CrossMark click for updates

Comparing measurement errors for formants in synthetic and natural vowels[a)]

Christine H. Shadle,[b)] Hosung Nam,[c)] and D. H. Whalen[d)]
Haskins Laboratories, 300 George Street, New Haven, Connecticut 06511, USA

(Received 4 August 2014; revised 4 December 2015; accepted 8 January 2016; published online 9 February 2016)

The measurement of formant frequencies of vowels is among the most common measurements in speech studies, but measurements are known to be biased by the particular fundamental frequency ($F0$) exciting the formants. Approaches to reducing the errors were assessed in two experiments. In the first, synthetic vowels were constructed with five different first formant ($F1$) values and nine different $F0$ values; formant bandwidths, and higher formant frequencies, were constant. Input formant values were compared to manual measurements and automatic measures using the linear prediction coding-Burg algorithm, linear prediction closed-phase covariance, the weighted linear prediction-attenuated main excitation (WLP-AME) algorithm [Alku, Pohjalainen, Vainio, Laukkanen, and Story (2013). J. Acoust. Soc. Am. **134**(2), 1295–1313], spectra smoothed cepstrally and by averaging repeated discrete Fourier transforms. Formants were also measured manually from pruned reassigned spectrograms (RSs) [Fulop (2011). *Speech Spectrum Analysis* (Springer, Berlin)]. All but WLP-AME and RS had large errors in the direction of the strongest harmonic; the smallest errors occur with WLP-AME and RS. In the second experiment, these methods were used on vowels in isolated words spoken by four speakers. Results for the natural speech show that $F0$ bias affects all automatic methods, including WLP-AME; only the formants measured manually from RS appeared to be accurate. In addition, RS coped better with weaker formants and glottal fry. © *2016 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4940665]

[DAB] Pages: 713–727

I. INTRODUCTION

Vowel production has often been characterized by measurements of the formants, especially the formant frequencies (Chiba and Kajiyama, 1941). These have been used for many purposes: for example, to characterize the differences in vowel space of men, women, and children (e.g., Peterson and Barney, 1952) and of the hearing-impaired (Monsen, 1976); to compare dialects (e.g., Clopper and Pierrehumbert, 2008); to compare speaking styles (Bradlow, 2002); and to provide basic data with which formant synthesizers can be specified (Allen *et al.*, 1987). The acoustic theory of speech production shows that the shape of the vocal tract determines the acoustic output we then perceive as speech (Fant, 1960). As vocal tract imaging has advanced, the first test of a new method often involves a comparison of vowel formants: those measured from a speaker's acoustic output, those predicted from that speaker's vocal tract shape as measured with a new technique, and those predicted using previous techniques (e.g., Baer *et al.*, 1991; Davies *et al.*, 1992; Story *et al.*, 1996).

However, it has also long been recognized that measuring formant frequencies is not straightforward. Early studies used manual measurements of narrowband spectral slices (Potter and Steinberg, 1950; Peterson and Barney, 1952); although they did not specify in detail how their subjects arrived at their measurements, they did study the consistency of both measurers and speakers. One graph of formant frequencies plotted against the fundamental frequency shows a correlation, but they could not ascertain whether this was inherent in the production or in the measurement method (Potter and Steinberg, 1950, Fig. 11). When linear prediction coding (LPC) analysis became widespread, formants were measured automatically using LPC, but its drawbacks were described often: bandwidths are consistently underestimated (Atal, 1975; Atal and Schroeder, 1978), and formant frequencies are biased by the fundamental frequency (Atal, 1975; Monsen and Engebretson, 1983; Klatt, 1986; Fulop, 2010), particularly when fundamental frequency is high and/or the first formant is low. Studies that have examined errors in formant estimation have generally been forced to report the range of errors rather than recommend ways in which the error can be reduced (Vallabha and Tuller, 2002; Mehta *et al.*, 2012; Burris *et al.*, 2014).

The ANSI/ASA standard of acoustic terminology defines the formant (ANSI, 2013, p. 62) as "a range of frequencies in which there is an absolute or relative maximum in the sound spectrum. Unit, hertz (Hz). The frequency at the maximum is the formant frequency." However, as noted recently, "…as speech analysis and synthesis have progressed in a half century, the definition has not been universally maintained. Fant (1960, pp. 20, 53) defined formants as the poles of the transfer function of the supraglottal vocal tract…. He was followed in

[a)]Portions of this work were presented at the ASA meetings in San Francisco, December 2013, and Providence, May 2014.
[b)]Electronic mail: shadle@haskins.yale.edu
[c)]Also at: Department of English Language and Literature, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 136-701, South Korea.
[d)]Also at: Program in Speech-Language-Hearing Sciences, CUNY Graduate Center, 365 Fifth Ave., New York, NY 10016, USA.

this path by many authors…" (Titze *et al.*, 2015, p. 3006). We follow Fant's definition also.

A central problem for assessing techniques is what to take as the gold standard; hand measurement by experts seemed to be the most obvious candidate, perhaps in part, because it predated automatic algorithms. In pitch-tracking, manual measurements, sometimes aided by machine estimates, have been used as the gold standard against which automatic algorithms are compared (Rabiner *et al.*, 1976), but detecting periodicity in a waveform by eye is a simpler task than determining the resonance frequencies. All the harmonics of the fundamental contribute to the signal, while a very small number (usually 1–3) contribute to any particular formant. Early studies (Potter and Steinberg, 1950; Peterson and Barney, 1952) used manual measurements of printed (actually burned) spectrograms made after training, with checks on the consistency of the measurers. However, only consistency, not accuracy, could be assessed. Monsen and Engebretson (1983) compared LPC to manual measurements made from spectrograms of synthetic speech, and found that errors were approximately ±60 Hz for both methods for $F1$ and $F2$ when $F0$ was between 100 and 300 Hz. For $F3$ in this $F0$ range, manual measurements had a larger error; LPC error remained the same. For higher $F0$, the error increased for both methods. More recently, Zhang *et al.* (2013) compared formant measurements of Standard Chinese vowels made by human "supervisors" correcting LPC-extracted formants using a spectrogram and speech synthesis as guides to aid their corrections. Within-supervisor reliability was from 16 to 22 Hz for $F1$; across-supervisor reliability, 25 Hz for $F1$, indicating that the supervisors had different biases. As in earlier studies, there was no way to verify the accuracy of the formants measured, only their consistency. For their application, forensic voice comparison, reliability of the formant measures was far more important.

The next method for finding a gold standard is to compare sets of measurements by two or more analysis techniques, again searching for consistency. Often LPC was one of the methods, and it has often implicitly been taken as the standard. For instance, Woehrling and Maureuil (2007) compared formants of two French dialects measured using Praat and SNACK, two software systems that are widely used. They found substantial differences (e.g., 63 Hz for $F1$, 113 Hz for $F2$), with Praat formant frequencies higher than SNACK ones, on average, but had no way to determine which (if either) was correct.

A third approach is to measure vocal tract resonances directly (Fujimura and Lindqvist, 1971; Castelli and Badin, 1988; Djeradi *et al.*, 1991; Pham thi Ngoc and Badin, 1994; Epps *et al.*, 1997; Joliveau *et al.*, 2004; Swerdlin *et al.*, 2010; Henrich *et al.*, 2011), which, if successful, would be the most straightforward method. However, these methods generally require the vowel to be sustained, and some require that no sound be produced by the person while the measurement is made (e.g., Castelli and Badin, 1988). Other studies have compared analysis of the acoustic signal to prediction of the formants from vocal tract shape measurements, but they are generally focused on validating the vocal tract shape measurements rather than testing acoustic analysis methods (Baer *et al.*, 1991; Story *et al.*, 1996; Narayanan *et al.*, 1997; Story *et al.*, 1998).

The final approach is to take the values used to synthesize speech as the standard. Such an approach is itself reliant on the accuracy of the synthesis algorithm. Although such algorithms are clearly largely successful, we have no direct way of verifying the results more finely other than to use the very techniques that we are interested in testing the accuracy of. Nonetheless, speech synthesis appears to be our best approximation to knowing the correct formant value ahead of time.

In a study by Klatt (1986), synthetic stimuli were used to assess the error of various formant analysis methods. An all-pole synthesizer used nine values of $F0$, ranging from 133 to 200 Hz, to excite, in turn, a set of four formants with fixed frequency and bandwidth similar to the vowel /ɪ/. He chose the $F0$ values to be in approximately equal logarithmic steps, and to include two cases where a harmonic would line up with $F1$ (at 400 Hz) precisely (the third harmonic for 133 Hz, the second for 200 Hz). There was no additive noise or voicing irregularities. Three analysis methods were used: (1) assigning $F1$ to the frequency of the strongest harmonic; (2) smoothing the discrete Fourier transform (DFT) spectrum using a 300-Hz wide Gaussian filter; and (3) LPC analysis, using the autocorrelation method, 14 poles, and a 25.6 ms Hamming window. Although the model on which LPC is based is especially appropriate for the synthetic stimuli, "since they were generated from an all-pole synthesizer and have virtually no noise or voicing source irregularities" (Klatt, 1986, p. 5), the error was still sizable, with the maximum error ranging from −4% to +9%. A bias toward the nearest harmonic of $F0$ was evident, not only in the harmonic method (which had maximum error from −15% to +16%) and wideband filter method (−8% to +7%), but also in the LPC analysis.

Vallabha and Tuller (2002) used both synthetic and real speech to investigate the accuracy of formant estimation by LPC, and how it depends on peak-picking vs root-solving, $F0$ quantization effects (that is, the sampling of the spectral envelope by the harmonics), the order of the LPC analysis, and the nearness of formants to each other. The synthetic speech included some deliberately designed test cases (e.g., with only two formants), and natural speech consisted of two sustained vowels spoken by two speakers. They found that the error increases linearly with $F0$, and the error range can be large for $F1$ because of its typically small bandwidth.

The optimal LPC order depends not only on the speaker, but on the vowel; back vowels require a higher order. Root-solving was found to be more error-prone when formants approach each other or their complex conjugates, and for higher-bandwidth formants; peak-picking using parabolic interpolation exhibited $F0$ bias, with errors higher for low-bandwidth formants.

Fulop (2010) used ten synthetic vowels in which the formants were stationary but $F0$ varied during the 300 ms signals from 120 Hz downward to about 75 Hz. He compared formants derived from LPC using the Burg algorithm (an asynchronous algorithm; Burg, 1967; Andersen, 1974, both reprinted in Childers, 1978; Press *et al.*, 1986), asynchronous

covariance, and closed-phase covariance to his own method of reassigned spectrograms (RSs) with pruning. A range of formant values was tabulated for the LPC methods, indicating how much the formant estimates varied with $F0$ change. The Burg method and asynchronous covariance generated $F1$ values that varied during the 300 ms vowel by 22–78 Hz; in nearly all cases that range did not include the true value. Closed-phase covariance results were worse. By contrast, the RS values for $F1$ had errors ranging from 0 to 8 Hz, and did not vary with $F0$.

