

EVALUATION OF THE MODIFIED GROUP DELAY FEATURE FOR ISOLATED WORD RECOGNITION

Leigh D. Alsteris and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University, Brisbane, Australia
e-mail: L.Alsteris@griffith.edu.au, K.Paliwal@griffith.edu.au

ABSTRACT

The results of our recent human perception experiments indicate that the short-time phase spectrum can significantly contribute to speech intelligibility over small window durations (i.e., 20–40 ms). This motivates us to investigate the use of the short-time phase spectrum to derive features for automatic speech recognition, which generally uses small window durations of 20–40 ms for spectral analysis. In this paper, we specifically investigate the frequency-derivative of the short-time phase spectrum (i.e., group delay function, GDF) from which to extract features. We demonstrate, with some simple examples, the volatility of the GDF to noise, pitch epochs and windowing effects. We summarise the work by Yegnanarayana and Murthy on the modified GDF (MGDF), which serves to remedy the problems of the GDF. We then implement Murthy and Gadde's MGDF-based features (MODGDF) to determine if they provide an improvement over the popular MFCC representation either by themselves or in combination with MFCCs on an isolated word recognition task.

1. INTRODUCTION

Automatic speech recognition (ASR) systems generally employ features derived purely from the short-time magnitude spectrum; the short-time phase spectrum is completely discarded (from herein, the modifier 'short-time' is implied when mentioning the phase spectrum and magnitude spectrum). This is due to the general belief that the phase spectrum does not contribute to speech intelligibility at small window durations used in short-time Fourier analysis [1]. In addition, from a signal processing viewpoint, the phase spectrum is difficult to interpret due to phase wrapping and other problems [2].

The results of some recently conducted human perception experiments [3], indicate that the phase spectrum can significantly contribute to speech intelligibility over small window durations (i.e., 20–40 ms). This finding provides motivation to investigate the use of the phase spectrum to derive features for ASR, which generally uses small window durations of 20–40 ms for spectral analysis. In its raw form, the phase spectrum is not amiable to ASR processing. Unlike the magnitude spectrum, the phase spectrum does not explicitly exhibit the system resonances. A phys-

ical connection between the phase spectrum and the structure of the vocal apparatus is not apparent. It is therefore necessary that the phase spectrum be transformed into a more physically meaningful representation. If such a representation can be found, we need to determine if it can be used to improve ASR recognition performance. The phase spectrum has two independent variables: frequency and time. Thus, while there may be many ways to represent the information present in the phase spectrum, two representations that first come to mind are those that can be obtained either by taking its frequency-derivative (group delay function, GDF) or its time-derivative (instantaneous frequency distribution, IFD). The focus of this paper is on the use of the GDF for ASR. Murthy and Gadde have recently proposed a feature set that is derived from a modified GDF [4]. In this paper, we conduct an experiment (independent of the authors in [4]) to determine if their proposed features provide an improvement over the popular MFCC representation either by themselves or in combination with MFCCs on an isolated word recognition task.

The paper outline is as follows: In Section 2, we review the GDF and highlight the problems when using it directly for ASR. In Section 3, we summarise the work by Yegnanarayana and Murthy [2] on the modified GDF (MGDF), which serves to remedy the problems of the GDF. In Section 4, we provide the implementation details of Murthy and Gadde's MGDF-based features [4] (MODGDF) then test the MODGDF features on an isolated word recognition task.

2. GROUP DELAY FUNCTION

The Fourier transform of a frame of digitised speech, $x(n)$ for $n = 0, 1, \dots, N - 1$, is given by:

$$X(\omega) = \sum_{n=0}^{N-1} x(n)e^{-j\omega n} = |X(\omega)|e^{j\theta(\omega)}, \quad (1)$$

where $|X(\omega)|$ is the magnitude spectrum and $\theta(\omega)$ is the phase spectrum. The GDF, $\tau(\omega)$, is defined as the negative derivative of the phase spectrum with respect to ω [5]:

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega} = -\left(\frac{d(\log X(\omega))}{d\omega}\right)_I \quad (2)$$

