Robustness of Group Delay Representations for Noisy Speech Signals

Sree Hari Krishnan P. [1]

IDIAP Research Institute, 1920 Martigny, Switzerland

R. Padmanabhan *

Department of Computer Science and Engineering, Indian Institute of Technology Madras, India

Hema A. Murthy

Department of Computer Science and Engineering, Indian Institute of Technology Madras, India

Abstract

This paper demonstrates the robustness of group delay based features to additive noise. First, we analytically show the robustness of group delay based representations. The analysis makes use of the fact that, for minimum-phase signals, the group delay function can be represented in terms of the cepstral coefficients of the log-magnitude spectrum. Such a representation results in the speech spectrum dominating over the noise spectrum, both at low and high SNRs. Further, we experimentally demonstrate the robustness of the representation on a voice activity detection (VAD) task, comparing a group delay based VAD algorithm with standard VAD methods as well as a magnitude-spectrum based method.

Key words: Voice activity detection, group delay functions, G.729 B, AMR VAD

* Corresponding author.
Email addresses: Hari.Parthasarathi@idiap.ch (Sree Hari Krishnan P.), padman@cse.iitm.ac.in (R. Padmanabhan), hema@cse.iitm.ac.in (Hema A. Murthy).

[1] Work was done at: Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. Joint first authors appear in random order.

1 Introduction

All practical speech processing systems must have robustness as an important characteristic. The system must be able to perform satisfactorily under conditions where the input speech signal is distorted by effects that may be acoustic, articulatory or phonetic. Distortions caused by acoustic variations include those by additive or convolutive noise, channel and environmental effects. Development of robust feature extraction methods form a crucial part of building reliable speech processing systems.

Most short-term feature representations for speech are derived from the magnitude spectrum [eg. Davis and Mermelstein (1980); Hermansky (1990)]. There has been a growing series of studies which show that the short-term phase spectrum contains important information [Alsteris and Paliwal (2006); Paliwal et al. (2011)]. However, extracting information from the phase spectrum is not as straightforward as extracting information from the magnitude spectrum. Unlike the magnitude spectrum, the phase spectrum does not explicitly exhibit the system resonances. Further, signal processing difficulties (such as wrapping of the phase spectrum) are encountered while working directly with the phase spectrum [Alsteris and Paliwal (2006)]. The information in the phase spectrum has been utilised mostly by processing its derivative, the group delay function. The group delay function has been used in applications like signal reconstruction [Yegnanarayana et al. (1984)], formant extraction [Murthy and Yegnanarayana (1991)] and spectrum estimation [Yegnanarayana and Murthy (1992)]. Researchers have also come up with conventional short time feature representations derived from the group delay function [Zhu and Paliwal (2004); Hegde et al. (2007b)]. Moreover, it is shown that for applications like speech and speaker recognition, magnitude-based and phase-based features achieve comparable performance, and combining them results in improvement [Zhu and Paliwal (2004); Hegde et al. (2007a)].

This paper demonstrates that a group delay based representation is robust against additive noise. Although this property has been observed before (see [Parthasarathi et al. (2007)]), no attempt has yet been made to provide a mathematical analysis of the robustness property. In this article, we first show analytically how the group delay spectrum tends to follow the speech spectrum (as opposed to the noise spectrum) when speech is corrupted by additive noise. For minimum phase systems, log-magnitude and phase spectra are related through cepstral coefficients (see Section 2.) Using this property, we can represent the group delay spectrum in terms of speech and noise components at different signal to noise ratios.

The result of the analysis is validated by an experimental voice activity detection (VAD) task. A simple group delay based VAD algorithm is compared

to different VAD methods in various environments with additive noise. The experiments further corroborate the noise robustness properties of group delay functions.

The rest of this paper is organised as follows. In section 2, we briefly describe group delay functions. We then show analytically in section 3 that group delay functions are robust to additive noise. In section 4 we briefly describe the various representations of group delay functions and how features are derived from them. In section 5, we describe the experimental evidence supporting the analysis. Finally, we conclude in Section 6.

2 A brief overview of group delay functions

The Fourier transform of a signal $x(n)$ can be represented as

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)} \tag{1}$$

The group delay function $\tau(\omega)$ of a signal is defined as the negative derivative of the continuous phase spectrum $\theta(\omega)$:

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \tag{2}$$

Starting from Equation 1, the group delay function can be computed directly from the signal as:

$$\tau(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_I(\omega)\hat{X}_I(\omega)}{|X(\omega)|^2} \tag{3}$$

where subscripts R and I respectively denote the real part and imaginary part, $x(n) \leftrightarrow X(\omega)$ and $\hat{x}(n) \leftrightarrow \hat{X}(\omega)$ are Fourier transform pairs, and $\hat{x}(n) = nx(n)$.

