (Sondhi, 1974) or pharyngeal anti-resonances (Dang and Honda, 1997). Reasons for not downsampling include (1) owing to advances in microelectronics, storage and computation are of greatly reduced cost, (2) the LPC calculation may be stabilized by matrix regularization (Makhoul, 1975), (3) anti-resonances departing from the all-pole model, including nasalization, occur across the entire vowel frequency range, and (4) anti-alias filtering and downsampling introduce distortion into the all-pole model, even if the filter corner frequency is well above the highest formant of interest.

With respect to the last reason, an analog all-pole model with a limited number of formants requires a high-frequency correction, which varies continuously with the vocal tract length (Olive, 1971). The digital all-pole model in LPC has an implicit correction from the low-frequency

[a)]ORCID: 0000-0003-0506-1497.

poles being periodically repeated over the infinite frequency range (Rabiner and Schafer, 1978). As the sampling rate increases, more of the higher poles need to be represented explicitly, which accounts for the need for proportionately more LPC coefficients. The effective vocal tract length varies in steps with integer changes in the number of LPC coefficients. Depending on the placement of the anti-alias filter corner frequency in relation to the LPC filter poles, an artifact in the way foldover or a spectrum gap results from the interaction of the natural frequency content of the vowel with the filter is introduced immediately below and also above the half-sampling frequency. Because the overall all-pole fit is not strictly local to the narrow frequency range about spectrum features (Olive, 1971), formants well below the Nyquist frequency could also be affected.

These considerations motivate comparing LPC formant estimates from a 48 kHz sampled signal with different downsampling conditions; 44.1 kHz is expected to give similar results. These concerns are especially pronounced for speakers whose formants extend to higher frequencies, as is the case for women and children. Although data are not abundant for the third and fourth formants, it is likely that the fourth formant of children's vowels reaches or exceeds 5 kHz (Kent and Vorperian, 2018). The specific aim of the present study is to quantify the effect of sampling rate and anti-alias filtering techniques on LPC formant estimates from recorded signals of three speaker groups (men, women and children). Our goal is to provide analysis guidelines to speech clinicians, clinical researchers and other users along with developers of acoustic-analysis software.

2. Methods

2.1 LPC analysis procedure

The first difference of the sampled acoustic signal gives $y[n] = s[n] - s[n-1]$. Under simplifying assumptions about the combined spectrum tilt of the voice source and lip radiation load, this *preemphasis* flattens the spectrum to better approximate the underlying vocal-tract frequency response. The *covariance method* of LPC finds coefficients giving a least-squares prediction from a linear combination of p prior samples $\hat{y}[n] = -a_1 y[n-1] \cdots - a_p y[n-p]$ (Markel and Gray, 1976). Such may also be expressed as an *inverse filter* $e[n] = y[n] + a_1 y[n-1] \cdots + a_p y[n-p]$ giving a prediction error signal of maximum spectral flatness. The reciprocal of the inverse filter frequency response supplies an estimate of the vocal tract acoustic frequency response where the complex values of z satisfying $1 + a_1 z^{-1} \cdots + a_p z^{-p} = 0$ are the *zeroes* of the inverse filter, as well as the *poles* of the forward filter representing the vocal tract response. A polynomial root solver—our study used Laguerre's method (Press, 1989)—finds these zeroes (and consequently, the poles). Each complex-valued pole is parameterized by the resonant frequency and bandwidth of a second-order digital filter. The narrowband poles provide candidate estimates for formant frequencies.

Factoring a *covariance matrix* computed from $y[n]$ is a step in solving for the LPC coefficients. *Regularizing* the covariance matrix by adding a small constant to its diagonal elements stabilizes the factorization when the matrix has a wide eigenvalue range. This condition can occur when the acoustic spectrum has a wide range of intensity (in dB) from spectrum roll off (Makhoul, 1975), either by the voice source, acoustic anti-resonances, or signal pre-filtering. The regularization constant in this study is equivalent to adding a low level (−30 dB) of white noise to $y[n]$. The structure of the covariance matrix allows Cholesky factorization (Rabiner and Schafer, 1978), which in order notation (Cormen *et al.*, 2009) requires $O(p^3)$ operations for p LPC coefficients. An alternative method (Hu, 1988) is $O(p^2)$.

