JASA ARTICLE

Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986)[a)]

D. H. Whalen,[1,b)] Wei-Rong Chen,[1] Christine H. Shadle,[1] and Sean A. Fulop[2]
[1]*Haskins Laboratories, New Haven, Connecticut 06511, USA*
[2]*Department of Linguistics, California State University Fresno, Fresno, California 93740, USA*

ABSTRACT:
Formants in speech signals are easily identified, largely because formants are defined to be local maxima in the wideband sound spectrum. Sadly, this is not what is of most interest in analyzing speech; instead, resonances of the vocal tract are of interest, and they are much harder to measure. Klatt [(1986). in *Proceedings of the Montreal Satellite Symposium on Speech Recognition, 12th International Congress on Acoustics*, edited by P. Mermelstein (Canadian Acoustical Society, Montreal), pp. 5–7] showed that estimates of resonances are biased by harmonics while the human ear is not. Several analysis techniques placed the formant closer to a strong harmonic than to the center of the resonance. This "harmonic attraction" can persist with newer algorithms and in hand measurements, and systematic errors can persist even in large corpora. Research has shown that the reassigned spectrogram is less subject to these errors than linear predictive coding and similar measures, but it has not been satisfactorily automated, making its wider use unrealistic. Pending better techniques, the recommendations are (1) acknowledge limitations of current analyses regarding influence of F0 and limits on granularity, (2) report settings more fully, (3) justify settings chosen, and (4) examine the pattern of F0 vs F1 for possible harmonic bias.
 https://doi.org/10.1121/10.0013410
(Received 9 March 2022; revised 24 June 2022; accepted 14 July 2022; published online 11 August 2022)
[Editor: Richard A. Wright] Pages: 933–941

I. INTRODUCTION

The realization that speech sounds have bands of energy whose frequencies signal different vowels was a major advance in our understanding of communication (e.g., Hermann, 1890; Russell, 1928; Chiba and Kajiyama, 1941). Showing that the frequency bands could be predicted from articulatory data (such as x rays) provided powerful support for the acoustic theory of speech production that was and remains foundational to speech research (Fant, 1960). The realization that the vocal tract is essentially a malleable tube with resonant frequencies has allowed us to measure, synthesize, and machine-recognize acoustic speech signals with wide-ranging practical benefits in speech theory and application.

Klatt (1986) may be one of the best three-page papers on speech science ever written. In its tight confines, the author shows that formant measurements are both highly skewed toward strong harmonics and away from the resonance. Even though Klatt's first experiment was substantial enough to stand on its own, he further showed, using synthetic speech, that listeners locate the resonance and ignore the measured formants (unless they happen to coincide, of course). Both results have stood the test of time and replication, but improvements to measurement methods have been incremental and the implications are routinely ignored.

Studies that depend on accurate resonance estimates may acknowledge that linear predictive coding (LPC) algorithms often go astray and attempt to counteract this by using hand measurements in a correction or verification step. Although LPC analysis results in the formant estimate corresponding to the pole configuration that will minimize the error (Atal and Hanauer, 1971), manual checking is generally aimed at excluding unrealistic or otherwise outlying values and not at counteracting the harmonic bias. Indeed, as we will show, manual methods do not counteract that bias.

When authors report "formant" values, they are typically being quite scrupulous in reporting what their numbers are, but they are assuming that those formants represent the truly important aspect of speech, i.e., the resonances. The fundamental need for this distinction was carefully laid out in Titze *et al*. (2015). While there may be some recognition of the importance of the distinction, most publications continue to report formant values as if they were telling us directly about resonances rather than estimating them. Researchers in speech technology largely bypass formants and resonances entirely by moving to Mel-frequency cepstral coefficients (MFCCs), and the success of technological approaches justifies that step (Gupta *et al*., 2018). However, researchers in basic speech science continue to report formant measurements widely.

The definition of accuracy must include the issue of what is being measured. With resonances as the desired object, it is difficult to find the "ground truth" from natural speech. Even if human listeners are attuned to the

[a)]This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.
[b)]Also at: City University of New York, New York, NY 10016, USA. Electronic mail: whalen@haskins.yale.edu

resonances, as Klatt (1986) demonstrated, deriving those resonances from the acoustic signal of typical speech is not a directly solvable problem. Our fullest measure of the vocal tract shape comes from three-dimensional (3D) magnetic resonance imaging (MRI), but those measures have limitations on accuracy that are at least as large as those found with LPC. In one study (Story, 2008), the average difference of calculated F1 from acoustically measured F1 was approximately 60 Hz (see Fig. 2 therein). There are many possible sources for such a mismatch, including limits on the resolution of the vocal tract volumes, inadequacies of the acoustic calculation from those volumes, and differences in the conditions of the acoustic recording and MRI (although that study took care to make the conditions as similar as possible). The conclusion must remain, however, that the determination of resonances in natural speech is not accurate enough to provide us with the ground truth needed to assess the accuracy of acoustic estimates of resonances.

Even hand measurements specifically designed to address the harmonic bias turn out to be ineffective (Shadle *et al.*, 2016). In that work, the formants of vowels synthesized with a range of F0s were measured by four experienced phoneticians from narrowband spectrograms. When queried about their methods, they stated that they were aware of the problem of harmonic bias, hence, they estimated what the source intensity was likely to be and adjusted their resonance measures accordingly. The harmonics of the source decline in amplitude as the harmonic number (and, thus, the frequency) increases, resulting in spectral tilt for the amplitude of the harmonics. Those amplitudes are then modified by the resonances of the vocal tract. Unfortunately, the radiated acoustic signal does not directly show the slope of the phonation source, therefore, it is impossible to know how much of a given harmonic's amplitude is due to source strength and how much is the result of its proximity to a resonance. As a result, these hand measurements were no more accurate than the Burg method of LPC as implemented within Praat (asynchronous, 30 ms window). Figure 1 plots the signed error in estimates of F1 vs the difference between the resonance and nearest harmonic for a phonetician's hand measurements (triangles) and a LPC algorithm (circles), based on values of Shadle *et al.* (2016). If F0 had no effect on the accuracy of either method of determining F1, error values would fall along the abscissa, above and below zero, indicating random error. However, errors align on diagonal lines for LPC and manual measurements, implying that the error is biased toward the nearest harmonic of F0.

