Speech Communication 105 (2018) 77–91

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Time-varying sinusoidal demodulation for non-stationary modeling of speech

Neeraj Kumar Sharma*, Thippur V. Sreenivas

Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India

ARTICLE INFO

Keywords:
Speech modeling
Sinusoidal modeling
Speech analysis
Speech synthesis
Harmonic demodulation
Subband modeling

ABSTRACT

Speech signals contain a fairly rich time-evolving spectral content. Accurate analysis of this time-evolving spectrum is an open challenge in signal processing. Towards this, we visit time-varying sinusoidal modeling of speech and propose an alternate model estimation approach. The estimation operates on the whole signal without any short-time analysis. The approach proceeds by extracting the fundamental frequency sinusoid (FFS) from speech signal. The instantaneous amplitude (IA) of the FFS is used for voiced/unvoiced stream segregation. The voiced stream is then demodulated using a variant of in-phase and quadrature-phase demodulation carried at harmonics of the FFS. The result is a non-parametric time-varying sinusoidal representation, specifically, an additive mixture of quasi-harmonic sinusoids for voiced stream and a wideband mono-component sinusoid for unvoiced stream. The representation is evaluated for analysis-synthesis, and the bandwidth of IA and IF signals are found to be crucial in preserving the quality. Also, the obtained IA and IF signals are found to be carriers of perceived speech attributes, such as speaker characteristics and intelligibility. On comparing the proposed modeling framework with the existing approaches, which operate on short-time segments, improvement is found in simplicity of implementation, objective-scores, and computation time. The listening test scores suggest that the quality preserves naturalness but does not yet beat the state-of-the-art short-time analysis methods. In summary, the proposed representation lends itself for high resolution temporal analysis of non-stationary speech signals, and also allows quality preserving modification and synthesis.

1. Introduction

Temporal evolution of the short-time spectrum is a key attribute of speech signals. The evolution is slow or fast depending on what is spoken, and how it is spoken (Rosen, 1992). Often short-time modeling of speech suffices for applications such as speaker/speech recognition, and speech coding/compression (Quatieri, 2008). However, such short-time modeling presents fundamental limitations for analyzing fine temporal aspects associated with acoustic attributes of speech (Smits, 1994). A repercussion is below par naturalness in quality of synthesized speech (Stylianou, 2009). Some of the applications which suffer owing to this are speech modifications and text-to-speech synthesis. This paper proposes an alternate and simplified time-varying modeling of the temporally evolving speech spectrum, devoid of any short-time modeling.

Consider a speech signal $x(t)$ depicted in Fig. 1(a). $x(t)$ can be represented as sum of two signal streams, namely, a voiced stream $x_v(t)$ and an unvoiced stream $x_u(t)$. Visualizing the spectrograms in Fig. 1(b,c), $x_v(t)$ and $x_u(t)$ have a markedly different time-frequency structure. In the spectrogram, $x_v(t)$ manifests as horizontal striations, and a time-domain representation can be obtained by explicitly modeling these striations as narrowband time-varying sinusoids. In contrast to $x_v(t)$, the unvoiced signal $x_u(t)$ is devoid of any narrowband spectral spread. Instead, as seen in the corresponding spectrogram, it has a localized temporal distribution of energy with wideband spectral spread (see Fig. 1(c)). $x_u(t)$ can be represented as a mono-component wideband time-varying sinusoid. Combined, we have a time-varying representation for $x(t)$ as follows.

$$x(t) = \underbrace{\sum_{k=1}^{N} a_k(t)\sin\phi_k(t)}_{x_v(t)\text{: voiced stream}} + \underbrace{a_u(t)\sin\phi_u(t)}_{x_u(t)\text{: unvoiced stream}} \tag{1}$$

where $a_k(t)$ and $\phi_k(t)$ denote the instantaneous amplitude (IA) and instantaneous phase (IP), respectively, of the k^{th} sinusoid in $x_v(t)$, Similarly, $a_u(t)$ and $\phi_u(t)$ denote the IA and IP, respectively, of $x_u(t)$. The first derivative[1] of IP is referred to as the instantaneous frequency (IF) (Boashash, 1992) of the sinusoid. Recent literature in speech modeling Kawahara et al. (1999) suggests a quasi-harmonic relationship between sinusoids in a voiced stream. We incorporate this aspect into the mode

[1] Derivative will be denoted by ′(prime).

* Corresponding author.
E-mail addresses: neerajs@iisc.ac.in (N.K. Sharma), tvsree@iisc.ac.in (T.V. Sreenivas).

https://doi.org/10.1016/j.specom.2018.10.008
Received 5 March 2018; Received in revised form 16 September 2018; Accepted 29 October 2018
Available online 1 November 2018

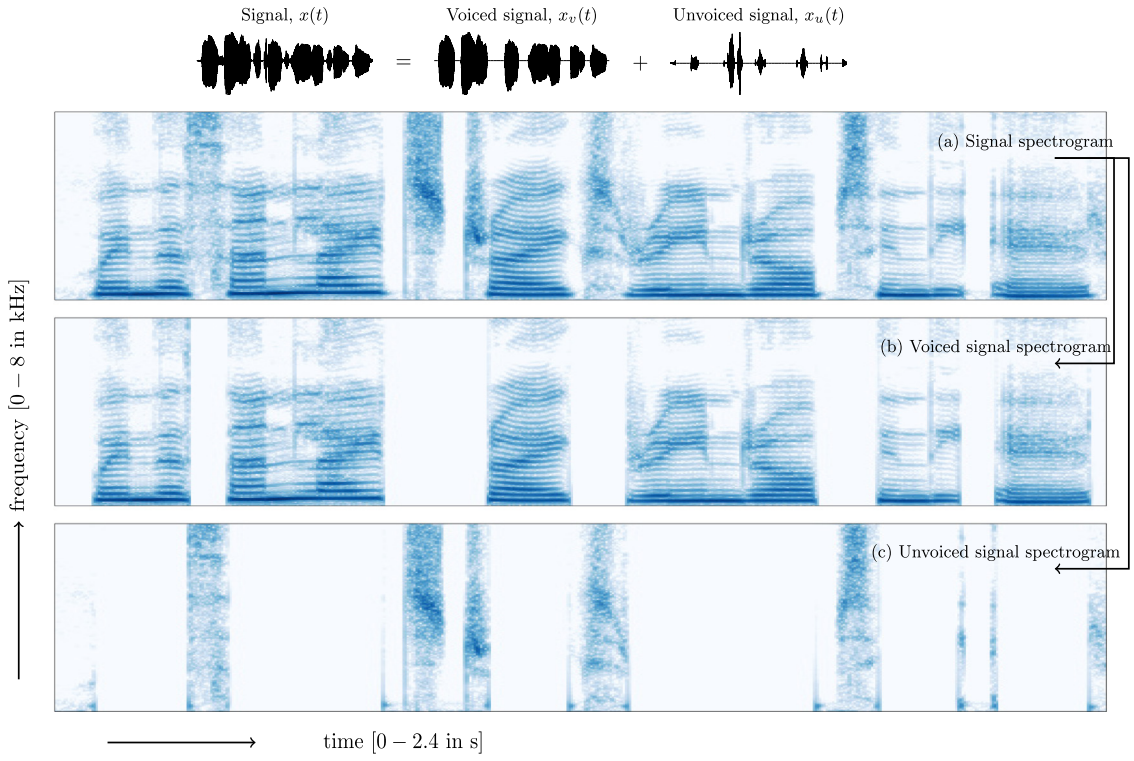


Fig. 1. Illustration of speech signal $x(t)$ as composed of a sum of voiced signal stream $x_v(t)$ and an unvoiced signal stream $x_u(t)$. The distinct spectro-temporal features in these streams can be observed in the spectrograms (of 10 ms hanning windowed short-time segments with 66% overlap). The utterance corresponds to "*When the sunlight strikes raindrops in the air*", spoken by a female speaker and lowpass filtered to 8 kHz.

l by relating the $\phi_k(t)$ as follows.

$$\phi_k(t) = k\phi_1(t) + \delta_k(t),\ k \geq 2. \tag{2}$$

where $\delta_k(t)$ models the deviations in harmonicity amongst the sinusoids in $x_v(t)$. As these deviations are small, we assume $\delta'_{\max} << \phi'_1(t)$ where, $\delta'_{\max} = \max_k \max_t |\delta'_k(t)|$ (here $\delta_1(t) = 0$). In a nutshell, Eq. (1) suggests a representation of speech using $N+1$ time-varying sinusoids with the following features. Firstly, the representation is a parsimonious model of the extremely rich time-frequency energy distribution of speech. Secondly, the temporal evolution of the short-time spectrum is captured in modulations of the individual sinusoids. This provides a means to capture the nonlinear processes in speech production (Teager and Teager, 1990), example, correlation between amplitude and frequency variations. Thirdly, there is an emerging evidence on selective analysis of harmonic complex by the auditory cortex (Wang, 2013). In this regard, explicit modeling of the time-varying sinusoids in the voiced stream provides a perceptually meaningful representation (Micheyl and Oxenham, 2010). The crux of the paper lies in obtaining this proposed representation without resorting to any short-time modeling.

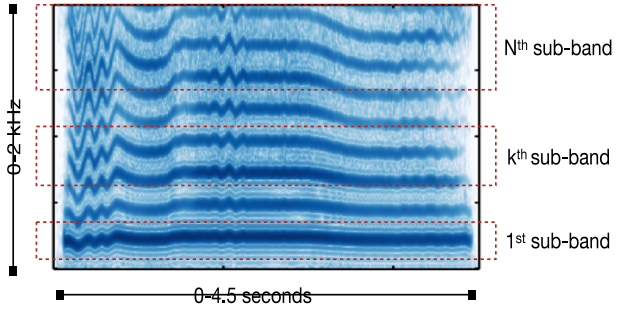


Fig. 2. [In color] The spectrogram depicts time-varying quasi-harmonic complex in a melodic vocal (*a Hindustani Alap*, opening vocals in typical North Indian classical singing performance). The sinusoids often "move in and out" of the subbands associated with a pre-defined filter bank. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1.1. Prior art

The proposed representation is essentially a specific kind of subband modeling of speech. Majority of subband modeling approaches use a pre-defined filter bank (example, gaussian and gammatone filter banks (Potamianos and Maragos, 1999; Drullman et al., 1994; Kumaresan and Rao, 1999). A popular choice for distribution of subband bandwidths is to partition the fullband signal into uniform mel-bandwidths (Zwicker et al., 1957; Patterson, 1976). However, as also discussed by Ghitza Ghitza (2001), the modulations captured in these subband signals also inherit partition induced artefacts. For example, as shown in Fig. 2, we can potentially have sinusoids "moving in and out" of the subbands. This will result in abrupt variations in IA and IP of the subband signal - reflecting the property of the filter bank and not the speech signal. Further, in mel-scale partitioning, the high frequency subbands (>1.5 kHz) have bandwidths which potentially can accommodate more than one sinusoid. This leads to poor modulation characterization in the higher subbands. Identifying this, Rao and Kumaresan (2000) propose using a bank of time-varying filters, centered around dominant energy regions (or the formants) in the time-frequency plane. Schimmel and Atlas (2008) propose centering the filters around the fundamental frequency and its harmonics. Both these proposals suggest that analysis benefits can be drawn via adapting the subband partitioning with the temporal evolution of the short-time segments.

An alternate approach to subband modeling is using a sequential decomposition of the signals into sinusoids, without any explicit use of filter bank. A popular such approach is empirical mode decomposition

(EMD) (Huang et al., 1998). However, its applicability to speech modeling is found to be limited due to the large number of harmonics (Sharma and Prasanna, 2016). Alternatively, Gianfelici et al. (2007) propose an iterative decomposition of the Hilbert envelope of speech to obtain a simplified time-varying representation. Although the approach is mathematically elegant however, the signals obtained from the decomposition have not provided insight into modeling of speech production or perception attributes. Exploiting the fact that a subband decomposition of a wideband signal is non-unique, probabilistic inference (Turner and Sahani, 2011) and convex optimization (Sell and Slaney, 2010) are also proposed in the literature but the limitations are similar.