Owing to the underlying approximations, the covariance method can generate negative bandwidths, resulting in an unstable forward filter unrepresentative of the passive acoustics of the vocal tract. Although filter instability poses a problem in speech synthesis, a negative bandwidth is not actually an impediment to estimating formant frequencies. Alternatively, the *autocorrelation* LPC method guarantees a stable filter, but at the expense of a time-bandwidth tradeoff in the analysis interval. The Burg method assures filter stability too by performing a nonlinear averaging operation between prediction in the forward and backward directions (Childers, 1978), but it does not benefit from Hu's fast factorization. For this study, the covariance method was used for evaluating the effect of sampling frequency on the estimation of formants.

To reduce the influence of analysis frame alignment with the voice source pulses, the least-squares LPC interval was aligned with glottal epochs marking maximal excitation. This was performed with the following automatic procedure that required no manual correction, judging by visual inspection of formant epoch intervals of the type in Fig. 1. Epochs were located in the acoustic signal after further anti-alias filtering and downsampling, LPC inverse filtering and finally smoothing of the inverse filter output. Feeding $s[n]$ through a fourth-order 2-kHz low-pass Butterworth filter and downsampling (by factor 6 from 48 kHz to 12 kHz, by factor 2 from 16 kHz to 8 kHz, no further downsampling at 10 kHz) generates $s_L[n]$. An LPC analysis not used for estimating formants, was applied to a 33 ms interval of $y_L[n] = s_L[n] - s_L[n-1]$. Signal $s_L[n]$ is input to an LPC inverse filter, which is followed by a fourth-order zero-phase low-pass filter

JASA EXPRESS LETTERS
https://doi.org/10.1121/10.0000824

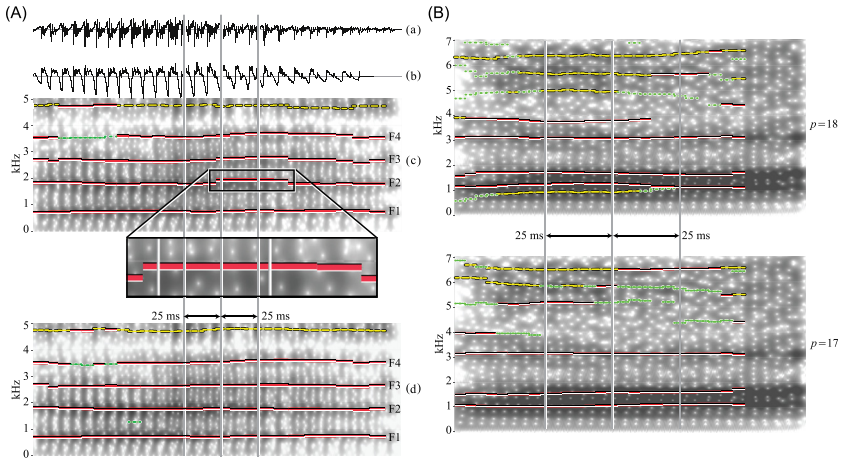

Fig. 1. (Color online) (A) Adult male utterance of the word "hat" sampled at 48 kHz showing (a) acoustic wave, (b) LPC inverse filter glottal flow derivative, (c) time-frequency spectrum overlaid with formant tracks (solid lines), F2 anomaly from log area ratio averaging (rectangular inset), and (d) no anomaly with formant tracks from averaging autocorrelation coefficients. (B) Adult female utterance "hot," downsampled to 16 kHz with "foldover" anti-alias filter and log area ratio averaging, showing sensitivity of formant tracks (solid lines) to change in number of LPC coefficients from $p = 18$ (top panel) to $p = 17$(bottom panel).

for waveform smoothing. This inverse filter suppressed formant oscillations for the purpose of epoch detection, without the need to identify formant frequencies. The corner frequency of the smoothing filter is set to the zero-crossing rate of its input.