There are hundreds of reports of speech formants based on LPC. For example, of nearly 300 papers published in *Journal of Phonetics* from 2015 to 2020, all but 2 of the 59 papers that reported formants used the legacy LPC algorithm (Burg, 1967); the majority (79%) of those were done with Praat. Even though large samples will serve to average out random error so that one might assume central tendencies are well represented, systematic biases will not average out (see discussion in Sec. II B and Fig. 3), leading to dubious claims of accuracy.

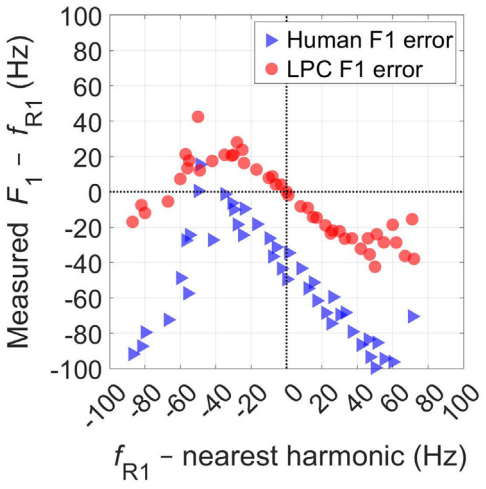

FIG. 1. (Color online) Parts of the original Figs. 2 and 4 in Shadle *et al.* (2016) have been replotted. Triangles indicate the manual measurement errors, and circles represent the LPC-Burg errors. The *x* axis is the distance from the resonance (f_{R1}) to the nearest harmonic, and the *y* axis is the signed error of formant measurement. Errors aligning on a diagonal line indicate strong harmonic bias.

A limitation of spectral analysis that can further limit the accuracy of acoustic reports derives from the interactions among the parameter settings themselves. To quote from the Praat manual's entry on "Sound: To Formant (burg)…" (Boersma and Weenink, 2019): "For instance, if the Window length is 0.025 s, the actual Gaussian window duration is 0.050 seconds. This window has values below 4% outside the central 0.025 seconds, and its frequency resolution (-3 dB point) is 1.298/(0.025 s) = 51.9 Hz, as computed with the formula given at Sound: To Spectrogram…. This is comparable to the bandwidth of a Hamming window of 0.025 seconds, which is 1.303/(0.025 s) = 52.1 Hz, but that window (which is the window most often used in other analysis programs) has three spectral lobes of about −42 dB on each side."

Although this quote applies to Praat's default settings, those are the settings that are most commonly used. Even before beginning to take the harmonic bias into account, these settings should make us cautious in interpreting the measurements. With the best resolution being 51.9 Hz, it is difficult to support differences less than that.

The reports on variability of formants necessarily rely on the accuracy of each individual measurement: if the measurements are inaccurate, the estimation of variability is compromised. For example, variability is sometimes reported in acquired apraxia of speech (Haley *et al.*, 2001; den Ouden *et al.*, 2018) and developmental apraxia of speech (Lenoci *et al.*, 2021), but some of this may be due to the granularity of the formant measurements. Speakers with variability within 60 Hz might appear to have no variability at all for a single resonance, whereas those with variability of 90 Hz would likely appear to show variable productions on 33% of tokens even if their resonance frequency was constant. Other studies report small effects that may be due to F0 variability (e.g., Hall, 2013; Han *et al.*, 2018; Turton and Baranowski,

https://doi.org/10.1121/10.0013410

2021), and the differences reported are rather low for the resolving power of LPC.

Formant variability could also be due to greater variability in F0. Because LPC is greatly influenced by F0 in addition to the resonances, production (reflected in the resonances) might be classified as variable when only F0, in fact, varied. Conversely, a tracking of a strong harmonic might lead to an underestimation of resonance variability if F0 is stable. To model these possibilities, Chen *et al.* (2019) synthesized vowels with known (artificial) resonances with a wide range of F0s. Figure 2 summarizes their original Fig. 4. In each of the two panels in Fig. 2, the (blue) circles on the left represent the distribution of designated synthetic resonances with the specified F0s. The (red) triangles on the right show the corresponding formant measurements by LPC. The top panel [Fig. 2(a)] shows the distribution that resulted when resonances were randomly sampled from a Gaussian distribution with mean = 400 Hz and standard deviation (SD) = 16 Hz. The lower panel [Fig. 2(b)] has the same mean resonance value for the central tendency but a narrower distribution with SD = 3 Hz. The underlying F0 distribution was identical in both panels (mean = 224 Hz; SD = 5.2 Hz). When LPC analysis was used on the 470 000 synthetic vowels, the results clearly differed from the actual resonance frequencies, and the LPC-measured formants increased with F0. Figure 2, thus, demonstrates that the LPC-measured formant variability can be simply overridden by F0 variability. In both panels, the LPC-measured F1s were attracted to the second harmonic (aligning on a diagonal line around 420–470 Hz); therefore, the measured F1 variability bore no resemblance to the underlying resonance variabilities but reflected exactly twice the F0 variability.