Sinusoidal modeling via short-time analysis has also been applied to model speech. The well-known approaches include harmonic parameter estimation from short-time spectrum (Almeida and Tribolet, 1983; Marques and Almeida, 1989; Laroche et al., 1993; Serra, 1989; Christensen et al., 2002; Schimmel and Atlas, 2008; Zubrycki and Petrovsky, 2007; 2010; Petrovsky et al., 2011; Petrovsky and Azarov, 2014), peak tracking in short-time spectrum (McAulay and Quatieri, 1986), matching pursuit Gribonval and Bacry (2003), elliptic filters (Kim et al., 2006), and parametrized model fitting on short-time segments (Pantazis et al., 2011; Degottex and Stylianou, 2013; Caetano et al., 2016). However, the estimation of the sinusoidal parameters using these methods is sensitive to the type of short-time window used and its duration (Goodwin, 1998; Caetano et al., 2016). Recently it has been found that adapting the window duration based on the time-varying fundamental frequency is beneficial (Nørholm et al., 2016). The approaches to model unvoiced speech stream, lacking harmonic structure, include using: white noise to excite the short-time spectral envelope estimated in unvoiced regions (Serra, 1989), dense sampling of spectral peaks (McAulay and Quatieri, 1986), iterative temporal magnitude envelope sinusoidal modeling (Shechtman, 2013), overlap-add phase randomization Macon and Clements (1997), and Fan-Chirp transform for short-time parametric modeling (Kafentzis and Stylianou, 2016). In summary, this suggests that the problem of sinusoidal modeling of speech is being actively pursued using varied kinds of approaches.

1.2. Contributions

This paper re-visits sinusoidal modeling (and subband modeling) of speech with two distinctions: (*i*) operates without the use of filter bank, and (*ii*) it does not require any short-time analysis. As a result, the estimated sinusoids better capture the inherent modulations (which was an issue of concern in prior art associated with subband modeling), and require no interpolation of signal estimates (and no overlap-add) as required in short-time modeling approaches.

A block diagram of the proposed methodology is shown in Fig. 3. Firstly, the IF associated with the fundamental frequency sinusoid (FFS) in speech is estimated by flipping the spectrum, successive differentiation, and following by analytic signal estimation (Gabor, 1946). We present this approach as the Flip-Diff algorithm (see Section 3). Secondly, using the IF estimate of the FFS, an approach is designed to extract the voiced stream and subsequently estimate the quasi-harmonic sinsuoids contained in it. We present the estimation of the quasi-harmonic sinusoids as the quasi-harmonic demodulation (QHD) algorithm (see Section 4). The QHD makes use of in-phase and quadrature phase (IQ) demodulation (Haykin, 1994), a technique used first in radio frequency communication. We implement a stage-wise variant of this with time-varying carrier frequency, and for the first time show its application in speech modeling. Thirdly, the unvoiced steam is extracted from speech and is modeled as a mono-component wideband sinusoid. Combinedly, the proposed methodology provides a synchronized analysis of speech, carried at varied time-scales, corresponding to the instantaneous time-periods of the constituent sinusoids. A detailed experimentation with impact of model parameters on the quality is carried out by synthesizing the speech from the estimated model (see Section 5).

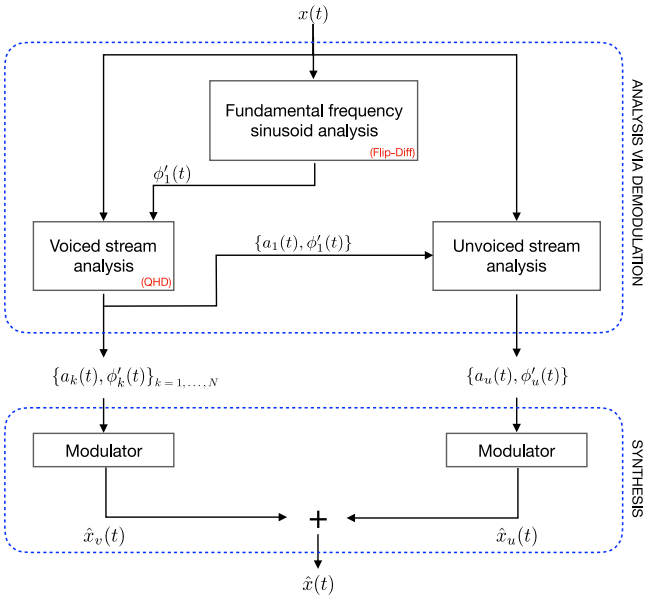


Fig. 3. Block diagram of proposed methodology for speech analysis-synthesis.

2. Notations

Let $y(t)$ denote the voiced stream $x_v(t)$ in Eq. (1), that is, $y(t) = \sum_{k=1}^{N} a_k(t) \sin \phi_k(t)$. We will denote the bandwidths of $a_k(t)$ and $\phi_k'(t)$ by $\mathcal{B}_{IA,k}$ and $\mathcal{B}_{IF,k}$, respectively. The rate of temporal evolution of IA and IF in a *pseudo*-time-period $T_k(t) := 1/2\pi\phi_k'(t)$ can be expressed as Flandrin (2001),

$$\kappa_{1,k}(t) = \frac{a_k'(t)}{a_k(t)\{\phi_k'(t)\}}, \; \kappa_{2,k}(t) = \frac{\phi_k''(t)}{\{\phi_k'(t)\}^2}. \tag{3}$$

The upper bound on $\kappa_{1,k}(t)$ and $\kappa_{2,k}(t)$ is dependent on the bandwidths and modulation indices of IA and IF (Sharma and Sreenivas, 2015). For simplifying the mathematical modeling of $y(t)$, we will make following three assumptions. Firstly, $a_k(t) > 0$ (Rice, 1982). Secondly, we will assume $\kappa_{1,k}(t)$ and $\kappa_{2,k}(t)$ are smaller than 1, implying that the IA and IF do not vary fast within $T_k(t)$. Thirdly, as these sinusoids are quasi-harmonically related hence, $\phi_k'(t) < \phi_{k+1}'(t)$ holds, that is, at any instant the sinusoids do not crisscross each other in the time-frequency plane.

3. Fundamental frequency sinusoid analysis

The lowest frequency sinusoid in $y(t)$ is our definition for fundamental frequency sinusoid (FFS). Here, the FFS is $y_1(t) = a_1(t)\sin\phi_1(t)$. $y(t)$ is a multi-component signal (Boashash, 1992). As shown in Wei and Bovik (1998), an amplitude dominance in one of the sinusoids in $y(t)$ can potentially make the analytic signal IF estimate converge to that of the amplitude dominating sinusoid. We use this concept for FFS analysis, and design an approach to boost the strength of FFS and subsequently estimate its IF.

Let $y_f(t)$ denote the signal obtained by flipping the spectrum of $y(t)$ about DC (zero frequency). This flipping can be done by modulating $y(t)$ with $\cos 2\pi f_c t$, where, $f_c = f_{\max}$ (the maximum frequency contained in $y(t)$) and lowpass filtering the output to f_c. The first sinusoid in $y(t)$ will be the last sinusoid in $y_f(t)$. Thus,

$$y_f(t) = \sum_{k=1}^{N} b_k(t) \sin \psi_k(t) \tag{4}$$

where $b_k(t) = a_{N+1-k}(t)$, and $\psi_k(t) = 2\pi f_c t - \phi_{N+1-k}(t)$. It can be noted that flipping the spectrum does not impact the IAs of the sinusoids, and the IFs of the sinusoids undergoes a linear shift (an invertible operation). Let $y_{f,k}(t) = b_k(t)\sin\psi_k(t)$ denote the kth sinusoid in $y_f(t)$. The pth

derivative of $y_{f,k}(t)$ can be expressed as follows.

$$y^{(p)}_{f,k}(t) := \frac{\mathrm{d}^p\{b_k(t)\sin\psi_k(t)\}}{\mathrm{d}t^p} = b_k(t)\{\psi'_k(t)\}^p \sin\left(\psi_k(t)+\frac{p\pi}{2}\right)+\epsilon_{k,p}(t) \quad (5)$$

$$\approx b_k(t)\{\psi'_k(t)\}^p \sin\left(\psi_k(t)+\frac{p\pi}{2}\right) \quad (6)$$

where $\epsilon_{k,p}(t)$ contains terms associated with derivatives of $b_k(t)$ and $\psi'_k(t)$. The goodness of the approximation in Eq. (6) is dependent on magnitude of $\epsilon_{k,p}(t)$ relative to $b_k(t)\{\psi'_k(t)\}^p$. For sinusoids with $\kappa_{1,k}(t)$, $\kappa_{2,k}(t)<1$, we have found $\epsilon_{k,p}(t)/b_k(t)\{\psi'_k(t)\}^p$ to be insignificant for subsequent analysis. The derivative of $y_f(t)$ is a linear summation of derivative of each sinusoid in $y_f(t)$, that is,

$$\begin{aligned} y^{(p)}_f(t) &:= \frac{\mathrm{d}^p y_f(t)}{\mathrm{d}t^p} = \sum_{k=1}^{N} y^{(p)}_{f,k}(t) \\ &= \sum_{k=1}^{N} b_k(t)\{\psi'_k(t)\}^p \sin\left(\psi_k(t)+\frac{p\pi}{2}\right) \\ &= b_N(t)\{\psi'_N(t)\}^p \sin\left(\psi_N(t)+\frac{p\pi}{2}\right) \\ &\quad + \sum_{k=1}^{N-1} b_k(t)\{\psi'_k(t)\}^p \sin\left(\psi_k(t)+\frac{p\pi}{2}\right). \end{aligned} \quad (7)$$

Now, as $\psi'_N(t) > \psi'_k(t)$ for $k = 1,\ldots,N-1$ hence, beyond some $p > P$ (an integer) we will have $\left\{\frac{\psi'_k(t)}{\psi'_N(t)}\right\}^p << 1$. Thus, there will always exist a P such that irrespective of $b_k(t)$s we will have an amplitude dominance of the Nth sinusoid in $y^{(p)}_f(t)$. That is.

$$y^{(p)}_f(t) = \underbrace{b_N(t)\{\psi'_N(t)\}^p}_{A(t)} \sin\underbrace{\left(\psi_N(t)+\frac{p\pi}{2}\right)}_{\Psi(t)} + \epsilon(t) \quad (8)$$

$$= A(t)\sin\Psi(t)+\epsilon(t). \quad (9)$$

where $\epsilon(t)$ contains the terms corresponding to derivatives of the rest of the $N-1$ sinusoids. To further attenuate the interference in $y^{(p)}_f(t)$ from $\epsilon(t)$ we apply empirical mode decomposition (EMD Rilling et al., 2003). The strength of the Nth sinusoid is amplified over successive derivatives. This benefits EMD because mode mixing artifact in extracting $A(t)\sin\psi(t)$ as the first intrinsic mode function (IMF) will be minimal. Let, the first IMF thus obtained be denoted by $z_1(t)$. Estimating $\psi(t)$ from $z_1(t)$ is now straightforward using the analytic signal approach. That is,

$$z_a(t) := z_1(t) + j\mathcal{H}[z_1(t)], \quad (10)$$

$$\hat{\Psi}(t) = \angle z_a(t),\ \hat{\Psi}'(t) = \frac{\mathrm{d}\hat{\Psi}(t)}{\mathrm{d}t} \quad (11)$$

where $\mathcal{H}[.]$ denotes the Hilbert transform and $\hat{\Psi}(t)$ denotes estimate of $\Psi(t)$. Subsequently, $\phi'_1(t)$ is estimated as $\hat{\phi}'_1(t) = 2\pi f_c - \hat{\Psi}'(t)$.

An illustration of the approach is provided in Fig. 4. We will refer to this by Flip-Diff (that is, Flip and Differentiate) method for estimating IF associated with the FFS (the algorithm is provided in Algorithm 1). It should be noted that $p > P$ is desired for accurate estimation of $\phi'_1(t)$. P will depend on $\beta_1 = \max_k\{\max_t\{\phi'_k(t)/\phi'_1(t)\}\}$ and $\beta_2 = \max_k\{\max_t\{a_k(t)/a_1(t)\}\}$, $k = 2,\ldots,N$. Small $\beta_1\times\beta_2$ will imply a small P. For harmonic sinusoids, β_1 is 0.5 Interestingly, for speech signals there is a spectral roll-off in magnitude spectrum, as well. Hence, β_2 is dependent mainly on the relative amplitude weighting across first few harmonics. From our empirical observations, we have found that $P = 7$ is sufficient to make the FFS dominant in clean speech signals.

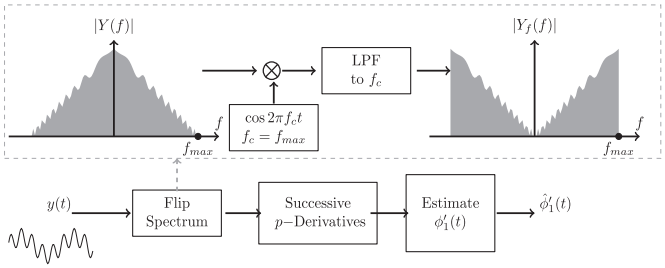


Fig. 4. Illustration of the proposed approach Flip-DIff for estimating the IF associated with the fundemental frequency sinusoid.

Algorithm 1 Flip-Diff: Instantaneous fundamental frequency estimation.