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

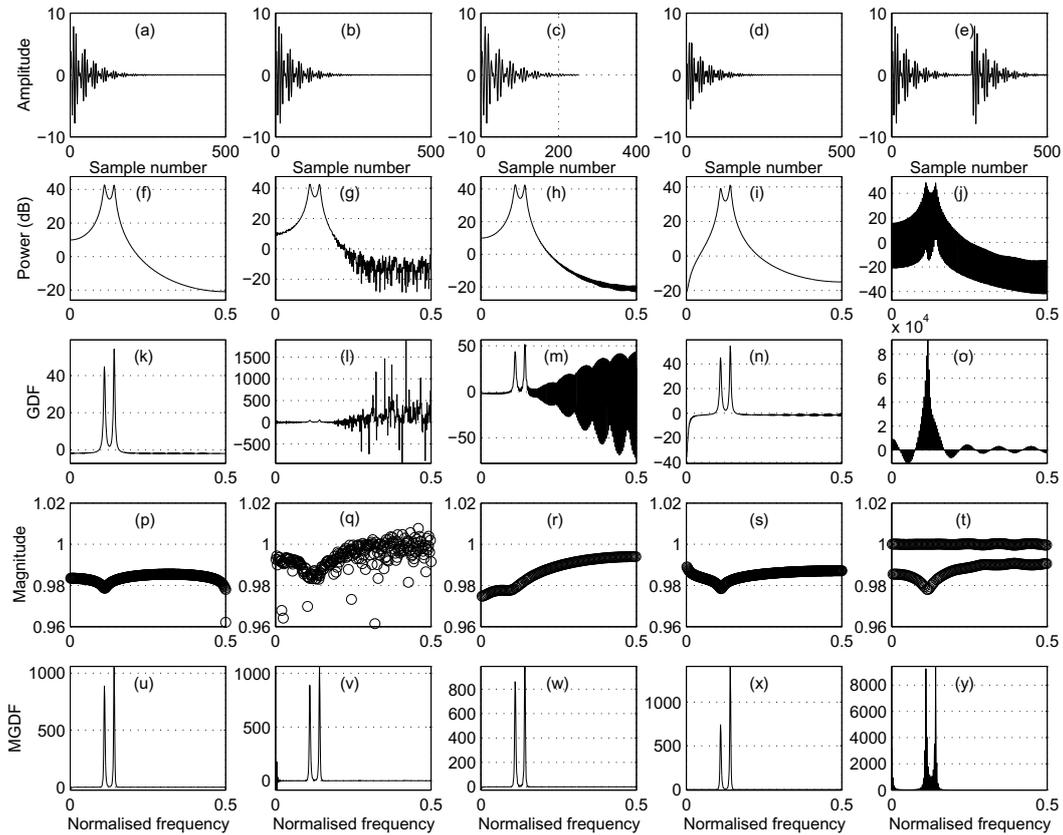


Fig. 1. The first row shows the (a) impulse response of an autoregressive process, (b) impulse response of the same autoregressive process in white noise (SNR=40dB), (c) truncated (or windowed) impulse response, (d) pre-emphasised impulse response (coeff. 0.97), and (e) the system response with an excitation of two impulses. The corresponding power spectra are shown in (f)–(j). The GDFs are shown in (k)–(o). The zero distributions are shown in (p)–(t). The MGDFs ($\alpha = 1$, $\gamma = 1$, and $s_w = 6$) are shown in (u)–(y). A rectangular analysis window is used in all cases.

where $Y(\omega)$ is the Fourier transform of $nx(n)$, and the subscripts R and I denote the real and imaginary parts respectively.

A theoretical analysis of the volatility of the GDF to the effects of noise, pitch epochs and windowing (or truncation) has been provided by other authors [2, 4]. Our intention in this section is not to repeat this theory, but rather to convey our own practical understanding of the GDF through a comprehensive set of simple examples.

