High-Pitch Formant Estimation by Exploiting Temporal Change of Pitch


http://dx.doi.org/10.1109/tasl.2009.2024732

Institute of Electrical and Electronics Engineers

Final published version

Sat Apr 06 05:06:21 EDT 2019

http://hdl.handle.net/1721.1/53522

Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Please share how this access benefits you. Your story matters.
High-Pitch Formant Estimation by Exploiting Temporal Change of Pitch

Tianyu T. Wang, Student Member, IEEE, and Thomas F. Quatieri, Fellow, IEEE

Abstract—This paper considers the problem of obtaining an accurate spectral representation of speech formant structure when the voicing source exhibits a high fundamental frequency. Our work is inspired by auditory perception and physiological studies implicating the use of pitch dynamics in speech by humans. We develop and assess signal processing schemes aimed at exploiting temporal change of pitch to address the high-pitch formant frequency estimation problem. Specifically, we propose a 2-D analysis framework using 2-D transformations of the time–frequency space. In one approach, we project changing spectral harmonics over time to a 1-D function of frequency. In a second approach, we draw upon previous work of Quatieri and Ezzat [1], [2], with similarities to the auditory modeling efforts of Chi et al. [3], where localized 2-D Fourier transforms of the time–frequency space provide improved source-filter separation when pitch is changing. Our methods show quantitative improvements for synthesized vowels with stationary formant structure in comparison to traditional and homomorphic linear prediction. We also demonstrate the feasibility of applying our methods on stationary vowel regions of natural speech spoken by high-pitch females of the TIMIT corpus. Finally, we show improvements afforded by the proposed analysis framework in formant tracking on examples of stationary and time-varying formant structure.

Index Terms—Formant estimation, high-pitch effects, linear prediction, spectrottemporal analysis, temporal change of pitch.

I. INTRODUCTION

ESTIMATING the formant frequencies of a speaker during vowel utterances is a fundamental problem of speech analysis. Current state-of-the-art formant estimation systems typically perform short-time analysis in conjunction with a tracking mechanism (Fig. 1). For example, in [4], short-time analysis is used to generate linear predictive cepstral coefficients (LPCC) which are used as observations in a Kalman filtering framework. This paper focuses on the analysis component (shaded, Fig. 1) of such systems in relation to formant estimation of high-pitch speakers (e.g., with pitch values ranging from 150 to 450 Hz). Under such conditions, short-time traditional and homomorphic linear prediction analysis (by which LPCC are obtained) are known to provide inadequate representations of the formant structure [5], [6]. This paper addresses this limitation in analysis by exploiting temporal changes in pitch.

The current work is inspired by psychophysical and physiological studies of the auditory system implicating its use of pitch dynamics in processing speech. For instance, McAdams showed in a series of concurrent vowel segregation tasks that subjects reported an increased “prominence” percept for vowels whose pitch was modulated relative to those that were not modulated [7]. In another study by Diehl et al. [8], although limited to a two-category vowel identification task, results indicated that a linearly changing pitch can improve human vowel identification accuracy. In both studies, the observed effects were greatest when the synthetic source was chosen to have a high pitch (e.g., ~250–400 Hz). Physiological models of the auditory system (e.g., [3]) have also proposed that high-level cortical processing analyzes an auditory-based time–frequency distribution in both the temporal and spectral dimensions, thereby exploiting temporal changes in speech such as pitch dynamics.

The auditory model proposed by Chi et al. [3] may be viewed as one realization of a generalized two-dimensional (2-D) processing framework. In our work, we develop and assess realizations of this framework for improving high-pitch formant estimation. The foundation for our framework was first introduced in [9], [10]. In one approach, we project changing pitch harmonics to a single function of frequency to improve the spectral sampling of a stationary formant envelope under high-pitch conditions. This initial step is similar to the multi-frame analysis method proposed by Shiga and King [11], in which iterative techniques were subsequently used to derive a cepstrum best fit to the collection of harmonic peaks in a least-squared-error sense. Alternatively, in this paper, we perform a simple interpolation across the harmonic peaks to generate a magnitude spectrum to be used with the autocorrelation method of linear prediction. In the second approach, we compute two-dimensional Fourier transforms of spectrottemporal local regions of the short-time Fourier transform magnitude as proposed by Quatieri [1] and extended by Ezzat.
et al. [2]. Ezzat et al. have made phenomenological observations regarding the source-filter separability of this transformed space. In our paper, we argue for these observations analytically and exploit this characteristic to improve high-pitch formant estimation.

We emphasize that this work focuses on improving the analysis framework used in formant estimation for high-pitch vowels rather than performing formant tracking (e.g., in conjunction with Kalman filtering [4]). We therefore obtain formant estimates directly from the output of the analysis framework (*, Fig. 1) through linear prediction analysis rather than through a tracking mechanism. Furthermore, our evaluation primarily focuses on the canonical problem of estimating formant frequencies under the assumption of a stationary vocal tract (i.e., monophthong vowels) but with the added constraint that pitch is changing throughout the duration of the stationary vowel. These conditions will be shown to be often present on natural vowels extracted from the standard TIMIT corpus [12]. The constraint of changing pitch will also be argued to be a modest one with regards to the degree of pitch change required to obtain improvements in formant estimates. Finally, we will demonstrate that the proposed 2-D framework can provide improvements in formant tracking for both the stationary and time-varying vocal tract conditions. As will be subsequently discussed, these improvements can be obtained whenever the pitch dynamics and formant trajectories of a vowel are moving in distinct directions. This more generalized set of conditions for improving formant estimates may be explored in future work towards a full formant-estimation system that incorporates a tracking mechanism.

This paper is organized as follows. Section II briefly reviews the problem of high-pitch formant estimation using a frequency-domain view often referred to as spectral undersampling. We propose a 2-D analysis framework in contrast to traditional short-time analysis for addressing this issue under certain conditions. Section III describes specific formant estimation methods motivated from our observations in Section II. Section IV presents a comparative evaluation of formant frequency estimates from these methods using synthetic vowels. In Section V, to demonstrate the feasibility of our analysis methods on natural speech, we present formant estimation results on monophthong vowels spoken by high-pitch females from the TIMIT corpus. In Section VI, we demonstrate the implications of our work for formant tracking of monophthong vowels, comparing standard LPCC analysis with the proposed 2-D framework in several tracking tasks. Section VII extends these efforts by demonstrating the feasibility of applying the proposed 2-D framework to formant tracking of a time-varying vocal tract configuration. Section VIII summarizes our conclusions and describes future directions.

II. TWO-DIMENSIONAL PROCESSING FRAMEWORK

A long-standing problem in formant estimation is that of obtaining accurate estimates for high-pitch speakers. Existing analysis methods are known to provide poor representations of the formant structure for a high-pitch source signal. For instance, traditional linear prediction suffers from aliased autocorrelation coefficients while cepstral analysis on high-pitch speech results in reduced source-filter separability in the frequency domain [5]. In this paper, we adopt an equivalent view of this difficulty in analysis referred to as spectral undersampling. Specifically, from the traditional source-filter production model for vowels, a periodic source signal can be viewed as harmonically sampling an underlying formant envelope in the frequency domain. As the pitch of the source signal increases, this harmonic sampling becomes more sparse, thereby leading to a poorer spectral representation of the underlying formant envelope in the resulting spectrum [9]. Herein we discuss two realizations of a (2-D) analysis framework to address the spectral undersampling problem under certain conditions. Specifically, for a stationary formant envelope and changing pitch, we argue that spectral sampling can be improved. Similarly, in a second realization, we argue for source-filter separability in a transformed 2-D space.