Published studies may, thus, have misattributed changes in F0 to changes in resonances. Heald and Nusbaum (2015), for example, found small but significant changes in F0 and F1 within speakers but at different times of the day. They attributed the F1 changes to changes in vowel articulation, but F0 and F1 tended to rise over the course of the day in their data. If most of the F1 variability was due to harmonic attraction, then there is no change in vowel articulation to be explained. Cochlear implant (CI) users are often reported to have a reduced vowel space (Bharadwaj and Assmann, 2013; Verhoeven *et al.*, 2016), but the amount of reduction may not be as great as reported because CI speakers often have higher than typical F0s (cf. Osberger and McGarr, 1982). A higher F0 increases the risk of poor resonance estimates by raising the F1 of high vowels and lowering the F1 of low vowels (Chen *et al.*, 2019). A vowel space difference might be underestimated as well: Infant-directed speech has been reported as having higher pitch and a larger vowel space, equivalent to hyperarticulation, compared to adult-directed speech (e.g., Cristia and Seidl, 2014; Burnham *et al.*, 2015). See Sec. II B for a detailed discussion of this effect.

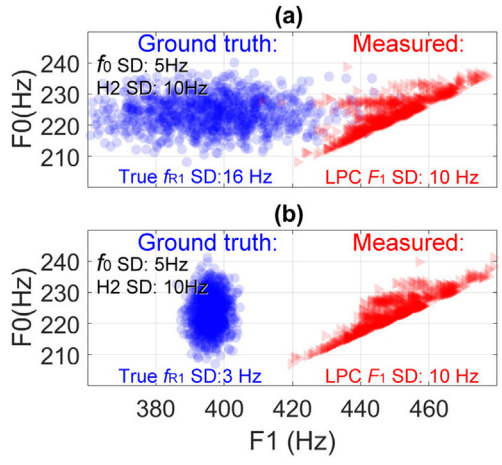

FIG. 2. (Color online) A summary of Fig. 4 of Chen *et al.* (2019) shows simulations of LPC-measured formant variability with 2000 synthesized productions of /i/ having the same variability in F0 [standard deviation (SD) = 5.2 Hz] and the second harmonic (H2, SD = 10.4 Hz) but different (high vs low) underlying variability of the resonance [(a) resonance SD = 16 Hz and (b) resonance SD = 3 Hz)]. In each panel, the (blue) circles on the left represent the ground truth and the (red) triangles on the right indicate the measurements.

Current time-frequency methods for formant measurement are not sufficiently developed. We reviewed all of the formant measurement methods published in three major journals (*IEEE Transactions*, *JASA*, and *Speech Communication*) from 2000 to 2020. Most of the methods were improvements in (1) LPC-based methods (e.g., Smit *et al.*, 2012; Alku *et al.*, 2013; Gowda *et al.*, 2017; Gowda *et al.*, 2020) or (2) discrete Fourier transform (DFT)-based methods (e.g., Medabalimi *et al.*, 2014; Daubechies *et al.*, 2016; Story and Bunton, 2016; Zhang *et al.*, 2020). However, most of the analysis methods are not publicly available, and none of them has been extensively (if at all) used in phonetic studies.

This article summarizes the aspects of previous studies and arguments that continue to be relevant to the kinds of measurements reported in the literature, outlines a more accurate system that nonetheless is not yet automated, and suggests ways of reporting resonance estimates more accurately with regard to the assessment of resonances rather than energy peaks.

II. EXAMPLES AND ISSUES

Various replications of Klatt (1986) have been performed, as will be discussed in Sec. II A. The reasons that averaging over large samples does not fully address the issue are reviewed in Sec. II B. And Sec. II C will discuss studies that try to improve the methods either directly or through formant tracking.

A. Replications

Several studies have undertaken the task of replicating Klatt (1986), always needing many more pages to do so, but all have had the same result. We will point out a prior study that is relevant as well.

In an earlier study with hand-measured spectrograms, Monsen and Engebretson (1983) found sizable errors in

resonance estimation by using a range of formant frequencies and F0 values. These were on the order of ±60 Hz for F1 and F2. Although they asserted that this magnitude was acceptable because this is within the just noticeable difference (jnd) for formants, it can easily represent a 20% error in F1, which can impair the analyses presented in many studies.

Vallabha and Tuller (2002) confirmed Klatt's result with synthetic and natural speech: Errors in resonance measurement increase with increasing F0. Various aspects of real speech, such as the closeness of two formants, further contribute to errors. They showed that the choice of LPC order and other parameters can affect the accuracy of the result, but in some cases, the best choice may depend on prior knowledge of the speech or speakers being analyzed.

Burris *et al.* (2014) compared four LPC-based programs (Praat, Wavesurfer, TF32, and CSL) commonly used to measure formants, applying them to measure synthetic and natural vowels. They selected eight tokens of synthesized corner vowels /i/, /u/, /ɑ/, /æ/, four with female F0s (F0 falling from 245 to 205 Hz) and four with male F0s (falling from 145 to 105 Hz), and measured formants at the midpoint with the default settings in the four programs. For F1, they found that all four programs produced values within ±5% (=25 Hz in F1, 75 Hz in F2) of reference values (F1 = 500 Hz, F2 = 1500 Hz) for synthetic vowels on the male F0 but only one did so for F1 with the female F0. Although using a percentage seems reasonable as a first approximation, it is the case that errors for F1 (as a percentage) were larger than those for F2 in Shadle *et al.* (2016, p. 718). Further, while the 25 Hz range is one that is reasonable for measurements of central tendencies, it is half the resolution possible in typical LPC analyses (see above) and, thus, does not have the accuracy needed for measurements of variability.