The resulting signal $d_L[n]$ gives a smoothed estimate of the first derivative of the glottal airflow signal. The pitch period is estimated from a cross correlation analysis of $d_L[n]$ for time shifts up to 25 ms. The cross correlation interval is the prior pitch period adjusted to no less than 5 ms and no more than 8 ms, defaulting to 8 ms on voice onset. The pitch-period candidate is the first correlation peak c_p occurring at time shift. t_p satisfying thresholds $c_P > 0.5\,c_{\max}$ and $c_P > \alpha\,c_0$ in relation to the highest peak $c_{\max}$ at shift $t_{\max}$. Coefficient α starts at 0.75 and declines linearly until it is fixed at 0.5 when time shift $t_p \geq 0.5\,t_{\max}$. These non-dimensional thresholds track the pitch period under conditions of varying excitation amplitudes. The Pearson correlation coefficient is $\rho_p = c_p/\left(\sqrt{c_0 c_{pp}}\right)$ where c_0 also gives the signal energy for the correlation interval at zero shift and c_{pp} gives the same at t_p. When $c_0 < 4c_{\max}$ or $\rho_p < 0.7$, the 33 ms interval is advanced past this spurious glottal epoch, and the pitch-period is reported as zero.

For frames reporting a non-zero pitch period, formants were estimated from a second LPC analysis applied to $y[n]$ on contiguous, non-overlapping intervals containing one glottal epoch each. Searching for the peak within an estimated pitch-period interval of $d_L[n]$ locates the epoch. Each interval was aligned so the epoch occurs after the first quarter of that estimated pitch-period interval. Any error in epoch finding will shift the position and length of this analysis interval over a restricted range, which is expected to have a minor effect on formant frequencies.

Selecting a "closed-glottis interval" (Wong *et al.*, 1979) or conducting "weighted linear prediction" (Alku *et al.*, 2013) identifies parameters of the vocal tract impulse response by observing only the tail of that impulse response. Whereas improved separation of the vocal-tract response from the voice source is observed on synthetic speech, these alternative methods can be sensitive to variability in natural speech not reproduced in the synthetic test cases along with changes with alignment of the analysis interval. This interpretation is supported by variability of these methods observed in natural speech (Shadle *et al.*, 2016).

Even with pitch-period intervals that are less sensitive to precise alignment than closed-glottis intervals, considerable period-to-period variability was observed, especially for the children with higher fundamental frequency. To mitigate this, non-linear averaging was applied to the epoch-delimited frames. The LPC coefficients from these pitch-period intervals were transformed into log area ratio coefficients, averaged, and then transformed back to LPC coefficients for root solving and calculating formant-frequency candidates. Log area ratios describe the shape of a highly idealized hard-walled acoustic tube having the same frequency response as the LPC forward filter (Wakita, 1979). The resulting shape may not realistically model the actual vocal tract (Sondhi, 1979), but interpolating that shape accurately tracks rapid changes in formant frequencies (Shadle and Atal, 1979).

An alternative procedure transforms the LPC coefficients to the low-order autocorrelation coefficients (ACFs), which are averaged and then transformed back to LPC coefficients by the autocorrelation LPC algorithm. For stable formants in a sustained vowel, this averaging does not have the bandwidth widening of the autocorrelation method where those coefficients are calculated from a truncated signal interval. Under rapid formant movement, however, this procedure is

equivalent to averaging the acoustic power spectrum that blurs formants (Shadle and Atal, 1979). Results for both averaging methods are presented in Sec. 3.

Finally, estimates of formant frequencies need to be selected from the LPC-generated candidates. An automatic procedure was used to assign LPC candidates to formants based on frequency ranges at each frame along with continuity between frames. This study, however, applied manual correction to the resulting assignments of formants to LPC poles. The automatic selection of formant frequencies along with plots of other LPC formant-frequency candidates overlay a wideband Fourier time-frequency spectrogram display—see Fig. 1 in Sec. 3. A human operator selects from these candidates when correcting the formant tracks. In this way, the formant frequencies all correspond to calculations by the LPC algorithm. This reduces formant assignment errors in the automatic procedure as a source of variability.

2.2 Acoustic recordings

The source material for this study included recordings from the University of Wisconsin-Madison Health Sciences IRB-approved Acoustics Database of the Vocal Tract Development Lab. The recordings were from 12 participants [four adult males mean age = 21:3 (years:months) SD = 0:9; four adult females mean age 22:7; SD = 2:5 and four 5-year-old children (2 male, 2 female)] producing four monosyllabic words each containing one of the four corner vowels (*eat, hoot, hat*, and *hot*). Recordings from ten of the twelve participants were formerly used in Derdemezis (Derdemezis *et al.*, 2016), and two additional recordings of five-year-old participants were selected from the same database. All recordings were made in a quiet room with a cardioid-pattern Shure SM 48 microphone (Shure Inc., Niles, IL) 15 cm from each subject's mouth, feeding a Marantz PMD660 digital audio recorder (Marantz Professional, Cumberland, RI). See Derdemezis *et al.* (2016) for additional detail on acoustic recording methodology. The recorder samples the microphone output at 48 kHz with 16-bit resolution. These data were transferred to a desktop computer for further processing and analysis.