A. Harmonic Projection and Interpolation

Fig. 2 shows a schematic of a short-time Fourier transform (STFT) for a fixed vocal tract (shaded regions) under 1) fixed high-pitch (225 Hz) and 2) changing high-pitch conditions (200 to 250 Hz). As previously noted, a fixed high pitch results in spectral undersampling; we highlight these spectral samples in Fig. 3(b).1 In contrast, consider now the STFT of the same vowel but with changing pitch from 200 Hz to 250 Hz; under a multiplicative source-filter model, the pitch harmonics sweep through the spectral envelope over time in a fan-like structure. To understand why this fan structure arises, consider a pitch $f_0$ with the $n$th and $(n + 1)$th harmonics as $n f_0$ and $(n + 1) f_0$, respectively. For a pitch change of $\Delta f_0$, the $n$th and $(n + 1)$th harmonics of the new pitch $f_0 + \Delta f_0$ become $(f_0 + \Delta f_0)n$ and $(f_0 + \Delta f_0)(n + 1)$, respectively. Whereas the $n$th harmonic is shifted by $\Delta f_0 n$, the $(n + 1)$th harmonic is shifted by $\Delta f_0(n + 1)$. Consequently, for a given $\Delta f_0$, higher-order harmonics are shifted more on an absolute frequency scale than lower-order harmonics. A modest change in pitch shift can therefore invoke a substantial shift in increasing harmonic frequencies. In the STFT, this results in fanning of the harmonic line structure at higher frequency regions [Fig. 2(b)].

Invoking now the spectral sampling from the pitch harmonics, multiple short-time spectral slices computed across time can therefore be viewed as a collection of nonuniform samples of the stationary spectral envelope for the condition of changing pitch. These samples can be projected [arrows, Fig. 2(b)] to the vertical frequency axis to provide improved sampling of

1Unless otherwise indicated, spectrograms plotted in this paper are on a linear scale.
the underlying envelope. Fig. 3(c) shows an example of this increased sampling due to changing pitch, contrasting the uniform sampling in Fig. 3(b). A spectrum derived from this increased sampling could provide an improved representation of the underlying formant envelope especially under conditions of high pitch relative to that of a single spectral slice. Observe also that low-frequency regions (e.g., ~500 Hz) exhibit narrower sampling than the broader sampling in high-frequency regions (e.g., ~2000 Hz) due to the fanning of harmonic lines in higher-frequency regions (i.e., higher order harmonics).

For the subsequent discussions, our results involving harmonic projection are largely empirical. Note, however, that if additional samples from projection were to provide an increased density of uniform spectral samples, aliasing in the signal’s corresponding autocorrelation function can be reduced and linear prediction all-pole modeling can be improved [5]. More generally, reconstruction methods from nonuniform samples exist in the literature (e.g., [13]), but are beyond the scope of this paper.

B. Grating Compression Transform

As another realization of the 2-D framework, we use an analysis scheme proposed by Quatieri in [1] and extended by Ezzat et al. in [2]. Specifically, the Grating Compression Transform (GCT) is defined as the 2-D Fourier transform computed over a spectrotemporally localized region of the STFT magnitude. The resulting 2-D space has been suggested to exhibit source-filter separability by Ezzat et al. through phenomenological analyses [2]. Herein we motivate analytically the GCT and propose a model for source-filter separability to support those observations.

1-D Modeling of Source: To motivate the GCT, we first consider a periodic impulse train in discrete time

$$p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kN_0]$$  \hspace{1cm} (1)

with $N_0 = 30$ as shown in Fig. 4(a). Windowing $p[n]$ with a symmetric window $w_s[n] = w_s[N - n]$ results in the signal shown in Fig. 4(a). $w_s[n]$ is chosen to be a Blackman window with duration four times the periodicity of $p[n]$ such that $N = 4N_0 + 1 = 121$. This short-time pitch-adaptive scheme ensures that the window captures three impulses when the window center is aligned with an impulse and at most four impulses when it is not aligned since the window has zeros at its endpoints. Fig. 4(a) shows the condition when the center of $w_s[n]$ is aligned with an impulse such that the windowed signal $p_w[n]$ contains three impulses. By symmetry of the window, and denoting $\alpha = w_s[N_0]$, $p_w[n]$ then becomes

$$p_w[n] = \delta[n - 2N_0] + \alpha \delta[n - N_0] + \alpha \delta[n - 3N_0].$$  \hspace{1cm} (2)

Fig. 4(b) shows the magnitude of $N_{\text{DFT}} = 2048$-point discrete Fourier transform (DFT) of $p_w[n]$ (denoted as $P_w[m]$)

$$P_w[m] = \sum_{n=0}^{N-1} p_w[n] e^{-j\frac{2\pi mn}{N_{\text{DFT}}}} = e^{-j\frac{2\pi m2N_0}{N_{\text{DFT}}}} + \alpha e^{-j\frac{2\pi mN_0}{N_{\text{DFT}}}} + \alpha e^{-j\frac{2\pi m3N_0}{N_{\text{DFT}}}}$$

$$= e^{-j\frac{2\pi mN_0}{N_{\text{DFT}}}} \left[ 1 + 2\alpha \cos \left( \frac{2\pi mN_0}{N_{\text{DFT}}} \right) \right]$$

such that

$$|P_w[m]| = 1 + 2\alpha \cos \left( \frac{2\pi mN_0}{N_{\text{DFT}}} \right)$$  \hspace{1cm} (3)

when $\alpha < 1/2$. Our analysis shows that under certain conditions, the short-time analysis scheme described results in a Fourier transform magnitude corresponding to a sinusoid resting on a DC pedestal. For our assumed values $N_0 = 30$ and $N_{\text{DFT}} = 2048$, the period of $|P_w[m]|$ (denoted as $\gamma_0$) is $\gamma_0 \approx (N_{\text{DFT}})/(N_0) \approx 68$.

Consider now the more general situation in which $w_s[n]$ is positioned such that it does not align with an impulse. In this
case, $p_w[n]$ will correspond to four impulses with varying amplitudes as shown in Fig. 4(c), i.e.,

$$p_w[n] = \sum_{i=0}^{3} \alpha_i \delta[n-iN_0-\Delta]$$  \hspace{1cm} (5)

where $0 < \Delta < N_0$ corresponds to an offset based on the position of the window and $\alpha_i$ are the distinct amplitudes. Because the spacing between impulses in $p_w[n]$ is $N_0$, the corresponding $P_w[m]$ exhibits periodicity with spacing $\gamma_0 \approx 68$. In the case shown in Fig. 4(c), observe that $\alpha_1 > \alpha_2 > \alpha_0 > \alpha_3$ such that $P_w[m]$ can be written as

$$P_w[m] = \alpha_0 e^{-j \frac{2\pi m}{N_{\text{DFT}}} \Delta} + \alpha_1 e^{-j \frac{2\pi m}{N_{\text{DFT}}} (\Delta+N_0)} + \alpha_2 e^{-j \frac{2\pi m}{N_{\text{DFT}}} (\Delta+2N_0)}$$
$$+ \alpha_3 e^{-j \frac{2\pi m}{N_{\text{DFT}}} (\Delta+3N_0)}.$$  \hspace{1cm} (6)

The previous derivation is similar to that of the aligned window but with the addition of two complex exponential terms. Fig. 4(d) shows $|P_w[m]|$ for this case; observe that despite the presence of these added terms, $|P_w[m]|$ is similar in shape to that of the ideal case with periodicity $\gamma_0 \approx 68$ as expected.

To quantitatively assess the effect of the exponential terms in (6), Fig. 5 compares the Fourier transform of the short-time spectra $P_w[m]$ for both the aligned and unaligned cases. We denote this transformed result as $P_w(\Omega)$

$$P_w(\Omega) = \sum_{m=0}^{N_{\text{DFT}}} P_w[m] e^{-j \frac{2\pi m}{\gamma_0}}.$$  \hspace{1cm} (7)

In both cases, $|P_w(\Omega)|$ exhibits peaks at $\Omega = (2\pi N_0)/(N_{\text{DFT}}) \approx 0.03 \pi$, consistent with $P_w[m]$ being derived from a periodic sequence. For the aligned case, a single peak [solid arrow, Fig. 5(a)] at $\Omega \approx 0.03 \pi$ can be observed in addition to the DC term in $|P_w(\Omega)|$, corresponding to a sinusoid resting on a DC pedestal. The unaligned case exhibits additional peaks at multiples of $\Omega \approx (0.03 \pi)$ as can be expected for a periodic sequence [e.g., dotted arrows, Fig. 5(b)]. Fig. 6(a) and (b) plot gives locations and amplitudes of the two largest non-DC peaks of $|P_w(\Omega)|$ for the unaligned case as a function of the window offset $\Delta$. Observe that the dominant peak is consistently located at $\Omega \approx 0.03 \pi$ while the secondary peak is located at $\Omega \approx 0.06 \pi$.

