Speech Analysis Using Instantaneous Frequency Deviation

Anthony P. Stark and Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane Queensland 4111, Australia
{a.stark, k.paliwal}@griffith.edu.au

Abstract

In this paper, our aim is to derive a phase spectrum representation computed via the short-time Fourier transform. Specifically, we are interested in developing a narrow-band speech representation – employing 20-40 ms analysis windows. Furthermore, this representation should be as physically meaningful as the magnitude spectrum. To achieve these ends, we concentrate on instantaneous frequency (IF) derived from the phase spectrum. In doing so, we introduce the IF deviation spectrum, and show that this spectrum exhibits pitch and formant structure similar to the magnitude spectrum. Lastly we demonstrate the advantages of the proposed IF deviation spectrum over the IF distribution spectrum proposed earlier in the literature.

Index Terms: phase spectrum, instantaneous frequency

1. Introduction

Most speech processing applications are based on the short-time magnitude spectrum, while relatively little attention is paid to the short-time phase spectrum. The aversion toward using the phase spectrum can be accounted for by two primary reasons. Firstly, the phase spectrum is difficult to interpret and process. To extract useful information, the phase spectrum requires significant processing. Furthermore it suffers from many tractability issues, including the phase unwrapping problem [1]. In contrast, the magnitude spectrum requires no such post-processing. Also, the visual cues that manifest within the magnitude spectrum correlate very well with our understanding of speech. Formant and pitch frequency are easily identifiable in the magnitude spectrum. The second reason for avoiding phase can be attributed to several well known perceptual experiments [2, 3, 4], which show marginal phase spectrum intelligibility over short (20-40 ms) window durations. Contrary to these findings, it has recently been shown that stimuli constructed from the short-time phase spectrum can convey intelligibility comparable to its magnitude-only counterpart [5]. This result is supported by many studies which highlight a strong relationship between the two spectra – notably the recovery of magnitude spectrum from phase [6] and vice versa [7]. The present paper is motivated by our desire to further pursue this link, and aims to derive a meaningful magnitude-like spectral representation from the phase spectrum.

The short-time phase spectrum is a function of time as well as frequency. While there can be many ways to derive meaningful representations from the phase spectrum, two possible ways that come to mind are those that are obtained by taking its frequency derivative or its time derivative. Differentiation along the frequency axis yields group delay and differentiation along the time axis gives instantaneous frequency (IF)[8]. Both group delay and instantaneous frequency are much more meaningful than the unprocessed phase, and both can be tied to physically relevant phenomena [9]. However, in this paper, we choose to restrict our focus to include only the IF spectrum branch of phase processing.

Though there are several methods (and interpretations) for calculating IF, we further restrict our focus to include only IF via the short-time Fourier transform (STFT) phase spectrum. The STFT of signal $x(t)$ is [10]:

$$X(\omega, t) = \int_{0}^{T} x(t + \tau)w(\tau)e^{-j\omega\tau}d\tau,$$  (1)

where $w(t)$ and $T$ are the window function and window duration, respectively. The short-time IF spectrum $\nu(\omega, t)$ is given as the time-derivative of the short-time phase spectrum as follows,

$$\nu(\omega, t) = \frac{\partial}{\partial t} \text{ARG}[X(\omega, t)].$$  (2)

For discrete-time signal processing, Eq. 2 can be given as:

$$\nu(\omega, n) = \text{ARG}[X(\omega, n + 1)] - \text{ARG}[X(\omega, n)].$$  (3)

To avoid some of the phase unwrapping problems, we compute IF by using the Kay’s method as follows [11]:

$$\nu(\omega, n) = \text{ARG}[X(\omega, n + 1)]X^*(\omega, n)].$$  (4)

This method offers a number of advantages over the direct discretization of Eq. 2. Firstly, it requires only a single complex phase (ARG) function. Secondly, it produces IF values within the range of $-\pi \leq \nu(\omega, n) \leq \pi$. The majority of other methods, do not guarantee this constraint – making interpretation of IF values falling outside this interval somewhat problematic.