Alku *et al.* (2013) used synthetic and natural vowels with $F0$ ranging from 100 to 450 Hz to test their weighted linear prediction with attenuated main excitation (WLP-AME) algorithm against five other linear prediction (LP) algorithms, all designed to measure the formants of high-$F0$ voices more accurately. The algorithm uses a temporal weighting function that attenuates the prediction error during the closed phase so that the times at which the residual error increases would not unduly warp the results of the analysis. The glottal closure instants are found either by using the electroglottograph (EGG) signal, if one exists, or by processing the speech signal itself. The synthetic vowels were generated using a physical modeling approach to create voices corresponding to adult men and women, and a child; this approach was chosen to avoid the circularity of using a synthesis model that closely matched the analysis model of LPC. Natural vowels were obtained from adult subjects sustaining vowels for 2 s at increasingly high $F0$'s. The WLP-AME algorithm was shown to have the smallest error of the six algorithms tested; its error values stayed relatively low as $F0$ increased, unlike the general pattern of the other algorithms. For the natural speech, the formant estimates were shown to exhibit less variability as $F0$ increased than conventional LP; a plausible explanation is that the WLP-AME algorithm reduces $F0$ bias.

Burris *et al.* (2014) used synthetic and natural speech to compare four acoustic analysis systems: Praat, Wavesurfer, TF32, and CSL. They found the results for all but the CSL system to be accurate and comparable—defined as within 5% of the synthesized value—for $F1$–$F4$ for most synthetic vowels, and comparable for adult male vowels. Results varied by vowel for adult female and children's vowels, however.

The goal of this paper is to evaluate automatic analysis methods that can be used on vowels in isolated words, so that large data sets can be analyzed in a realistic amount of time but with the greatest accuracy. Post-processing methods using, e.g., adaptive Kalman filtering (Deng *et al.*, 2007), which are designed for use with running speech, are thus not appropriate. Collection of an ancillary signal such as EGG is also assumed not to be feasible. Methods that depend on sustained vowels are not appropriate. If a manual method can be shown to be much more accurate, it can be argued that it is worth the cost in terms of time spent on analysis, but the ideal is to have an automatic method. In this study we first compare analysis methods using synthetic speech in a replication of Klatt's (1986) classic study, in which the ground truth is relatively well known, so that error patterns can be measured. We then compare the same analysis methods using a speech corpus in which some of the formants are likely to be relatively constant, though unknown, and any error patterns related to $F0$ bias are likely to be apparent.

II. EXPERIMENT 1: SYNTHETIC SPEECH

A. Method

1. Stimuli

Synthetic stimuli were generated for the parameters given by Klatt (1986) using his $F1$ value of 400 Hz, as well as four additional values (350, 375, 425, and 450 Hz). All used a first formant bandwidth of 50 Hz, with the other formants held constant [$F2$ ($B2$) = 1800 (140) Hz; $F3$ ($B3$) = 2900 (240) Hz; $F4$ ($B4$) = 3800 (350) Hz]. Klatt's nine values of $F0$, 133, 139, 145, 152, 160, 169, 179, 189, and 200 Hz, were used. Praat's Klattgrid commands were used to generate the 45 signals. A sampling rate of 10 kHz was used; each signal was 1 s long, with $F0$ and the four formants constant throughout. As in Klatt's (1986) study, there was no additive noise or voicing irregularities, no jitter or shimmer, no tracheal or nasal formants or antiformants, and no frication.

2. Measurers

Four colleagues were recruited to measure the formants manually, using narrowband spectra. All four were speech scientists who had measured formants manually before: the least experienced, Ph3, for 2 years; the others had from 20 to 60 years' experience.

3. Procedure

For the hand measurements, a single token was measured for each of the 45 $F0$/$F1$ combinations. The order of the stimuli was randomized before they were sent to the measurers (all measurers received stimuli in the same random order). They were asked to measure $F1$ from narrow-band spectral cross-sections and to detail their methods after sending their measurements. The details varied among panel members; three of the subjects used Praat, and its default pre-emphasis of the signal; the fourth used Macquirer and did not use pre-emphasis. Window lengths with which the narrowband spectra were generated were 30 ms for two, 40 ms for one, and ten glottal cycles (which would vary from 50 to 75 ms) for the fourth measurer. The descriptions of their estimation methods were very similar: they used three or four harmonics surrounding a peak. If three harmonics formed a symmetric peak, with the highest- and lowest-frequency harmonics being equal in amplitude, they estimated the formant frequency to be the same as the frequency of the central, highest amplitude harmonic. An asymmetry shifted their estimate toward the higher-amplitude harmonic; one measurer noted that the lowest-frequency harmonic was "discounted somewhat to allow for source spectral tilt."

Five semiautomatic methods were selected (after some exploration of their parameters), three of which consist of different types of linear prediction analysis. First, we used the Burg method of LPC analysis, which is that recommended by

the Praat manual for obtaining formants (Boersma and Weenink, 2013). A 30 ms window was used; an order of 14 was specified for this asynchronous LPC analysis, which matches the order used by Klatt (1986) for his autocorrelation-based LPC analysis, and also that used by Shue *et al.* (2011). An order of 14 actually would be sufficient to specify 5 formants plus 4 poles to match glottal and radiation spectral characteristics, 1 more formant than that used by Klatt; we note the inconsistency. Formants were found by manual peak-picking.

Second, we used closed-phase LP covariance analysis (hereafter, LP-CP or CP). The closed phase was identified automatically using a search for the minimum amplitudes in the waveform (see, e.g., Holmes, 1976, for an explanation of this method). The MATLAB command arcov was used specifying an order of 10. The output was processed semiautomatically to identify the poles corresponding to formants rather than identifying peaks in the general spectral shape.

Third, WLP-AME (Alku *et al.*, 2013) was used with an order of 14, matching that of the Burg method. The algorithm depends on determining the glottal closure instant, using either the EGG signal (as reported in Alku *et al.*, 2013) or the speech signal (as in the code we obtained from Manu Airaksinen). For this study, we computed the glottal closure instants from the speech signal in order to create a fair comparison between synthetic and natural speech. The LP-closed-phase analysis was then redone using the same glottal closure instants. Formants were found by both peak-picking and root solving.

Two other analysis methods were implemented from Harrington and Cassidy (1999): averaged DFT analysis (hereafter, AVG), and cepstral analysis (hereafter, CEPS). For the AVG analysis, DFTs were computed using 6 ms windows zero-padded to 1024 points and overlapped by 1 ms; 6 such windows were located within 30 ms beginning at 330 ms of the synthetic vowel signals. The six DFTs were averaged, and peak-picking was used to determine the formant frequencies in the final spectrum. CEPS also used a 30 ms window beginning at 330 ms in the synthetic signals. The highest 25 cepstral coefficients were removed to filter out the harmonics. Formant frequencies were then found in the resulting spectral envelope by peak-picking.

Finally, as a sixth method, we used a manual method: RSs following a pruning procedure (Fulop 2011). We programmed a graphical user interface (GUI) that allowed the frequency range being viewed and the pruning thresholds to be varied easily. Default thresholds for the phase derivative of 0.1 (for line components) and 0.2 (for impulses) were set up, as suggested by Fulop (2011, pp. 136–137). The entire waveform and a zoomed-in portion of it were visible simultaneously, and 40 ms of the RS was visible centered on the cursor in the zoomed-in waveform. The RS was computed using 60-sample frames (6 ms) zero-padded to 1024 points, with a frame advance of 2 samples; these parameters are within ranges useful for speech (Fulop 2007, 2011).

In the RS, each cycle could be easily identified; the particular single part of the track within each cycle that should be used for a single formant measurement was somewhat open to interpretation, as described by Fulop (2011). After independent measurements by the first two authors, followed by conferring with each other to establish consistency, the highest-amplitude part of the track occurring after the initial impulse was chosen as the formant value to be used.

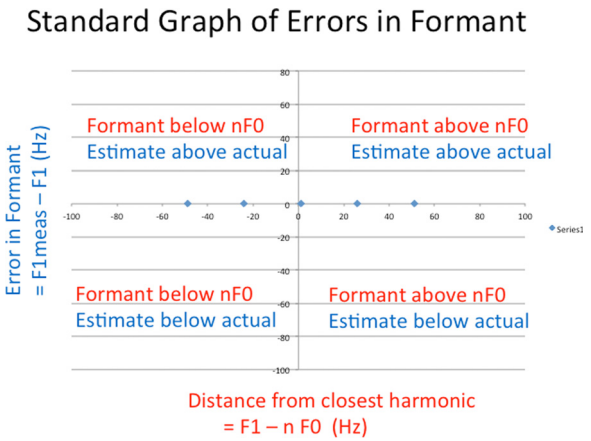

FIG. 1. (Color online) Graph format used for Figs. 2–5 and 7. $F1$ indicates actual $F1$ frequency (known for synthetic vowels). $F1$meas indicates $F1$ as measured by the algorithm specified with each graph.

B. Results

Given that errors in formant measurement have been most closely linked to $F0$, we measured accuracy relative to $F0$. A way of displaying the measurements that would illustrate such a relationship was therefore devised that consisted of plotting as ordinate the difference in Hz between the estimated and actual formant, $F1_{est} - F1$, and as abscissa, the difference between the actual formant frequency and that of the nearest harmonic, $F1 - nF0$, as shown in Fig. 1. The particular method of measurement is indicated as a subscript on the measured quantity, for example, $F1_{Burg}$. Correct measurements should lie on the x axis. For all graphs in this format, the aspect ratio is kept the same to facilitate comparison. Thus, in all graphs of $F1$ error (in Figs. 2–5, and 7), y ranges over 180 Hz (either −100 to +80 or −120 to +60 Hz); in graphs of $F2$ error, y ranges from −25 to +20 Hz.

The manual measurements, shown in Fig. 2, showed similar patterns of errors in formant measurements for all four phoneticians. The strongly linear pattern with a negative slope indicates that when the harmonic nearest to $F1$ is below it in frequency, the estimated $F1$ is below actual $F1$ (right lower quadrant); when the nearest harmonic is above $F1$, estimated $F1$ is above actual $F1$ (left upper quadrant). When a harmonic coincides with $F1$, one might predict that the $F1$ measured would be identical to the target $F1$; this would correspond to data points at (0,0). In fact, only one subject, Ph4, followed the predicted pattern, while subjects Ph1, Ph2, and Ph3 tended to underestimate $F1$ in this case. This is likely due to the fact that the amplitude of the source harmonics falls off with frequency, so that when a harmonic coincides exactly with $F1$, the harmonic below it will have a higher amplitude than the harmonic above it. The differences between participants may be due to differences in whether

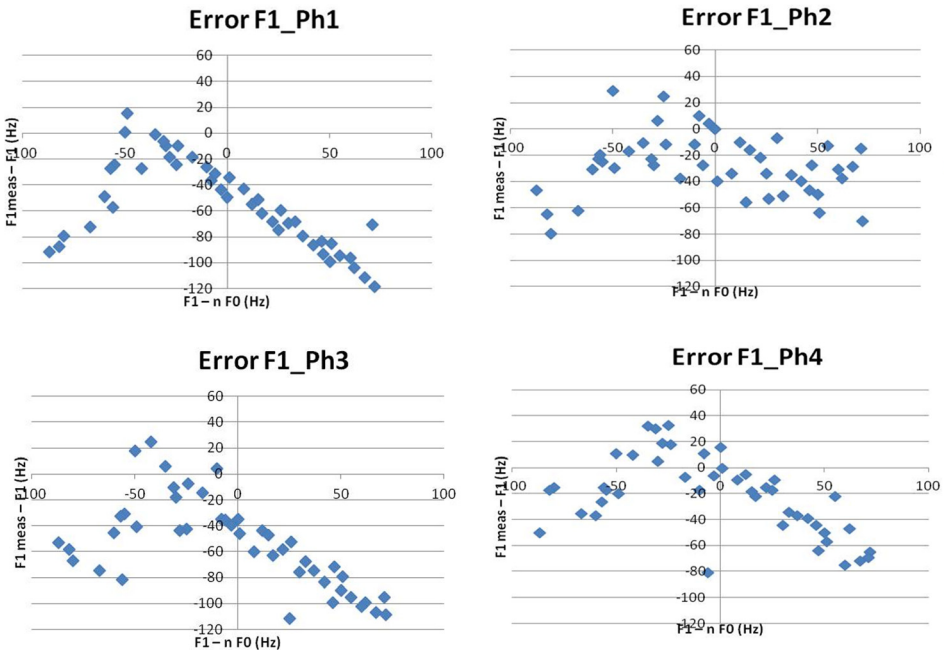

FIG. 2. (Color online) Errors in $F1$ measurement of 45 synthetic vowels for each of the 4 phonetician subjects.

they used pre-emphasis, and whether they allowed for that in their interpretation of the spectra. Errors ranged from a minimum of 32 Hz (8%, for $F1 = 400$ Hz) to a maximum of 118 Hz (26%, for $F1 = 450$ Hz).