Our analysis demonstrates that the maximum non-DC peak of $|P_w(\Omega)|$ is consistently located at a value corresponding to the pitch period of the signal, and that its magnitude tends to dominate $|P_w(\Omega)|$, independent of the window alignment. In our subsequent discussion, we approximate $|P_w(\Omega)|$ of the unaligned case as a sinusoid resting on a DC pedestal as in the aligned case.

One-Dimensional Modeling of Vocal Tract: The previous discussion has shown that for an appropriate choice of window, the Fourier transform of a windowed impulse train resembles a sinusoid resting on a DC pedestal. In the traditional source-filter production model, this impulse train is convolved with both glottal source and formant structure components to generate the speech signal. In accordance with this model, we view the sinusoid on a DC pedestal as corresponding to the periodic source component; in addition, for a localized frequency region, we view the source component as being modulated by a slowly-varying function corresponding to the local glottal source and formant structure spectral components. Let us denote $|X[m]|$ as the short-time magnitude spectrum computed using the previously described pitch-adaptive Blackman window. Denoting $a[m]$ as the envelope and $u[m]$ as a window along the frequency dimension, we model a local portion of the short-time magnitude spectrum as

$$|X_w[m]| = u[m]a[m] \left[ 1 + \cos \left( \frac{2\pi m}{\gamma_0} \right) \right]$$
$$= u[m]a[m] + a[m]u[m] \cos \left( \frac{2\pi m}{\gamma_0} \right)$$  \hspace{1cm} (8)
where \( \gamma_0 \) is inversely proportional to the pitch. Denoting the Fourier transform of \([X_w[m]]\) as \(X(\Omega)\), we have that

\[
X(\Omega) = H(\Omega) + H \left( \Omega - \frac{2\pi}{\gamma_0} \right) + H \left( \Omega + \frac{2\pi}{\gamma_0} \right) \\
H(\Omega) = \frac{1}{2\pi} A(\Omega) *_{\Omega} W(\Omega)
\]

(9)

where \(A(\Omega)\) and \(W(\Omega)\) are the Fourier transforms of \(a[m]\) and \(w[m]\), respectively, and \(H(\Omega)\) is their convolution in the \(\Omega\) domain.

In Fig. 7, we analyze a portion of the vowel /ae/ spoken by a female talker from the TIMIT corpus with pitch \(\approx 300\) Hz; the sampling rate of the waveform is 8 kHz such that \(N_0 \approx 27\). Fig. 7(a) shows the short-time magnitude spectrum \([X_w[m]]\) computed using a pitch-adaptive Blackman window and \(N_{\text{FFT}} = 2048\). Fig. 7(b) shows a localized frequency region of \([X_w[m]]\) to be analyzed. In absolute frequency, this region has size of \(\approx 700\) Hz to capture approximately a single formant peak region as shown in Fig. 7(a) (rectangle). By low-pass filtering the corresponding \(X(\Omega)\), we may recover and divide out the envelope \(a[m]\) (Fig. 7(b), red solid) from \([X_w[m]]\). The filtering operation is done using a Hamming window with length corresponding to a low-pass\(^2\) cutoff of \(\Omega = (2\pi)/(\gamma_0) = (2\pi N_0)/(N_{\text{FFT}}) \approx 0.026\). The result of dividing out \(a[m]\) from \([X_w[m]]\) is shown in Fig. 7(b) (magenta dashed). Observe that this result approximates the previously described sinusoid on a DC pedestal. In addition, Fig. 7(d) shows the Fourier transform magnitude computed for \([X_w[m]]/(a[m])\) after windowing with a Hamming window, which we denote as \(P_{X_w}(\Omega)\). Observe that a dominant peak at \(\Omega \approx 0.026\) can be observed, consistent with the sinusoidal nature. As previously discussed, \(P_{X_w}(\Omega)\) may in general contain additional lower-amplitude peaks at multiples of \(\Omega \approx 0.026\) due to imperfection alignment of the short-time analysis window. Nonetheless, as will be subsequently discussed in Section III-B, these second peaks will have negligible effect on the GCT analysis method used in formant estimation.

Two-Dimensional Modeling of Source and Vocal Tract: The previous discussions invoke a 1-D modulation view of the short-time magnitude spectrum along the frequency dimension. This framework can be extended to include the time dimension of the STFT magnitude. Consider now a localized spectrotemporal region of the STFT magnitude obtained using a 2-D window [e.g., Fig. 8(a)]. Specifically, denote \(s[n, m]\) as the STFT magnitude, \(w[n, m]\) as the 2-D window, and the localized region centered at \(n = n_0\) and \(m = m_0\) as

\[
s_w[n, m] = s[n, m]w[n - n_0, m - m_0].
\]

(10)

While a changing \(f_0\) invokes a fanned harmonic line structure in \(s[n, m]\), a localized region exhibits a harmonic line structure that is angled and approximately parallel as shown in Fig. 8(b) (solid lines). Likewise, a stationary pitch will invoke a set of horizontally-oriented parallel harmonic lines (dashed lines). Analogous to the 1-D case, we model the localized region as the product between a 2-D sinusoid resting on a DC pedestal with a slowly-varying envelope (denoted as \(a[n, m]\)) such that \(s_w[n, m]\) is

\[
s_w[n, m] = w[n, m](1 + \cos(\omega_0 \Phi[n, m]))a[n, m]
\]

(11)

where \(\omega_0\) is the spatial frequency of the 2-D sinusoid, and \(\Phi[n, m]\) is defined as

\[
\Phi[n, m] = m \cos(\theta) + n \sin(\theta)
\]

(12)

corresponding to a term representing its spatial orientation. In relation to the traditional source-filter production model, we interpret \(1 + \cos(\omega_0 \Phi[n, m])\) as a 2-D “source” component corresponding to harmonic line structure while \(a[n, m]\) corresponds to a localized (in time and frequency) portion of the underlying vocal-tract formant excitation envelope. Expanding (11), we obtain

\[
s_w[n, m] \approx w[n, m]a[n, m] + u[n, m]a[n, m] \cos(\omega_0 \Phi[n, m])
\]

(13)

such that the localized region is the sum of a slowly-varying component corresponding to the localized portion of the formant envelope with a modulated version of itself.

Applying the 2-D Fourier transform to \(s_w[n, m]\), results in the GCT. Specifically, let \(N\) and \(M\) denote the time and frequency widths of the 2-D window, respectively; likewise, let \(\omega\) and \(\Omega\) denote the frequency variables corresponding to \(n\) and \(m\), respectively. The GCT of \(s_w[n, m]\) is defined as

\[
S_w(\omega; \Omega) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} s_w[n, m] e^{-j\omega n} e^{-j\Omega m}.
\]

(14)

Substituting (13) into (14), we obtain

\[
S_w(\omega; \Omega) \approx \hat{H}(\omega; \Omega) + \hat{H}(\omega + \omega_0 \sin\theta, \Omega - \omega_0 \cos\theta) + \hat{H}(\omega - \omega_0 \sin\theta, \Omega + \omega_0 \sin\theta)
\]

\[
H(\omega; \Omega) \approx \frac{1}{2\pi} W(\omega; \Omega) *_{\omega, \Omega} A(\omega; \Omega).
\]

(15)

\(^2\)Although there are a number of methods for computing the spectral envelope, we do so here by low-pass filtering.
away from the origin at an angle $\theta$ off the $\Omega$-axis when $\omega_0$ is fixed. For changing $\omega_0$, the rotational nature in transforming rotated harmonic lines maps them off the $\Omega$-axis at a nonzero angle $\theta$ (solid ellipses). Our formulation therefore illustrates analytically the GCT’s ability to invoke source-filter separability observed phenomenologically in [2]. Specifically, for fixed $f_0$, source-filter separability is invoked along the $\Omega$-axis when $\omega_0$ and the $\Omega$-bandwidth of $H(\omega, \Omega)$ are such that there is minimal interaction between $H(\omega, \Omega)$ and $H(\omega, \Omega + \omega_0)$. For changing pitch, separability may be improved since the modulated version of $H(\omega, \Omega)$ is rotated off the $\Omega$-axis, thereby reducing its interaction with the unmodulated $H(\omega, \Omega)$ oriented along the $\Omega$-axis.

