INTERSPEECH 2009 BRIGHTON

Group-Delay-Deviation Based Spectral Analysis of Speech

Anthony Stark and Kuldip 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, we investigate a new method for extracting useful information from the group delay spectrum of speech. The group delay spectrum is often poorly behaved and noisy. In the literature, various methods have been proposed to address this problem. However, to make the group delay a more tractable function, these methods have typically relied upon some modification of the underlying speech signal. The method proposed in this paper does not require such modifications. To accomplish this, we investigate a new function derived from the group delay spectrum, namely the group delay deviation.We use it for both narrowband analysis and wideband analysis of speech and show that this function exhibits meaningful formant and pitch information.

Index Terms: group delay deviation, spectral analysis, speech processing.

1. Introduction

The majority of speech processing applications are based on the short-time magnitude spectrum, while relatively little attention is paid to the short-time phase spectrum. This is true for both automatic speech / speaker recognition (ASR), as well as speech enhancement. 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. The visual cues that manifest within the magnitude spectrum correlate very well with our understanding of speech. Formant and pitch frequency are both readily seen 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]. While the short-time magnitude and phase spectrums contain similar information, they manifest such information in very different ways. Because of this, we investigate the phase spectrum as a possible source of additional speech information to be leveraged.

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 obtained by taking its derivative. Differentiation along the frequency axis yields group delay (GD) and differentiation along the time axis gives instantaneous frequency [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]. It has been shown previously, that magnitude-like information can be derived from the short-time instantaneous frequency spectrum [10]. In this paper, we are interested in deriving a meaningful spectral representation from the short-time group delay spectrum.

Given the signal $x(n)$, $n = 0, 1, ..., N-1$, the group delay spectrum is defined as the negative frequency-derivative of the phase spectrum.

$$\begin{aligned}\tau(\omega) &= -\frac{\partial\theta(\omega)}{\partial\omega}, \\ &= -\frac{\partial}{\partial\omega}\mathrm{Im}\log\left[X(\omega)\right], \\ &= -\mathrm{Im}\left[\frac{X'(\omega)}{X(\omega)}\right],\end{aligned} \tag{1}$$

where signal spectrum $X(\omega)$ is given by,

$$X(\omega) = \sum_{n=0}^{N-1} w(n).x(n)e^{-j\omega n}. \tag{2}$$

Here $w(n)$ is the analysis window of length N and the phase spectrum $\theta(\omega)$ is given as the argument of $X(\omega)$. The group delay spectrum may also be given by [11]

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{3}$$

where $Y(\omega)$ is defined as the Fourier Transform of the signal $n.x(n)$, and $_R$ and $_I$ are the real and imaginary component modifiers. Typically, group delay is used to characterize digital filter design and communication channel characteristics. For information carrying signals – such as speech, the definition is less clear cut. Furthermore the group delay of speech is often observed to behave poorly, having many impulsive regions. From Eq. 3, we can see that zeros in $X(\omega)$ correspond to poles in the group delay. In the regions near these zeros, the group delay becomes both large and chaotic – masking more useful features. While some research has gone into choosing appropriate window functions for GD speech analysis, in practice it is difficult to predict where the group delay poles will appear. Because of this, the majority of group delay methods do not use Eq. 3 directly, but rather compute modified spectrums.

The rest of this paper is organized as follows. In section 2, we cover the main approaches to using group delay covered in past literature. In section 3, we introduce our proposed group

delay function, and show its characteristics for both narrowband and wideband speech analysis. Finally, in section 4, we present concluding remarks.

2. Group delay processing for speech

2.1. Modified group delay spectrum

Since zeros within the magnitude spectrum $|X(\omega)|$ are a direct cause of GD intractability, Yegnanarayana and Murthy [11] replaced $|X(\omega)|$ by a cepstrally smoothed magnitude spectrum $S(\omega)$. The updated group delay function is given as follows.