Figure 3 shows the error pattern for averaged and cepstral measurements. The patterns differ somewhat from the manual measurements, but still exhibit $F0$ bias. The errors are still large: for AVG, the range is -87 to $+69$ Hz, with a standard deviation of 42.1 Hz; for CEPS, the range is -108 to $+55$ Hz, standard deviation of 29.5 Hz. We do not specify a mean error because the large positive and negative errors would nearly cancel, as can be seen in Fig. 3.

Figure 4 shows the error pattern for LPC-Burg and LP-closed-phase covariance measurements, both using peak-picking, plotted with the same axis ranges as the plots of manual measurements. As with the manual measurements, the LPC-Burg measurements fall on a main diagonal, indicating that the error in the formant measurements is biased by $F0$. The slope is not as large as for the manual measurements, however, indicating that the errors are not as large. When the difference between the specified formant and the nearest harmonic (the x axis value) is >40 or <-20, the pattern breaks down somewhat; that is, the error decreases. The asymmetry

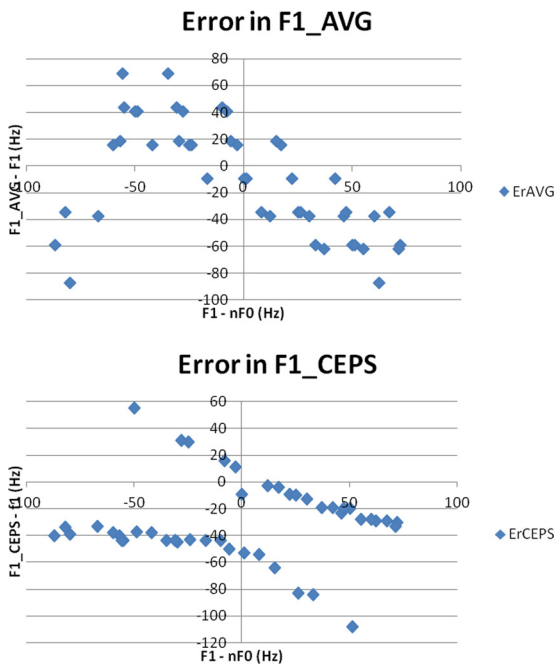

FIG. 3. (Color online) Errors in $F1$ measurement of 45 synthetic vowels for (top) spectral averaging and (bottom) cepstral algorithms.

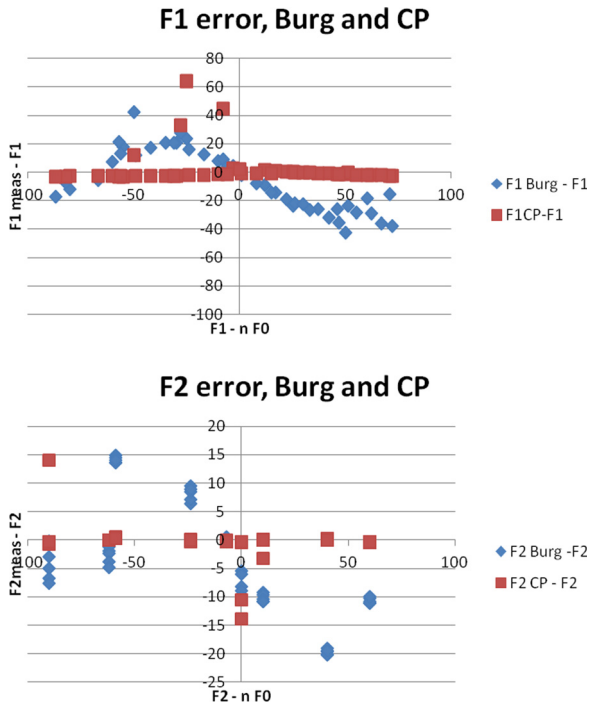

FIG. 4. (Color online) (Top) Errors in $F1$ and (Bottom) F2 measurement of 45 synthetic vowels using LPC-Burg with peak-picking (diamonds) and LP-closed-phase covariance (or LP_CP) with peak-picking (squares) methods.

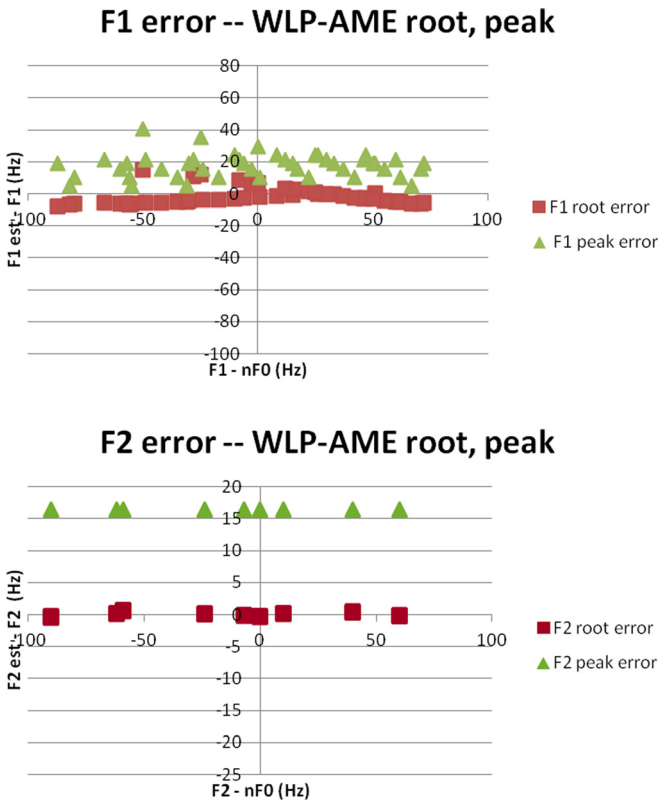

FIG. 5. (Color online) (Top) Errors in $F1$ and (bottom) F2 measurement of 45 synthetic vowels with WLP-AME, using peak-picking (triangles) and root-solving (squares) methods in both graphs.

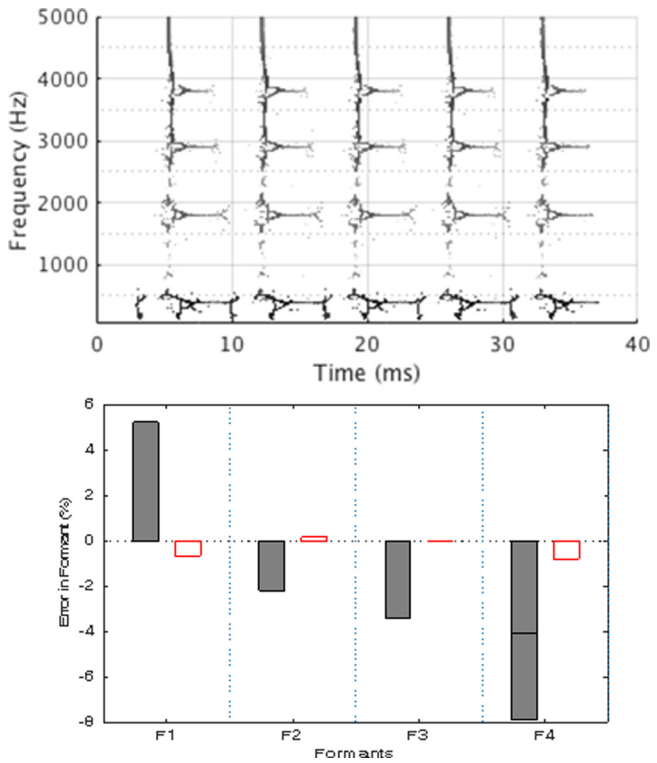

FIG. 6. (Color online) (Top) The RS for the synthetic vowel with $F0 = 145$ Hz, $F1 = 400$ Hz. (Bottom) The errors in each formant measurement for this synthetic vowel as a percentage of the actual frequency of each formant for LPC-Burg (peak-picking) (shaded) and RS (white).

in the breakpoint indicates that a harmonic below the frequency of the formant distorts the measured value more than a harmonic above the frequency of the formant, likely because the tilt of the source spectrum means the harmonics on the lower side of a formant will have higher amplitudes than harmonics with the same frequency difference on the higher side. The estimates based on LP closed phase are, in general, more accurate than those based on LPC-Burg, but the four outliers have large errors. LP closed phase appears to be very sensitive to the exact placement of the analysis window, which defines the closed phase; this extreme sensitivity is mentioned by Klatt (1986) and Fulop (2011, p. 174).

Figure 5 shows the error pattern for WLP-AME, using both peak-picking and root-solving. The root-solving results for $F1$ show a pattern in the shape of a "Y," with the upper branch exhibiting $F0$ bias similar to all other methods, but the lower branch exhibiting $F0$ bias in the opposite direction; in both cases, the error is much smaller, ranging from -8 to 15 Hz for $F1$. The peak-picking results appear to be more randomly scattered, but with a positive bias; the error in $F1$ ranges from $+4.7$ to $+40.6$ Hz. The $F2$ results show this difference more strikingly, with root-solving errors ranging from -0.5 to $+0.6$ Hz, and peak-picking errors all equal to $+16.4$ Hz. Vallabha and Tuller (2002) found that peak-picking results were more sensitive to $F0$ bias than root-solving.

Figure 6 shows (top) a typical RS for one synthetic vowel (with $F0 = 145$ Hz, $F1 = 400$ Hz); the vertical streaks indicate the impulse-like nature of the beginning of the closed phase in each cycle. The horizontal markings indicate the instantaneous frequencies present in the signal at the formant frequencies. Measurements are made in the part of each cycle where the pruned spectrum becomes a single, horizontal line. At the bottom, a bar graph indicates the errors in formant measurement using LPC-Burg and RS for this particular signal. The percentage error is highest for $F4$ for both methods, with $F1$ a close second; for each formant, the RS error is smaller than that of LPC-Burg.