Observe that the separability invoked by the GCT is similar to that of the cepstrum [5]. Specifically, both analysis methods transform a multiplicative source-filter spectral model to an alternate domain where the source and filter are additive components; however, the GCT does so without invoking a homomorphic framework. In contrast to the cepstrum, the GCT can provide additional separability when $f_0$ is changing even under high-pitch conditions. Finally, observe that for changing $f_0$, harmonic lines in the STFT tend to fan out with greatest slope in high-frequency regions. The increased fanning corresponds to larger values of $\theta$ when mapped in the GCT. We therefore expect increased separability for these regions relative to lower frequency regions. This is analogous to the broader harmonic sampling discussed in Section II-A.

Our observations motivate a simple method for improving spectral representations of the formant structure in high-pitch speech to address spectral undersampling. Specifically, filtering in the GCT domain followed by an inverse 2-D Fourier transform can be used to generate a spectrum distinct from that obtained from a single spectral slice.

### III. FORMANT ESTIMATION METHODOLOGY

This section describes the methodology used in evaluating our 2-D processing framework for formant estimation. As discussed in Section I, we aim to directly assess the value of spectral representations obtained from different analysis methods and therefore obtain formant frequency estimates directly from the results of analysis (Fig. 1, *) rather than with a tracking mechanism. Herein we develop specific methods for obtaining spectral representations motivated from our observations presented in Section II. The resulting magnitude spectra are then used in linear prediction to estimate formant frequencies. In the development of these estimation methods, and later in Section IV, we use a series of synthesized vowels with varying amounts of linear pitch shifts. Finally two standard analysis methods, traditional and homomorphic linear prediction, are described, to be used as baselines in our comparative study in Section IV.

#### A. Vowel Synthesis

In this portion of our development, we use for a source signal in synthesis periodic impulse trains with varying degrees of pitch dynamics, thereby excluding the glottal pulse shape and its contribution to the speech spectrum. This is done to assess the benefits of exploiting pitch dynamics in high-pitch formant estimation with adequate compensation for this component. In our subsequent discussion of processing natural speech (Section V), we illustrate the feasibility of applying these methods to real and resynthesized speech using a simple compensation method.

Periodic impulse-train source signals with starting $f_0(f_{0s})$ ranging from 80–200 Hz (males), 150–350 Hz (females), and 200–450 Hz (children) were synthesized with linear pitch increases ($df_0$) ranging from 10 to 50 Hz. $f_{0s}$ and $df_0$ varied in 5-Hz steps. Starting and ending pitch values were linearly interpolated across the duration of each synthesized utterance. To generate the impulse train with time-varying pitch, impulses were spaced according to this interpolation across the desired duration. Initially, we synthesized the source at 8 kHz; however, with this sampling rate, sharp transitions were observed in the spectrogram at pitch transition points, indicative of insufficient temporal resolution. To generate a smoother pitch track, we used instead a 96-kHz sampling rate for generating the impulse train and downsampled the result to 8 kHz, thereby minimizing these
transitions. Due to the time-varying nature of the desired source signal and downsampling, the resulting source signal is not a periodic impulse train; however, the source signal was not observed to invoke significant spectral shaping to the vowel.

Source signals were filtered with sixth-order all-pole models [15] corresponding to the vowels /ah/ (v = 1), /iy/ (2), /ey/ (3), /ae/ (4), /oh/ (5), and /oo/ (6). Formant frequencies (F1, F2, F3) were set to average measurements reported by Hillenbrand et al. for males, females and children [16]. Formant bandwidths (B1, B2, B3) for all three genders were set to the average vowel-specific measurements made by Dunn listed in Table I [17]. Similarly, synthesized utterance durations for all three genders were set to average measurements from the Switchboard corpus by Greenberg and Hitchcock and are listed in Table II [18].

### B. Method of Harmonic Projection and Interpolation

To generate a magnitude spectrum using the projection and interpolation of harmonics (Fig. 2), we first used peak picking to obtain the harmonic peaks. Initially, we performed short-time analysis using a fixed 20-ms Hamming window. Fig. 9 (top) illustrates short-time spectra for the female vowel /ae/ across the minimum, average, and maximum pitch values used in synthesis. Observe that distinct mainlobes of the analysis window are located at harmonic frequencies of the pitch values for the 290- and 500-Hz pitch conditions; this is not the case for the 80-Hz pitch value due to the close proximity of the corresponding harmonics in the frequency domain. To address this, we used instead the pitch-adaptive short-time analysis scheme described in Section II such that larger pitch periods were analyzed with longer windows. Fig. 9 (bottom) shows spectra derived in this manner. Specifically, the Blackman window of length four times the pitch period was observed to invoke the same absolute magnitudes across spectral slices, independent of window length (Step 2). In Fig. 9(a), we show the pitch-adaptive spectrum while Fig. 9(b) illustrates peak-picking using the SEEVOC algorithm. Interpolation across the harmonic peaks using a shape-preserving piecewise cubic interpolator [20] was performed to obtain a smooth spectrum. Fig. 9(c) shows the collection of harmonic samples obtained across the entire vowel and the interpolated magnitude spectrum. We denote this method as $m_3$. Finally, to remove confounds of the interpolation method itself, interpolation was also performed on harmonic spectral samples from a single spectral slice of STFT1 extracted from the middle of the utterance ($m_4$).

The resulting single-slice interpolation is shown in Fig. 9(b). In addition to $m_3$, we also implemented a simpler method of projection and interpolation by computing the average (across time) of all spectral slices of STFT1 which we denote as $m_5$. Fig. 9(d) shows the result of averaging all magnitude spectral slices in STFT1 (solid) along with two sample spectral slices (dashed) used in computing this average. This method can also be thought of in the context of the GCT; specifically, it can be easily shown that this method is equivalent to extracting the $\Omega$-axis of the GCT computed for localized regions of STFT1 followed by an inverse Fourier transform.

### C. Filtering the Grating Compression Transform

For method $m_6$, we implemented filtering of the GCT based on the model of Section II-B. As previously discussed in Section II, the pitch-adaptive short-time analysis used for $m_3$ through $m_5$ provides a reasonable correspondence to the model invoked by the GCT. Specifically, localized frequency regions approximately resemble a sinusoid on a DC pedestal modulated by a slowly-varying function (as in Fig. 7) across pitch conditions in Fig. 9. For method $m_6$, localized regions (denoted

---

**TABLE I**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>/ah/</th>
<th>/iy/</th>
<th>/ey/</th>
<th>/ae/</th>
<th>/oh/</th>
<th>/oo/</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>60</td>
<td>38</td>
<td>42</td>
<td>65</td>
<td>47</td>
<td>50</td>
</tr>
<tr>
<td>B2</td>
<td>50</td>
<td>66</td>
<td>72</td>
<td>90</td>
<td>50</td>
<td>58</td>
</tr>
<tr>
<td>B3</td>
<td>102</td>
<td>171</td>
<td>126</td>
<td>156</td>
<td>98</td>
<td>107</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>Vowel Duration (ms)</th>
<th>/ah/</th>
<th>/iy/</th>
<th>/ey/</th>
<th>/ae/</th>
<th>/oh/</th>
<th>/oo/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>95</td>
<td>100</td>
<td>125</td>
<td>135</td>
<td>135</td>
<td>105</td>
</tr>
</tbody>
</table>

Fig. 9. Short-time analysis based on a fixed 20-ms Hamming window (top); versus pitch-adaptive Blackman window (bottom); observe that harmonic resolution is preserved in the latter method.
as \( s[n, m] \) were extracted from STFT\(_1\) with a separable 2-D window \( w[n, m] \). In time, we chose a rectangular window spanning the full duration of the vowel; in frequency, we used a 700-Hz Hamming window. Since the filtering operation is adaptive across frequency regions, a 350-Hz overlap across regions was invoked to reduce the effects of abrupt changes in the filter.