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}, \tag{4}$$

where $\gamma \approx 0.9$ is a tuning paremeter. The modified group delay function (MGDF), then has its dynamic range altered

$$MGDF(\omega) = \text{sign}\left[\tau_p(\omega)\right].|\tau_p(\omega)|^{\alpha}, \tag{5}$$

where $\alpha \approx 0.4$. For values of $\gamma \neq 1$, the modified GD function directly incorporates magnitude spectrum information. In fact, setting $\gamma = 0$, yields a direct multiplication of the power and group delay spectrums – the so called product spectrum [12]. Since the dynamic range of the magnitude spectrum is comparatively larger than that of the group delay, the MGDF and product spectrum exhibit primarily magnitude spectrum information. Even when $\gamma = 1$, the modified group delay function implicitly leverages the power spectrum through use of the smoothed magnitude $S(\omega)$. As the level of smoothing increases, the modified group delay approaches a product spectrum. Because of this, we view the MGDF as leveraging group delay information to modify / enhance the magnitude spectrum, rather than vice-versa.

2.2. Chirp transform group delay

Since spectral zeros lying on, or near the unit circle are the primary cause of group delay volatility, it can be beneficial to compute group delay characteristics with a modified Fourier transform. The chirp transform group delay (CGD) analysis [13] involves two stages. First, the explicit elimination of zeros outside of the unit circle – thereby reducing a speech signal to its minimum phase form. And secondly, evaluation of the resulting Z transform over a circle whose radius is greater than unity (typically radius $\rho \approx 1.1$). Unlike the MGDF, the chirp group delay function does not primarily leverage magnitude information – using only the phase spectrum of the modified signal. However, CGD has high computational overhead required for extraction of the signal zeros.

3. Group delay deviation

Previous studies [11] have already characterized many GD spectrum characteristics – notably its ability to detect useful speech features such as formants and harmonics. However, to show such information special attention needs to be paid to reducing the volatility introduced by spectral zeros. In the previous sections, we have shown two methods used to overcome this problem. This study takes a different approach. Unlike the MGDF (and related product spectrum), we attempt to derive a GD function free from the influence of the power spectrum. Also, we seek a function that can process group delay without alteration to the underlying signal – as is done with CGD. To accomplish this, we direct our attention away from the group delay spectrum, and instead focus on the related group delay deviation – how much the group delay deviates from an expected value. We follow an approach similar to the one used in our earlier paper, where we derived meaningful, magnitude spectrum like information from the instantaneous frequency spectrum [10].

For a non-causal symmetric window (defined between $-T/2 \leq t \leq T/2$), high power spectral components are generally observed to have group delays close to zero. However, these group delays are often hard to see, due to the surrounding noisy components. Taking the inverse however, allows these regions to be pushed above the surrounding noise. For practical cases, a causal, symmetric window is used for analysis. In this case, high power regions are observed to have a group delay close to $(N-1)/2$, where N is the length of the analysis frame. We can thus define a new quantity $\eta(\omega)$ as the inverse group delay deviation (IGDD) spectrum.

$$\eta(\omega) = \left|\tau_w - \tau(\omega)\right|^{-1}, \tag{6}$$

or with log compression,

$$\log \eta(\omega) = -\log\left|\tau_w - \tau(\omega)\right|, \tag{7}$$

where τ_w is the expected group delay, which for a symmetric N point window is given by $(N-1)/2$. For practical applications, it should be noted that while the proposed function does remove some group delay noise, but it does not eliminate it altogether. In speech, formant and harmonic peak regions typically produce small GD deviations. However, low power regions are generally unpredictable and noisy. Because of this, we apply smoothing to the group delay spectrum. The inverse group delay deviation is smoothed with filter $H(\omega)$,

$$\hat{\eta}(\omega) = \eta(\omega) * H(\omega), \tag{8}$$

where $*$ is the convolution operator. The following sections detail some of the properties of the group delay deviation.