Figure 7 shows RS measurements of all 45 synthetic vowels for the first 2 formants. For $F1$, there is a single outlier value with an error of 27 Hz; all other estimates have an error of ± 4 Hz. For $F2$, all estimates have an error of ± 1 Hz or less.

Finally, Fig. 8 shows the error in $F1$ for each measurement method, expressed as both the total range of errors and the standard deviation of the error across all 45 synthetic signals. We discuss this further in Sec. II C.

C. Discussion

It is clear that manual measurement of formant frequencies based on narrow-band sections does not offer a gold standard for comparison with automatic methods. Participants varied somewhat in their accuracy and in their precise methods, but all four were biased by the frequency of the nearest harmonic of the fundamental.

LPC-Burg is recommended by the manual for the widely used Praat software, but it is not really any more accurate than the LPC autocorrelation method used by Klatt (1986). It is more accurate than manual measurements,

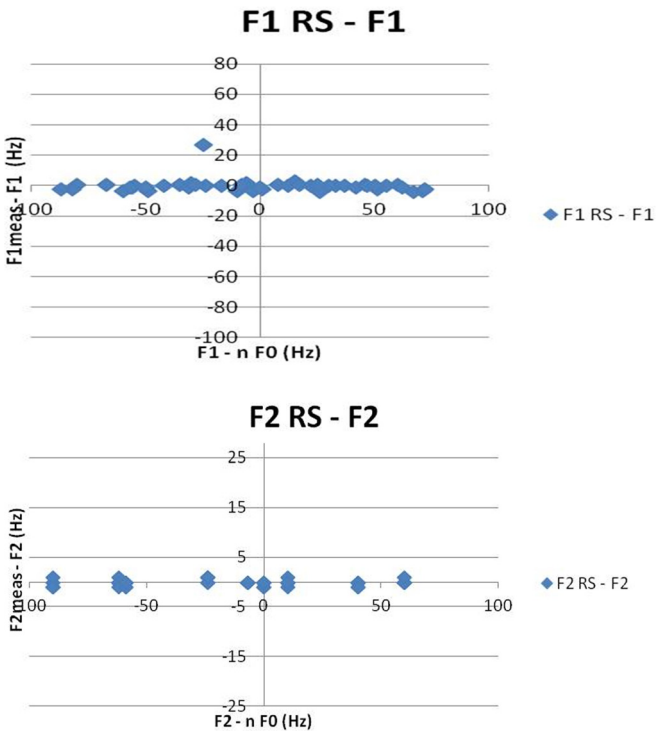

FIG. 7. (Color online) (Top) Errors in $F1$ and (Bottom) $F2$ measurement of 45 synthetic vowels using the RS.

especially when a harmonic coincides exactly with a formant frequency. The AVG and CEPS methods also perform poorly. AVG might be more useful in speech where formants and $F0$ were changing, but would only produce correct results if $F0$ happened to be changing so that a harmonic swept through $F1$ during the analysis interval.

LP closed phase is on average more accurate, but is more sensitive to certain conditions, with four outliers having large errors (from 15 to 64 Hz). Of the LPC methods, WLP-AME (developed by Alku *et al.*, 2013) is the most accurate, exhibiting two patterns of $F0$ bias but with a maximum error of 15 Hz for $F1$. Peak-picking for WLP-AME exhibited a bias (mean positive error); root-solving did not. RS is the most accurate of the methods investigated for these synthetic vowels, with the error on $F1$ within 4 Hz except for one outlier.

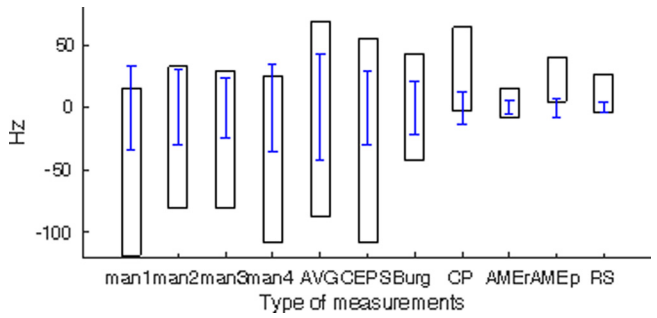

FIG. 8. (Color online) Errors in $F1$ measurement of 45 synthetic vowels by method. "man1" indicates manual measurements by phonetician 1. AMEr and AMEp mean the WLP-AME algorithm was used with, respectively, root-solving and peak-picking. Rectangles indicate the range of errors; whiskered bars indicate the standard deviation for each method.

In sum, the WLP-AME method had the greatest accuracy for the least sensitivity to parameter selection (i.e., closed-phase estimation) among the automatic techniques. RS had the least error of all, but requires (at present) manual estimation of the most likely value among the many presented (see Fig. 6). Hand measurements by experts were not noticeably more accurate than the LPC measures.

III. EXPERIMENT 2: NATURAL SPEECH

In natural speech, the true values of the formant frequencies cannot be known. We chose the speaking tasks to constrain the formants and elicit $F0$-bias errors by using vowels that were not diphthongs and requesting declarative intonation. Thus, the formant frequencies should be relatively constant and $F0$ should be decreasing over the duration of the vowel.

A. Method

1. Stimuli, speakers, and pre-processing

Five speakers were recorded in an anechoic chamber with a microphone and EGG, although the EGG signal was not used in the comparisons reported here. The microphone was a Bruel and Kjaer 4190 (half-inch condenser microphone) with pre-amplifier B&K 2669, powered by a 4-channel Nexus 2690 conditioning amplifier (Bruel & Kjaer Sound and Vibration Measurement, Naerum, Denmark). The filtered output was input to a Powerlab 8/35 running LabChart 7.0 (AD Instruments, Colorado Springs, CO). A sampling rate of 40 kHz was used.

Three of the speakers were male and two were female; they were native speakers of American English, ranging in age from 23 to 59 years. The corpus consisted of a list of words to be spoken three times with neutral intonation, then three times with declarative intonation. The words "heed, hod, had, who'd" were included in the list. Other items included in the word lists served as fillers to avoid list intonation effects.

On initial observation of the recordings, it was noted that one of the male speakers used a small $F0$ range and made no particular difference between neutral and declarative intonation. That speaker was not considered further. The remaining speakers had different $F0$ ranges: they were numbered accordingly, from lowest to highest, M2, M1, W2, W1. All used vocal fry at times; tokens with excessive fry were analyzed but not used for the examples discussed below.

The recordings were segmented by downsampling to 10 kHz, applying forced alignment (Yuan and Liberman, 2008), and then correcting the segment boundaries manually. Audio files of the excised vowels of the four speakers are available.[1]

2. Analysis procedure

The speech recordings were analyzed via the six techniques of experiment 1, with some small changes. Closed-phase intervals, needed for both LP closed phase and WLP-AME, were calculated by using a method of determining the glottal closure instants by Drugman *et al.* (2012). A MATLAB

wrapper was used to call all of the automatic algorithms; the function arburg was used for the Burg method, with order 14 and a 30 ms window as before, and arcov for the closed-phase covariance method, with order 10. For the determination of the formant frequencies from the analysis results, both peak-picking (using findpk) and root-solving were used for LPC-Burg, LP closed phase, and WLP-AME. Peak-picking alone was used for AVG and CEPS. For all automatic methods, one set of formant values was generated every 10 ms throughout the vowel; since the initial window was 30 ms long, the first set of values was found 15 ms after the beginning of the vowel, and the last set 15 ms from the end. For the closed-phase covariance, each set of formant values was based on analysis of the glottal cycle at the center of the 30 ms frame. The other methods used the entire 30 ms. For RS, formant values were measured for every glottal cycle during the vowel, so that the RS formant tracks extended slightly beyond all other sets. All formant values were graphed against time from the beginning of each vowel token so that they could be compared easily.

The formant tracks on the RSs were much less regular for natural than synthetic speech, as noted by Fulop (2011, pp. 140–156). In order to regularize the measurements made and thus improve reliability, a standard procedure was followed by the first author, who made all the RS measurements of natural speech. For a given 40 ms frame, the frequency range was zoomed in to include only one or two formants. Each formant was measured for each glottal cycle in the entire frame before proceeding to the next formant, so that the same part of the pattern in each glottal cycle would be chosen. Values were collected in a spreadsheet but not graphed until all formants for the entire vowel had been measured in order to avoid experimenter bias. The pruning thresholds were varied from the defaults given above when needed to clarify the best place at which to measure. This was particularly important for higher formants and other regions of lower energy.

B. Results

In the following, formant tracks through the vowel are compared for different algorithms. While the four corner vowels were analyzed for each speaker, three examples of [i] in "heed" are presented first for three of the four speakers because of its low $F1$, with its low bandwidth, and clustered $F2$–$F4$ present difficulties for formant estimation. One token of [u] in "who'd" is then discussed for the fourth speaker. Other vowels have been analyzed, but are not shown here. Plots of the error relative to RS estimates of $F1$ are then shown for these four tokens, and different ways of characterizing the error by method are discussed. The examples shown represent a small subset of all those analyzed. Plots of $F0$ and $F1$ are shown because $F1$ exhibits the most bias with changing $F0$. Plots of formants $F1$–$F4$ are shown because the set of estimated formant frequencies relative to the elicited vowel, and their steadiness over the vowel, can be assessed more readily. We have attempted to show at least one example of each combination of analysis methods, and did not show examples that seemed atypical for that method. For example, although Vallabha and Tuller (2002) predict that peak-picking should work better than root-solving when formants are close together, root-solving is shown in some [i] examples when peak-picking resulted in noticeably more missed (and therefore misassigned) formants.

Figure 9 shows formant tracks for /i/ in "heed" for M2, the lower-pitched of the male participants. $F1$ is shown on an expanded scale with $F0$; the LPC-Burg algorithm with peak-picking, which was the most consistent of the automatic algorithms for this token, and LP-closed-phase covariance (CP) with peak-picking, are contrasted with RS. The $F1$ tracks appear similar in shape, with a decrease mid-vowel, but $F1_{\mathrm{Burg}}$ is higher than $F1_{\mathrm{RS}}$ by 20–30 Hz throughout. $F1_{\mathrm{CP}}$ is slightly lower than $F1_{\mathrm{RS}}$ for the first 60 ms; thereafter $F1_{\mathrm{CP}}$ and $F1_{\mathrm{RS}}$ are very similar. The graphs of $F1$–$F3$ for the three methods, on a 0–4 kHz scale, appear more similar to each other, and vary little until nearing the /d/, which is plausible for a monophthong. LP-CP estimates for $F2$ and $F3$ vary a bit more than either Burg or RS. $F4$ is visible for LP-CP with peak-picking, but not with LP-CP root-solving, consistent with the findings of Vallabha and Tuller (2002). For WLP-AME (both peak-picking and root-solving, not shown) $F2$ estimates are very discontinuous. $F0$ range for this token was 124–101 Hz; the third harmonic is above $F1_{\mathrm{RS}}$ throughout. We would predict from the results on synthetic speech that the Burg estimate of $F1$ would tend toward the nearest harmonic and thus would be higher than $F1$; the relationship of $F1_{\mathrm{Burg}}$ to $F1_{\mathrm{RS}}$ thus increases the likelihood that $F1_{\mathrm{RS}}$ is closer to the true value.