Required: $\mathcal{D} = \{y(t), f_c, p, \mathcal{B}_{IF,1}\}$

function FFS($\mathcal{D}$)
- $h_{IF}(t) \leftarrow$ Lowpass filter with cut-off $\mathcal{B}_{IF,1}$
- $y_f(t) \leftarrow$ SSB-SC[$y(t)$] with carrier f_c
- $y^{(p)}_f(t) \leftarrow$ Take p^{th} derivative of $y_f(t)$
- $z_1(t) \leftarrow$ First mode from EMD[$y^{(p)}_f(t)$]
- $z_a(t) \leftarrow$ Analytic signal of $z_1(t)$
- $\hat{\Psi}'(t) \leftarrow \frac{\mathrm{d}\angle z_a}{\mathrm{d}t}$
- $\hat{\phi}'_1(t) \leftarrow 2\pi f_c - \hat{\Psi}'(t)$
- $\hat{\phi}'_1(t) \leftarrow h_{IF}(t) * \hat{\phi}'_1(t)$ ▷ Smoothing
- **return** $\hat{\phi}'_1(t)$ ▷ Output

end function

4. Quasi-harmonic demodulation

In this section, we aim at estimating the rest of the $N-1$ sinusoids in $y(t)$. As $y(t)$ models $x_v(t)$, there is a separation of at least $\phi'_1(t) - 2\delta'_{\max}$ between any two sinusoids in $y(t)$. Basing on this an initial estimate of the IF associated with the mth sinusoid can be expressed as $\hat{\phi}_m(t) = m\hat{\phi}_1(t)$. To simplify, $\hat{\phi}_m(t)$ can be expressed as $\hat{\phi}_m(t) = \phi_m(t) + \eta(t)$ where, $\eta(t)$ is the error in the estimate. Below we design an approach to reduce the error $\eta(t)$ and iteratively refine the estimate. This provides for estimating the IF and IA associsted with each of the sinusoid in $y(t)$.

A coherent modulation of $y(t)$ using in-phase and quadrature phase signals obtained with $\hat{\phi}_m(t)$ as the carrier estimate can be expressed as follows.

$$y_I(t) := y(t)\sin\hat{\phi}_m(t) = \{a_m(t)\cos\eta(t) + z_I(t)\}/2 \quad (12)$$

$$y_Q(t) := y(t)\cos\hat{\phi}_m(t) = \{-a_m(t)\sin\eta(t) + z_Q(t)\}/2 \quad (13)$$

where,

$$\begin{aligned} z_I(t) &= -a_m(t)\cos[\phi_m(t)+\hat{\phi}_m(t)] \\ &\quad + \sum_{k=1,\neq m}^{N} a_k(t)\{\cos[\phi_k(t)-\hat{\phi}_m(t)] - \cos[\phi_k(t)+\hat{\phi}_m(t)]\} \\ z_Q(t) &= a_m(t)\sin[\phi_m(t)+\hat{\phi}_m(t)] \\ &\quad + \sum_{k=1,\neq m}^{N} a_k(t)\{\sin[\phi_k(t)-\hat{\phi}_m(t)] + \sin[\phi_k(t)+\hat{\phi}_m(t)]\}. \end{aligned}$$

We assume $\phi'_1(t) > (\delta'_{\max} + |\eta'(t)|)$, that is $m\hat{\phi}_1(t)$ is a good initial estimate of $\phi_m(t)$. Then, (*i*) $\frac{a_m(t)}{2}\cos\eta(t)$ and $\frac{a_m(t)}{2}\sin\eta(t)$ serve as baseband signals, and (*ii*) spectrum of $y_{I,m}(t)$ and $y_{Q,m}(t)$ lies beyond these baseband signals. This allows isolating $\frac{a_m(t)}{2}\cos\eta(t)$ and $\frac{a_m(t)}{2}\sin\eta(t)$ using a lowpass filter. Conventional IQ demodulation can now be carried out using ideal

lowpass filter $h_{IA}(t)$ choosen such that $y_I(t) * h_{IA}(t) = a_m(t)\cos\eta(t)$ and $y_Q(t) * h_{IA}(t) = a_m(t)\sin\eta(t)$ ($*$ denotes convolution). This will provide a one step estimation of $a_m(t)$ and $\eta(t)$ using Eqs. (14) and (15), respectively, as follows.

$$\hat{a}_m(t) = \sqrt{\{y_I(t) * h_{IA}(t)\}^2 + \{y_Q(t) * h_{IA}(t)\}^2} \tag{14}$$

$$\hat{\eta}(t) = \arctan \frac{y_Q(t) * h_{IA}(t)}{y_I(t) * h_{IA}(t)}. \tag{15}$$

However, designing such a filter is non-trivial because of the time-varying nature of the IF and the error signal. Instead, practically it can be assumed that the modulation bandwidth of IA is restricted within certain range, below $\phi_1'(t)$. We design $h_{IA}(t)$ such that its lowpass cut-off frequency f_l is $f_l \leq \phi'(t)/2\pi$. Using $h_{IA}(t)$ we lowpass filter $y_I(t)$ and $y_Q(t)$, and obtain estimates of $a_m(t)$ and $\alpha(t)$, as stated in Eqs. (14) and (15), respectively. Next, using $\hat{\eta}(t)$ the initial carrier estimate $\hat{\phi}_m(t)$ can be updated to $\hat{\phi}_m(t) = \hat{\phi}_m(t) - \hat{\eta}(t)$. The IQ demodulation of $y(t)$ can then be repeated using this new carrier estimate for $\phi_m(t)$. Interestingly, from our observations we find that the IA and IF estimates obtained via this iterative-IQ demodulation carried with refined carrier estimates are significantly close to the ground truth. The convergence is found to occur provided in the first iteration the bandwidth of $h_{IA}(t)$ captures part of a spectrum of $a_m(t)\cos\eta(t)$. The proof for this is beyond the scope of this paper.

The above approach to demodulate the mth sinusoid in $y(t)$ holds for any $m \in \{1, \ldots, N\}$. Each sinusoid can be demodulated independently (hence, in parallel). We will refer to this iterative-IQ demodulation approach by quasi-harmonic demodulation (QHD). It should be noted that the refined carrier estimate contains the inharmonicity factor $\delta_m(t)$. The QHD is summarized in Algorithm 2. Note that, the algorithm uses filters

Algorithm 2 QHD: Quasi-harmonic demodulation (QHD).

Required: $D = \{y(t), \hat{\phi}_1'(t), \mathcal{B}_{IA}, \mathcal{B}_{IF}, N, \text{ and }, K\}$

function QHD(D)
  $h_{IA}(t) \leftarrow$ Lowpass filter with cut-off $\mathcal{B}_{IA}$
  $h_{IF}(t) \leftarrow$ Lowpass filter with cut-off $\mathcal{B}_{IF}$
  $\hat{\phi}_1(t) \leftarrow$ Integrate $\hat{\phi}_1'(t)$
  for $m \leftarrow 1$ to N **do** ▷ Analyze each sinusoid
    $\hat{\phi}_m(t) \leftarrow m\hat{\phi}_1(t)$
    $\hat{\eta}(t) \leftarrow 0$
    for $k \leftarrow 1$ to K **do** ▷ Stage-wise IQ
      $\hat{\phi}_m(t) \leftarrow \hat{\phi}_m(t) + \hat{\eta}(t)$ ▷ Error Correction
      $\hat{\phi}_m(t) \leftarrow h_{IF}(t) * \hat{\phi}_m(t)$ ▷ Smoothing
      $\{\hat{a}_m(t), \hat{\eta}(t)\} \leftarrow$ IQ Demodulation of $y(t)$ using $\hat{\phi}_m(t)$ and $h_{IA}(t)$
    end for
    $\hat{\phi}_m'(t) \leftarrow$ Differentiate $\hat{\phi}_m(t)$
  end for
  return $\hat{a}_m(t),\ \hat{\phi}_m'(t), \forall m \in \{1, \ldots, N\}$ ▷ Output
end function

$h_{IA}(t)$ and $h_{IF}(t)$. Alongside providing smoothing of the estimates, these also allow to limit the bandwidth of the IA and IF signals. We will experiment the impact of these on speech signals in Section 5. Next, we illustrate the application of Flip-Diff and QHD on synthetic signals.

4.1. Illustration on synthetic signals

Consider $y(t)$ as a multi-component time-varying sinusoid, with $N = 10$ sinusoids. The δ_k's, indicting deviation from harmonicity, are chosen randomly to lie between 0.1 and 30 Hz. $a_k(t)$, $\forall k \in \{1, 2, \ldots 10\}$, are chosen arbitrarily with spectral support between 0 and 20 Hz. $\phi_1'(t)$ lies between 70 and 150 Hz, and the variation is arbitrary with a spectral support of $0 - 20$ Hz. With these settings, we have $\delta'_{\max} = 30$ Hz and $\min_t \phi_1'(t) = 70$ Hz. Fig. 5(e,f) shows the $\phi_k'(t)$ and $a_k(t)$ associated with each sinusoid. The signal duration is 1 s and the sampling rate is 16 kHz. We use Algorithms 1 and 2 with parameters $\{f_c = 1600$ Hz, $p = 5,\ \mathcal{B}_{IA} = \mathcal{B}_{IF} = 30$ Hz, $N = 10,\ K = 2\}$.

Observations: The obtained estimates after two stage IQ demodulation are overlaid on the actual IAs, and IFs in Fig. 5(e,f). The estimates closely follow the corresponding true IAs and IFs. To highlight the benefit of two stage IQ, Fig. 5(a) shows the IF estimates associated with the eighth sinusoid ($m = 8$) in $y(t)$. The estimate after stage-1, denoted by $\hat{\phi}'_{8,1}(t)$ follows the trend in $\phi_8'(t)$ but has a time-varying offset in amplitude. The offset is specifically high at instants corresponding to fast rate of temporal variations in $\phi_8'(t)$. The poor estimate $\hat{\phi}'_{8,1}(t)$ results in an inaccurate IA estimate as seen in Fig. 5(b). The refined IF estimate after stage-1, denoted by $\hat{\phi}'_{8,2}(t)$, has no time-varying offset and overlaps with the true IF (shown in Fig. 5(a)). The improved accuracy of $\hat{\phi}'_{8,2}(t)$ gives significant improvement in IA estimate as well. This is seen in Fig. 5(b) where, $\hat{a}_{8,2}(t)$ (obtained from stage-2) tracks $a_8(t)$ better than $\hat{a}_{8,1}(t)$ (obtained from stage-1). The estimation accuracy can be quantified using signal-reconstruction-error-ratio (SRER) defined as,

$$\text{SRER} = 20\log_{10}\frac{\|s(t)\|_2}{\|s(t) - \hat{s}(t)\|_2} \text{ (in dB)} \tag{16}$$

where, $s(t)$ is the true signal (such as the IA or IF) and $\hat{s}(t)$ is the estimate. The variation in SRER of IF and IA estimates as a function of sinusoid index is shown in Fig. 5(c,d). We can see a significant improvement in SRER after stage-2 IQ.

To further evaluate the improvement in SRER over stages of IQ, we experiment with 100 instances of different signals. The signals are modeled as quasi-harmonic complexes, and the δ_k's and the spectral distribution of the IA (and also IF) for each instance are chosen randomly within the range $0.1 - 30$ Hz and $0 - 20$ Hz, respectively. The obtained SRER over successive stages of IQ is shown in Fig. 6. The SRER improves over the first 3 stages and then it begins to saturate. Further, the SRER (for both IA and IF) is better for first five compared to that of last five sinusoids. A contributor to this is the error in the initial carrier estimate. A small error in the $\phi_1(t)$ will get magnified when used to obtain the initial carrier estimate for the higher frequency sinusoids.

To summarize, QHD improves the initial estimate of IF for each of the sinusoids. Interestingly, this also provides for improving the IF estimate of the FFS obtained from Flip-Diff. Combinedly, the Flip-Deriv and the QHD algorithms allow estimating the IAs and IFs without any short-time windowing and use of filter bank.

5. Analysis-synthesis of speech

In this section, we apply Flip-Diff and QHD to speech signals for obtaining the representation in Eq. (1). We will follow the methodology shown in Fig. 3 Given $x(t)$, first step is anaysis of FFS (Section 5.1). The FFS is used to obtain the voiced ($x_v(t)$) and unvoiced ($x_u(t)$) streams. The voiced stream is analyzed using QHD (Section 5.2) and the unvoiced stream is analyzed using mono-component sinusoidal modeling (Section 5.3). Although the representation in Eq. (1) is non-parametric, the bandwidths of IA and IF impact the quality of estimates. A detailed analysis of this is pursued in these sections using recordings corresponding to a female and a male speaker. Subsequently, in Section 5.4, an evidence for the effectiveness of the representation in capturing inherent modulations in the signal is presented. A comparison of the proposed approach with two short-time harmonic modeling approaches is presented in Section 5.6.