Our own study (Shadle *et al.*, 2016) included some newer methods as well as hand measurements by phoneticians. One variant of the LPC method, weighted linear prediction with attenuated main excitation (WLP-AME), was explicitly designed to perform more accurately with high F0s (Alku *et al.*, 2013) with the help of electroglottograph (EGG) signals. Another used an entirely different algorithm, the reassigned spectrogram (RS; Fulop and Fitz, 2006; Fulop, 2010). The hand measurements were made on narrowband spectrograms, which show individual harmonics (for the most part) rather than broad areas of intensity (as in a wideband spectrogram). Judgments of resonance location, therefore, relied on the phonetician's separation of source and filter characteristics. To our surprise, the phoneticians' top-down knowledge that the harmonics will be sparsely represented within a resonance did not improve the hand measurements for F1 of differing F0s; it was still difficult to ignore the location of the most intense harmonic. With the new algorithmic measures, the WLP-AME did significantly better than typical LPC analyses but gave unrealistic results for some of the natural stimuli, possibly due to inaccurate estimations of glottal closing instances without EGG signals. Manual estimation of the formants from the RS display provided the most accurate results but was more time-intensive than any LPC-based analysis.

The previous studies, therefore, are consistent with Klatt's original result: Automatic formant measurements are inaccurate in predictable ways. Previous attempts at an automatic method to avoid those pitfalls have fallen short. The RS measurements are the most promising, but they require, at present, manual measurement. Despite expectations that later improvements would be developed, Klatt's caveats still apply.

B. Inadequacy of increasing sample size

Systematic errors may not be resolved by increasing sample size. Vallabha and Tuller (2002) provided mathematical details showing that some formant measurement errors (by using LPC) are systematic and not random. The F0 effects in their data are, as expected, larger as F0 increases. They point out that because formants may have different signs to their errors, methods that combine formants into new measures will, in some cases, have the errors of both added together (Vallabha and Tuller, 2002, p. 146). Their other reported sources of errors (incorrect filter order, exclusive reliance upon root-solving, and the parabolic interpolation method) are less likely to be corrected and could have different effects on vowels with various qualities.

For the F0 effect, Chen *et al.* (2019) found that even a thousand tokens were not sufficient to avoid bias. Figure 3 summarizes part of the results in Chen *et al.* (2019). Each datapoint (circle, triangle, or diamond symbol) represents a simulation of central tendency error over 1000 synthesized vowels with a fixed first resonance and 1000 randomly sampled F0s from a realistic normal distribution. Age and gender effects on resonances and F0 were approximated from values in the literature. The central tendency error was defined by subtracting the resonance value from the mean of formant measurements over multiple samples: $[(1/n)\Sigma F1(i)] - f_{R1}$, where n (=1000) is the total number of samples of the

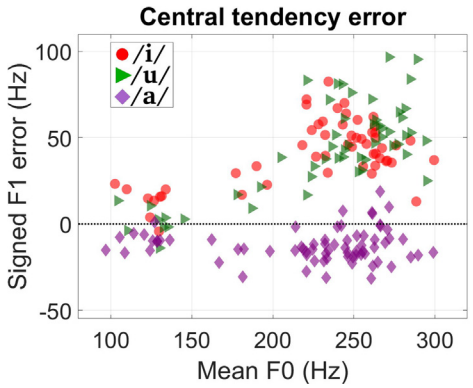

FIG. 3. (Color online) A plot summarizing the original Fig. 2 in Chen *et al.* (2019) with each symbol (circle, triangle, or diamond) representing the central tendency error of a simulated age-gender-vowel group. The central tendency error was calculated as the resonance minus the mean of the F1 measurements of 1000 synthesized vowels with fixed F1 and varying F0.

https://doi.org/10.1121/10.0013410

same vowel, F1(i) is the LPC-measured F1 for the ith sample, and f_{R1} is the first resonance. This simulates an experimenter's attempt to mitigate formant measurement errors by increasing sample size. As shown in Fig. 3, even when averaging 1000 samples, F1 of low vowels tended to be underestimated while F1 of high vowels were increasingly overestimated as F0 increased. Although random errors will average out as N increases, systematic errors will remain.

As mentioned in the Introduction, the simulatons by Chen *et al.* (2019) implied that when speech is produced at a high F0 (e.g., F0 > 200 Hz), formant measurements are more centralized than the resonances truly are. As can be seen in Fig. 3, the F1 estimates for the high vowels /i/ and /u/ are increasingly overestimated as F0 increases. Thus, if vowel spaces for voices with high F0 are calculated, they may be smaller than the resonances would show. One register that includes non-centralized (hyperarticulated) vowels and high F0 is "motherese" or infant-directed-talk (IDT; e.g., Bernstein Ratner, 1984). The bias in the formant measurements might, therefore, underestimate the amount of hyperarticulation as the resonances of vowels may be more extreme than the analysis indicates. Here, to replicate and extend the simulatons by Chen *et al.* (2019), we synthesized 6000 corner vowels to demonstrate the effect of F0 on the LPC-measured vowel space area (VSA). For each of the three corner vowels /i/, /a/, and /u/, 2 sets of 1000 synthesized vowels were created with fixed underlying resonances (/i/: f_{R1} = 350, f_{R2} = 2400, f_{R3} = 3200; /a/: f_{R1} = 891, f_{R2} = 1400, f_{R3} = 2900; /u/: f_{R1} = 400, f_{R2} = 1100, f_{R3} =3050) and varying F0s. The F0s of the first set were randomly sampled from a Gaussian distribution with mean = 143 Hz and SD = 15 Hz; the F0s of the second set were likewise randomly sampled from a Gaussian distribution with mean F0 = 234 Hz and SD = 15 Hz. Figure 4 plots the LPC-measured F1 and F2, averaged across the 1000 tokens for each of the corner vowels in low (green dashed lines) and high F0 (red dotted lines) settings, superimposed on the ground truth (blue solid lines). The LPC-measured VSA for the high F0 setting (F0 = 234 Hz) was underestimated by 15% as compared to the true VSA.