Figure 10 shows formant tracks for /i/ in "heed" for W2, the lower-pitched of the female participants. Formant tracks for all three LPC-based methods, all with root-solving, are shown. $F1_{\mathrm{Burg}}$ moves up to nearly 400 Hz, then down to 300, ending at 250 Hz; the wavy track indicates the effect of $F0$ bias. $F1_{\mathrm{CP}}$ is less extreme, but shows some up and down movement at the same times in the vowel. Upper formants ($F3$,$F4$) are not very continuous for either of these methods. $F1_{\mathrm{Alku}}$ is even more extreme than $F1_{\mathrm{Burg}}$, and all formants above $F1$ are discontinuous for this algorithm. For all three LP algorithms, $F1$ estimates for root-solving and peak-picking are similar; higher formants appear to be more consistent during the vowel for root-solving, which is not consistent with the findings of Vallabha and Tuller (2002). By contrast, all formants estimated from RS are continuous and fairly constant during the vowel; $F1_{\mathrm{RS}}$ begins just above $F0$, which ranges from 214 to 134 Hz, rises slightly and decreases near the end, which is plausible for /i/. RS showed a low $F2$, but it is not as high in amplitude as its $F3$, which the other algorithms labeled as $F2$. This is an instance in which RS found energy in the spectrum, but if higher amplitude had been a part of the constraint set, its $F2$ would have agreed with the other methods. Previous research in both measurement and synthesis suggests that this low amplitude $F2$ is indeed spurious.

Figure 11 shows formant tracks for /i/ in "heed" for W1, the higher-pitched of the female participants. $F0$ ranges from 242 to 164 Hz in this token. For both the Burg and LP-closed-phase methods, $F1$ estimates are similar for

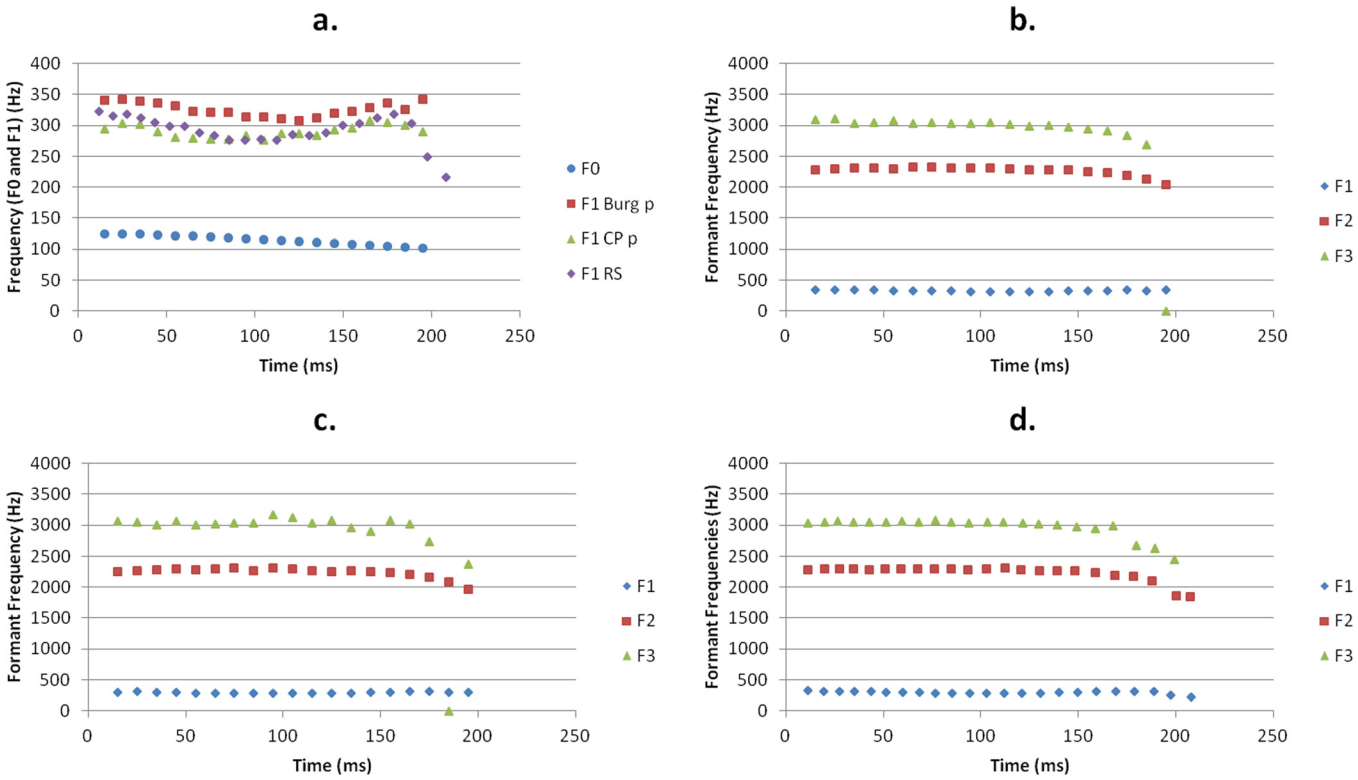

FIG. 9. (Color online) Estimated formants vs time for [i] in "heed" spoken by M2. (a) $F0$ (circles), $F1$_Burg peak (squares), $F1$_CP peak (triangles), and $F1$_RS (diamonds). In (b)–(d), estimated formants are shown with $F1$ (diamonds), $F2$ (squares), and $F3$ (triangles) for (b) Burg peak-picking, (c) CP peak-picking, and (d) RS, respectively.

peak-picking and root-solving, but root-solving works better for higher formants, contrary to the findings of Vallabha and Tuller (2002). The LP-closed-phase method produces some fairly plausible formants, with a slight rise in $F1$ and a bit of difficulty with higher formants near the beginning and end. The WLP-AME algorithm with root-solving (not shown) was even more problematic; in the first 150 ms many $F1$ values were missed altogether. WLP-AME with peak-picking (shown) has $F1$ veering from ∼150 Hz, below $F0$, to 400 Hz within the first half of the vowel; the higher formants are plausible in the last but not the first half of the vowel. $F1_{RS}$ is more constant than $F1_{CP}$; higher formants estimated by RS are not quite as continuous, especially $F3$ and $F4$ near the end; in the last 100 ms the WLP-AME-generated $F3$ and $F4$ appear to be more plausible. Comparison to the narrowband spectrogram, however, indicates that the RS-derived higher formants track regions of higher energy correctly. What is not as clear is whether the energy at 3 kHz should be thought of as constituting another formant. It is rather close to the ones identified as $F3$ and $F4$ by the other methods, and it may be included in the skirts of the transfer function if we assume rather large bandwidths at that frequency.

Figure 12 shows formant tracks for /u/ in "who'd" for M1, the higher-pitched of the male participants. The $F0$ range for this token is 172–101 Hz. $F1$ tracks again differ substantially by method, with noticeable discontinuities observed in the second half of the vowel for both closed-phase and WLP-AME methods and, to a lesser extent, cepstral. Even in the first half of the vowel, the LPC-based $F1$ estimates are all significantly higher than $F1_{RS}$. $F2$, however, is even more problematic in the third quarter of the vowel; it disappears in Burg and WLP-AME, with $F3$ identified as $F2$, and dips down toward $F1$ in closed phase. For the Burg method, formants found by peak-picking are slightly smoother and, so, are shown; for LP-closed-phase and WLP-AME root-solving results are slightly smoother and so are shown. The results of Vallabha and Tuller (2002)predict that root-solving should be more accurate than peak-picking for this vowel. The cepstrally estimated $F2$ is smooth and continuous, but $F3_{CEPS}$ is quite discontinuous. Only for RS do all of the formants appear to be continuous and plausible. The narrowband spectrogram shows that $F2$ becomes more weakly excited in the third quarter, though it is still visible; the lowered amplitude apparently caused mistracking for all methods except RS.

For these monophthong examples and others not shown, the RS formant estimates do not always form perfectly continuous formant tracks, particularly for the higher formants, but they are generally the most plausible. As an attempt to quantify the error in formant estimates for natural speech, when the ground truth is not known, we computed the difference between $F1_{RS}$ and $F1$ estimates by the other methods, and plotted the differences against time. One example for each speaker is shown in Fig. 13. Figure 13(a) shows the same token as in Fig. 12; Figs. 13(b)–13(d) show [i] as in "heed" spoken by M2, W1, and W2. The M1 and M2 graphs both show the surprisingly large range of $F1$ estimates, even for these low-$F0$ tokens that should provide optimal conditions for the measurement. The W1 and W2 graphs also show a large range of $F1$ estimates, including those for