5.1. Estimating the IF and IA associated with the FFS of speech

Being linked with the fundamental frequency variation, the FFS in speech signals will most likely reside in the frequency support 70 – 600 Hz. Depending on the fundamental frequency in the signal, this

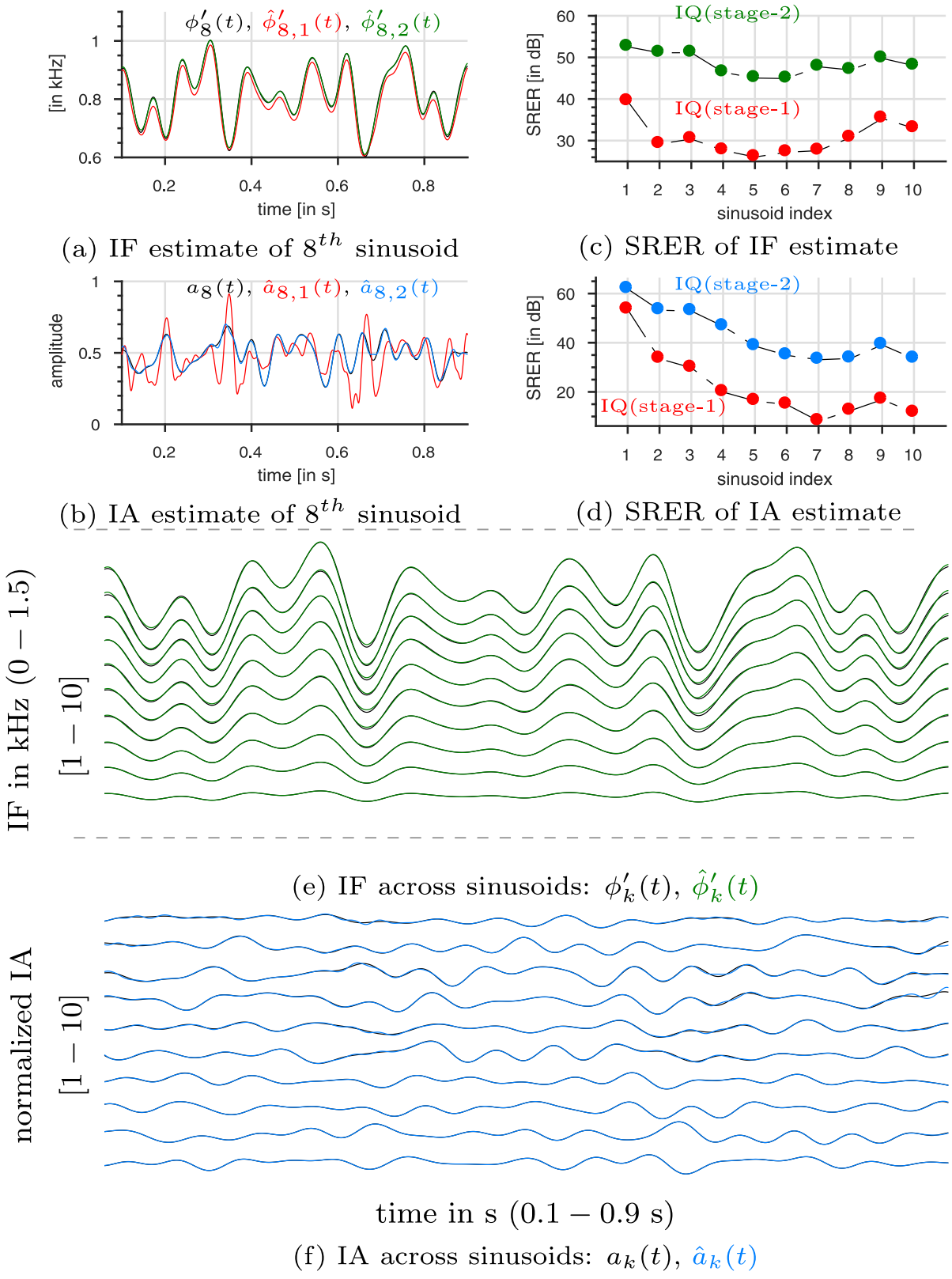


Fig. 5. [In color] Illustration of QHD of a signal composed of sum of ten quasi-harmonically related time-varying sinusoids. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

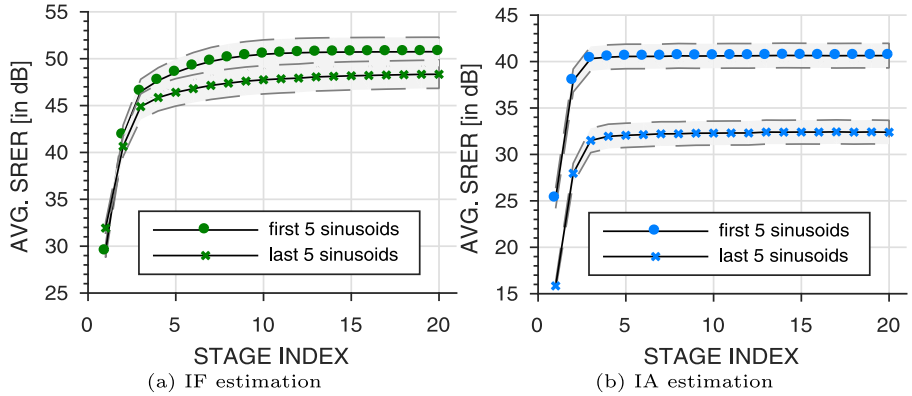


Fig. 6. Illustration of improvement in estimation SRER over successive stages of IQ refinement. The shown avg. SNR is computed over 100 instances of $y(t)$, a quasi-harmonic complex composed of ten sinusoids. The avg. SNR is further averaged separately on first five (from lower to higher IFs) and last five sinusoids. The shaded regions indicated 95% confidence interval.

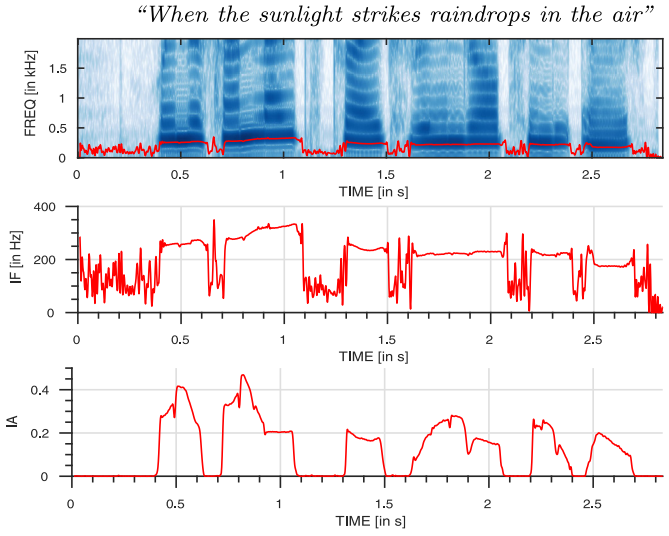


Fig. 7. [In color] Application of Flip-Diff on a speech signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

frequency support will at least have the FFS in it, and hence, suffices for FFS analysis using Flip-Diff.

Consider a female speech signal shown in Fig. 7. We input this signal to the Flip-Diff algorithm with parameters chosen as $\{f_c = 600$ Hz, $p = 3$, $\mathcal{B}_{IA,1} = \mathcal{B}_{IF,1} = 100$ Hz$\}$. The obtained IF estimate $\hat{\phi}'_1(t)$ is shown in Fig. 7. There is no ground truth for confirming the accuracy of the estimate however, we can still make some observations, and these will be strengthened when we attempt analysis-synthesis in Section 5.2.

Observations: Comparing with time-frequency trajectories in the spectrogram (in Fig. 7), we can see that the overlaid $\hat{\phi}'_1(t)$ follows the temporal variations in the frequency trajectory of the first sinusoid in voiced regions. As FFS is associated with voicing, $\hat{\phi}'_1(t)$ is erratic in the unvoiced regions. We also obtained the IA estimate of the FFS using one stage of IQ demodulation. The estimate $\hat{a}_1(t)$ can be interpreted as capturing the strength of the FFS, and this is low in unvoiced regions. By choosing an appropriate threshold (denoting by ν), $\hat{a}_1(t)$ can be used to segregate the voiced and unvoiced signal streams from $x(t)$. As an example, Fig. 1 shows a voiced/unvoiced stream segregation obtained using FFS of $x(t)$. The ν was set to $0.1 \max[\hat{a}_1(t)]$ (chosen empirically), and an instant qualifies as voiced if $\hat{a}_1(t) > \nu$, else it is considered as unvoiced. The spectrograms corresponding to $x_v(t)$, and $x_u(t)$ show that segments containing the quasi-harmonic structure are captured in the voiced stream.

Coming back to Fig. 7, the Pearson correlation co-efficient between $\hat{a}_1(t)$ and $\hat{\phi}'_1(t)$ (considering only voiced regions) was found to be 0.44. This is interesting, as it implies temporal variation in $\phi'_1(t)$ (capturing intonations) reflect in a correlated temporal variation in the strength of the FFS.

5.2. QHD Of voiced signal stream

Consider a voiced utterance - *"I owe you a yo yo."*, spoken by a female speaker (referred by Subject-1, average pitch of 220 Hz), sampled at $F_s = 8$ kHz (shown in Fig. 8(a,f)). This can practically be modeled as an only voiced stream, $x_v(t) = \sum_{k=1}^{N} a_k(t) \sin \phi_k(t)$. The goal is to estimate $a_k(t)$ and $\phi_k(t)$ for all $k \in \{1, \ldots, N\}$ and evaluate the accuracy via synthesis. N is chosen such that: $N \min_t \phi'_1(t) < F_s/2$. To avoid aliasing during estimation, the signal is re-sampled to F_{rs} such that: $N \max_t \phi'_1(t) < F_{rs}/2$. In experiments in this paper, $F_s = 8$ kHz, and $F_{rs} = 16$ kHz, and post synthesis the signal is re-sampled back to F_s. The estimation accuracy is evaluated by synthesizing $x_v(t)$ from the estimates, that is[2],

$$\hat{x}_v(t) = \sum_{k=1}^{N} \hat{a}_k(t) \sin \hat{\phi}_k(t) \tag{17}$$

and comparing it with $x_v(t)$. The objective measures used for comparison are: (*i*) average frequency weighted signal-to-noise ratio (avg. FW-SNR) (Hu and Loizou, 2008), and (*ii*) perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) score. FW-SNR (higher the better) is analogous to SNR but the spectrum distance (on short-time segments of 25 ms) is computed on a bark-scale. This makes it a perceptually correlated measure. The PESQ (ranging between 0 and 4.5, higher the better) is an ITU-T standard for automated measurement of speech quality, as experienced by a telephony system user.

Analysis: We apply the QHD algorithm on $x_v(t)$. The algorithm parameters are chosen as: $\mathcal{B}_{IA} = \mathcal{B}_{IF} = 100$ Hz, and $p = 7$. The IA, and IF estimates obtained are shown in Fig. 8(b,c) (the IF estimate of the FFS is obtained using Flip-Diff). The temporal glides in the IF estimate indicate intonations introduced by the speaker. Interestingly, IF and IA estimates of stage-1, and stage-2 IQ are similar. This implies that, rate of variation in the underlying $\phi_1(t)$ is small hence, and $\hat{\phi}'_1(t)$ tracks it accurately (or, $\eta(t)$, introduced in Section 4, is small). The energy of IA estimates (in Fig. 8(d)) depicts a roll-off in magnitude as we move from lower frequency to higher frequency sinusoids. The correlation matrix obtained by computing Pearson correlation between the IA estimates across sinusoid is shown in Fig. 8(e). Not all IAs shows high correlation however, IAs corresponding to $\{k_1, k_2\} = \{6, 7\}$, $\{8, 9\}$, $\{10, 11\}$ have a correlation

[2] The IP estimates are obtained from the IF estimates by intergration.