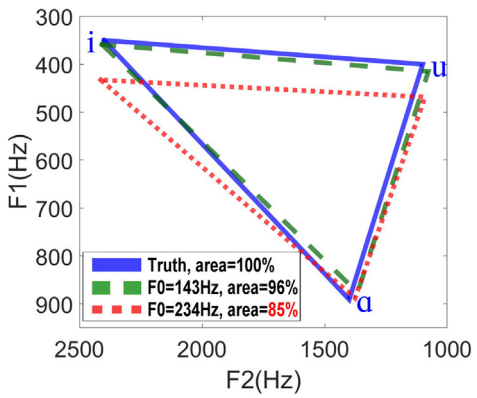

FIG. 4. (Color online) The model using synthesized data in which the blue lines indicate the ground truth of the vowel space area (VSA). The green dashed line indicates the VSA resulting from LPC analysis of vowels synthesized with a relatively low F0 = 143 Hz. When the F0 is 234 Hz (red dotted lines), the LPC-measured VSA was underestimated by 15% as compared to the true VSA.

C. Inadequacy of methods improving existing algorithms

Recent pitch-synchronous approaches for LPC (e.g., Alku *et al.*, 2013; Gowda *et al.*, 2020) were specifically designed to overcome F0 bias but require very accurate glottal pulse information, usually provided by an EGG signal (Alku *et al.*, 2013). However, requiring EGG to be recorded along with the audio signal severely restricts the usefulness of pitch-synchronous LPC.

Most formant tracking algorithms attempt to correct formant selection error (i.e., formant jumps; e.g., wrongly selecting F3 as F2) and usually result in smoother sets of formant values but cannot correct for bias toward the nearest strong harmonic; one can end up with very smooth formant tracks at very wrong frequencies. The default tracking function in Praat implements the Viterbi algorithm (Viterbi, 1967) to select the optimal (smoother) paths through formant candidates, but it does not change the LPC-estimated formant value itself (i.e., F0 bias remains). By contrast, a recent study (Dissen *et al.*, 2019) presented more accurate (than LPC) formant estimates predicted by two deep learning architectures, trained on human-annotated formants from wideband spectrograms. Such an approach was not limited to the parameters of the LPC, but it was limited by the low frequency resolution (smearing) in the wideband spectrogram [see Fig. 5(a) for an example of frequency smearing] and human annotator disagreements (their reported inter-annotator differences = 78 Hz for F1, and 100 Hz for F2, averaged across six phonetic categories). Thus, its accuracy against the true underlying resonances is unknown (cf. Shadle *et al.*, 2016). Dissen *et al.* (2019) reported that the accuracy degraded when the network was applied to data that had not been used for training, indicating a further limitation to interpreting their accuracy results.

As an alternative to LPC, the RS has been developed over a period of many years, beginning in 1976 (Kodera *et al.*, 1976). It is a new version of the spectrogram that plots *instantaneous* frequency and time in lieu of the standard short-time Fourier transform (STFT). The computation leverages the information in the complex phase of the STFT, which is discarded in the conventional spectrogram, to compute the instantaneous times and frequencies of, respectively, the impulsive events (e.g., glottal cycles) and line components (e.g., vocal tract resonances) in the signal (Fulop and Fitz, 2006). Going a step further, it is also possible to *prune* the RS so that only impulses and/or components remain (Fulop and Fitz, 2007). In recent years, the RS has found a wide range of applications and is generating considerable interest in the signal processing community (Kusano *et al.*, 2020; Averbuch, 2021), being applied to, among other things, beamforming (Averbuch, 2021) and component localization in underwater signals (Cho *et al.*, 2019). Nevertheless, it remains underutilized in speech science despite being demonstrated as a superior technique for

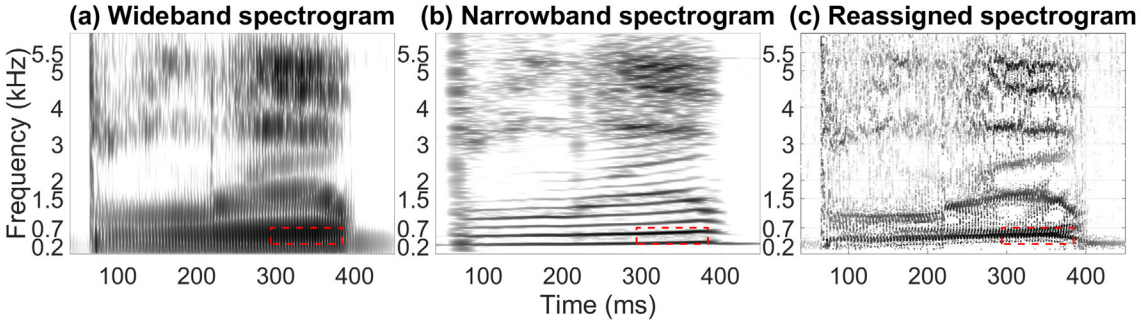

FIG. 5. (Color online) A comparison of (a) wideband, (b) narrowband, and (c) RSs of the same utterance (a 9-year-old girl saying the word "blue") is shown. The utterance was taken from 'fabm2ad2_' in the CMU kids corpus (Eskenazi *et al.*, 1997). The red dashed boxes indicate the zoomed-in region in Fig. 6(a).

resonance frequency estimation. The chief reason for this, aside from the usual sociology and inertia of an established and large scientific field, is likely because it has not been automated, whereas LPC is easily computed from a variety of platforms.