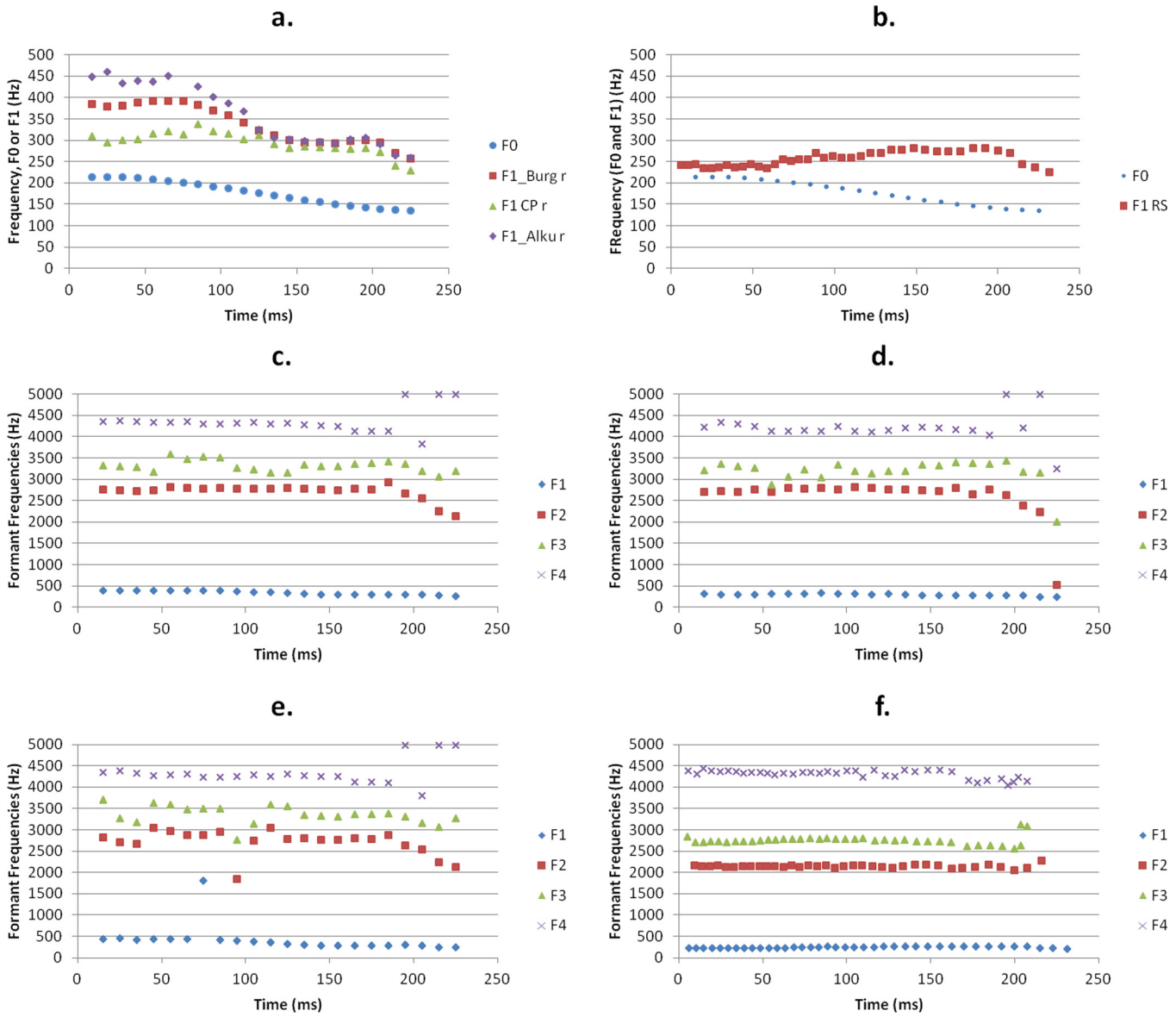

FIG. 10. (Color online) Estimated formants vs time for [i] in "heed" spoken by W2. (a) $F0$ (circles), $F1$_Burg root (squares), $F1$-CP root (triangles), and $F1$_WLP-AME root (diamonds). (b) $F0$ (circles) and $F1$_RS (squares). (c) $F1$–$F4$ for Burg root. (d) $F1$–$F4$ for CP root. (e) $F1$–$F4$ for WLP-AME root. (f) $F1$–$F4$ for RS.

WLP_AME (developed by Alku *et al.*, 2013), although that is designed for higher-$F0$ voices.

It would be useful to know whether there is a best time at which to measure formants during a vowel; that is, is there a time at which the errors tend to be smallest? It is clear that the start and end are less stable [see, for example, the last point in Fig. 13(b), for M2]. The largest variation across methods occurs in the first half for three of the four examples here; it may be that $F0$ changes faster at first, contributing to this variation.

For a given speaker, the closed-phase estimates are less than the Burg estimates at any given time in all four cases. The closed-phase estimates are based on the closed phases only, whereas the Burg estimates use the entire glottal cycle; the increased loss at the glottis has been shown to increase the formant value during open phase (Fulop, 2011, p. 144; Quatieri, 2002, pp. 158–159), so this effect may help explain the difference.

The graphs in Fig. 13 also allow a quick assessment of the effect of voice bar on $F1$ estimates. In the RS of natural speech (not shown here, but see Fulop and Disner, 2012), voice bar is noticeable, generally occurring earlier in time and lower in frequency than the $F1$ track in each glottal cycle. The energy in the voice bar would, however, be included along with $F1$ in most of the analysis methods used, with the exception of LP-closed-phase analysis, because they are asynchronous and have analysis windows longer than a single glottal cycle. However, one would then predict that the voice bar would lower the estimate of $F1$ below the "true" value. Yet, in Fig. 13, nearly all of the error differences in $F1$ are positive. Exceptions include WLP-AME near the beginning of W1's "heed," already discussed above and more likely related to $F0$ bias, and the $F1_{AVG}$ values at the beginning of M2's "heed," which may in fact be due to voice bar being included. Clearly the voice bar cannot explain the majority of the errors noted; the RS provides a way to study voice bar more systematically.

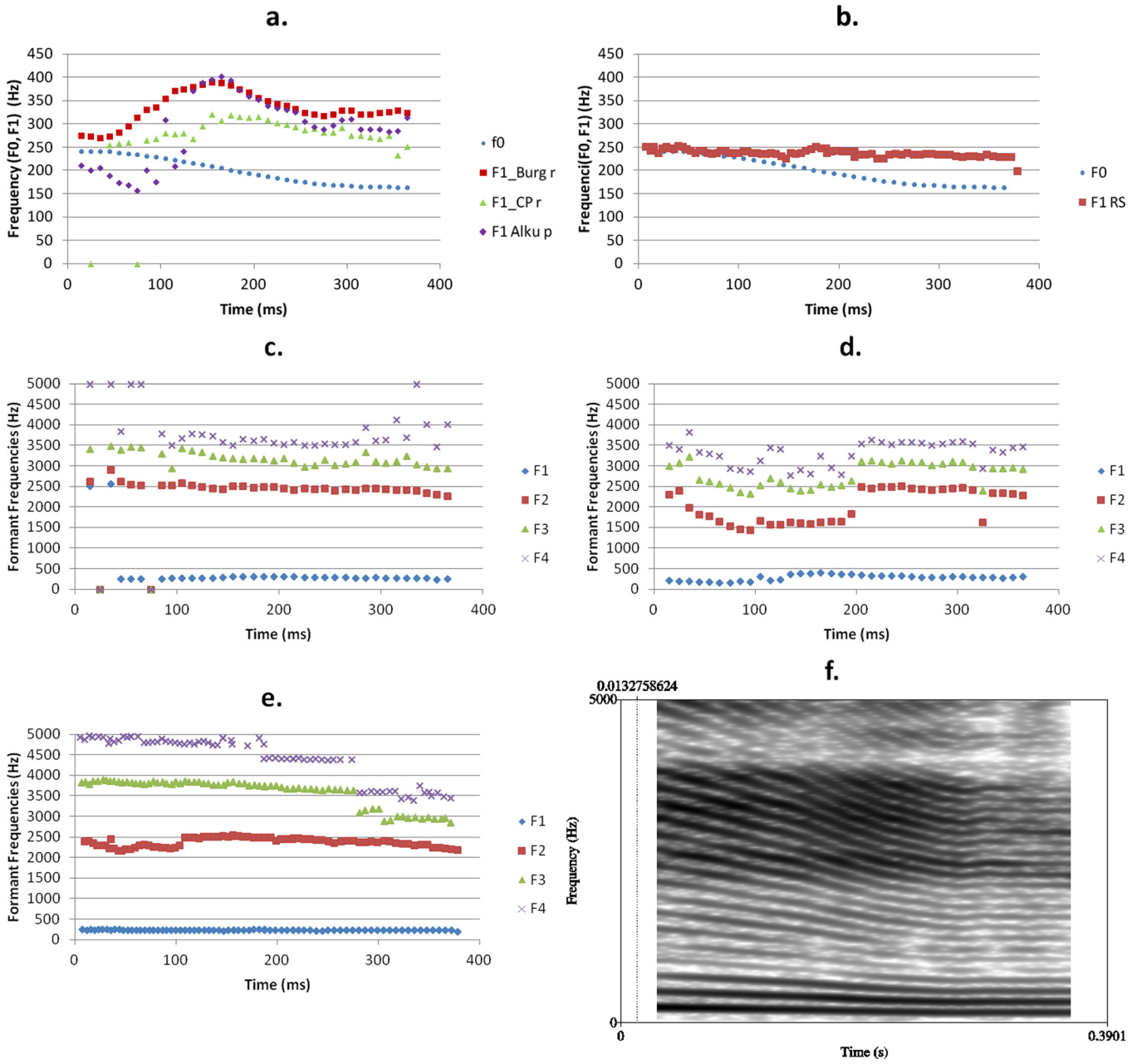

FIG. 11. (Color online) Estimated formants vs time for [i] in "heed" spoken by W1. (a) *F*0 (circles), *F*1_Burg root (squares), *F*1_CP root (triangles), *F*1_WLP-AME peak (diamonds). (b) *F*0 (circles), *F*1_RS (squares). (c) *F*1–*F*4 for CP root. (d) *F*1–*F*4 for WLP-AME peak. (e) *F*1–*F*4 for RS. (f) Narrowband spectrogram.

With the different patterns of variation over time it is difficult to arrive at a single way of quantifying the difference between each method of estimating *F*1. Tables I–IV show various error measures by method computed over the entire vowel for Figs. 13(a)–13(d); outliers are excluded in a few cases (Tables II and III). These illustrate the ways in which any single measure can disguise severe estimation problems.

C. Discussion

Formant trackers typically include preset constraints on the number of formants and, sometimes, frequency ranges for each formant (as used by, e.g., Alku *et al.*, 2013), and post-processing smoothing procedures. The post-processing can be quite complex, especially for algorithms designed to be used on large databases of continuous speech (e.g., Mehta *et al.*, 2012). Such methods can eliminate brief errors in one formant affecting adjoining formants and render the formant tracks smoother. However, making a formant track look smooth and consistent with itself does not necessarily guarantee that it is within even 10 Hz of the "right" value. The actual resonance frequencies can be determined by direct methods if it is possible to sustain a vowel for 2 s or more, but that is not a reasonable alternative for our goal of accurately determining the variability of formant frequencies in vowels spoken in isolated words.

It is surprising that the WLP-AME algorithm did not perform better, especially for the female speakers. As used by Alku *et al.* (2013), the frequency ranges within which