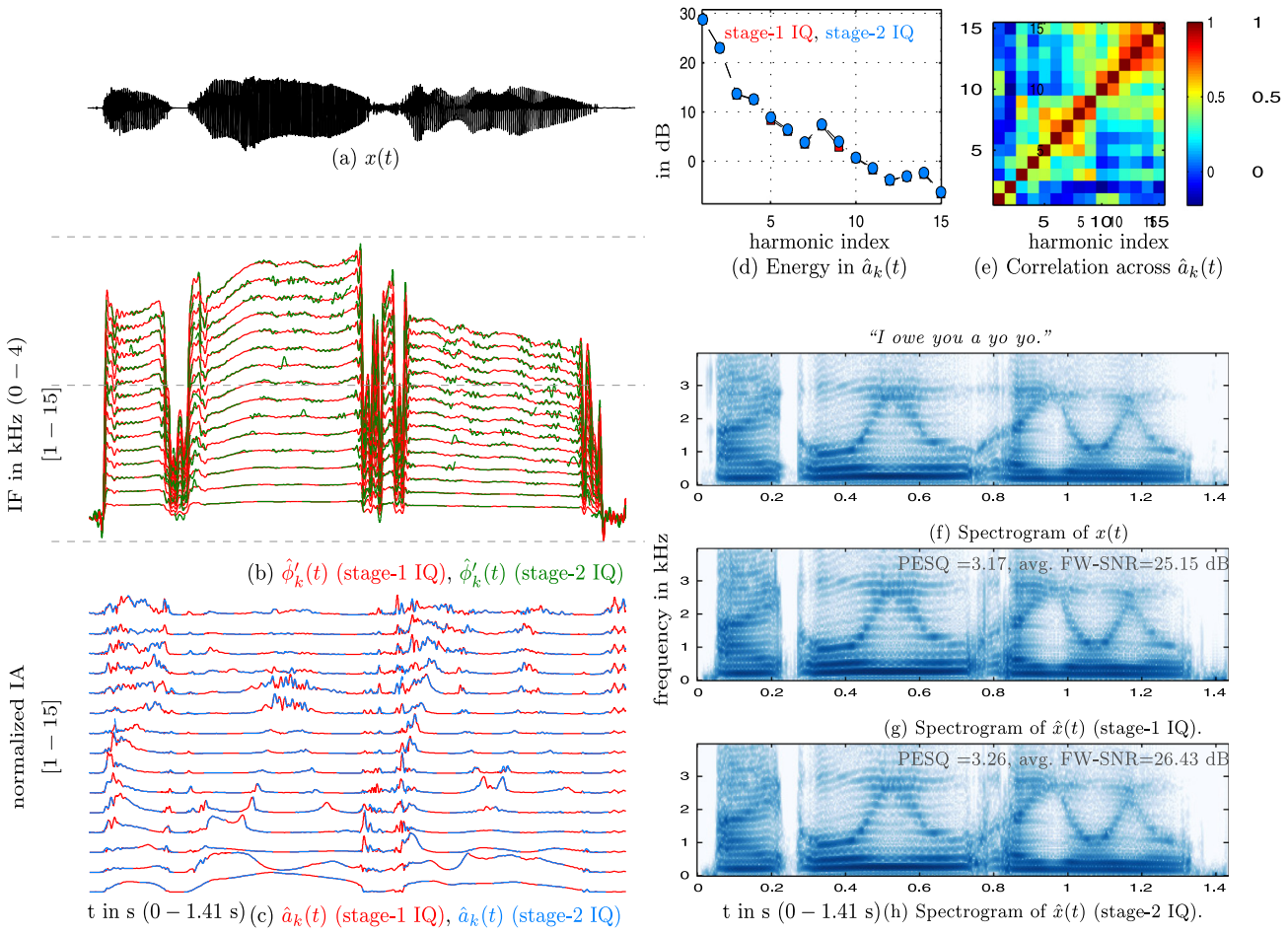


Fig. 8. [In color] Illustration of analysis-synthesis using QHD on a voiced speech signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

more than 0.6. This is likely because the corresponding sinusoids fall under same time-varying formant bandwidths.

Synthesis: The IA and IF estimate are plugged into Eq. (17) to obtain $\hat{x}_v(t)$. The narrowband (10 ms short-time segment) spectrograms of $x_v(t)$ obtained from stage-1, and stage-2 estimates are shown in Fig. 8(g,h). Both preserve the spectro-temporal structures present in the original signal spectrogram (shown in Fig. 8(f)). There is a (local) spectro-temporal mismatch around 0.8 s in the spectrogram. Owing to the low strength of voicing around this instant, the IF estimate is erratic around this instant and this locally impacts the IA estimate around this instant, as well. The objective measures (indicated in the plot) gave similar value for synthesis from stage-1 and stage-2, that is FW-SNR > 25 dB and PESQ > 3. A listening test (conducted using a GUI providing a slider for rating) for quality comparison between reconstructed and original signal in a scale of $0-5$ ($0-1$: Bad, $1-2$: Poor, $2-3$: Good, $3-4$: Fair, and $4-5$: Excellent) was conducted. Six subjects took part and an average score of 3.3 was obtained for stage-1 signal. This implies that the synthesized quality is Fair in naturalness. Also, the intelligibility was intact in the synthesis.

Experimenting with bandwdths of IA and IF: The choice of $\mathcal{B}_{IA}$, and $\mathcal{B}_{IF}$ made during the estimation will impact synthesis quality. Further, a good choice may depend on gender. We explore this aspect, along with the female all voiced utterance used earlier, we also use a male utterance (Subject-2). The average pitch of Subject-1 (female) is 220 Hz and that of Subject-2 (male) is 110 Hz. To be concise, we present results with estimate of $\hat{x}_v(t)$ obtained from stage-1 IQ only.[3] Fig. 9 depicts the

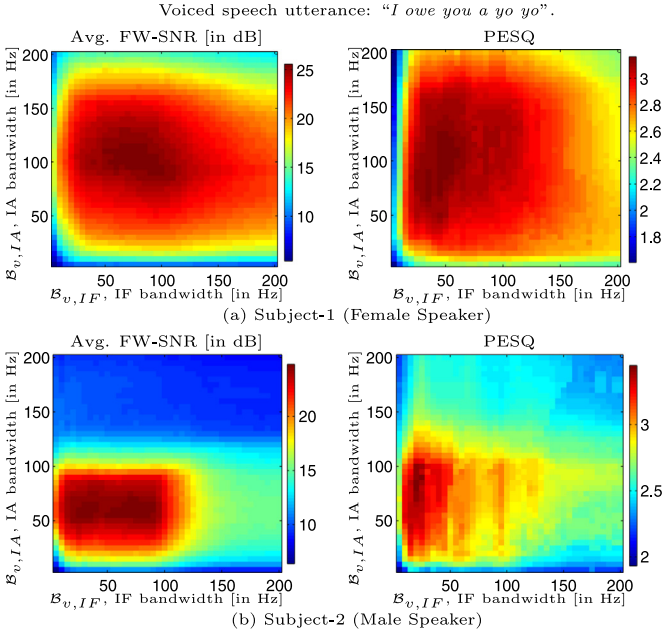


Fig. 9. [In color] Performance evaluation for analysis-synthesis of an all voiced speech utterance. The synthesis is carried for different analysis bandwidths of IA ($\mathcal{B}_{v,IA}$), and IF ($\mathcal{B}_{v,IF}$) choosen from the range $5-100$ Hz. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[3] We found that the update from stage-1 to stage-2 was perceptually insignificant for speech signals analyzed here.

objective measures as a function of $\{\mathcal{B}_{v,IF}, \mathcal{B}_{v,IA}\}$ (the subscript v denotes voiced stream analysis). We see that there exists a region with a good objective score, that is, avg. FW-SNR > 22 dB and PESQ > 3, in all the plots. This region is relatively small for the male speaker when compared to that of the female speaker. Analysing the variation in objective score as a function of $\mathcal{B}_{v,IF}$ we see that, there is a gradual increase and then a gradual decrease. The decrease in objective measure beyond a certain $\mathcal{B}_{v,IF}$ is due to interference in estimation of IF of FFS from the nearby sinusoid in Flip-Diff. Analysing the variation in performance as a function of $\mathcal{B}_{v,IA}$ we see that, the performance is poor for $\mathcal{B}_{v,IA} < 35$ Hz. Beyond this range, the objective measure improves for both the subjects. It starts falling for $\mathcal{B}_{v,IA} > 170$ Hz for Subject-1, and $\mathcal{B}_{v,IA} > 100$ Hz for Subject-2. In hindsight, this seems related to the average fundamental frequency of the two subjects. Opting for $\mathcal{B}_{v,IA}$ greater than the fundamental frequency will introduce interference from adjacent harmonics during QHD. Interestingly, such a difference in frequency support between the two subjects is not evident for IF. Overlapping the avg. FW-SNR performance and PESQ performance plots for each subject, we see that a choice of bandwidths for obtaining good quality analysis-synthesis is $\{\mathcal{B}_{v,IA} = 100$ Hz, $\mathcal{B}_{v,IF} = 100$ Hz$\}$, and $\{\mathcal{B}_{v,IA} = 60$ Hz, $\mathcal{B}_{v,IF} = 40$ Hz$\}$ for Subject-1, and Subject-2, respectively.

5.3. Sinusoid modeling of unvoiced signal stream

We analyze an utterance - *"When the sunlight strikes raindrops in air they act as a prism and form a rainbow."*, spoken by the same two subjects considered in the previous subsection. A voiced/unvoiced stream segregation is carried out using the IA estimate of the FFS. $\nu = 0.05 \max[a_1(t)]$ is chosen for Subject-1, and $\nu = 0.01 \max[a_1(t)]$ is chosen for Subject-2. These threshold values are chosen empirically, and a validation of good segregation is done by listening the streams and visualizing the spectrograms for indication of presence/absence of voicing in respective voiced/unvoiced streams. For illustration, a speech signal segment highlighting the segregation for Subject-1 is shown in Fig. 1. The analysis of $x_v(t)$ is carried out using the QHD algorithm (as in Section 5.2). We choose $\{\mathcal{B}_{v,IA}, \mathcal{B}_{v,IF}\}$ as {100 Hz, 100 Hz} and {60 Hz, 40 Hz} for Subject-1 and Subject-2, respectively. These choices are based on the observations from the previous subsection. We model the unvoiced stream $x_u(t)$ as $a_u(t)\sin\phi_u(t)$; as discussed in Section 1, this is a wideband signal. However, a unique decomposition to $a_u(t)$ and $\sin\phi_u(t)$ can be obtained using the analytic signal approach (Picinbono, 1998). Let, $\hat{a}_u(t)$, and $\hat{\phi}_u(t)$ denote the estimates of $a_u(t)$, and $\phi_u(t)$, respectively.

Experimenting with bandwidths of IA and IF: We experiment with choice of frequency support for $a_u(t)$ (denoted by $\mathcal{B}_{u,IA}$) and $\phi'_u(t)$ (denoted by $\mathcal{B}_{u,IF}$). $x_u(t)$ is synthesized for each frequency support as follows,

$$\hat{x}_u(t) = \hat{a}_u(t)\sin\hat{\phi}_u(t). \tag{18}$$

The full speech signal is synthesized as $\hat{x}(t) = \hat{x}_v(t) + \hat{x}_u(t)$, and this is compared with $x(t)$ using the objective measures. The quality of $\hat{x}_u(t)$ is dependent on choice of $\{\mathcal{B}_{u,IF}, \mathcal{B}_{u,IA}\}$. It should be noted that update in $\{\mathcal{B}_{u,IF}, \mathcal{B}_{u,IA}\}$ does not alter $\hat{x}_v(t)$. Fig. 10 depicts the analysis-synthesis performance obtained based on the objective measures. There exist a region in each plot signifying a good synthesis performance. Unlike for the voiced sentence in Fig. 9, this region is similar for both the subjects. The avg. FW-SNR improves as $\mathcal{B}_{u,IF}$, and $\mathcal{B}_{u,IA}$ are increased. We observe that, to obtain an avg. FW-SNR > 23 dB, we need a high frequency support for both IF and IA, that is, $\mathcal{B}_{u,IF} > 1500$ Hz, and $\mathcal{B}_{u,IA} > 1500$ Hz. Interestingly, PESQ does not show much variation for $\mathcal{B}_{u,IF}$ and $\mathcal{B}_{u,IF}$ beyond 700 Hz. This is likely because the synthesized voiced stream dominates the PESQ measure.

Segment-wise Analysis: A segment-wise (25 msec) analysis of FW-SNR and log-likelihood ratio (LLR) measure for the utterance corresponding to Subject-1 (female) is shown in Fig. 11(c,d). Based on the observations from Figs. 9 and 10, we have chosen $\mathcal{B}_{v,IA} = \mathcal{B}_{v,IF} = 100$ Hz, and $\mathcal{B}_{u,IA} = \mathcal{B}_{u,IF} = 1950$ Hz so as to obtain a good quality analysis-synthesis for this speaker. In voiced segments, the FW-SNR is always above 20 dB and often goes beyond 35 dB. Also, the LLR < 0.05 in these segments. This implies a good match of short-time spectral envelope between the original and synthesized signal. In unvoiced segments, FW-SNR drops and often lies between 10 – 25 dB. The LLR > 1, implying a poor match of spectral envelope for unvoiced segments. A listening test was conducted with protocol same as described in Section 5.2. An average score of 4.2 was obtained with the same six participants. This implied an excellent match in naturalness with the original signal.