Two distinct GCTs were computed for each localized region: GCT\(_1\) is used to perform peak-picking in estimating the local spatial frequency \( (\omega_n) \) of the sinusoid due to harmonic structure, while GCT\(_2\) is used to generate a spectral representation by filtering in the GCT domain. For GCT\(_1\), a 2-D gradient operator of the matrix form \([14]\)

\[
\begin{bmatrix}
0 & 2 & 2 \\
-2 & 0 & 2 \\
2 & -2 & 0
\end{bmatrix}
\]  

was applied to the entire STFT followed by removal of the DC component, prior to windowing and computing the 2-D Fourier transform. These operations were observed to reduce the magnitude of components near the GCT origin such that \( \omega_n \) could be estimated using a simple peak picker \([1], [21]\). For GCT\(_2\), the DFT was computed directly from the windowed region without application of the 2-D gradient operation and DC removal

\[
p_2[n, m] = w[n - n_0, m - m_0]s[n, m]
\]

\[
\text{GCT}_2(\omega, \Omega) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} p_2[n, m]e^{-j\omega n}e^{-j\Omega m}.
\]

GCT\(_2\) is filtered with an adaptive 2-D elliptic filter and reconstructed to generate a time–frequency distribution as illustrated through the example in Fig. 11. The 2-D filter was designed by taking the product of two linear-phase low-pass filters in frequency. In time, the pass and stop band edges were fixed to \( 0.25\omega_l \) and \( 0.5\omega_l \), respectively. \( \omega_l \) corresponds to the \( \omega_0 \) estimate derived from the lowest frequency region of STFT\(_1\) with center frequency of 350 Hz (arrow, Fig. 11(a)). In frequency, we used pass and stop band edges of \( 0.25\omega_l \) and \( \omega_l \), respectively, with \( \omega_0 \) corresponding to the local estimate for each region. The filter cutoffs were motivated from empirical observations showing that \( \omega_0 \) tended to increase with frequency region. This effect is caused by the increased fanning of the harmonic line structure towards high-frequency regions; because the harmonic lines are no longer strictly parallel, \( \omega_0 \) was observed to be slightly overestimated in these regions. Using the described filter cutoffs, we therefore obtain an adaptive low-pass elliptical filter that becomes more permissive along the \( \Omega \)-direction for regions with increasing frequency, thereby allowing an improved recovery of \( a[n, m] \) in each localized region. This is consistent with the improved source-filter separation in high-frequency regions previously discussed in Section II-B.

As shown in Fig. 11(e), filtering of GCT\(_2\) removes the modulated versions of the localized formant envelope \( H(\omega \pm \omega_0 \sin \theta, \Omega \mp \omega_0 \cos \theta) \), thereby leaving only the unmodulated version \( H(\omega, \Omega) \). Note that while we show only the magnitudes of the GCT in Fig. 11, filtering was done on the complex GCT\(_2\). To generate the magnitude spectrum for comparison with other methods, a reconstructed time–frequency distribution (Fig. 11(f)) was computed using overlap-add, and a spectral slice was extracted corresponding in time to the middle of the utterance. Finally, recall from Section II-B that the approximate sinusoidal model can exhibit harmonic peaks in the GCT domain at multiples of the local spatial frequency. Nonetheless, since the current method extracts formant information near the GCT origin using the low-pass filter, these peaks will have negligible effect on the resulting reconstruction.

**D. Baseline Methods**

Two baseline methods were implemented for comparison with those aiming to exploit temporal change of pitch. Specifically, a magnitude STFT (denoted as STFT\(_0\)) was computed for each utterance using a 20-ms Hamming window, 1-ms
frame interval, and 2048-point DFT. A single spectral slice located in the middle of the utterance (denoted as \(X_{\text{STFT0}}[k]\)) was then extracted for use with linear prediction; we refer to this as traditional linear prediction (m1).

As another reference, we performed homomorphic linear prediction on \(X_{\text{STFT0}}[k]\). Specifically, Rahman and Shimamura have suggested the computation of the real-cepstrum

\[
c_{\text{STFT0}}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \log|X_{\text{STFT0}}[k]|e^{j \frac{2\pi kn}{N}}
\]  

(18)

followed by liftering with an ideal lifter with cutoff \((0.6)/(f_0)\) for \(f_0 \leq 250\) Hz and \((0.7)/(f_0)\) for \(f_0 > 250\) Hz \((N = 2048)\) [22]. As in method m3, \(f_0\) was obtained from the pitch contour used in synthesis. The DFT was then computed to generate a magnitude spectrum to be used with linear prediction.

**E. Autocorrelation Method of Linear Prediction**

For formant estimation, the one-sided magnitude spectra resulting from methods m1 through m6 (denoted as \(|X[k]|\)) are used to obtain autocorrelation estimates for use in linear prediction. Specifically, \(|X[k]|\) is appended by a frequency-reversed version, thereby resulting in a two-sided zero-phase spectrum denoted as \(X_2[k]\). The inverse DFT of \(X_2[k]\) provides the autocorrelation estimate \(r_x[n]\). \(r_x[n]\) is then used to generate the normal equations which are then solved using the Levinson–Durbin recursion [23]. Recall from our discussion in Section III-A that we used as the vocal tract source signal a downsampled periodic impulse train to eliminate the effects of spectral tilt due to the glottal waveform present in natural speech. The model order was therefore set to 6 to correspond directly to the three synthesized formants. Finally, the roots of the resulting coefficients from linear prediction are solved to obtain the formant frequency estimates for all spectral representations.

**IV. RESULTS ON SYNTHETIC VOWELS**

This section compares the results of formant estimation using methods m1 through m6 in the experimental setup described in Section III on synthetic vowels. Table III lists the methods to be compared. We use as a metric the absolute percent error between an estimated \(\hat{F}(F)\) versus the true \(F\) formant-frequency value:

\[
\text{%error} = 100 \frac{|\hat{F} - F|}{F}.
\]  

(19)

**A. Raw Percent Error**

For the \(i\)th formant of the \(v\)th vowel with \(d_f\) pitch shift, starting pitch \(f_0\), and gender \(g\), the raw percent error using method \(m\) is defined as

\[
E_g(i,v,d_f,f_0; m) = 100 \frac{|\hat{F}_g(i,v,d_f,f_0; m) - F_g(i,v)|}{F_g(i,v)}.
\]  

(20)

\(F_g(i,v)\) corresponds to the true \(i\)th formant frequency of the \(v\)th vowel for gender \(g\) while \(\hat{F}_g(i,v,d_f,f_0; m)\) corresponds to the estimate using method \(m\). Figs. 12–14 show representative results across methods and formant number for the female vowel.
$/ae/$ with $d f_0 = 25$ Hz; in this case, (20) can be viewed as a function of the pitch start ($f_{0h}$).

Some preliminary observations can be made for the traditional linear prediction baseline ($m1$, Figs. 12–14). Recall that we use for analysis a 20-ms Hamming window and a sixth-order linear prediction estimate. We observe that errors exhibit an oscillatory behavior across $f_{0h}$; in addition, the rate of oscillations tends to increase with formant number, with the fastest oscillations occurring for F3. These observations are consistent with those observed by Vallabha and Tuller and may be explained by the alignment of pitch harmonics near (for local error minima) and away from (for local error maxima) formant peaks [24]; we refer to this explanation as “fortuitous sampling.” In accordance with this explanation, pitch changes will invoke greater absolute changes in harmonic positions for higher frequency regions than lower regions such that F3 errors would be expected to oscillate more than F1 errors. Finally, observe that the size of oscillations increases with $f_{0h}$, consistent with the effect of spectral undersampling for higher-pitch formants.