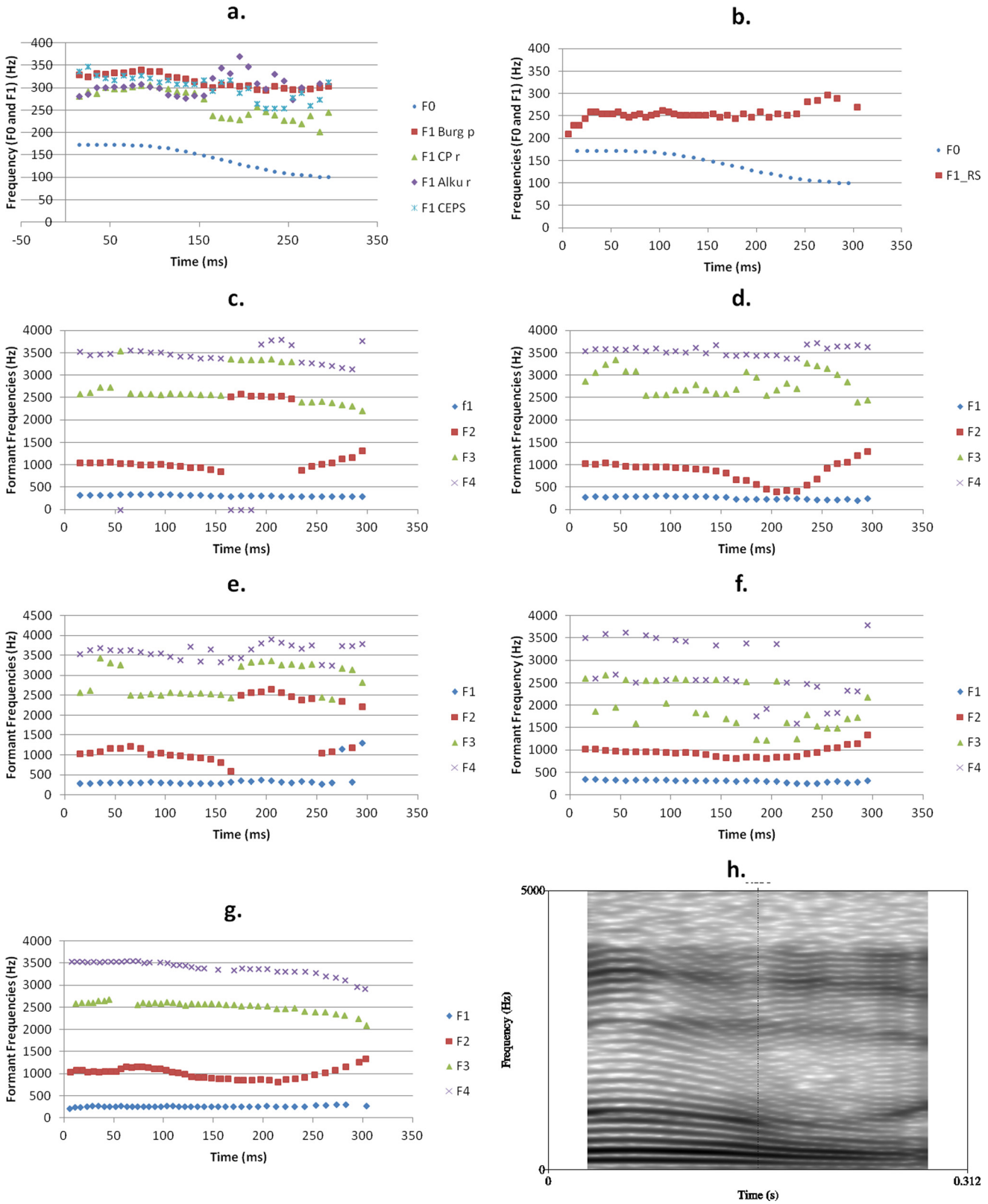

FIG. 12. (Color online) Estimated formants vs time for [u] in "who'd" spoken by M1.(a) *F*0 (diamonds), *F*1_Burg peak (squares), *F*1_CP root (triangles), *F*1_WLP-AME root (diamonds), *F*1_CEPS (X's). (b) *F*0 (circles), *F*1_RS (squares). (c) *F*1–*F*4 for Burg peak. (d) *F*1–*F*4 for CP root. (e) *F*1–*F*4 for WLP-AME root. (f) *F*1–*F*4 for CEPS. (g) *F*1 to *F*4 for RS. (h) Narrowband spectrogram.

formants could be identified for the natural vowels were specified; such constraints would have limited some of the minima in *F*1 tracks, but would not have limited the maxima (e.g., *F*1 changing from 400 to 200 Hz in [i] would have been allowed with their range of 200–600 Hz for *F*1 in /i/). The corpus in the study by Alku *et al.* (2013) consisted of vowels sustained for 2 s or more at pitches on a diatonic scale. Is a sustained vowel easier to analyze? Fulop (2010)

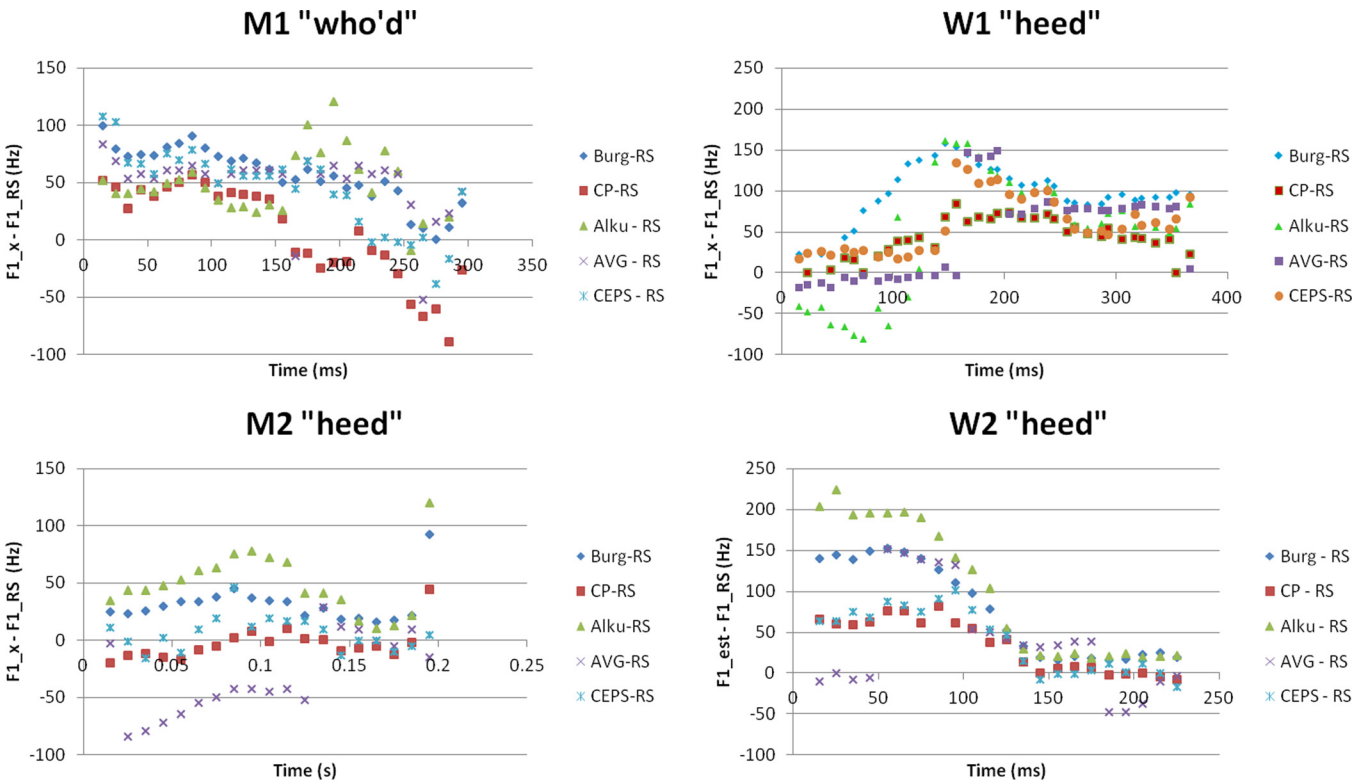

FIG. 13. (Color online) Difference in $F1$ estimates for each of six methods from the RS $F1$ estimate graphed vs time during vowel for: (a) [u] in "who'd," speaker M1; (b) [i] in "heed," speaker M2; (c) [i] in "heed," speaker W1; (d) [i] in "heed," speaker W2. For male speakers [results in (a) and (b)], formants found by peak-picking (Burg) and root-solving (closed-phase and WLP-AME). For female speakers [results in (c) and (d)], formants found by root-solving (Burg and closed-phase) and peak-picking (WLP-AME).

used a descending $F0$ for synthetic stimuli, which resulted in a 20–70 Hz range of $F1$ estimates from LPC algorithms while RS varied within 8 Hz; it seems that not only high $F0$ but also changing $F0$ poses difficulties for formant analysis, but in addition, changing $F0$ reveals the extent to which formant analysis methods are biased by $F0$. Finally, though WLP-AME performed best among algorithms assessed by Alku *et al.* (2013), in synthetic vowels their error measure, which combined the errors in the first three formants, ranged from 33 to 78 Hz, and it still seemed sensitive to $F0$: the variance in its measures of natural vowels did, in general, increase with $F0$, and was higher for women than men.

Burris *et al.* (2014) performed a comparison of four analysis techniques on synthesized and spontaneous speech and reported that three of the systems provided "generally consistent and fairly accurate" (p. 26) formant values. Although $F0$ was falling by 40 Hz over the course of a syllable in their synthetic stimuli, they only reported a single error value. Thus, errors induced by harmonics above and below the synthesized formant value are likely to have been averaged together. Their natural utterances were those of Hillenbrand *et al.* (1995). The measurements by Burris *et al.* (2014) for these were compared with the published values from Hillenbrand *et al.* (1995). Those values, in turn, were based on hand-edited LPC values that were manually checked against a spectrogram, a peak display, and the measurer's phonetic knowledge. Not surprisingly, the two types of LPC analyses gave similar results.

TABLE I. Subject M1, vowel in "who'd": error parameters combining differences in $F1$ estimates for measures made every 10 ms through vowel. For instance, for leftmost column, Avg of (Burg−RS) is the average of all the differences between $F1$_Burg and $F1$_RS; "St dev" is the standard deviation of those differences; range is the maximum $F1$ difference minus the minimum $F1$ difference.

	Difference between $F1$ estimates				
Error parameter	Burg−RS	CP−RS	Alku−RS	AVG−RS	CEPS−RS
Avg (Hz)	57.0	7.0	50.0	50.0	57.0
St dev (Hz)	25.2	41.5	27.6	27.0	25.2
Range (Hz)	99.0	145.0	110.0	135.0	146.0

TABLE II. Subject M2, vowel in "heed." See caption for Table I. Differences on the final value get very large [see Fig. 13(b)], so average and range are also shown omitting the last value (as Avg 2 and Range 2).

	Difference between $F1$ estimates				
Error parameter	Burg−RS	CP−RS	Alku−RS	AVG−RS	CEPS−RS
Avg (Hz)	31.8	−2.6	49.9	−30.6	6.3
St dev (Hz)	16.9	14.1	26.8	34.4	14.7
Range (Hz)	76.9	30.0	66.6	113.5	61.0
Avg 2 (Hz)	28.4	−5.2	45.9	−31.5	6.4
Range 2 (Hz)	29.2	30.0	66.6	113.5	61.0

IV. GENERAL DISCUSSION

Formant measurements are the basis for much work in phonetics, yet the field has not addressed the issue of

TABLE III. Subject W1, vowel in "heed." See caption for Table I. Differences between $F1$_CP and $F1$_RS are very large on the first three values (LP-closed-phase estimates were in $F2$ range initially), so for CP−RS, the average, st dev, and range are also computed omitting the first three values (Avg 2, St dev 2, Range 2).

	Difference between $F1$ estimates				
Error parameter	Burg−RS	CP−RS	Alku−RS	AVG−RS	CEPS−RS
Avg (Hz)	96.3	179.0	50.0	48.7	60.1
St dev (Hz)	36.5	536.6	74.6	53.8	35.1
Range (Hz)	136.4	2326	242.3	166.9	116.9
Avg 2 (Hz)		47.0			
St dev 2 (Hz)		21.4			
Range 2 (Hz)		83.5			

accuracy until quite recently. Certainly, for distinguishing vowel categories, we can obtain reliable results even from very minimal data (Peterson and Barney, 1952). As we explore issues that require finer distinctions, the issue of accuracy becomes more critical. Despite the evidence of the large influence of $F0$ on most formant measurement techniques, only recently have methods been developed to avoid this influence. Here, we reported on two such methods that greatly reduce error, both in synthesized and natural speech.