For the following illustrations, we employ an autoregressive system:

$$H(z) = \frac{1}{1 + \sum_{i=1}^4 a_i z^{-i}}, \quad (4)$$

with coefficient values: $a_1 = -2.760$, $a_2 = 3.809$, $a_3 = -2.654$ and $a_4 = 0.924$ (these values are the same as those used in [2]). We compute the system impulse response and retain a sufficient number of its initial samples such that the impulse response has fully decayed (Fig. 1(a)). This version of the truncated impulse response is, for all intents and purposes, representative of the complete impulse response. The power spectrum of this signal, shown in Fig. 1(f), exhibits two resonances. The GDF, shown in Fig. 1(k), also clearly conveys the two resonances. The zeros of this signal are shown in Fig. 1(p).

Fig. 1(b) shows the impulse response with additive white noise, such that the signal-to-noise ratio (SNR) is 40 dB. The resonance peaks are still clearly discernible in the associated power spectrum of Fig. 1(g). The resonances, previously conveyed by the GDF in Fig. 1(k), are non-existent in the GDF for the noisy signal (Fig. 1(l)). The additive white noise introduces zeros close to the unit circle (Fig. 1(q)) which results in very small power spectral values at the frequency locations of these zeros. These small values of the power spectrum, $|X(\omega)|^2$, in the denominator of Eq. 3, subsequently result in large GDF values.

Now consider the same impulse response, but this time we only have half the amount of samples, such that the full decay of the impulse response is not captured (Fig. 1(c)). The windowing results in zeros being closer to the unit circle (Fig. 1(r)). Thus, a severe amount of distortion is introduced into the GDF (Fig. 1(m)). Note that windowing does not distort the power spectrum (Fig. 1(h)) as much as the GDF. In fact, by applying different window types (e.g., Hamming, Hanning, Gaussian, Blackman), the distortion can be reduced somewhat, in exchange for diminished resolving capability. The choice of window has a large impact on the resulting GDF.

Fig. 1(d) presents a pre-emphasised impulse response

(coeff. 0.97). The power spectrum for this signal is shown in Fig. 1(i). The effect of pre-emphasis on the GDF is shown in Fig. 1(n). Pre-emphasis introduces a zero near the unit circle causing a negative peak at $\omega = 0$. The associated zero distribution is shown in Fig. 1(s).

Considering that speech can be approximately modeled as the output of an autoregressive system excited by a periodic train of impulses, we now examine a signal obtained by exciting the autoregressive system with two impulses (Fig. 1(e)). Although pitch harmonics locally dominate the power spectrum, the resonance peaks are still globally discernible (Fig. 1(j)). However, no such peaks are visible in the GDF (Fig. 1(o)). This is due to the fact that the excitation introduces zeros extremely close to, if not on, the unit circle (Fig. 1(t)).

These simple examples demonstrate the volatility of the GDF to noise, pitch epochs and windowing effects. In all cases, it is the presence of zeros close to the unit circle that corrupt the GDF. Therefore, the GDF (given by Eq. 3) needs modification for it to be useful in ASR feature extraction.

3. MODIFIED GROUP DELAY FUNCTION

If we assume that speech is produced by a source-system model, the speech power spectrum, $|X(\omega)|^2$, can be expressed as the multiplication of the system component of the power spectrum, $S(\omega)^2$, with the source (or excitation) component of the power spectrum, $E(\omega)^2$:

$$|X(\omega)|^2 = S(\omega)^2 E(\omega)^2. \quad (5)$$

As demonstrated in the previous section, the excitation contributes zeros near the unit circle which cause meaningless peaks in the GDF. The modified group delay function (MGDF), $\tilde{\tau}(\omega)$, proposed by Yegnanarayana and Murthy [2], is formed by multiplying the GDF by the source component of the power spectrum:

$$\tilde{\tau}(\omega) = \tau(\omega) E(\omega)^2. \quad (6)$$

This operation gives less weight to peaks in the GDF which are the result of excitation-induced zeros near the unit circle. This is equivalent to replacing the denominator in Eq. 3 with the system component of the power spectrum, $S(\omega)^2$:

$$\tilde{\tau}(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^2}. \quad (7)$$

$S(\omega)^2$ is obtained by cepstral smoothing of $|X(\omega)|^2$. In practice, the cepstral smoothing operation not only smooths out zeros introduced by excitation, but also those contributed by noise and windowing. In fact, the cepstral smoothing removes the effect of any zeros that are close to the unit circle.