2.3 Downsampling procedures

The Nyquist frequency of half the sampling frequency determines the maximum signal frequency range and must be considered prior to downsampling: 0–24 kHz at 48 kHz, 0–8 kHz at 16 kHz and 0–5 kHz at a 10 kHz sampling frequency. For example, a signal downsampled from the original 48 kHz to 16 kHz will have a Nyquist frequency of 8 kHz. This means that spectrum components at 9 and 7 kHz in the original signal become indistinguishable after downsampling. The 9 kHz component from the original signal "folds over" the Nyquist frequency (8 kHz) and appears incorrectly at 7 kHz. Prior to downsampling, a digital low-pass filter may be applied to the signal sampled at 48 kHz to attenuate frequency components above 8 kHz to suppress this effect commonly called aliasing.

For many speech-acoustics software packages, and the cited papers on LPC analysis of formants, detail of the downsampling procedure is neither disclosed nor obvious from context. For example, measurements of the amplitude and frequency of sine-wave signals conducted on the CSL software package (Computerized Speech Laboratory model 4500, Version 2.7.0; Kay Elemetrics, 1996) show spectrum foldover consistent with downsampling without a low-pass filter. Downsampling was therefore conducted in our study using the following MATLAB commands implementing well-documented algorithms that are readily replicable. The MATLAB variables *S48*, *F16*, *G16*, *S16*, *F10*, and *G10* contain the signal sampled at 48, 16, and 10 kHz where, in each case, *S* denotes no rate-conversion anti-alias filter, *F* denotes the "foldover" configuration for a filter tolerating foldover over a narrow frequency range and G denotes the "(spectrum) gap" configuration that more rigorously suppresses foldover, where

F16 = resample(S48, 1, 3),
G16 = resample(S48, 1, 3, fir1(60, (8–1.35)/24, kaiser(61, 5))),
S16 = downsample(S48, 3),
F10 = resample(S48, 5, 24), and
*G10 = resample(S48, 5, 24, 5*fir1(480, (5- 0.85)/120, kaiser(481, 5))*.

Whereas the *resample* command applies an anti-alias filter allowing the rate to change by the ratio of the supplied integer factors, the *downsample* command omits this filter and only allows downsampling by an integer factor. Hence the S16 condition relies entirely on the natural frequency roll-off of the audio signal to limit foldover. There is no S10 condition because 10 kHz and the original 48 kHz are not related by an integer factor.

The MATLAB *resample* command with integer parameters $P = 1$ for the upsample factor and $Q = 3$ for downsample factor converts 48 kHz to 16 kHz; $P = 5$ and $Q = 24$ upsamples 48 kHz to 240 kHz and downsamples back down to 10 kHz. Downsampling thus follows upsampling, as needed, to change the rate by the ratio of integers P/Q. Upsampling is accomplished by filling in zero-valued samples between samples of the original signal. A digital anti-alias filter applied to this

https://doi.org/10.1121/10.0000824

upsampled signal is given a corner frequency appropriate to the frequency range of the signal after downsampling. A multirate filter (Crochiere and Rabiner, 1983) uses a minimum number of calculations by skipping filter taps for the zero-fill samples on the input side and by calculating only the samples to be retained after downsampling on the output side.

The default filter for *resample* gives the "foldover" (*F*) configuration placing the *corner frequency*, marking the start of the filter roll-off frequency band, at the Nyquist frequency for the downsampled rate. Considering this default setting in MATLAB as the de facto standard, the present study makes its settings explicit. The default frequency width of its roll-off band is nominally f_s/N with $N = 10$ or 1.6 kHz at 16 kHz, 1 kHz at 10 kHz sampling rate. The default filter type uses a Kaiser window with a stop-band response of −50 dB. Measuring its frequency response finds the default width closer to 1.35 kHz at 16 kHz, 0.85 kHz at 10 kHz. When downsampling to 16 kHz, for example, the default anti-alias filter will therefore "fold over" frequency components from 8 up to 9.35 kHz to the range 8 down to 6.65 kHz. Owing to this roll-off behavior of the anti-alias filter, the foldover artifact in this case is minimal at or below 6.65 kHz and increases to its largest amount at 8 kHz.