Our results for homomorphic linear prediction ($m2$) are consistent with those reported in [22] in providing gains over traditional linear prediction under some conditions (e.g., $F1 - f_{0h} = 25$ Hz). Nonetheless, observe that harmonic projection and interpolation ($m3$) affords substantial error reductions over $m1$ and $m2$. In addition, the similarity between the errors invoked with single-slice interpolation ($m4$) and $m1$ in this particular case suggest that the error reduction via $m3$ is due to exploiting temporal change of pitch rather than the interpolation method itself. Similarly, $m5$ and $m6$ also afford reductions in the error magnitude under certain conditions (e.g., $F2, m6$). These results suggest that exploiting temporal change of pitch can improve formant estimation for both low- and high-pitch conditions.

Observe that $m2, m3, m5$, and $m6$ exhibit some oscillatory behavior similar to $m1$; however, the local maxima of these oscillations can be lower than those of $m1$ (e.g., $m3$, all formants). For $m2, m3$, and $m5$, we interpret this effect in relation to the “fortuitous sampling” explanation of the oscillatory behavior as increasing the chances of harmonic peaks to align with formant peaks. Nonetheless, this is achieved differently between the methods. Whereas cepstral lifting smooths a spectrum across frequencies and can therefore distribute energy towards the true formant peaks, the projection and interpolation of harmonics does so across time. It appears that the latter method outperforms the former for this purpose (e.g., compare the peak errors for $F2$ between $m2$ and $m3$). For $m6$, we attribute the reduction in oscillation amplitude (e.g., for $F2$ and $F3$) to the improved source-filter separability invoked in transformed 2-D space.

### B. Global Average

To assess the performance of methods across all synthesis conditions (i.e., vowel, pitch shifts, pitch starts), we computed a global average metric defined as

$$F_{i,m} = \frac{100}{SDV} \sum_{s=1}^S \sum_{d=1}^D \sum_{v=1}^V |\hat{F}_{i,m,s,d,v} - F_i|^2 / F_i^2$$  \hspace{1cm} (21)

with S, D, and V corresponding to the total number of $f_{0h}, df_0$, and vowels, respectively. In addition, $F_i$ corresponds to the true $i$th formant frequency for the $v$th vowel while $\hat{F}_{i,m,s,d,v}$ is its corresponding estimate for $s$th starting $f_{0h}$ and $d$th $df_0$.

Table IV gives values of this global average for males, females, and children using the traditional linear prediction ($m1$). Observe that the average error tends to increase from males, females, to children, consistent with the effects of increased spectral undersampling for the higher pitch females and children. Figs. 15–17 show the relative gains of all other methods with respect to this baseline and is computed as

$$RG_{i,m} = 100 \frac{F_{i,m1} - F_{i,m'}}{F_{i,m1}}$$  \hspace{1cm} (22)

for each gender; here, $m'$ denotes the methods $m2$ through $m6$.

Overall, these results are consistent with those observed for the example discussed in Section IV-A. Harmonic projection and interpolation ($m3$) exhibits the best performance with relative gains up $\sim$83% for $F3$; in addition, relative losses incurred by single-slice interpolation ($m4$) are consistent with the role of changing pitch in improving formant estimation (rather than the interpolation method itself). Spectral slice averaging ($m5$) and filtering in the GCT ($m6$) also exhibit similar gains over $m1$ in all conditions. In addition, they provide gains over $m2$ for $F2$ and $F3$ in males and females and $F3$ in children.

The similar performance of $m5$ and $m6$ is not entirely surprising, as we previously saw that the averaging method can be thought of in the context of the GCT (Section III-B). We believe that $m5$ and $m6$’s relatively smaller gains with respect to $m2$ for $F1$ stems from the reduced spreading of harmonic lines in lower frequency regions of the STFT. These are more likely to be mapped along the $\Omega$-axis in the GCT, thereby reducing source-filter separability. In conjunction with the exceptionally high-pitch source signals used for synthesizing children’s speech, this effect could also account for the relatively smaller gains in $F2$ for children using $m5$ and $m6$. Conversely, $m3, m5$, and $m6$ generally exhibit the larger gains for higher formants (e.g., $F3$) than lower formants, presumably due to broader harmonic sampling for $m3$ and $m5$ and the increased source-filter separability in the GCT for $m6$ in high-frequency regions.

### V. Monophthong Vowels in Natural Speech

In this section, we illustrate the feasibility of applying the proposed methods for natural speech. As in Section III, we focus on the analysis framework for estimating formants rather than perform formant tracking. Specifically, we apply the proposed formant estimation methods discussed in Section III to real vowels spoken by high-pitch female talkers of the TIMIT corpus [12]. The vowels analyzed were extracted from TIMIT based on the available phone transcriptions and correspond
to monophthong vowels (i.e., those with stationary or near stationary vocal tract configurations). While it is likely that the TIMIT transcriptions correspond to vowel regions in which the vocal tract is not strictly stationary, we believe this to be a reasonable approximation to demonstrate that 1) sufficient pitch variations are available in natural speech and 2) can occur in regions of stationary or near-stationary vocal tract configurations to be exploited by our methods. Two sets of analyses were performed for each vowel. In the first, we applied our methods directly to the natural speech. In the second, we estimate a set of reference formant frequencies from the natural speech and apply our methods to a resynthesized vowel. Herein we discuss the motivation, methodology, and results of our analyses.

A. Real Vowels

Monophthong vowel regions were extracted from TIMIT waveforms based on the available phone transcriptions. The autocorrelation-based pitch tracker of the Praat software package was employed using a 1-ms frame interval to estimate pitch values across each vowel’s duration [25]. As a preprocessing step to methods m1 through m6, we applied a simple method of spectral tilt compensation to reduce the glottal source contribution to the speech spectrum present in natural speech. Specifically, the real cepstrum of STFT_1 was computed and the first (non-DC) cepstral coefficient was set to zero. This operation has the effect of removing a spectral tilt as derived from the real cepstrum at quefrency n = 1 [5]. The modified STFT_1 was then obtained using the inverse transform and methods m1 through m6 were invoked as previously described in Section III. For the analysis of natural speech, reference formant frequencies have been proposed by Deng et al. in [26]. In that work, an initial set of formant values were estimated using an automatic formant tracking system. These values were then manually corrected to compensate for errors made by the formant tracker, particularly in regions of rapid formant transitions. However, because the ground truth formant values were derived in part using linear predictive cepstral coefficients (LPCC) in conjunction with a tracking mechanism, we are not able to use these as an appropriate reference for comparing against non-LPCC analysis methods (e.g., harmonic projection, GCT-based filtering). The results of our analysis for natural speech therefore serve primarily to illustrate differences in estimation results between the proposed methods and those of standard techniques. In contrast, our use of resynthesized vowels provide reference formant values not derived from LPCC-based estimation.

B. Resynthesized Vowels

In our second set of analyses, we aimed first to obtain a set of reference formant frequency values for use in resynthesis followed by the application of m1 through m6 on the resynthesized vowel. As was previously discussed, one aim of this experimental setup is to assess whether sufficient pitch dynamics are present in natural speech during stationary vowels such that the proposed methods can provide improvements over traditional methods. As a secondary aim for assessing feasibility of our methods for natural speech, recall from Section IV that our synthetic experiments used a periodic impulse train for a source signal, thereby excluding the glottal pulse shape’s contribution to the speech spectrum in natural speech. This was done to illustrate the benefits of exploiting pitch dynamics in formant estimation if this component can be adequately removed. In this section, we instead incorporate a glottal pulse shape in resynthesis followed by the previously discussed cepstrum-based method of compensation in Section VI-A. While a variety of techniques in the literature (e.g., [27]) exist for spectral tilt compensation, our aim here is to assess the feasibility of employing the proposed methods in
formant estimation in conjunction with a simple, though likely incomplete, compensation method.