The IF spectrum has been extensively studied in the literature; it has been applied to formant extraction [10, 12], pitch extraction [13, 14] and speech recognition [15, 16, 17]. For the computation of IF spectrum from the STFT, two types of analysis procedures have been reported in the literature: narrow-band analysis and wide-band analysis. For narrow-band analysis, the duration of the analysis window is taken to be 20 to 40 ms. Since the duration is much longer than the typical pitch period of speech, the narrow-band analysis is generally used for pitch-asynchronous analysis (where the beginning of the analysis frame is needed to be synchronized with the pitch epoch). It has been shown in the literature [14, 18] that the IF spectrum obtained from narrow-band analysis contains information about the excitation source, but not about the vocal tract system. As a result, the narrow-band IF spectrum has been
used in the past for pitch extraction [13, 14]. The narrow-band magnitude spectrum, on the contrary, contains information about the excitation source as well as the vocal tract system and has been used for both pitch and formant extraction [19]. For wide-band analysis, the duration of the analysis window is typically taken to be 2 to 4 ms. Since it is much smaller than the typical pitch period expected in normal speech, wide-band analysis has to be used in a pitch-synchronous manner, otherwise the placement of analysis window with respect to pitch epochs will introduce significant frame-to-frame variability. The resulting IF spectrum contains information about the vocal tract system, but not about the excitation source [10, 20]. Friedman [10] has derived a novel representation (known as the IF distribution representation) from the wide-band IF spectrum and shown that this representation provides information about the vocal tract system (i.e., formant structure).

In this paper, our aim is to derive a representation from the narrow-band IF spectrum in such a way that it contains information about the source excitation as well as the vocal tract system – similar to the narrow-band magnitude spectrum. With this aim in mind, we examine the properties of IF deviation – the measure of how far each IF value strays from its center frequency. By taking the inverse absolute value of the IF deviation, we have developed a novel phase-based spectral representation. This new representation (referred here as the IF deviation spectrum) gives clear visual indication of underlying pitch and formant structure, and is more correlated to the narrow-band magnitude spectrum than the current IF distribution representations.

The rest of this paper is organized as follows. In section 2, we briefly describe Friedman’s IF distribution representation. In section 3, we describe our proposed a IF deviation based spectral representation. And lastly in section 4, we present some concluding remarks and direction for future research.

2. The IF distribution representation

By noting the fact that the IF values tend to cluster around dominant frequencies of the signal, Friedman [10] computed the histogram of IF values (i.e., IF density) across the frequency axis. The IF density as a function of the bin’s center frequency, gives rise to the IF distribution spectrum (or, spectral representation). This representation when used with for wideband analysis (2-4 ms window duration) was shown to display formant detection. However, for narrow-band analysis (20-40 ms window duration), the IF distribution spectrum loses the formant structure, displaying only the pitch information in the form of fine harmonic detail.

To illustrate this property of the narrow-band IF distribution spectrum, we take a 32 ms long speech signal of the vowel ay uttered by a male speaker. We show the magnitude spectrum, the IF spectrum and the IF distribution spectrum in Fig. 1 (a), (b) and (c), respectively. We can see from Fig. 1(a) that the magnitude spectrum provides both pitch and formant information. The IF spectrum shown in Fig. 1(b) has a diagonal staircase shape with horizontal stairs occurring at pitch harmonics (indicating high density of IF values near dominant harmonic frequencies). Lastly, the IF distribution spectrum shown in Fig. 1(c) has information about pitch harmonics, but does not exhibit any formant structure. In order to illustrate the spectro-temporal properties of the IF distribution based spectral representation, we use a speech utterance of the sentence “He knew the skill of the great young actress.” spoken by a male speaker sampled at 8 kHz and perform a narrow-band analysis with a 32 ms Hann window and 4 ms frame shift. The magnitude and the IF distribution spectrograms of this utterance are shown in Fig. 3 (a) and (b), respectively. Again, note that the magnitude spectrogram displays pitch as well as formant information. On the other hand, the IF distribution spectrogram shows only pitch harmonics and does not exhibit formant structure.