Murthy and Gadde [4] recently expanded on this expression, proposing the addition of two variables, γ and α . The role of γ is to vary the contribution from the system component of the power spectrum, as follows:

$$\tilde{\tau}_\gamma(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}. \quad (8)$$

The second additional variable, α , is a compression factor, such that the final expression for the MGDF is:

$$\tilde{\tau}_{\alpha,\gamma}(\omega) = \frac{\tilde{\tau}_\gamma(\omega)}{|\tilde{\tau}_\gamma(\omega)|} |\tilde{\tau}_\gamma(\omega)|^\alpha. \quad (9)$$

This parameter does not add any information, but rather represents the information already present in a more favourable form for ASR (as does the application of logarithmic compression for filter-bank energy coefficients). Please refer to [4] for a detailed discussion of these additional variables.

The bottom row of Fig. 1 shows the MGDFs for each of the cases previously examined in Section 2 (a cepstral smoothing window of size $s_w = 6$ is used, with $\alpha = 1$ and $\gamma = 1$). In each case, the resonance peaks are now clearly discernible in the MGDF.

4. EXPERIMENT

We perform an ASR experiment on the isolated letter database (ISOLET). ISOLET is an isolated-word, speaker-independent task, with speech sampled at 8 kHz. The vocabulary consists of 26 English letters. Two repetitions of each letter are recorded for each speaker. Speakers are divided into two sets: 90 for training, 30 for testing. Each word is modeled by a HMM with 5 emitting states and 5 Gaussian mixtures per state. We use the Cambridge Hidden Markov model (HMM) Toolkit (HTK) to train and test the HMMs. HMMs are tested on data with white noise added at several SNRs. We use an isolated word task since it eliminates extra parameters such as language model weight (because the likelihood of each word is the same) and word insertion penalty (since only one word can occur per utterance). Thus, when we change the length of the parameter vectors, there is no need to re-tune these heuristic parameters. Also, we test with additive white noise in order to gain some insight into the robustness of the MODGDF features (only matched-condition testing was done in [4]).

The details for constructing the features are given in the following paragraphs. In all cases, speech is pre-emphasised before analysis (coeff. 0.97) and a Hamming analysis window of duration 25 ms is used, with 10 ms frame-shift. As a baseline for recognition performance, we test with MFCCs. These are derived from the magnitude spectrum. For each frame: (i) compute the discrete Fourier transform (DFT) of $x(n)$, denoted by $X(k)$, (ii) compute the power spectrum $|X(k)|^2$, (iii) apply a Mel-warped filter bank (0 – 4 kHz) to $|X(k)|^2$ to obtain 24 filter-bank energies (FBEs), (iv) compute the discrete cosine transform (DCT) of the log FBEs, and (v) keep 12 cepstral coefficients, not including $c(0)$ (i.e., keep $c(n)$ for $n = 1, 2, \dots, 12$).

The MODGDF features are computed as follows. For each frame: (i) compute the DFT of $x(n)$ and $nx(n)$, denoted by $X(k)$ and $Y(k)$ respectively, (ii) compute the cepstrally smoothed spectrum of $|X(k)|$, denoted by

Table 1. ISOLET word recognition scores: white noise

Feature type (E–energy, D–delta , A–acceleration.)	SNR (dB)				
	∞	30	20	15	10
1.MFCC	78.27	74.87	67.44	56.67	37.50
2.MODGDF	76.79	63.01	32.82	16.22	7.37
3.MFCC+MODGDF	78.59	69.55	51.67	33.33	18.65
4.MFCC+E	79.62	76.73	66.03	52.50	36.41
5.MODGDF+E	78.65	65.51	37.82	20.58	8.65
6.MFCC+MODGDF+E	79.42	70.45	51.86	33.59	19.74
7.MFCC+E+D+A	90.83	89.04	79.36	68.91	52.95
8.MODGDF+E+D+A	89.29	82.37	70.58	56.79	35.32
9.MFCC+MODGDF+E+D+A	91.41	85.19	75.06	64.49	46.86
10. ISOLET-tuned case 9	92.31	85.64	76.79	66.41	46.22