The "foldover" anti-alias filter indeed introduces artifact into a restricted frequency band of the downsampled signal. An alternative avoids foldover artifact by lowering the filter corner to 6.65 kHz at 16 kHz and to 4.15 kHz at the 10 kHz sampling rate, in this manner placing the upper edge of the roll-off band at the Nyquist frequencies 8 and 5 kHz, respectively. Such a method substitutes a spectrum gap from the roll-off of the anti-alias filter for the aliasing artifact, both conditions affecting the frequency range downward from the Nyquist frequency by the extent of the roll-off band. In this "gap" configuration, default settings of the MATLAB anti-alias filter need to be overridden, including the default *filter order* $n = 2 \cdot N \cdot Q$ ($N = 10$) supplied to the MATLAB *fir1* command and total number of filter coefficients $n + 1$ supplied to the *Kaiser* command as shown above. The lowered corner frequency is the fraction of the Nyquist frequency at the highest sampling rate occurring in the upsample-followed-by-downsample chain.

Although the manufacturer of the digital audio recorder for the S48 condition does not disclose its internal anti-alias filter, analog-to-digital modules intended for use in such recorders implement the "foldover" condition of a cutoff frequency centered at 24 kHz (PCM1801 single-ended analog-input 16-bit stereo analog-to-digital converter, single-supply 16-bit Sigma-Delta stereo ADC AD1877). The anti-alias protection at 24 kHz already afforded by the roll-off of the microphone and along with the natural roll-off for vowels in natural speech, however, makes differences in anti-alias protection between recorders of secondary importance.

3. Results

Table 1 lists the numbers of LPC coefficients p applied to each of three speaker groups along with different sampling rates. Regarding the adult males as having an average formant spacing of 1 kHz, with each formant resonator modeled by a pair of LPC coefficients, and with the Nyquist frequency range being half the sampling frequency, these speakers are assigned one LPC coefficient per kHz of sampling rate to represent the vocal-tract frequency response. To account for reduced vocal tract lengths giving more widely spaced formants, the adult females are assigned 90% of the adult-male value, and both the male and female children aged 5 years are assigned 70% of the adult-male value, rounded to the nearest whole number of coefficients.

The number of coefficients p in each speaker category was increased to represent the spectrum shaping of the voice source and lip radiation load. The baseline analyses were conducted by assigning 4 extra coefficients for all speaker groups and sampling rates. Formant tracks are computed after averaging log area ratio values from LPC analysis applied to epoch-aligned pitch-period intervals of the speech waveform. The running average centered on each pitch-period interval includes an equal number of prior and past intervals. The starting times of each interval are limited to ±25 ms from the start of the central interval. Each token is for a speaker, a vowel and a

Table 1. Number of LPC coefficients p by subject group and sampling rate: Baseline p keeps the number of coefficients for source spectrum shaping fixed; adjusted p varies these coefficients in proportion to the sampling rate.

		Adult males			Adult females			Children age 5		
		Tract	Source	Total	Tract	Source	Total	Tract	Source	Total
48 kHz	baseline p	48	4	52	43	4	47	34	4	38
	adjusted p	48	9	57	43	9	52	34	9	43
16 kHz	baseline p	16	4	20	14	4	18	11	4	15
	adjusted p	16	3	19	14	3	17	11	3	14
10 kHz	baseline p	10	4	14	9	4	13	7	4	11

downsampling condition. Formant values are sampled from these tracks for each token at a location centered in the steady-state portion of the vowel. For example, formant values for the two conditions in Fig. 1(A) are each from 5 frames centered on the left cursor spanning 45.4 ms, in Fig. 1(B) from 11 such frames spanning 45.6 ms.

Each entry in Table 2 gives the maximum-absolute formant-frequency difference across the four vowels produced by the four speakers in a subject group (16 tokens). Between-rate differences S48-F16 and F16-F10 were higher than within-rate differences F16-G16, F10-G10, and F16-S16, especially for the adult male group. For adult females, comparing the "foldover" with "gap" anti-alias conditions gave increased differences at the higher formants, especially at the 10 kHz rate where the spectrum gap is closer to those formant frequencies. The effect is more pronounced for the child group—comparisons with 10 kHz do not list F4 owing to the large number of tokens where that formant exceeds the 5 kHz Nyquist frequency.