To demonstrate the RS method, Fig. 5 compares conventional wideband and narrowband spectrograms and RS, calculated on a real speech sample of a 9-year-old girl saying "blue," with a mean F0 around 280 Hz. The speech sample was taken from the CMU kids corpus (Eskenazi *et al.*, 1997). As expected, the wideband spectrogram [Fig. 5(a)] is smeared in frequency; the narrowband spectrogram [Fig. 5(b)] shows the harmonic structure clearly up to approximately 3 kHz and harmonics mixed with noise above that (increased harmonic spacing in the final 100 ms indicates a rising pitch); and the RS [Fig. 5(c)] manifests clear and crisp resonance components. Figure 6(a) zooms into the region of the first resonance of the vowel /u/ in "blue" as indicted by the red dashed boxes in Fig. 5. As shown in Fig. 6(a), the LPC-measured F1 (white dotted line in the middle panel) was attracted to the second harmonic (rainbow-colored band just above the white dotted line) and the RS-measured F1 (maroon solid line in the right panel) was unaffected by the harmonics. Figure 6(b) further demonstrates a comparison of the same three types of spectrograms on a synthesized vowel /ʌ/ with the true resonance set at 614 Hz and F0 set at 416 Hz; again, the LPC-measured F1 (middle panel) was attracted to the second harmonic with an error of 282 Hz, whereas the RS-measured F1 (right panel) was very accurate (error = 33 Hz).

Currently, formant measurement by RS still involves manual tracing, but Fulop and Shadle (2018) have shown some initial success of algorithmically tracking the maximum-energy time-frequency ridges on RS. However, more work is needed to develop a fully automated method of RS formant measurement and a system of acquiring ground truth resonance measurements with which automated methods can be tested. While synthetic signals have their place, they embody simplified models of the acoustics in the vocal tract (Zhang *et al.*, 2020). The use of mechanical, 3D-printed models bypasses such synthesis models and provides an experimental system whereby articulatory variables can be manipulated in a controlled way, the acoustic effects (including resonance frequencies) can be measured to obtain the ground truth, and both can be replicated, unlike with natural speech.

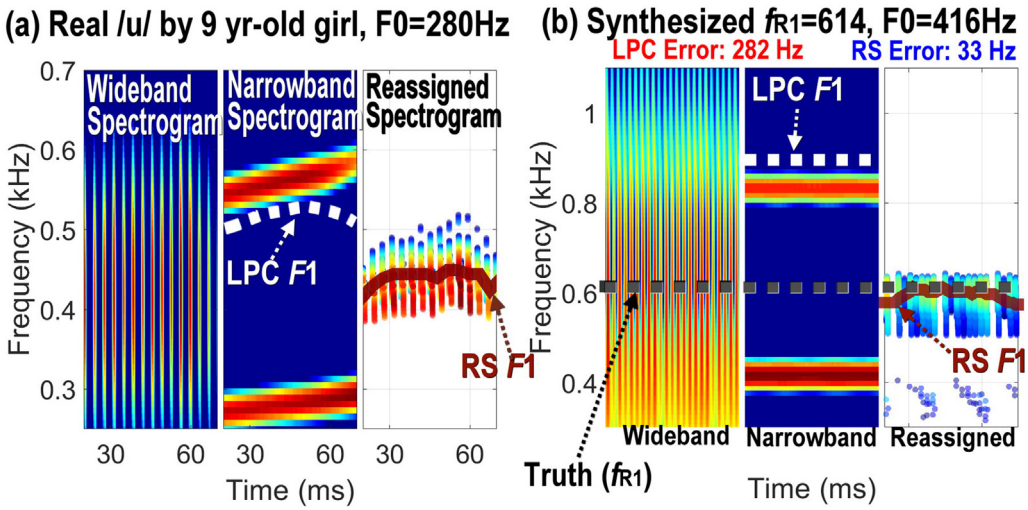

FIG. 6. (Color online) A comparison of three forms of acoustic analysis, showing (a) a child saying /u/, from the region marked with a box in Fig. 5, and (b) synthesized /ʌ/.

https://doi.org/10.1121/10.0013410

III. DISCUSSION AND CONCLUSION

Much about the articulation of speech can be revealed by the acoustic signal, and current measurement techniques can give us, and have given us, valuable insights. The limitations that Klatt (1986) pointed out remain to this day despite efforts to improve analysis techniques. Our own replication (Shadle *et al.*, 2016) found that one spectral representation, the RS, was an improvement and demonstrated that measurements made on it were more accurate than LPC, cepstral, or hand measures of narrowband spectrograms. However, use of RS is not viable for large projects until the method is automated; that work is under way, but success is not guaranteed.

None of the methods discussed so far account for possible zeroes in the spectrum. This is a serious deficit given that zeroes are components of nasality (Fujimura, 1962) and side channels (Stevens, 1998, p. 549), and they occur during the open phases of the glottis (Ananthapadmanabha and Fant, 1982; Plumpe *et al.*, 1999). LPC analysis on speech containing zeroes can lead to serious errors in locating poles (e.g., Gutowski *et al.*, 1978). Atal and Hanauer (1971, p. 638) discuss the issue at length, acknowledging that the all-pole model intrinsic to LPC cannot model zeroes, and antiresonances do occur in many speech sounds. However, they argue that "the location of a pole is considerably more important perceptually than the location of a zero." Cepstral analysis (Schafer and Rabiner, 1970) allows for separating source and filter components and, thus, retaining zeroes, but finding the correct cut-off boundary between source and filter in the quefrency domain can sometimes be difficult. The autoregressive moving average (ARMA) method was specifically designed to include poles and zeroes in the model. However, it requires postulating the correct number of zeroes; in practice, the resultant formant estimations may not show improvement over LPC (Fulop, 2011, p. 125). It may well be that we will have to solve the all-pole issues before we can move on to the important but more challenging issue of spectral zeroes. RS does not have the "no-zero" assumption that LPC does and may allow for a fuller treatment of the entire spectrum. As is the case with LPC, the accuracy of RS measurements of poles is unaffected by the presence of zeroes unless the zero is extremely close to the location of a pole.