$S(k)$, (iii) compute the MGDF as:

$$\tilde{\tau}_{\alpha,\gamma}(k) = \text{sign} \cdot \left| \frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{S(k)^{2\gamma}} \right|^\alpha \quad (10)$$

where sign is the sign of $\frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{S(k)^{2\gamma}}$, (iv) compute the DCT of $\tilde{\tau}_{\alpha,\gamma}(k)$, and (v) keep 12 cepstral coefficients, which includes $c(0)$ (i.e., keep $c(n)$ for $n = 0, 1, \dots, 11$). Results of experiments by Murthy and Gadde [4] indicate that $s_w = 6$ is best for smoothing. They also recommend that $\alpha = 0.3$ and $\gamma = 0.9$. Best recognition performance is obtained when keeping $c(0)$. Therefore, we employ all of these settings.

We do not use cepstral mean subtraction (CMS) because the ISOLET utterances are too short (empirical evidence suggests that, for increased recognition performance from CMS, utterances must be longer than 2-4 seconds). Also note that Murthy and Gadde weight the MODGDF cepstral values, $c(n)$, by n for $n > 0$; this is liftering and makes no difference to recognition performance in a HMM framework. Therefore, no liftering is performed on any feature set.

Word recognition scores are provided in Table 1. Bold font denotes the best word recognition score for each SNR. We observe that MODGDFs perform worse than MFCCs in all SNRs (cases 1 and 2). The same is true when energy and deltas are attached (cases 4, 5, 7, and 8). When the MODGDFs are concatenated with the MFCCs (case 3), a slight performance improvement over using MFCCs alone is observed in matched conditions; however, this improvement in matched conditions is at the expense of a reduced performance in unmatched conditions. This is also the case when deltas are attached (compare case 7 to 9).

Murthy and Gadde conducted a line search on the SPINE database to determine the best values for α , γ , and s_w [4]. Using the same feature size and configuration as in case 9, we conduct a line search on ISOLET. Optimal values for the MODGDF feature, such that the matched condition score for MFCC+MODGDF+E+D+A is maximised, were found to be $\alpha = 0.3$, $\gamma = 0.9$, and $s_w = 8$. Note that we do not determine l_w (see [4] for definition) as we are ignoring channel effects. The matched recognition

score improves slightly (case 10), but again at the expense of reduced performance in lower SNRs.

5. CONCLUSION

We demonstrated with some simple examples the volatility of the GDF to noise, pitch epochs and windowing effects. We summarised the work by Yegnanarayana and Murthy on the MGDF, which serves to remedy the problems of the GDF. We implemented Murthy and Gadde's MODGDF features and compared the recognition performance to MFCCs on the ISOLET task. MFCCs provided better performance than MODGDFs in all SNRs (with additive white noise). Concatenating MODGDFs to MFCCs seems to provide a slight increase in performance for matched conditions, but at the expense of reduced performance in lower SNRs.

6. REFERENCES

- [1] A.V. Oppenheim and J.S. Lim, "The importance of phase in signals" Proc. IEEE, Vol. 69, pp. 529-541, May 1981.
- [2] B. Yegnanarayana and H.A. Murthy, "Significance of group delay functions in spectrum estimation", IEEE Trans. Signal Processing, Vol. 40, No. 9, pp. 2281-2289, Sept. 1992.
- [3] K.K. Paliwal and L.D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests", Speech Communication, Vol. 45, No. 2, pp. 153-170, Feb. 2005.
- [4] H.A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. I-68-I-71, Apr. 2003.
- [5] A.V. Oppenheim and R.W. Schaffer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1975.