Large formant differences are seen between sampling rates, especially for formants F3 and F4 and even for adult males, whose lower fundamental frequency and other voice source characteristics are considered favorable to the LPC acoustic model. In investigating the origin of these differences, high sensitivity of higher formants to the number of LPC coefficients p at 16 kHz was seen. In light of this observation, p was adjusted to make the source-shaping coefficients proportionate to the sampling rate as with the vocal tract coefficients. The adjustment reduced the source coefficients from 4 to 3 at 16 kHz and increased them from 4 to 9 at 48 kHz as reported in Table 1. The row in Table 2 labeled "baseline-adjusted p" hence varies the number of coefficients by one at 16 kHz, giving a measure of formant sensitivity to number of coefficients. The difference in this row are comparable to the first-row differences between the S48 and F16 condition at the baseline p. This is consistent with formant shifts with change in sampling rate being influenced by coefficient sensitivity at a given sampling rate.

The two rows for "adjusted p" compare 48 kHz with 16 kHz at the adjusted number of coefficients, first for averaging log area ratios, next for averaging the ACF. The first of these rows shows a pronounced reduction in formant shifts, except for one entry listing 158 Hz. In Fig. 1(A), this high difference occurred for one vowel of one adult male subject, exhibiting a noticeable departure of the track for formant F2 from its expected position on a time-frequency spectrogram. The lower panel of Fig. 1(A) shows formant tracks generated by averaging LPC-derived autocorrelation coefficients. Not only does this remove the F2 anomaly for this subject (rectangular inset of upper panel), it reduces the formant differences for all three subjects apart from two entries with minor increases (below 14 Hz) as seen in the second of the two "adjusted p" rows in Table 2.

Figure 1(B) shows formant overlays for the vocalic portion of "hot" from an adult female speaker giving an ambiguous F4. The comparison is between two values of the number of LPC coefficients using log area ratio averaging. This particular speech token did not contribute to a large formant shift for F4 in Table 2 on account of a formant dropout—the $p = 17$ condition lacks an appropriate LPC formant candidate near the position where $p = 18$ shows F4. The spectrogram also shows F4 vanishing, possibly by cancellation by a tract anti-resonance or a voice-source null.

Figures 1(A) and 1(B) also show the alignments of the epoch-aligned pitch-period analysis frames with the durations of the horizontal bars showing a constant formant value (solid line) or LPC formant candidate (broad dashes for a pole bandwidth below 500 Hz, fine dashes for above 500 Hz). Visual inspection showed accurate automatic epoch labeling for all vowels of all

Table 2. Maximum absolute formant differences (Hz) of all vowels by subject group (log area ratio averaging apart from last row using ACF averaging). Blank entries occur for conditions not reliably giving F4 values. S = no rate-conversion anti-alias filter, F = foldover filter, and G = spectrum gap filter. Baseline p and adjusted p given in Table 1.

	Adult males				Adult females				Children age 5			
Baseline p:	|ΔF1|	|ΔF2|	|ΔF3|	|ΔF4|	|ΔF1|	|ΔF2|	|ΔF3|	|ΔF4|	|ΔF1|	|ΔF2|	|ΔF3|	|ΔF4|
S48-F16	63	34	207	261	178	103	195	303	214	90	327	178
F16-F10	15	65	199	270	43	63	98	392	103	239	294	
F16-G16	8	47	68	71	27	26	82	205	24	45	183	89
F10-G10	7	13	63	91	21	30	227	290	23	102	547	
F16-S16	37	11	34	48	22	53	244	189	16	138	165	141
baseline-adjusted p:												
F16	62	41	59	292	194	96	142	204	62	176	134	116
adjusted p:												
S48-F16	20	158	72	69	67	77	121	127	37	61	61	95
ACF: S48-F16	20	6	35	31	67	80	68	119	29	59	74	65

https://doi.org/10.1121/10.0000824

subjects with the exception of one vowel of a male subject—that one condition did not contribute to a maximum formant frequency shift reported in the tables.