3.1. Group delay deviation for synthetic signals

A complex sinusoid signal $x(n) = A_0 e^{j\omega_0 n}$, windowed by an N-point symmetric window can be shown to have group delay equal to $(N-1)/2$. Therefore, we start our analysis by analyzing a signal that consists of two complex sinusoids. If an N-point window $w(n)$ is applied to this signal, its short-time Fourier transform is given by

$$\begin{aligned} X(\omega) &= X_0(\omega) + X_1(\omega) \\ &= A_0 W(\omega - \omega_0) + A_1 W(\omega - \omega_1). \end{aligned} \tag{9}$$

Where A_k is a complex number representing the magnitude and initial phase of the k'th sinusoidal component, ω_k is the sinusoid frequency and $W(\omega)$ is the Fourier transform of the analysis window. Solving for the group delay yields

$$\begin{aligned} \tau(\omega) &= -\text{Im}\frac{\partial}{\partial\omega}\log\left[X_0(\omega) + X_1(\omega)\right] \\ &= -\text{Im}\frac{\partial}{\partial\omega}\log\left[X_0(\omega)\right] - \text{Im}\frac{\partial}{\partial\omega}\log\left[1 + V(\omega)e^{j\psi}\right], \end{aligned} \tag{10}$$

where,

$$V(\omega) = \left|\frac{X_1(\omega)}{X_0(\omega)}\right|, \tag{11}$$

and

$$\psi(\omega) = \angle\left[\frac{X_1(\omega)}{X_0(\omega)}\right]. \tag{12}$$

Looking at group delay of Eq. 10, we can see that the first term simply evaluates to $(N-1)/2$. Thus we can think of the second term as a deviation measure – how much the actual group delay is pushed off from the expected group delay. We define the group delay deviation as follows

$$\begin{aligned}\tau_w - \tau(\omega) &= \frac{N-1}{2} - \tau(\omega) \\ &= \frac{\partial}{\partial\omega}\mathrm{Im}\log\left[1 + V(\omega)e^{j\psi(\omega)}\right].\end{aligned} \tag{13}$$

When the term $V(\omega) \ll 1$, the bracketed term in Eq. 13 is almost constant. This means the scope for introducing group delay deviation becomes small. $V(\omega)$ itself becomes small whenever $|X_0(\omega)| \gg |X_1(\omega)|$ or, (rearranging equation 10), $|X_1(\omega)| \gg |X_0(\omega)|$. Because of this, the GD deviation can be thought of as a crude measure of spectral purity – the smaller the GD deviation, the more likely the energy in $X(\omega)$ originated primarily from a single sinusoid.

This assumption also holds when analyzing multi-component signals, where a signal is given by K sinusoids

$$X(\omega) = \sum_{k=0}^{K-1} X_k(\omega), \tag{14}$$

where $X_k(\omega)$ is the k'th sinusoidal component. It can be useful to look at regions of the spectrum where a single sinusoid is dominant. That is, in the region where $X_0(\omega)$ is dominant, the group delay deviation is then given by

$$\begin{aligned}\tau_w - \tau(\omega) &= \frac{\partial}{\partial\omega}\mathrm{Im}\log\left[1 + \frac{\sum_{k=1}^{K-1} X_k(\omega)}{X_0(\omega)}\right] \\ &= \frac{\partial}{\partial\omega}\mathrm{Im}\log\left[1 + V(\omega)e^{j\psi(\omega)}\right].\end{aligned} \tag{15}$$

It is of particular interest to look at the ratio that comprises $V(\omega)$

$$V(\omega) = \left|\frac{\sum_{k=1}^{K-1} X_k(\omega)}{X_0(\omega)}\right|. \tag{16}$$

We can see that as $V(\omega)$ becomes small (as a result of $X_0(\omega)$ becoming dominant), the group delay deviation is again pushed toward zero. This means that we can expect that at frequencies where a single sinusoid is dominant (i.e. at sinusoid center frequencies), the group delay deviation should be small.