The ease of measurements in current systems makes it tempting to think that the issues relating to accuracy of resonance measurement have been solved, and many authors appear to take that (implicit) stance by reporting without qualification such values as given by the programs. Reporting formant values to three (or more) decimal points, for example, reflects a serious lack of understanding of the limits of the measurements even if the program's algorithm generates such numbers. Although standards at various journals may differ, it would make sense to round formant values to the nearest 10 Hz and note that any LPC-measured formant differences smaller than 50 Hz may be the result of errors. Work on babbling or toddlers' speech is particularly prone to error, given the high F0s of productions of young children, but formants and even bandwidths are sometimes reported (e.g., Robb *et al.*, 1997).

For results that seem to test the limits of LPC analysis, researchers may want to double check their results by performing manual RS measurements. (The MATLAB code is available online.[1]) This is a very accurate means of obtaining resonance measures, as we have said, but it is time-consuming. Therefore, only small samples can be expected to be checked in this way.

In general, until a viable alternative to the automaticity available in LPC appears, it is recommended that authors (1) acknowledge the limitations of current analyses with regard to influence of F0 and limits on frequency granularity, (2) report analysis settings more fully, and (3) justify choices made in those settings. A good example of explicit description, and the reasons for various selections, is found in Hillenbrand *et al.* (1995).

A fourth and final check on reliability of formant measurements can be done by a visualization test, a scatterplot of F0s and estimated formant values of the same tokens, superimposed by hypothetical harmonic lines (F0 range vs integer multiples of F0 range) to check if the measurements fall unduly on the harmonic lines (as in Fig. 2). If there is a strong relationship with the harmonics, the formant values must be treated with greater caution. If not, then there is less likelihood of harmonic attraction. Our own work (Whalen and Chen, 2019) used a more complex calculation for the influence of F0, and we hope to provide an open source version of that algorithm in the future. The visualization test, in the meantime, is a reasonable alternative.

In the 36 years since Klatt (1986), speech science has made many advances, but our methods of obtaining formant measurements continue to be treated as resonance measurements, which will obscure some results. With greater attention to parameter settings, more cautious interpretations, and renewed attention to tool development, even more advances can be envisioned, especially given the current emphasis on large datasets.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) Grant No. DC-002717 to Haskins Laboratories. We thank Laura L. Koenig and Kevin D. Roon for helpful comments.

[1]See http://zimmer.csufresno.edu/~sfulop/SpeechSpecmfiles.zip (Last viewed August 9, 2022).

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.

Ananthapadmanabha, T. V., and Fant, G. (**1982**). "Calculation of true glottal flow and its components," Speech Commun. **1**, 167–184.

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.

Averbuch, G. (**2021**). "The spectrogram, method of reassignment, and frequency-domain beamforming," J. Acoust. Soc. Am. **149**, 747–757.