3. Proposed IF deviation representation

As mentioned earlier, our aim in this paper is to derive a representation from the IF spectrum computed via the narrow-band (20-40 ms long analysis window) STFT. Like the magnitude spectrum, this representation should display the pitch as well as the formant structure. The IF distribution spectrum described in the preceding section exhibits pitch structure but not the formant structure. In this section, we introduce a new representation based on IF deviation to achieve this aim. We define the IF deviation ψ(ω, t) as:

$$
\psi(\omega, t) = \nu(\omega, t) - \omega.
$$

(5)

For discrete-time signals, it becomes:

$$
\psi(\omega, n) = \nu(\omega, n) - \omega.
$$

(6)

Or following Kay’s method (Eq. 4):

$$
\psi(\omega, n) = \operatorname{ARG} [X(\omega, n + 1)X^*(\omega, n)] - \omega
= \operatorname{ARG} \left[ X(\omega, n + 1)X^*(\omega, n)e^{-j\omega} \right].
$$

(7)
We have seen earlier from Fig. 1(b) that the IF spectrum behaves like a diagonal stair-case, with horizontal stairs occurring at pitch peaks. The IF values basically track the frequencies of pitch harmonic peaks. A closer inspection of this figure reveals that the accuracy of the tracking (the IF deviation magnitude) is inversely proportional to the spectral magnitude; i.e., the higher the magnitude of a harmonic peak, the better the IF value tracks its corresponding harmonic frequency. In other words, the absolute value of IF deviation at a particular frequency is inversely proportional to the value of magnitude spectrum at that frequency; i.e.,

\[
\frac{1}{|\psi(\omega, t)|} \propto |X(\omega, t)|.
\]

(8)

Therefore, we use the inverse, absolute IF deviation value to define a new function,

\[
\alpha(\omega, t) = |\psi(\omega, t)|^{-1}.
\]

(9)

We use the function \(\alpha(\omega, t)\) at a given time \(t\) to define a new spectral representation and call it the IF deviation spectrum. We should point out that the relationship given by Eq. 8 is not influenced by scaling \(|X(\omega, t)|\) globally. Instead it refers to the relative spectral magnitudes within a given analysis frame. Figure 2(b) shows proposed spectrum. We can see from this figure that the IF deviation spectrum shows both pitch and formant structure similar to that shown by the magnitude spectrum in Fig. 2(a).

Figure 4(b) shows the IF deviation spectrogram for the same speech utterance as used in the preceding section. We can see that this spectrogram exhibits pitch and formant structure in a manner similar to the magnitude spectrogram. Again, comparisons with Fig. 3 show its advantage over the IF distribution representation in terms of formant structure preservation.

---

2For computing the IF deviation spectrum, we use the the Chebyshev equiripple windows [22].

4. Conclusion

In this paper we have developed a new spectral representation derived from the STFT phase spectrum. Past IF distribution representations when used for narrow-band analysis were often limited – being ill-suited for formant extraction. By focusing our attention on the IF deviation quantity, we have been able to develop a narrow-band IF representation that captures information for both excitation source as well vocal tract – mirroring much of the information provided by the magnitude spectrum. Visual inspection of the IF deviation spectrum suggests its potential to many ASR applications in which the short-time magnitude spectrum is traditionally used.

As a final note, while this paper has focused on creating a magnitude-like representation from phase, it still remains to be seen if phase can provide useful features not possible from the magnitude spectrum. Deriving such phase-exclusive features remains an interesting and challenging problem for the future.

5. References

Figure 3: Illustration of IF distribution spectrogram for a speech utterance sampled at 8 kHz from a male speaker using narrow-band analysis with 32 ms Hann window with 4 ms frame shift. (a) Magnitude spectrogram and (b) IF distribution spectrogram representation.

Figure 4: Illustration of IF deviation spectrogram for a speech utterance sampled at 8 kHz from a male speaker using narrow-band analysis with 32 ms Chebyshev 50 dB window with 4 ms frame shift. (a) Magnitude spectrogram and (b) IF deviation spectrogram representation.