One method was the RS developed by Fulop (2010, 2011). The spectrum in these displays is impulselike, showing energy at all frequencies, at the onset of closed phase, but it quickly settles into a pattern around the formants. In general, these patterns accurately represent the formants. On occasion, weak resonances that would typically not be classified as formants are detected and would have to be removed because they are of lower amplitude than formants higher in frequency. At present, unfortunately, no automatic means of extracting those values has been found to be reliable. Human estimation (via visual averaging of a large number of values in a RS) results in very accurate measurements. It is to be hoped that a reliable automatic means for obtaining the formant values will be developed so that this technique can be more widely used.

The WLP-AME algorithm was the best automatic method for the synthetic speech, with an $F1$ error range of 22 Hz (root-solving) or 36 Hz (peak-picking) and standard deviations of 5.4 and 7.5 Hz, respectively. The results of this algorithm were not always plausible for the natural speech, however. For automatic measurements, WLP-AME is likely to give the most accurate measurements. It may be that the best overall solution is to examine outliers with RS by hand.

It is somewhat discouraging that hand measurement by experts is not the "gold standard" for obtaining formant values, but humans are still able to obtain accurate formants when it counts: in perception. As Klatt (1986) directly demonstrated, and various synthesis experiments have shown incidentally, human listeners do indeed react to synthesized stimuli as if they heard the intended formant, not the formant that our algorithms generally measure. This shows the tight link between production and perception in speech (Liberman and Whalen, 2000): The plausible transfer function for a combination of $F0$ and related harmonics is perceived, not just a direct representation of the acoustic energy. Human listeners are little affected by large changes in $F0$ despite the sparse sampling at higher $F0$'s (e.g., Assmann and Katz, 2000). One promising approach is the "missing information" matching of spectral templates, in which only frequencies at which a harmonic is present contribute to the output (de Cheveigné and Kawahara, 1999). In this way, entire formants can be unrepresented by harmonics, but their absence does not affect identification rates. Whether or not this is the way that human listeners perform remains to be demonstrated.

TABLE IV. Subject W2, vowel in "heed." See caption for Table I.

Difference between $F1$ estimates	Burg−RS	CP−RS	Alku−RS	AVG−RS	CEPS−RS
Error parameter					
AVG (Hz)	77.2	34.9	101.1	39.5	41.5
St dev (Hz)	57.2	32.1	81.8	63.9	39.1
Range (Hz)	136.4	90.0	206.3	199.9	118.3

ACKNOWLEDGMENTS

We thank the participants who made manual measurements of the stimuli, and explained their methods; we thank the speakers of the natural speech corpus. Thanks to Mark Tiede for providing alternate analysis methods, to Sean Fulop for help with the reassigned spectrogram, and to Paavo Alku and Manu Airaksinen for help and their code for the WLP-AME method. Thanks to the four anonymous reviewers for their comments. This work was supported by National Institutes of Health-National Institute on Deafness and Other Communication Disorders (NIH-NIDCD) Grant No. DC-002717 to Haskins Laboratories. C.H.S. and H.N. contributed equally to this study.

[1]https://yale.box.com/s/66fug8e0hgq52bvduiox 6a1stqiut96l.

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**(2), 1295–1313.
Allen, J., Hunnicutt, M. S., and Klatt, D. H. (**1987**). *From Text to Speech* (Cambridge University Press, Cambridge), pp. 108–122.
Andersen, N. (**1974**). "On the calculation of filter coefficients for maximum entropy spectral analysis," Geophysics **39**, 69–72.
ANSI (**2013**). ANSI/ASA S1.1-2013, *American National Standard Acoustical Terminology* (Acoustical Society of America, Melville, NY).
Assmann, P. F., and Katz, W. F. (**2000**). "Time-varying spectral change in the vowels of children and adults," J. Acoust. Soc. Am. **108**, 1856–1866.
Atal, B. S. (**1975**). "Linear prediction of speech—Recent applications to speech analysis," in *Speech Recognition*, edited by R. D. Reddy (Elsevier, New York), pp. 221–230.
Atal, B. S., and Schroeder, M. R. (**1978**). "Linear prediction analysis of speech based on a pole-zero representation," J. Acoust. Soc. Am. **64**, 1310–1318.
Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," J. Acoust. Soc. Am. **90**(2), 799–828.
Boersma, P., and Weenink, D. (**2013**). "Praat: Doing phonetics by computer (version 5.3.56) [computer program]," http://www.praat.org.
Bradlow, A. (**2002**). "Confluent talker- and listener-oriented forces in clear speech production," in *Lab Phonology7*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin, Germany), pp. 241–273.

Burg, J. P. (**1967**). "Maximum entropy spectral analysis," in *Proc. 37th Meeting of the Society of Exploration Geophysicists*, October 31, Oklahoma City, OK.
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. Res. **57**, 26–45.
Castelli, E., and Badin, P. (**1988**). "Vocal tract transfer functions with white noise excitation—Application to the naso-pharyngeal tract," in *Proc. 7th FASE Symp.*, Edinburgh, pp. 415–422.
Chiba, T., and Kajiyama, M. (**1941**). *The Vowel: Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo), pp. 115–154.
Childers, D. G. (**1978**). *Modern Spectrum Analysis* (IEEE, New York), pp. 34–41, 252–255.
Clopper, C., and Pierrehumbert, J. (**2008**). "Effects of semantic predictability and regional dialect on vowel space reduction," J. Acoust. Soc. Am. **124**(3), 1682–1688.
Davies, P. O. A. L., McGowan, R. S., and Shadle, C. H. (**1992**). "Practical flow duct acoustics applied to the vocal tract," in *Vocal Fold Physiology: Frontiers in Basic Science*, edited by I. R. Titze (Singular Pub. Group, Inc., San Diego), pp. 93–142.
de Cheveigné, A., and Kawahara, H. (**1999**). "Missing-data model of vowel identification," J. Acoust. Soc. Am. **105**, 3497–3508.
Deng, L., Lee, L. J., Attias, H., and Acero, A. (**2007**). "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," IEEE Trans. Audio Speech Lang. Process. **15**(1), 13–23.
Djeradi, A., Guerin, B., Badin, P., and Perrier, P. (**1991**). "Measurement of the acoustic transfer function of the vocal tract: A fast and accurate method," J. Phonetics **19**, 387–395.
Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (**2012**). "Detection of glottal closure instants from speech signals: A quantitative review," IEEE Trans. Audio, Speech, Lang. Process. **20**(3), 994–1006.
Epps, J., Smith, J. R., and Wolfe, J. (**1997**). "A novel instrument to measure acoustic resonances of the vocal tract during phonation," Meas. Sci. Technol. **8**, 1112–1121.
Fant, G. (**1960**). *The Acoustic Theory of Speech Production* (Mouton, The Hague), pp. 20, 53.
Fujimura, O., and Lindqvist, J. (**1971**). "Sweep-tone measurements of vocal tract characteristics," J. Acoust. Soc. Am. **49**, 541–557.
Fulop, S. A. (**2007**). "Phonetic applications of the time-corrected instantaneous frequency spectrogram," Phonetica **64**, 237–262.
Fulop, S. A. (**2010**). "Accuracy of formant measurement for synthesized vowels using the reassigned spectrogram and comparison with linear prediction," J. Acoust. Soc. Am. **127**(4), 2114–2117.
Fulop, S. A. (**2011**). *Speech Spectrum Analysis* (Springer, Berlin), pp. 127–201.
Fulop, S. A., and Disner, S. F. (**2012**). "Examining the voice bar," POMA **14**, 060002.
Harrington, J., and Cassidy, S. (**1999**). *Techniques in Speech Acoustics* (Kluwer Academic, Dordrecht, The Netherlands), pp. 174–177, 222–225.
Henrich, N., Smith, J., and Wolfe, J. (**2011**). "Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones," J. Acoust. Soc. Am. **129**(2), 1024–1035.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111.
Holmes, J. N. (**1976**). "Formant excitation before and after glottal closure," in *Proc. International Conf. Acoust. Speech and Signal Proc.*, pp. 39–42.
Joliveau, E., Smith, J., and Wolfe, J. (**2004**). "Vocal tract resonances in singing: The soprano voice," J. Acoust. Soc. Am. **116**(4), 2434–2439.
Klatt, D. (**1986**). "Representation of the first formant in speech recognition and LF models of the auditory periphery," in *Proc. Montreal Satellite Symposium on Speech Recognition, 12th International Cong. on Acoust.*, edited by P. Mermelstein, Toronto (July).
Liberman, A. M., and Whalen, D. H. (**2000**). "On the relation of speech to language," Trends Cognit. Sci. **4**(5), 187–196.
Mehta, D. D., Rudoy, D., and Wolfe, P. J. (**2012**). "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," J. Acoust. Soc. Am. **123**(3), 1732–1746.
Monsen, R. B. (**1976**). "The production of English vowels by deaf adolescents," J. Phonetics **4**, 189–198.
Monsen, R. B., and Engebretson, A. M. (**1983**). "The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction," J. Speech Hear. Res. **26**(3), 89–97.
Narayanan, S., Alwan, A., and Song, Y. (**1997**). "New results in vowel production: MRI, EPG, and acoustic data," in *Proc. Eurospeech 97*, Rhodes, Greece, Vol. 2, pp. 1007–1010.
Pham Thi Ngoc, Y., and Badin, P. (**1994**). "Vocal tract acoustic transfer function measurements: Further developments and applications," J. Phys. IV C **5**, 549–552.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**(2), 175–184.
Potter, R. K., and Steinberg, J. C. (**1950**). "Toward the specification of speech," J. Acoust. Soc. Am. **22**(6), 807–820.
Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1986**). *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge), pp. 430–435.
Quatieri, T. F. (**2002**). *Discrete-Time Speech Signal Processing* (Prentice-Hall, Upper Saddle River, NJ), pp. 158–159.
Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., and McGonegal, C. A. (**1976**). "A comparative study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Process. **24**, 399–417.
Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "Voicesauce: A program for voice analysis," in *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, Hong Kong, pp. 1846–1849.
Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**(1), 537–554.
Story, B. H., Titze, I. R., and Hoffman, E. A. (**1998**). "Vocal tract area functions for an adult female speaker based on volumetric imaging," J. Acoust. Soc. Am. **104**(1), 471–487.
Swerdlin, Y., Smith, J., and Wolfe, J. (**2010**). "The effect of a whisper and creak vocal mechanisms on vocal tract resonances," J. Acoust. Soc. Am. **127**(4), 2590–2598.
Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., Švec, J. G., Ternström, S., and Wolfe, J. (**2015**). "Toward a consensus on symbolic notation of harmonics, resonances and formants in vocalization," J. Acoust. Soc. Am. **137**(5), 3005–3007.
Vallabha, G. K., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**, 141–160 (2002).
Woehrling, C., and de Maureuil, P. B. (**2007**). "Comparing Praat and Snack formant measurements on two large corpora of northern and southern French," in *Proc. Interspeech 2007*, August, Antwerp, Belgium, pp. 1006–1009.
Yuan, J., and Liberman, M. (**2008**). "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics'08*, Paris, France, pp. 5685–5688.
Zhang, C., Morrison, G. S., Ochoa, F., and Enzinger, E. (**2013**). "Reliability of human-supervised formant-trajectory measurement for forensic voice comparison," J. Acoust. Soc. Am. **133**(1), EL54–EL60.