To obtain a set of reference formant frequencies from the extracted vowel, we reasoned that the spectral slice corresponding to the lowest $f_0$ value in the vowel would be less likely to suffer from spectral undersampling. From the previous section, the spectral slice corresponding to the lowest $f_0$ of the modified STFT was therefore used to estimate a set of reference formant frequencies via traditional linear prediction (i.e., $m_1$). These values were used in conjunction with the formant bandwidths of Table I to specify an all-pole model as described in Section III; therefore, in the resynthesized vowels, we define these reference formant frequency values as ground truth. To generate the source signal for resynthesis, we used the estimated pitch contour to determine pulse periodicity as in Section IV; however, instead of a periodic impulse as in Section III, we used in this section the derivative of the Rosenberg model to invoke a glottal pulse shape [28]; the derivative was invoked to account for the radiation characteristics at the lips [29]. The resynthesized vowel was then analyzed by methods $m_1$ through $m_6$ in conjunction with the cepstrum-based method of spectral tilt compensation.

C. Results

In Fig. 18, we show an example of a real and resynthesized vowel /ah/ from a female TIMIT speaker. Both the narrowband (a,b) and log broadband (c,d) spectrograms are shown to emphasize harmonic and formant structure, respectively. Observe from the broadband spectrogram for the real vowel that $F_3$ exhibits some movement at the beginning of vowel (e.g., near 0.02 s). Nonetheless, the formant structure appears stationary for the majority of the vowel duration. The pitch variation throughout the vowel increases from 251 to 278 Hz. Table V (top) lists the formant frequency estimates of the real vowel using $m_1$ through $m_6$ as well as the raw percent errors computed with respect to the reference formant frequencies on the resynthesized vowel for $m_1$ through $m_6$ (bottom). Similar sets of results are presented in Tables VI–VIII for /iy/, /ey/, and /ae/.

The results of our analysis on natural speech show that the proposed methods provide estimates distinct from traditional and homomorphic linear prediction ($m_1$ and $m_2$). In addition, observe for the resynthesized vowels that the methods $m_3$, $m_5$, and $m_6$ exhibit smaller percent errors from the reference formant frequencies than $m_1$ and $m_2$. These results illustrate that sufficient pitch variations are available in natural speech which can be exploited in improving formant estimation accuracy over traditional methods. In addition, the results illustrate the feasibility of applying the proposed methods in conjunction with a relatively simple spectral tilt compensation method in the presence of a glottal pulse shape.

TABLE V

<table>
<thead>
<tr>
<th>Estimates of $m_1$ through $m_6$ on natural speech utterance (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
<tr>
<td>Raw % errors of $m_1$ through $m_6$ on resynthesized vowel.</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
</tbody>
</table>

TABLE VI

<table>
<thead>
<tr>
<th>Estimates of $m_1$ through $m_6$ on natural speech utterance (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
<tr>
<td>Raw % errors of $m_1$ through $m_6$ on resynthesized vowel.</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
</tbody>
</table>

TABLE VII

<table>
<thead>
<tr>
<th>Estimates of $m_1$ through $m_6$ on natural speech utterance (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
<tr>
<td>Raw % errors of $m_1$ through $m_6$ on resynthesized vowel.</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
</tbody>
</table>

TABLE VIII

<table>
<thead>
<tr>
<th>Estimates of $m_1$ through $m_6$ on natural speech utterance (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
<tr>
<td>Raw % errors of $m_1$ through $m_6$ on resynthesized vowel.</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
</tbody>
</table>
VI. IMPLICATIONS FOR FORMANT TRACKING

In this section, we demonstrate and discuss implications of the proposed analysis framework for the formant tracking problem. As previously noted, current state-of-the-art systems typically employ linear predictive cepstral coefficients (LPCC) in conjunction with a tracking mechanism (e.g., Kalman filtering [4], [30]) to obtain formant estimates across time. In particular, LPCCs are modeled as noisy observations corresponding to an unobserved state (the desired formant frequencies) to be estimated. Our previous results, however, have shown that traditional and homomorphic linear prediction, from which LPCCs are obtained, invoke poorer representations of stationary formant structure compared to analysis methods aiming to exploit temporal pitch dynamics. Herein we demonstrate through examples the effect that such representations have in formant estimation in stationary vowels in the context of formant tracking (i.e., the final output in the scheme of Fig. 1).

For our experimental framework, we applied the baseline extended Kalman filter (EKF) proposed originally in [4] and extended in [30] to synthetic vowels using different analysis methods for generating observations. State variables were chosen to be the three formant frequencies of the synthesized vowel and initialized to the reference formant frequencies. Formant bandwidths were also initialized and held fixed to the true bandwidths across state estimates in applying the EKF equations.

As a baseline observation type, we used 15th-order LPCCs computed for a 20-ms window of the speech signal computed at a 10-ms frame interval as in [30]. To exploit temporal pitch dynamics in generating observations, we applied the harmonic projection, averaging, and GCT-based analysis methods previously described to generate short-time spectral representations; these estimates were subsequently used to obtain 15th-order LPCCs to be used as observations in the EKF. For averaging, collections of spectral slices from STFT spanning 20-ms durations (20 slices total) were extracted at 10-ms intervals and averaged. For harmonic projection, these same spectral slices were used in peak-picking, projection to a single axis, and interpolation to generate a spectral representation as described in Section III; in addition, a control estimate of peak-picking and interpolation using peaks of a single slice was computed. For GCT-based analysis, the adaptive 2-D elliptical filter described in Section III was applied across regions of STFT of size 700 Hz by 20 ms using a 2-D Hamming window. For overlap-add, overlaps of 350 Hz and 10 ms were invoked in frequency and time, respectively. This is in contrast to the GCT-based analysis in Section III that used the full duration of the vowel. Finally, spectral slices were extracted from the filtered spectrogram at 10-ms intervals to generate the desired LPCC observations.

Figs. 19 and 20 show results of the tracker for a synthetic female vowel /æ/ with pitch ranging from 150 to 200 Hz and 250 to 300 Hz, respectively. The baseline LPCC observation results in errors in frequency estimates ranging from ~1 Hz (e.g., F3 for the 150 to 200 Hz pitch variation) up to ~50 Hz (e.g., F1 for the 250 to 300 Hz pitch variation). For both vowels, the variation in errors across time likely reflects the “fortuitous sampling” effect previously noted in Section IV for baseline linear prediction analysis. Qualitatively, the proposed methods that exploit temporal pitch dynamics reduce the magnitude of this variation in error across time. To quantitatively compare the performance of the tracker using different observation types, we show in Tables IX–X the root mean-squared errors (RMSE) in absolute error across time. To quantitatively compare the performance of the tracker using different observation types, we show in Tables IX–X the root mean-squared errors (RMSE) in absolute frequency computed across the full durations of the vowels. The proposed analysis methods provide RMSE reductions ranging from ~2 Hz up to ~25 Hz with respect to the LPCC baseline for the cases considered. In addition, consistent with our observations in Section IV, the relatively poorer performance of the control method suggests that the gains from the projection method are due to its use of temporal pitch dynamics rather than the interpolation method itself. Finally, observe that the overall errors of the LPCC baseline are smaller for the vowel synthesized in the 150 to 200 Hz case relative to the 250 to 300 Hz case. This is consistent with our observations for the traditional linear prediction baseline (Figs. 12–14) showing that errors in formant frequency estimates increase with pitch.
TABLE IX
RMSE IN Hz FOR PITCH 150-TO-200-Hz CASE

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPCC Baseline</td>
<td>17.08</td>
<td>9.8</td>
<td>6.83</td>
</tr>
<tr>
<td>Harmonic Projection</td>
<td>7.83</td>
<td>2.32</td>
<td>3.18</td>
</tr>
<tr>
<td>Control</td>
<td>16.06</td>
<td>12.03</td>
<td>13.94</td>
</tr>
<tr>
<td>Averaging</td>
<td>16.39</td>
<td>9.12</td>
<td>3.19</td>
</tr>
<tr>
<td>GCT-based Filtering</td>
<td>16.9</td>
<td>4.89</td>
<td>4.91</td>
</tr>
</tbody>
</table>

TABLE X
RMSE IN Hz FOR PITCH 250-TO-300-Hz CASE

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPCC Baseline</td>
<td>34.12</td>
<td>29.36</td>
<td>27.36</td>
</tr>
<tr>
<td>Harmonic Projection</td>
<td>8.35</td>
<td>5.78</td>
<td>8.48</td>
</tr>
<tr>
<td>Control</td>
<td>32.7</td>
<td>29.87</td>
<td>28.15</td>
</tr>
<tr>
<td>Averaging</td>
<td>33.85</td>
<td>29</td>
<td>27.16</td>
</tr>
<tr>
<td>GCT-based Filtering</td>
<td>21.23</td>
<td>20.66</td>
<td>8.91</td>
</tr>
</tbody>
</table>

Our findings illustrate limitations of standard LPCC analysis for use in the formant tracking task. Observe that absolute errors in formant frequency estimates up to ~50 Hz can result when using LPCCs as observations in a Kalman filtering framework (Figs. 19–20). These effects are due to the inability of LPCCs to accurately represent formant structure under conditions of high pitch from spectral undersampling. Furthermore, our results suggest that by explicitly exploiting temporal pitch dynamics in analysis for generating observations, tracking performance can be improved.