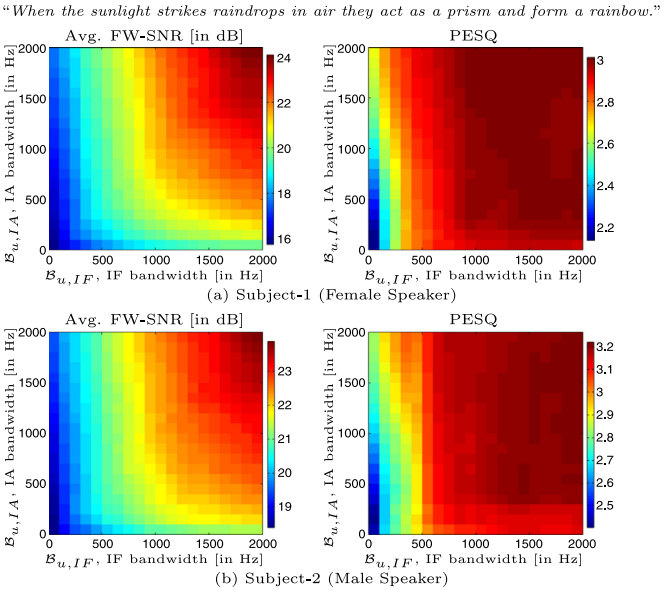


Fig. 10. [In color] Performance evaluation for analysis-synthesis of a speech utterance. The unvoiced stream is synthesized chosing the bandwidths of IA ($\mathcal{B}_{u,IF}$), and IF ($\mathcal{B}_{u,IF}$) in the range 50 – 1950 Hz. The voiced stream is synthesized with a fixed frequency support for IA and IF that is, $\{\mathcal{B}_{v,IA} = 100$ Hz, $\mathcal{B}_{v,IF} = 100$ Hz$\}$ for Subject-1, and $\{\mathcal{B}_{v,IA} = 60$ Hz, $\mathcal{B}_{v,IF} = 40$ Hz$\}$ for Subject-2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Stage-wise analysis: Fig. 12 shows the variation in avg. FW-SNR over successive stages of IQ demodulation in QHD. The avg. FW-SNR increases by 1.7 dB from stage-1 to stage-2, and over subsequent stages it fluctuates within 0.7 dB. The PESQ also fluctuates within 0.8 of the initial value. The low fluctuating but non-monotonic variation in avg. FW-SNR is contrasting to that observed for synthetic signals (see Fig. 6). In the case of speech the harmonics are more in number (example, 15 – 40 within a full signal bandwidth of 4 kHz), and owing to the spectral roll-off the strength of higher harmonics is usually very weak. This adversely impacts the stage-wise update to the higher harmonics.

Can QHD be used independent of Flip-DIff? We experimented with using instantaneous fundamental frequency (F0) estimates derived from quasi-stationary approaches to initialize $\hat{\phi}'_1(t)$ in QHD. For this, we used three different estimators for $\hat{\phi}'_1(t)$, (a) Flat: a constant contour at 100 Hz signal, (b) YIN: using the F0 trajectory estimate obtained using the YIN (de Cheveigné and Kawahara, 2002) algorithm, and (c) STRAIGHT: using the F0 trajectory obtained using the STRAIGHT (Kawahara et al., 1999) algorithm (a state-of-the-art estimator), and compared the analysis-synthesis performance. The speech utterance used is same as used in Fig. 11. The obtained avg. FW-SNR and PESQ scores are shown in Fig. 13. The estimate from FLAT is poor and this reflects in poor performance in stage-1. The YIN operates on 33 ms short-time segments of speech, and the output F0 estimate is a quasi-stationary approximation of the time-varying F0 trajectory. This is much better than FLAT and thereby provides better performance. The STRAIGHT's F0 estimate as finer temporal resolution than YIN (Kawahara et al.,

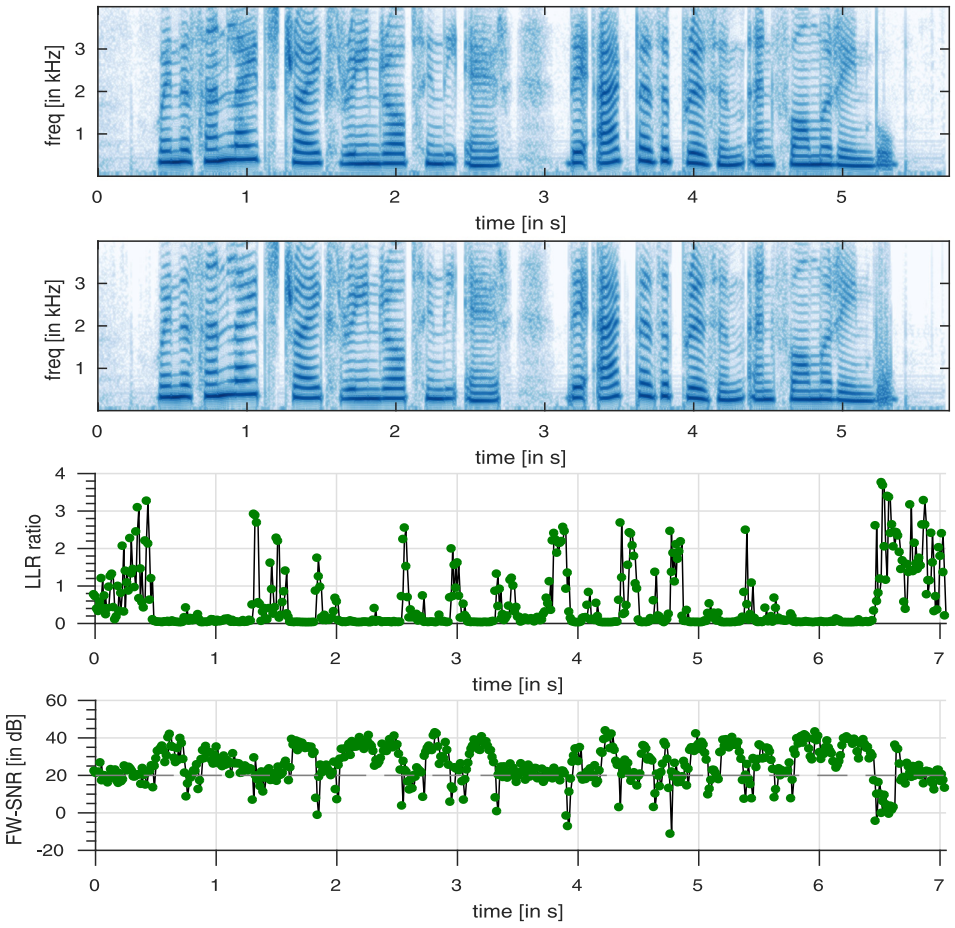


Fig. 11. Illustration of segment-wise performance evaluation of speech analysis-synthesis using the proposed approach. The spectrograms are obtained with 25 ms hanning short-time segments with 87.5% overlap.

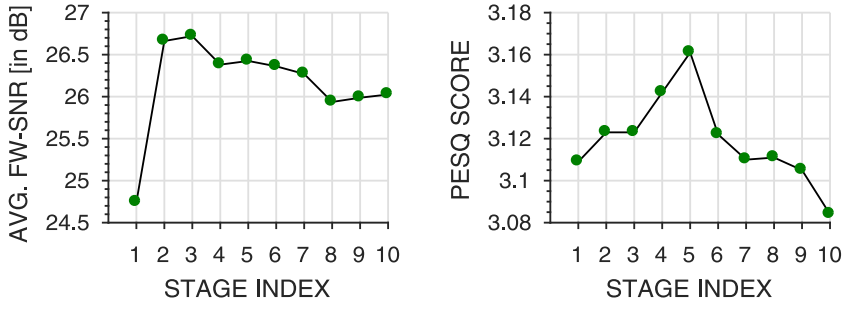


Fig. 12. Variation in objective measures over successive stages of IQ demodulation for the signal in Fig. 11.

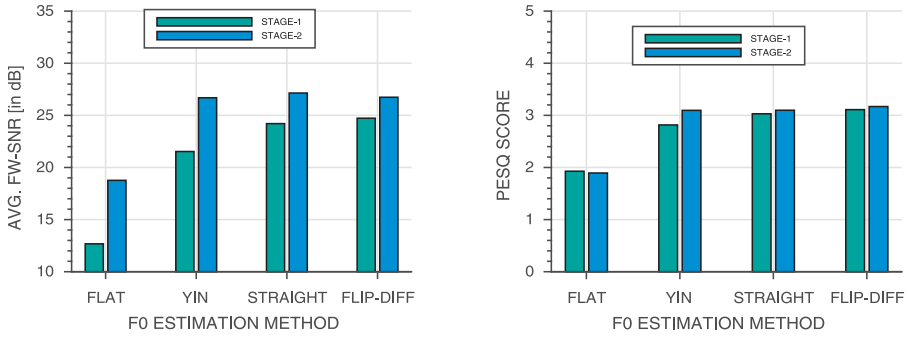
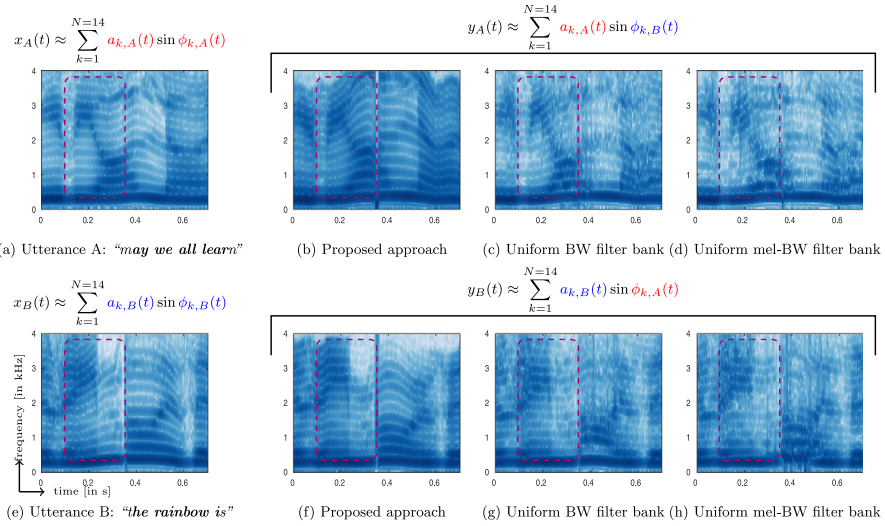


Fig. 13. Dependence of synthesis performance on initial instantaneous fundamental frequency estimate (F0). The speech signal used is same as in Fig. 11.



(a) Utterance A: *"may we all learn"* (b) Proposed approach (c) Uniform BW filter bank (d) Uniform mel-BW filter bank

(e) Utterance B: *"the rainbow is"* (f) Proposed approach (g) Uniform BW filter bank (h) Uniform mel-BW filter bank

Fig. 14. Illustration of spectrograms of signals obtained by swapping the IAs and IPs of two speech signals. The IAs and IPs are obtained by modeling $x_A(t)$ and $x_B(t)$ using the proposed approach or as combination of subband IAs and IPs of pre-defined filter banks. Here, we consider two kinds of filter banks, namely, uniform-linear bandwidth (BW) filter bank and uniform mel-BW filter bank. The red box highlights a comparison of the glides in the harmonics between $x_A(t)$ and $x_B(t)$, and subsequently, a comparison of the effectiveness in their exchange in $y_A(t)$ and $y_B(t)$, obtained using the three approaches.

1999), and this also provides better performance in analysis-synthesis. Interestingly, the performance is significantly improved with stage-2 of QHD for FLAT and YIN. The improvement is marginal for STRAIGHT and FLIP-DIFF as the input estimates from these are already accurate. These observations suggests that QHD can take input from any instantaneous fundamental frequency estimator, and the performance over stages is dependent on the accuracy of the estimator.

5.4. Information content in IFs and IAs

We analyzed the effectiveness of the proposed representation in capturing the perceived attributes of voiced speech. For this considered two utterances, $x_A(t)$ and $x_B(t)$, and swap their IFs. The spectrograms of the signals corresponding to these utterances are shown in Fig. 14(a,e). The spectrograms of the signals obtained after swapped the IFs are shown in Fig. 14(b,f). As can be seen, swapping the IFs transfers the IF glides in $x_A(t)$ to $x_B(t)$, and vice versa. On listening, this gave a perception of the intonation getting swapped between the two utterances. However, the phonetic content was intact as the (non-swapped) IAs are overlaid on the (swapped) IF contours. This small experiment highlights that IFs are associated with intonation and IAs are associated with phonetic content.

We performed similar swapping experiment with other subband representations, example, those obtained with fixed filter bank, namely, uniform-linear bandwidth and uniform-mel bandwidth filter banks. To contrast with the proposed representation, we set the number of subbands in the filter bank to the number of harmonics used in the proposed approach, that is $N = 14$. The IAs and IFs of each subband signal are computed using the analytic signal approach. The obtained subband IFs for $x_A(t)$ and $x_B(t)$ are then swapped, as earlier. The spectrograms of the signals obtained after swapping the IFs are shown in Fig. 14. In comparison to the swapping results from the proposed approach, the results with fixed filter bank representations are not satisfactory. Neither the intonation nor the phonetic content seem to be preserved very well. This is due to the poor capture of inherent modulations in the signal when using fixed sub-band filter banks.