4. Discussion

The preceding quantification of the effect of sampling rate and anti-alias filter condition on LPC formant estimates reveals that (1) the anti-alias filter used for downsampling to 16 kHz perturbs formant frequency estimates, (2) the no-anti-alias-filter condition extends this effect to lower formants, (3) the "foldover" anti-alias filter gives the least shift, (4) the 10 kHz rate does not include adequate frequency range to measure fourth formant in children, (5) the differences with sampling rate are least when the total number of coefficients is made proportional to the sampling rate, (6) the undownsampled 48 kHz rate offers finer control over the acoustic tube length, such as when optimizing the number of coefficients based on the shape of the LPC acoustic tube model (Vallabha *et al.*, 2004; Vallabha and Tuller, 2002), and (7) averaging autocorrelation coefficients instead of the log area ratios should be considered when analyzing stable vocalic segments.

From these observations, we conclude that downsampling from 48 kHz is not needed for LPC formant analysis. Should downsampling be used, as with a large data set where storage and calculation cost remain a consideration, we recommend using the "foldover" anti-alias filter with a target rate no lower than 16 kHz. Because sampling at 44.1 kHz is similarly oversampled as 48 kHz in relation to the frequency range of the formants to be measured, we expect similar results for acoustic recordings at 44.1 kHz using the recommended scaling of the number of LPC coefficients.

Acknowledgments

This work was supported, in part, by NIH-NIDCD Research Grant No. R01 DC6282. Also, NIH-NICHD core Grant Nos. P30 HD03352 and U54 HD090256. We thank Courtney A. Miller for assistance with figures.

References and links

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., and Story, B. H. (**2013**). "Formant frequency estimation of high-pitched vowels using weighted linear prediction," J. Acoust. Soc. Am. **134**, 1295–1313.

Atal, B. S., and Hanauer, S. L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am. **50**, 637–655.

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., and Bolt, D. M. (**2014**). "Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements," J. Speech Lang. Hear. R. **57**, 26–45.

Childers, D. G. (**1978**). *Modern Spectrum Analysis* (IEEE Computer Society Press, New York).

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (**2009**). *Introduction to Algorithms* (MIT Press, Cambridge).

Crochiere, R. E., and Rabiner, L. R. (**1983**). *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).

Dang, J., and Honda, K. (**1997**). "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101**, 456–465.

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., and Bolt, D. M. (**2016**). "Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups," Am. J. Speech-Lang. Path. **25**, 335–354.

Hu, Y. H. (**1988**). "A rotation based method for solving covariance and related linear systems," *ICASSP-88, International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Piscataway, NJ), pp. 1659–1662.

Kent, R. D., and Vorperian, H. K. (**2018**). "Static measurements of vowel formant frequencies and bandwidths: A review," J. Commun. Disorders **74**, 74–97.

Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.

Markel, J. D., and Gray, A. H. J. (**1976**). *Linear Prediction of Speech* (Springer, New York).

Olive, J. P. (**1971**). "Automatic formant tracking by a Newton-Raphson technique," J. Acoust. Soc. Am. **50**, 661–670.

Plichta, B. (**2002**). "Best practices in the acquisition, processing, and analysis of acoustic speech signals," Univ. Penn. Work. Pap. Ling. **8**, 16.

Press, W. H. (**1989**). *Numerical Recipes in Pascal* (Cambridge University Press, New York).

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).

Shadle, C. H., and Atal, B. S. (**1979**). "Speech synthesis by linear interpolation of spectral parameters between dyad boundaries," J. Acoust. Soc. Am. **66**, 1325–1332.

Shadle, C. H., Nam, H., and Whalen, D. H. (**2016**). "Comparing measurement errors for formants in synthetic and natural vowels," J. Acoust. Soc. Am. **139**, 713–727.

Sondhi, M. (**1979**). "Estimation of vocal-tract areas: The need for acoustical measurements," IEEE Trans. Acoust. Speech Sign. Proc. **27**, 268–273.

Sondhi, M. M. (**1974**). "Model for wave propagation in a lossy vocal tract," J. Acoust. Soc. Am. **55**, 1070–1075.

Vallabha, G., Tuller, B., Slifka, J., Manuel, S., and Matthies, M. (**2004**). "Choice of filter order in LPC analysis of vowels," Sound Sense **50**, B148–B163.

Vallabha, G. K., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**, 141–160.

Wakita, H. (**1979**). "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," IEEE Trans. Acoust. Speech Sign. Proc. **27**, 281–285.

Wong, D., Markel, J., and Gray, A. (**1979**). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust. Speech Sign. Proc. **27**, 350–355.