3.2. Group delay deviation for voice signals

We apply the proposed group delay representation (Eq. 7) to a short spoken sentence. We compare the proposed GD representation against an instantaneous frequency deviation function and minimum-phase chirp group delay function. Since the product spectrum and modified group delay functions are primarily magnitude-based, they are not shown here. For narrowband analysis, the spectrograms derived from the magnitude spectrum, the instantaneous frequency deviation spectrum [10], the minimum-phase chirp group delay spectrum (CGD) [13], and the proposed inverse group delay deviation (IGDD) spectrum are shown in figure 1 a) through d), respectively. Corresponding wideband analysis is shown in figure 2 a) through d), respectively. The magnitude, instantaneous frequency deviation and IGDD spectra use a 50 dB Chebyshev window, while CGD spectrum uses the Blackman window suggested by the authors [13]. We can see that both the CGD and IGDD display speech information – though less than both the magnitude spectrum and instantaneous frequency derived spectrum. For wideband analysis, both group delay functions exhibit formant structure, though for narrowband analysis, IGDD gains harmonic structure at the expense of formant structure. In comparison to the instantaneous frequency representation proposed earlier by us [10], the group delay deviation spectrum does not exhibit vocal tract information as clearly.

4. Conclusion

In this paper, we have developed a new spectral representation derived from the short-time phase spectrum. In particular, we have focused on a phase derived quantity – the group delay deviation. We have also shown that many speech features are readily derived from the group delay spectrum. In particular, the harmonic peaks are easily identified. Unlike previous studies on speech-based group delay, our representation does not require a minimum phase signal, or supplementary magnitude information.

While this paper has focused on studying the characteristics of group delay for speech based signals, it remains to be seen if the proposed GD information can be used to derive features that are complementary to existing magnitude-based cepstral features. Furthermore, the proposed GD based function appears to manifest less speech information than the instantaneous frequency representation proposed earlier – both in its narrowband and wideband forms. Because of this, more promising avenues of utilizing GD information could include signal generation, combined magnitude / phase enhancement and magnitude-only signal reconstruction.

5. References

[1] H. Al-Nashi, "Phase unwrapping of digital signals," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 37, no. 11, pp. 1693–1702, Nov 1989.

[2] M. Schroeder, "Models of hearing," *IEEE*, vol. 63, pp. 1332–1350, 1975.

[3] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[4] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Commun.*, vol. 22, no. 4, pp. 403–417, 1997.

[5] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

[6] M. Hayes, J. Lim, and A. Oppenheim, "Phase-only signal reconstruction," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, vol. 5, pp. 437–440, Apr 1980.

[7] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236–243, Apr 1984.

[8] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice–Hall, 1975.

[9] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–568, 1992.