5.5. Plosive analysis-synthesis

Plosives in speech occur as transient bursts, and these bursts do not have a well-defined mathematical structure (Lisker, 1985; Quatieri, 2008). In this section, we specifically analyze the representation of plosive sounds with Eq. (1). We carried analysis-synthesis of ten vowel-consonant-vowel (VCV) sound samples spoken by Subject-1 (female). In the ten VCVs, *aba, ada, adha, aja*, and *aga* contain voiced plosive as the core phoneme, and *apa, atha, ata, aca* and *aka* contain unvoiced plosive as the core phoneme. The IPA notations of these phonemes is given in Fig. 16. We used the Flip-Diff and QHD combinedly (with parameters same as in previous subsection) to obtain the representation in Eq. (1) for each VCV sound sample. The observations on analysis-synthesis are described below.

Observations. The avg. FW-SNR was found to be greater than 25 dB for all ten VCV sound samples. A listening test was designed to assign labels to the synthesized signals. The test comprised of three trials. Trial-1 used original samples ($x(t)$), Trial-2 used the reconstructed samples ($\hat{x}(t)$ that is, $\hat{x}_u(t) + \hat{x}_v(t)$) and the Trial-3 used the reconstructed voiced stream samples ($\hat{x}_v(t)$). The same six listeners took part in the test. Trial-1 had no confusion. The confusion matrices obtained for Trial-2, and Trial-3 are shown in Fig. 15. There is no confusion in Trial-2. This indicates that, the proposed representation perfectly captures the perceived attributes of these VCV sound samples. In trail-3, majority of samples are classified correctly. This indicates that, most of the perceived attributes of these sound samples are carried in the voiced stream. Amongst the confusions, we see that the plosive /c/ is often assigned to /ʈ/. This is likely because /c/ has unvoiced frication. The spectral energy associated with this frication is absent in the voiced stream. To have a more careful look at the time-varyingness in the IAs and IFs of the voiced stream corresponding of the VCV sound samples, we illustrate the estimates in Fig. 16. The trajectory of $\hat{\phi}'_1(t)$ gives an indication of voiced or unvoiced plosives. For unvoiced plosives, $\hat{\phi}'_1(t)$ falls to zero for a considerable time interval prior to burst. In all sound samples, the $\hat{\phi}'_1(t)$ has glides before and after the plosive burst. This hints at some kind of adaptation in voicing during the course of utterance of these sounds. There are spurious rapid variations in IF around the plosive burst instants. This will result in increase of local bandwidth in $x(t)$ around these instants. Observing the normalized IAs, we see that for each sound sample, the effect of transition from closure to burst on $a_k(t)$ is different across the sinusoids. Likely, the time-varyingness in IF and IA together capture the identity of the plosives. Further investigation is required to access the importance of individual sinusoids in capturing the perceived attributes.

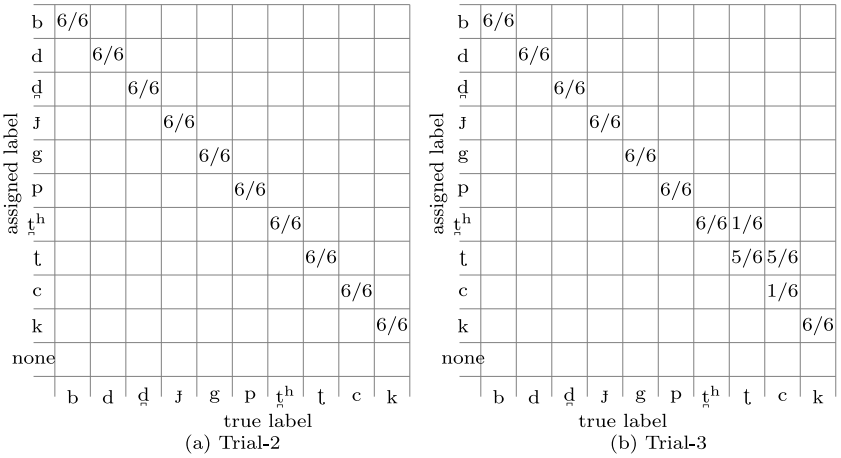


Fig. 15. Confusion matrix from listening test for classification of VCV sound samples. Presented samples: (a) $\hat{x}(t) = \hat{x}_v(t) + \hat{x}_u(t)$, and (b) $\hat{x}(t) = \hat{x}_v(t)$.

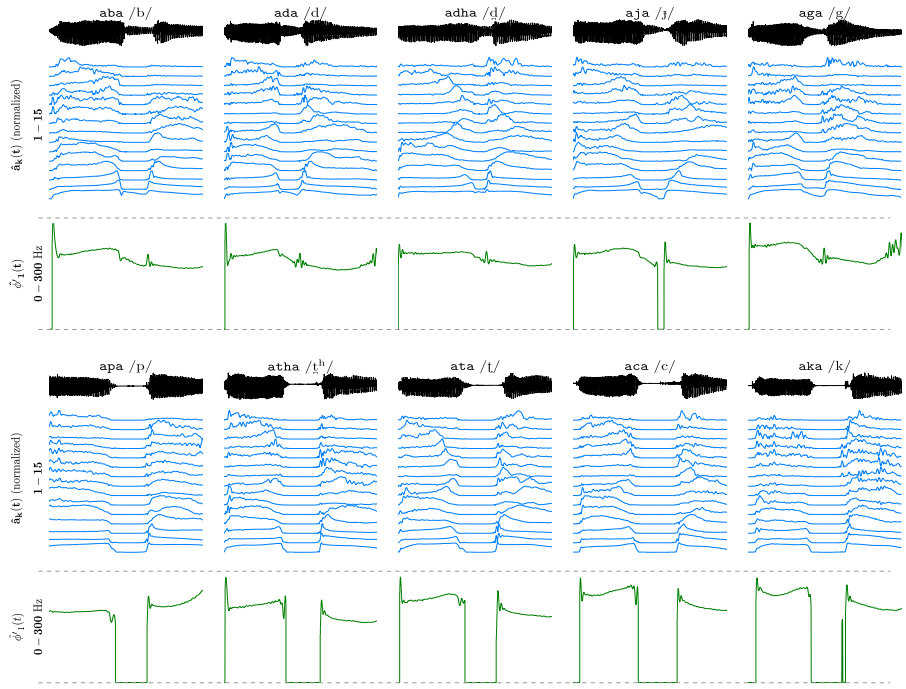


Fig. 16. Analysis of VCV sound samples. The original recordings correspond to utterances from Subject-1 (female).

5.6. Comparison with short-time analysis approaches

We compared the performance of the proposed approach with two other well-established approaches for sinusoidal modeling of speech, namely, sinusoidal modeling (SM McAulay and Quatieri, 1986) and adaptive harmonic modeling with adaptive iterative refinement (aHM-AIR Degottex and Stylianou, 2013). Both these approaches are short-time moving window analysis methods; SM estimates the sinusoidal parameters by spectral peak picking and aHM-AIR does it by solving a least-squares cost function over temporal segments. For synthesis, these either overlap-add or interpolation of the estimated parameters is used. Due to limitation in space we omit the theory of these two methods, an excellent review is presented in Caetano et al. (2016). The implementation of SM and aHM-AIR is done using the COVAREP (Degottex et al., 2014) (v1.4.1) (shared by the authors in Degottex and Stylianou, 2013). The temporal analysis window is chosen as 3 pitch periods and the hop size is set to 5 ms (same as used in Degottex and Stylianou, 2013). All the three methods are sensitive to input fundamental frequency estimate hence, for a fair comparison of SM and aHM-AIR with the proposed approach, we initialize all three approaches (including QHD) with the instantaneous fundamental frequency estimate obtained using the STRAIGHT algorithm (Kawahara et al., 1999). The fundamental frequency estimate in the unvoiced regions is filled in by using linearly interpolated values from adjoining voiced regions. This was found to be required for SM and aHM-AIR.

Dataset: The speech signals used are drawn from the Starkey open access speech dataset (Starkey Hearing Technologies, 2013). This is made up of high-quality anechoic chamber speech recordings from 16 American speakers (8 males and 8 females). All recordings are well articulated, and corresponds to 60 s of speech read from the standard rainbow passage (Fairbanks, 1960) by each speaker. We carry analysis-synthesis on (average) 6 s of speech (a different sentence) for each speaker. The parameter values used for the proposed approach are tabulated in Table 1; these are chosen by generalizing the observations from the previous subsections.

Objective evaluation: Fig. 17 depicts the avg. FW-SNR (grand average) and PESQ scores (average) obtained for the 8 male and 8 female speech recordings. The average number of harmonics are markedly different in

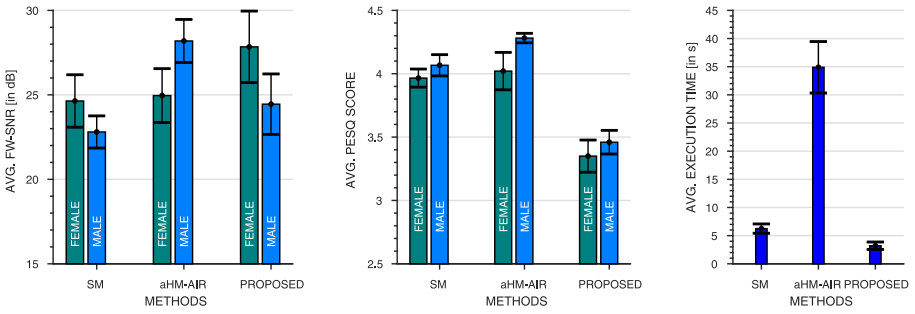


Fig. 17. Comparison between existing and proposed methods using objective evaluation. The error bars indicate 95% confidence interval.

Table 1
Choosen parameters for proposed approach.

Gender	Δ	$\mathcal{B}_{v,IA}$	$\mathcal{B}_{v,IF}$	$\mathcal{B}_{u,IA}$	$\mathcal{B}_{u,IF}$	K
Male	0.05max $[a_1(t)]$	60 Hz	60 Hz	1950 Hz	3500 Hz	3
Female	0.1max $[a_1(t)]$	100 Hz	100 Hz	1950 Hz	3500 Hz	3

male and female speech, and this is found to impact the performance of all three methods. The proposed method works better for female speech due to greater separation between the harmonics, and this reflects as 3 dB better avg. FW-SNR than SM and aHM-AIR. For male speech, the avg. FW-SNR is better than SM but falls below aHM-AIR. The proposed approach has an avg. PESQ between 3.4 and 3.5 (both male and female). This is good from the prespective that the proposed approach uses only 100 Hz (and 60 Hz) spectral bandwidth around each harmonic for female (and male) speech recordings. This is a subsampling of the full 4 kHz bandwidth of the spectrum of the signal. However, owing to this the PESQ is low compared to SM and aHM-AIR. The avg. execution time[4] for each method depicts the execution cost of short-time analysis. The aHM-AIR solves an optimization problem every 5 ms and this lends to increase execution time during analysis. The SM analysis uses peak picking in every 5 ms and this is computationally less demanding. The proposed approach requires no short-time analysis and hence, is faster than the other two approaches. It should be noted that the execution time of aHM-AIR can be reduced by increasing the hop size but this was found to adversely impact the FW-SNR and PESQ.

Subjective Evaluation: For subjective evaluation, a listening test was done using the sound recordings from the objective evaluation. The duration of all recordings ranged between 4 and 8.2 s with sampling rate of $F_s = 8$ kHz. The graphical user interface (GUI) for the listening test was designed in MATLAB. For each original speech recording (considered as reference), the GUI asked the listener to rate the quality of 4 test stimuli in comparison to the reference and also amongst themselves. There was no restriction on number of listening attempts. The 3 test stimuli comprised of synthesis from SM, aHM-AIR, and proposed approaches, and the fourth one was the original recording itself. The ordering of these stimuli was randomized for each reference recording comparison. For quality scoring a slider was provided on a continous scale $0-5$, partioned as shown in Fig. 18. Sennheiser HD 265 Linear headphones, with flat frequency response between 0.02 and 4 kHz, were used for listening. In total, each listener provided ratings for 10 reference recordings (5 female and 5 male), drawn randomly from the whole set of 16 recordings. The total sound files listened to by each listener is 50 $(10 + 5 \times 4 + 5 \times 4)$. Fourteen listeners, in the age group of $21-32$ (all university students), participated in the listening test. A listener on average took 15 min for completing the test. The avg. MOS score obtained by pooling trials from all listeners is shown in Fig. 18. The aHM-AIR, proposed in literature as a

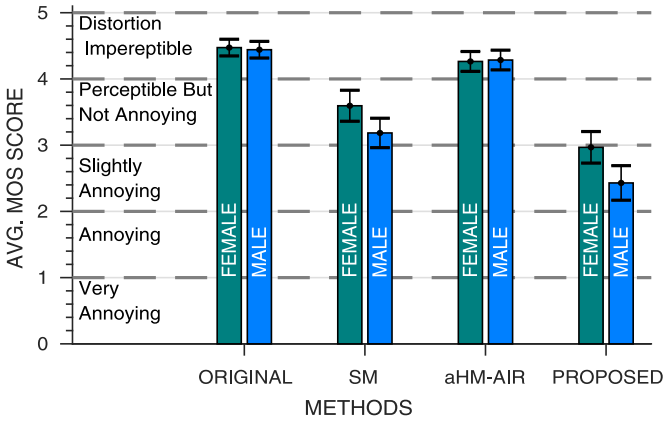


Fig. 18. Comparison between existing and proposed methods using subjective evaluation. The error bars indicate 95% confidence interval.

state-of-the-art, closely matched the scores for ORIGINAL. The SM was identified as having "disortions perceptible but not annoying". These distortions were identified as the phasiness artifacts (Laroche and Dolson, 1997). The proposed approach was rated as "slightly annoying", and female recordings were rated better than male recordings. On taking feedback from listeners and careful listening analysis we found that the "slightly annoying" disortion is different from the phasiness artifact in SM approach, and is attributable to temporally localized noisy bursts at certain instants in synthesis for some sound files. This happens at instants of low vocing, and present usually towards the end of the sentence. At these instants the IF (subsequently, the IA) estimation and updation is poor. As these artifacts are temporally localized, the intelligibility of the constitent words remain unaffected. The original sound recordings and the output recordings from all the methods have been hosted for public listening[5].