VII. ANALYSIS AND TRACKING OF TIME-VARYING FORMANTS

A limitation of the present work is that it imposes the constraint of a stationary vocal tract throughout analysis. As we have seen in Section IV, this constraint may be reasonable for portions of monophthong vowels; however, this condition is not satisfied for speech sounds where the formant envelope is changing (e.g., diphthongs, glides) as well as transitions from distinct speech sounds (e.g., stop-to-vowel transitions). For the analysis of speech sounds where the formant envelope is changing, the harmonic projection methods (e.g., \( m_3 \) and \( m_5 \)) are evidently not directly applicable. A potential extension to these methods could be to project harmonic samples across time along each formant trajectory; however, this approach would require a priori estimates of the formant trajectories. Alternatively, observe that the GCT can inherently maintain source-filter separability when the vocal tract is changing. Specifically, consider a formant transition in a localized region that is increasing in frequency as illustrated in Fig. 21(a). Under this condition, the corresponding \( \mathcal{A}(\omega, \Omega) \) centered at the origin of the GCT will be rotated at an angle off the \( \Omega \)-axis as can be shown from standard properties in image processing [14]. Let us denote this direction as a “principle axis,” analogous to the \( \Omega \)-axis when the formant structure is stationary (as in Fig. 8). From the model proposed in Section II, the modulated versions of \( \mathcal{A}(\omega, \Omega) \) will be located off of this principle axis when pitch is either stationary or decreasing (dashed and dotted lines in Fig. 21) [14]. These conditions are analogous to that of increasing and decreasing pitch for the stationary formant case, thereby allowing for improved source-filter separability.

Conversely, when the pitch harmonics are moving in the same direction as the transition (e.g., increasing, solid lines in Fig. 21), separability can only be obtained when the width of \( \mathcal{A}(\omega, \Omega) \) along the principle axis is less than the spatial frequency of the 2-D harmonic structure. This is analogous to the condition of stationary pitch and formant envelope illustrated in Fig. 8.

To illustrate application to a time-varying formant structure, we show in Fig. 22(a) a pitch-adaptive spectrogram (Sections II, III) computed for a synthetic female diphthong /ae/ \( \rightarrow \) /iy/. The vowel was synthesized using a 135-ms duration source signal as described in Section III with pitch variation from 325 to 275 Hz. A time-varying all-pole (order 6) model with linear formant and bandwidth trajectories was applied to the source signal. The starting and ending formant frequencies of the filter were set to the female monophthongs /ae/ (\( F_1 = 135 \) Hz, \( F_2 = 2349 \) Hz, \( F_3 = 2972 \)) and /iy/ (\( F_1 = 437 \) Hz, \( F_2 = 2761 \) Hz, \( F_3 = 3372 \) Hz), respectively. Similarly, the starting and ending bandwidths of the filter were set to the values in Table I for /ae/ and /iy/. In Fig. 22(a), we indicate a local region that contains the third formant transition from 2972 to 3372 Hz (rectangle). Consistent with the model of Fig. 21, the corresponding GCT is shown in Fig. 22(c), and contains components near the origin with orientation off the \( \Omega \)-axis for some positive value of \( \phi \). Observe that the modulated versions of this slowly varying envelope are also oriented in this direction (dashed).

To assess the value of analyzing diphthongs using the GCT, we employed GCT-based filtering and the EKF framework (Section IV) for formant estimation of the diphthong shown in Fig. 22(a). For GCT-based analysis, our steps are identical to those described in Section III with the exception of the choice of filter. Specifically, we extracted regions of size 700 Hz by 135 ms (i.e., the full duration of the vowel) with a 350 Hz overlap in frequency. An adaptive 2-D filtering was applied across regions, and overlap-add was used to reconstruct a time–frequency distribution. Observe from Fig. 21 that the 2-D filter for a diphthong would ideally be a low-pass filter along the new “principle axis” corresponding to the orientation of the moving formant structure. This is in contrast to being along the \( \Omega \)-axis as in the case of monophthongs. In our current development, we approximate this design by using a separable linear-phase low-pass filter as in Section III; however, the filter is designed to capture similar low-frequency content in both
the $\omega$- and $\Omega$-directions [Fig. 22(d)]. The pass- and stop-bands in both the $\omega$ and $\Omega$ directions were therefore set to $0.2\pi\omega_0$ and $0\omega_0$. For the localized region analyzed in Fig. 22(a)–(c), the resulting GCT is shown in Fig. 22(e). Observe that the harmonic components have been removed, and the slowly varying components near the origin are preserved. Fig. 22(f) shows the reconstructed time–frequency distribution from which spectral slices are extracted every 10 ms to generate linear predictive cepstral coefficients (LPCC) as observations to the tracker. This was compared against the baseline 20-ms short-time analysis LPCCs used in Section IV.

Fig. 23 and Table XI show the results of the output of the tracker for both analysis schemes. Fig. 23(a) shows a broadband log-spectrogram of the diphthong with the reference formant frequency trajectories overlain (lines). Observe that due to the high pitch values of the source signal, the formant trajectories do not exhibit smooth transitions across time as may be expected for a low-pitch source. Similar to the observations of Section IV, the baseline tracker exhibits deviations around the true formant frequencies while the GCT-based analysis appears to reduce the magnitude of these deviations [Fig. 23(b)–(d)]. This is quantitatively demonstrated in Table XI, with the GCT-based analysis providing reductions in RMSEs up to $\sim$30 Hz.

VIII. CONCLUSION AND FUTURE WORK

This work has proposed a 2-D processing framework to address formant estimation of high-pitch speakers by exploiting temporal change of pitch. We have shown quantitatively for synthetic signals that our methods outperform traditional and homomorphic linear prediction in formant estimation under conditions of a stationary vocal tract and changing pitch. In addition, we have illustrated the feasibility of the proposed methods for use on natural speech with examples of high-pitch monophthong vowels from female talkers of the TIMIT corpus. We have further demonstrated benefits of the proposed framework in relation to the formant tracking problem in providing improved representations of formant structure under high-pitch conditions for both stationary and time-varying formants. Our results show that the 2-D processing framework is a promising approach for addressing the spectral undersampling problem in formant estimation.

Several future directions are motivated from the current work. The harmonic projection method ($m3$, Section III-B) may be used in conjunction with a method for detecting stationarity of the vocal tract in formant analysis such as in [31]. This may involve incorporating prior information of vowels as proposed by Toda and Tokuda in [32]. Alternatively, an improved filtering method in the GCT domain may be used to better isolate localized formant structure (for both monophthong and diphthong vowels) from its modulated versions. These analysis methods may be used in conjunction with a tracking mechanism towards a full formant estimation system for running speech.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for helpful comments in improving this manuscript. They would also like to thank N. Malyska for use of his software in extracting vowel regions from the TIMIT corpus and D. Rudoy for use of his formant tracking software. Both colleagues also contributed helpful suggestions throughout the course of this work.

REFERENCES