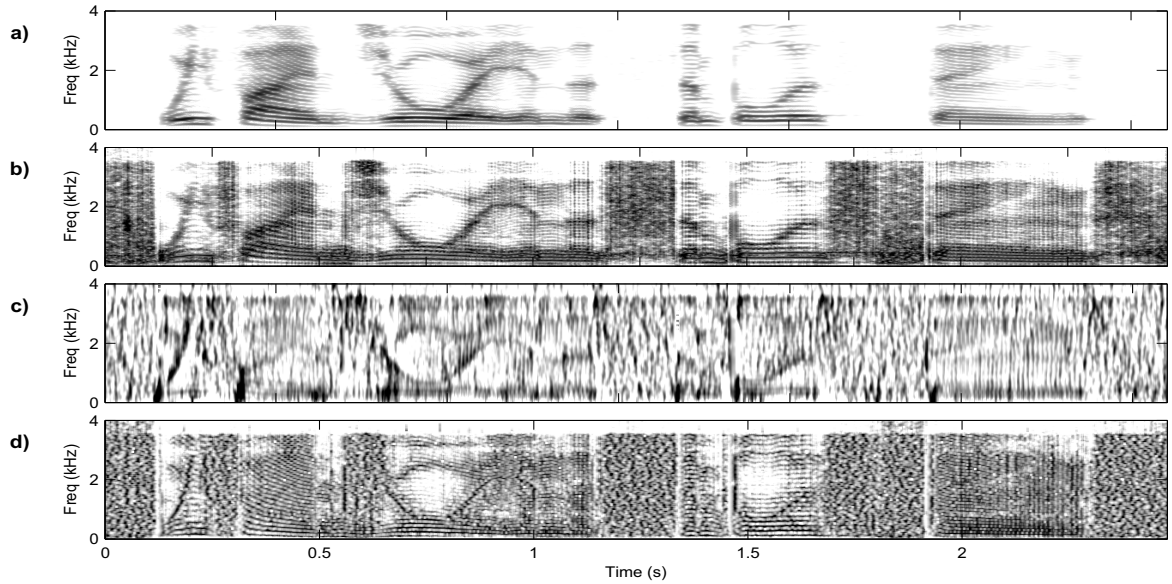

Figure 1: Illustration of narrowband (32 ms frames) GD based spectrograms for a speech utterance. (a) Magnitude spectrogram, (b) instantaneous frequency deviation, (c) chirp group delay spectrogram and (d) inverse group delay deviation spectrogram. Stimulus is an 8 kHz sentence: 'we find joy in the simplest things', spoken by a male speaker.

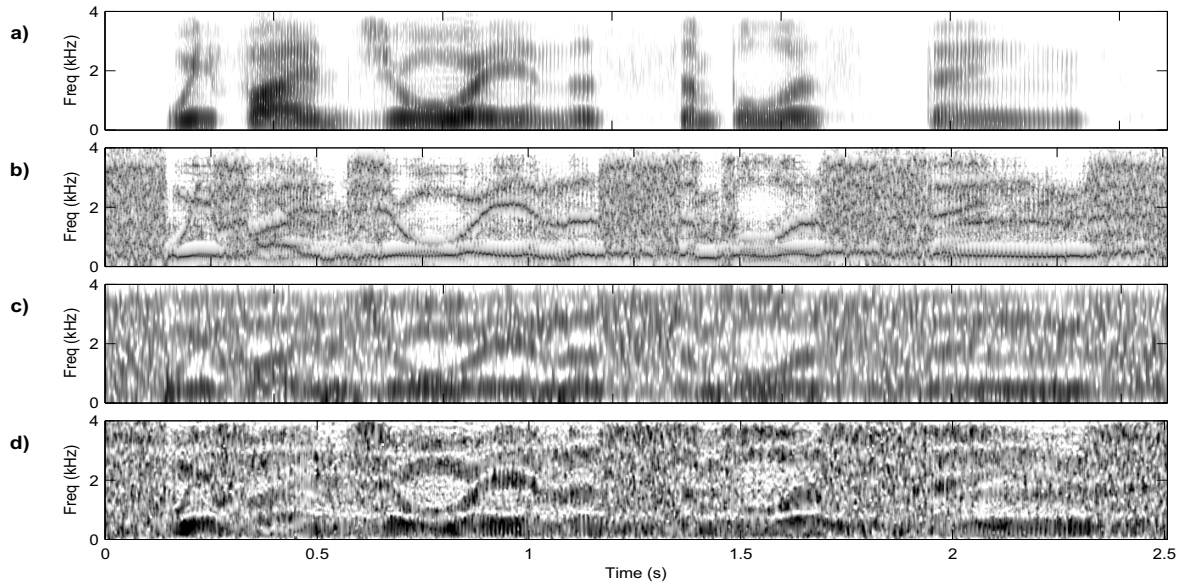

Figure 2: Illustration of wideband (4 ms frames) GD based spectrograms for a speech utterance. (a) Magnitude spectrogram, (b) instantaneous frequency deviation, (c) chirp group delay spectrogram and (d) inverse group delay deviation spectrogram. Stimulus is an 8 kHz sentence: 'we find joy in the simplest things', spoken by a male speaker.

[10] A. Stark and K. Paliwal, "Speech analysis using instantaneous frequency," in *Interspeech 2008*, 2008, pp. 2602–2605.

[11] B. Yegnanarayana and H. Murthy, "Significance of group delay functions in spectrum estimation," *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 40, no. 9, pp. 2281–2289, Sep 1992.

[12] K. Paliwal and D. Zhu, "Product of power spectrum and group delay function for speech recognition," *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 125–128, 2004.

[13] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Commun.*, vol. 49, no. 3, pp. 159–176, 2007.

[14] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, Jul 1989.