6. Conclusion and (future) perspective

The paper focused on analysis of time-varying sinusoids embedded in a quasi-harmonic complex. The introduced FFS and making use of its instantaneous frequency track for full signal analysis. Two algorithms, namely Flip-Diff and QHD, were designed towards this. Subsequently, the modeling was applied to speech signals. The proposed analysis-synthesis methodology illustrate the effectiveness of the simplified approach, devoid of any short-time parameter estimation (and subsequent interpolation), in modeling the specto-temporally rich non-stationarity in speech. The estimated IA and IF were found to capture the inherent modulations in the signal better than the fixed filter bank based approaches. A comparison with short-time approaches depicted

[4] computed on a Intel Core i7, 3.5 GHz 16 GB RAM MacBookPro machine, and all implementations done in MATLAB 2017.

[5] https://neerajww.github.io/preprint/demo/modeling/tvnm.html.

the decent objective and subjective scores compared to the state-of-the-art methods. We also found that the proposed approach has scope for improvement, specifically, on choice of parameters such as modulating signal bandwidths. We foresee useful application of the proposed approach to speech analysis, a field which has been dominated with short-time modeling techniques. The QHD algorithm can be used with any instantaneous fundamental frequency estimate. This is especially useful for noisy signals, where the robustness of Flip-Diff can be questionable. The time-varying sinusoidal modeling provides a high-resolution temporal modeling of the speech signals, with additional benefit of parsimony. Pursuing psychoacoustics with such model can provide for probing perceived speech attributes, in manner different from stimuli synthesized using traditional source-filter modeling.

Acknowledgment

We would like to thank Rakshita Joshi and Shreepad Potadar for taking part in recording sessions, all the volunteers who participated in the listening tests, and Chandra Sekhar Seelamantula, Sai Gunaranjan Pelluri and K. V. Vijay Girish for discussions on the content of this manuscript.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2018.10.008.

References

Almeida, L., Tribolet, J., 1983. Nonstationary spectral modeling of voiced speech. IEEE Trans. Acoust. Speech Signal Process. 31 (3), 664–678. doi:10.1109/TASSP.1983.1164128.

Boashash, B., 1992. Estimating and interpreting the instantaneous frequency of a signal-Part I: fundamentals. Proc. IEEE 80 (4), 520–538. doi:10.1109/5.135376.

Caetano, M., Kafentzis, G.P., Mouchtaris, A., Stylianou, Y., 2016. Full-band quasi-harmonic analysis and synthesis of musical instrument sounds with adaptive sinusoids. Appl. Sci. 6 (5), 127. doi:10.3390/app6050127.

de Cheveignè, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111 (4), 1917–1930.

Christensen, M.G., Albøge, C., Jensen, S.H., Rødbro, C.A., 2002. A harmonic exponential sinusoidal speech coder. In: Nordic Signal Processing Symposium I, Vol. 10, pp. 1–5.

Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S., 2014. COVAREP-a collaborative voice analysis repository for speech technologies. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., pp. 960–964. doi:10.1109/ICASSP.2014.6853739.

Degottex, G., Stylianou, Y., 2013. Analysis and synthesis of speech using an adaptive full-band harmonic model. IEEE Trans. Audio Speech Lang. Process. 21 (10), 2085–2095. doi:10.1109/TASL.2013.2266772.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95 (2), 1053–1064.

Fairbanks, G., 1960. The Rainbow Passage, Vol. 2. Addison-Wesley Educational Publishers, New York, NY, pp. 124–139.

Flandrin, P., Mar. 2001. Time-frequency and chirps. In: Proc. SPIE 4391, Wavelet Applications VIII March, pp. 161–175.

Gabor, D., 1946. Theory of communication. IEE J. Commun. Eng. 93, 429–457.

Ghitza, O., 2001. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J. Acoust. Soc. Am. 110 (3), 1628–1640. https://doi.org/10.1121/1.1396325.

Gianfelici, F., Biagetti, G., Crippa, P., Turchetti, C., 2007. Multicomponent AM-FM representations: an asymptotically exact approach. IEEE Trans. Speech Audio Process. 15 (3), 823–837. doi:10.1109/TASL.2006.889744.

Goodwin, M., 1998. Multiresolution sinusoidal modeling using adaptive segmentation. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., vol. 3, pp. 1525–1528. doi:10.1109/ICASSP.1998.681740.

Gribonval, R., Bacry, E., 2003. Harmonic decomposition of audio signals with matching pursuit. IEEE Trans. Signal Process. 51 (1), 101–111. doi:10.1109/TSP.2002.806592.

Haykin, S., 1994. An Introduction to Analog & Digital Communications. John Wiley & Sons, Asia.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (1), 229–238. doi:10.1109/TASL.2007.911054.

Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C., Liu, H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. London Ser. A 454 (1971), 903–995.

Kafentzis, G.P., Stylianou, Y., 2016. High-resolution sinusoidal modeling of unvoiced speech. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., pp. 4985–4989. doi:10.1109/ICASSP.2016.7472626.

Kawahara, H., Masuda-Katsuse, I., de Cheveignè, A., 1999. Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun 27 (34), 187–207.

Kim, K., Ahn, B., Chung, Y., Nam, T., Yi, S., 2006. Sinusoidal modeling using elliptic filter for analysis and synthesis of speech signals. In: SICE-ICASE Intl. Joint Conf., pp. 1705–1708. doi:10.1109/SICE.2006.315667.

Kumaresan, R., Rao, A., 1999. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. J. Acoust. Soc. Am. 105 (3), 1912–1924.

Laroche, J., Dolson, M., 1997. Phase-vocoder: about this phasiness business. In: Proc. of IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA), pp. 1–4. doi:10.1109/ASPAA.1997.625603.

Laroche, J., Stylianou, Y., Moulines, E., 1993. HNS: Speech modification based on a harmonic+noise model. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., 2, pp. 550–553. doi:10.1109/ICASSP.1993.319365.

Lisker, L., 1985. The pursuit of invariance in speech signals. J. Acoust. Soc. Amer. 77 (3), 1199–1202.

Macon, M.W., Clements, M.A., 1997. Sinusoidal modeling and modification of unvoiced speech. IEEE Trans. Speech Audio Process. 5 (6), 557–560. doi:10.1109/89.641301.

Marques, L.S., Almeida, L.B., 1989. Frequency-varying sinusoidal modeling of speech. IEEE Trans. Acoust. Speech, Signal Process. 37 (5), 763–765. doi:10.1109/29.17571.

McAulay, R., Quatieri, T., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. 34 (4), 744–754. doi:10.1109/TASSP.1986.1164910.

Micheyl, C., Oxenham, A.J., 2010. Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. Hear. Res. 266 (12), 36–51. Special Issue: Annual Reviews 2010.

Nørholm, S.M., Jensen, J.R., Christensen, M.G., 2016. Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (12), 2354–2367. doi:10.1109/TASLP.2016.2608948.

Pantazis, Y., Rosec, O., Stylianou, Y., 2011. Adaptive AM-FM signal decomposition with application to speech analysis. IEEE Trans. Audio Speech Lang. Process. 19 (2), 290–300. doi:10.1109/TASL.2010.2047682.

Patterson, R.D., 1976. Auditory filter shapes derived with noise stimuli. J. Acoust. Soc. Am. 59 (3), 640–654.

Petrovsky, A., Azarov, E., 2014. Instantaneous harmonic analysis: techniques and applications to speech signal processing. In: Ronzhin, A., Potapova, R., Delic, V. (Eds.), Speech and Computer. Springer International Publishing, Cham, pp. 24–33.

Petrovsky, A., Azarov, E., Petrovsky, A., 2011. Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding. Signal Process. 91 (6), 1489–1504. doi:10.1016/j.sigpro.2010.09.005.

Picinbono, B., 1998. Some remarks on instantaneous amplitude and frequency of signals. In: Proc. IEEE-SP Intl. Symposium on Time-Frequency and Time-Scale Analysis, pp. 293–300. doi:10.1109/TFSA.1998.721419.

Potamianos, A., Maragos, P., 1999. Speech analysis and synthesis using an AM-FM modulation model. Speech Commun. 28 (3), 195–209.

Quatieri, T., 2008. Discrete-Time speech signal processing: Principles and practice, Vol. 1st Edition. Prentice Hall.

Rao, A., Kumaresan, R., 2000. On decomposing speech into modulated components. IEEE Trans. Speech Audio Process. 8 (3), 240–254.

Rice, S., 1982. Envelopes of narrow-band signals. Proc. IEEE 70 (7), 692–699.

Rilling, G., Flandrin, P., Goncalves, G., 2003. On empirical mode decomposition and its algorithms. In: IEEE-EURASIP Workshop on Nonlinear Signal Image Processing (NSIP), Vol. 3, pp. 8–11.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality PESQ-a new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., Vol. 2, pp. 749–752. doi:10.1109/ICASSP.2001.941023.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Phil. Trans. R. Soc. London B 336 (1278), 367–373. doi:10.1098/rstb.1992.0070.

Schimmel, S.M., Atlas, L.E., 2008. Target talker enhancement in hearing devices. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., pp. 4201–4204.

Sell, G., Slaney, M., 2010. Solving demodulation as an optimization problem. IEEE Trans. Audio Speech Lang. Process. 18 (8), 2051–2066.

Serra, X., 1989. A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition. Stanford University.

Sharma, N.K., Sreenivas, T.V., 2015. Event-triggered sampling using signal extrema for instantaneous amplitude and instantaneous frequency estimation. Signal Process. 116, 43–54.

Sharma, R., Prasanna, S.M., 2016. A better decomposition of speech obtained using modified empirical mode decomposition. Digit Signal Process. 58, 26–39.

Shechtman, S., 2013. ttransient modeling for overlap-add sinusoidal model of speech. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., pp. 8189–8192. doi:10.1109/ICASSP.2013.6639261.

Smits, R., 1994. Accuracy of quasistationary analysis of highly dynamic speech signals. J. Acoust. Soc. Am. 96 (6), 3401–3415.

Starkey Hearing Technologies, 2013. Open access stimuli for the creation of multi-talker maskers. https://www.starkeyevidence.com.

Stylianou, Y., 2009. Voice transformation: a survey. In: Proc. IEEE Intl. Conf. Acoust. Speech Signal Process., pp. 3585–3588. doi:10.1109/ICASSP.2009.4960401.

Teager, H.M., Teager, S.M., 1990. Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract. Springer Netherlands, Dordrecht, pp. 241–261.

Turner, R., Sahani, M., 2011. Demodulation as probabilistic inference. IEEE Trans. Audio Speech Lang. Process. 19 (8), 2398–2411.
Wang, X., 2013. The harmonic organization of auditory cortex. Front. Syst. Neurosci. 7, 114.
Wei, D., Bovik, A., 1998. On the instantaneous frequencies of multicomponent AM-FM signals. IEEE Signal Process. Lett. 5 (4), 84–86. doi:10.1109/97.664173.
Zubrycki, P., Petrovsky, A., 2007. Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform. In: European Signal Processing Conf., pp. 2336–2340.
Zubrycki, P., Petrovsky, A., 2010. Quasi-periodic signal analysis using harmonic transform with application to voiced speech processing. In: IEEE Intl. Symposium on Circuits and Systems, pp. 2374–2377. doi:10.1109/ISCAS.2010.5537180.
Zwicker, E., Flottorp, G., Stevens, S.S., 1957. Critical band width in loudness summation. J. Acoust. Soc. Am. 29 (5), 548–557.