Bernstein Ratner, N. (**1984**). "Patterns of vowel modification in mother-child speech," J. Child Lang. **11**, 557–578.
Bharadwaj, S. V., and Assmann, P. F. (**2013**). "Vowel production in children with cochlear implants: Implications for evaluating disordered speech," Volta Rev. **113**, 149–169.
Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer (version 6.0.49) [computer program]," http://www.praat.org (Last viewed February 20, 2019).
Burg, J. P. (**1967**). "Maximum entropy spectral analysis," in *Proceedings of the 37th Meeting of the Society of Exploration Geophysicists*, Oklahoma City, OK.
Burnham, E. B., Wieland, E. A., Kondaurova, M. V., McAuley, J. D., Bergeson, T. R., and Dilley, L. C. (**2015**). "Phonetic modification of vowel space in storybook speech to infants up to 2 years of age," J. Speech, Lang., Hear. Res. **58**, 241–253.
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. Res. **57**, 26–45.
Chen, W.-R., Whalen, D. H., and Shadle, C. H. (**2019**). "F0-induced formant measurement errors result in biased variabilities," J. Acoust. Soc. Am. **145**, EL360–EL366.
Chiba, T., and Kajiyama, M. (**1941**). *The Vowel: Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo).
Cho, H., Kim, W. J., and Hong, W. (**2019**). "Underwater signal analysis in the modulation spectrogram with time-frequency reassignment technique," IEICE Trans. Fundam. Elec., Commun, Comp. Sci. **102**(11), 1542–1544.
Cristia, A., and Seidl, A. (**2014**). "The hyperarticulation hypothesis of infant-directed speech," J. Child Lang. **41**, 913–934.
Daubechies, I., Wang, Y., and Wu, H.-t. (**2016**). "ConceFT: Concentration of frequency and time via a multitapered synchrosqueezed transform," Philos. Trans. R. Soc. A **374**, 20150193.
den Ouden, D.-B., Galkina, E., Basilakos, A., and Fridriksson, J. (**2018**). "Vowel formant dispersion reflects severity of apraxia of speech," Aphasiology **32**, 902–921.
Dissen, Y., Goldberger, J., and Keshet, J. (**2019**). "Formant estimation and tracking: A deep learning approach," J. Acoust. Soc. Am. **145**, 642–653.
Eskenazi, M., Mostow, J., and Graff, D. (**1997**). "The CMU kids corpus," in *Linguistic Data Consortium* (Linguistic Data Consortium, Philadelphia).
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).
Fujimura, O. (**1962**). "Analysis of nasal consonants," J. Acoust. Soc. Am. **34**, 1865–1875.
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**, 2114–2117.
Fulop, S. A. (**2011**). *Speech Spectrum Analysis* (Springer Science and Business Media, Berlin).
Fulop, S. A., and Fitz, K. (**2006**). "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," J. Acoust. Soc. Am. **119**, 360–371.
Fulop, S. A., and Fitz, K. (**2007**). "Separation of components from impulses in reassigned spectrograms," J. Acoust. Soc. Am. **121**, 1510–1518.
Fulop, S. A., and Shadle, C. H. (**2018**). "Automated formant tracking using reassigned spectrograms," J. Acoust. Soc. Am. **143**, 1870.
Gowda, D., Airaksinen, M., and Alku, P. (**2017**). "Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation," J. Acoust. Soc. Am. **142**, 1542–1553.
Gowda, D., Kadiri, S. R., Story, B. H., and Alku, P. (**2020**). "Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals," IEEE/ACM Trans. Audio, Speech, Lang. Process. **28**, 1901–1914.
Gupta, D., Bansal, P., and Choudhary, K. (**2018**). "The state of the art of feature extraction techniques in speech recognition," in *Speech and Language Processing for Human-Machine Communications*, edited by S. S. Agrawal, A. Devi, R. Wason, and P. Bansal (Springer Singapore, Singapore), pp. 195–207.
Gutowski, P. R., Robinson, E. A., and Treitel, S. (**1978**). "Spectral estimation: Fact or fiction," IEEE Trans. Geosci. Electron. **16**, 80–84.
Haley, K. L., Ohde, R. N., and Wertz, R. T. (**2001**). "Vowel quality in aphasia and apraxia of speech: Phonetic transcription and formant analyses," Aphasiology **15**, 1107–1123.
Hall, N. (**2013**). "Acoustic differences between lexical and epenthetic vowels in Lebanese Arabic," J. Phon. **41**, 133–143.
Han, Y.-m., Wang, F., Huang, X.-x., and Wang, M.-h. (**2018**). "A comparison of resonant peaks and dental resonance in children with spastic cerebral palsy and normal children," Chin. Sci. J. Hear. Speech Rehab. **16**, 133–135.
Heald, S. L. M., and Nusbaum, H. C. (**2015**). "Variability in vowel production within and between days," PLoS One **10**, e0136791.
Hermann, L. (**1890**). "Phonophotographische Untersuchungen III" ("Phonophotographic investigations III"), Arch. Gesamte Physiol. Menschen Tiere **47**, 347–391.
Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Klatt, D. H. (**1986**). "Representation of the first formant in speech recognition and in models of the auditory periphery," in *Proceedings of the Montreal Satellite Symposium on Speech Recognition, 12th International Congress on Acoustics*, edited by P. Mermelstein (Canadian Acoustical Society, Montreal), pp. 5–7.
Kodera, K., De Villedary, C., and Gendrin, R. (**1976**). "A new method for the numerical analysis of non-stationary signals," Phys. Earth Planet. Inter. **12**, 142–150.
Kusano, T., Yatabe, K., and Oikawa, Y. (**2020**). "Maximally energy-concentrated differential window for phase-aware signal processing using instantaneous frequency," in *ICASSP 2020— 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5825–5829.
Lenoci, G., Celata, C., Ricci, I., Chilosi, A., and Barone, V. (**2021**). "Vowel variability and contrast in childhood apraxia of speech: Acoustics and articulation," Clin. Linguist. Phon. **35**, 1011–1035.
Medabalimi, A. J. X., Seshadri, G., and Bayya, Y. (**2014**). "Extraction of formant bandwidths using properties of group delay functions," Speech Commun. **63-64**, 70–83.
Monsen, R. B., and Engebretson, A. M. (**1983**). "The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction," J. Speech. Lang. Hear. Res. **26**, 89–97.
Osberger, M. J., and McGarr, N. S. (**1982**). "Speech production characteristics of the hearing impaired," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), pp. 221–283.
Plumpe, M. D., Quatieri, T. F., and Reynolds, D. A. (**1999**). "Modeling of the glottal flow derivative waveform with application to speaker identification," IEEE Trans. Speech Audio Process. **7**, 569–586.
Robb, M. P., Chen, Y., and Gilbert, H. R. (**1997**). "Developmental aspects of formant frequency and bandwidth in infants and toddlers," Folia Phoniatr. Logop. **49**, 88–95.
Russell, G. O. (**1928**). *The Vowel: Its Physiological Mechanism as Shown by X-Ray* (Ohio State University Press, Columbus, OH).
Schafer, R. W., and Rabiner, L. R. (**1970**). "System for automatic formant analysis of voiced speech," J. Acoust. Soc. Am. **47**, 634–648.
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.
Smit, T., Türckheim, F., and Mores, R. (**2012**). "Fast and robust formant detection from LP data," Speech Commun. **54**, 893–902.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).
Story, B. H. (**2008**). "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," J. Acoust. Soc. Am. **123**, 327–335.
Story, B. H., and Bunton, K. (**2016**). "Formant measurement in children's speech based on spectral filtering," Speech Commun. **76**, 93–111.
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**, 3005–3007.
Turton, D., and Baranowski, M. (**2021**). "Not quite the same: The social stratification and phonetic conditioning of the FOOT–STRUT vowels in Manchester," J. Ling. **57**, 163–201.
Vallabha, G. K., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**, 141–160.

Verhoeven, J., Hide, O., De Maeyer, S., Gillis, S., and Gillis, S. (**2016**). "Hearing impairment and vowel production. A comparison between normally hearing, hearing-aided and cochlear implanted Dutch children," J. Commun. Disord. **59**, 24–39.
Viterbi, A. J. (**1967**). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. Inform. Theory **13**, 260–269.
Whalen, D. H., and Chen, W.-R. (**2019**). "Variability and central tendencies in speech production," Front. Commun. **4**, 1–9.
Zhang, Z., Honda, K., and Wei, J. (**2020**). "Retrieving vocal-tract resonance and anti-resonance from high-pitched vowels using a rahmonic subtraction technique," in *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7359–7363.