For a minimum-phase signal, it can be shown that the log magnitude and continuous phase spectra are related as follows [Oppenheim and Schafer (2000)]:

$$\ln|X(\omega)| = \frac{1}{2}c(0) + \sum_{n=1}^{\infty} c(n)\cos(n\omega) \tag{4}$$

$$\theta(\omega) = -\sum_{n=1}^{\infty} c(n)\sin(n\omega) \tag{5}$$

where $c(n)$ are the cepstral coefficients. Taking the negative derivative of Equation 5, we get the group delay function for minimum phase signals:

$$\tau(\omega) = \sum_{n=1}^{\infty} nc(n)\cos(n\omega) \tag{6}$$

Equations 4 and 5 show that for minimum-phase signals, the log magnitude and phase are related through cepstral coefficients. Also, from Equation 6, we find that the group delay function is the Fourier transform of the weighted cepstrum.

3 Robustness of group delay functions to additive noise

In this section, we analytically show that group delay functions for minimum-phase signals are robust to additive, uncorrelated noise.

Let $x[n]$ denote a clean, minimum-phase speech signal, which is degraded by uncorrelated, zero-mean, additive noise $v[n]$ with variance $\sigma^2(\omega)$. Then, the noisy speech, $y[n]$, can be expressed as,

$$y[n] = x[n] + v[n] \tag{7}$$

Taking the Fourier transform, we have

$$Y(\omega) = X(\omega) + V(\omega) \tag{8}$$

Multiplying by corresponding complex conjugates and taking the expectation, we have the power spectrum

$$P_Y(\omega) = P_X(\omega) + \sigma^2(\omega) \tag{9}$$

where $P_Y(\omega) = |Y(\omega)|^2$, $P_X(\omega) = |X(\omega)|^2$ and we have used the fact that the expectation of noise is zero. The power spectra of the resulting noisy speech signal can be related to noise power and (clean) speech power in one of three mutually exclusive frequency regions:

(1) High noise power regions where $P_X(\omega) \ll \sigma^2(\omega)$
(2) High signal power regions where $P_X(\omega) \gg \sigma^2(\omega)$ and
(3) Equal power case where $P_X(\omega) \approx \sigma^2(\omega)$.

The power spectrum of the noisy speech signal in each case is represented by $P_Y^n(\omega)$, $P_Y^s(\omega)$ and $P_Y^e(\omega)$ respectively. We analyse the group delay representation of noisy speech in the three cases mentioned above.

3.1 High noise power spectral regions ($P_Y^n(\omega)$)

In this subsection, we consider frequencies ω such that $P_X(\omega) \ll \sigma^2(\omega)$, i.e., regions where the noise power is higher than signal power. From Equation 9 we have

$$\begin{aligned} P_Y^n(\omega) &= P_Y(\omega) \quad \forall\, \omega \quad \text{s.t.} \quad P_X(\omega) \ll \sigma^2(\omega) \\ &= P_X(\omega) + \sigma^2(\omega) \\ &= \sigma^2(\omega) \left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right) \end{aligned}$$

Taking logarithms on both sides, using the Taylor series expansion of $\ln(1 + \frac{P_X(\omega)}{\sigma^2(\omega)})$, and ignoring the higher order terms,

$$\begin{aligned} \ln\left(P_Y^n(\omega)\right) &= \ln\left[\sigma^2(\omega)\left(1 + \frac{P_X(\omega)}{\sigma^2(\omega)}\right)\right] \\ &\approx \ln\left(\sigma^2(\omega)\right) + \frac{P_X(\omega)}{\sigma^2(\omega)} \end{aligned} \tag{10}$$

Expanding $P_X(\omega)$ as a Fourier series ($P_X(\omega)$ is a periodic, continuous, function of ω with a period $\omega_0 = 2\pi$),

$$\ln\left(P_Y^n(\omega)\right) \approx \ln\left(\sigma^2(\omega)\right) + \frac{1}{\sigma^2(\omega)}\left[\frac{d_0}{2} + \sum_{k=1}^{\infty} d_k \cos\left(\frac{2\pi}{\omega_0}\,\omega\, k\right)\right] \tag{11}$$

where, d_k are the Fourier series coefficients in the expansion of $P_X(\omega)$. Since $P_X(\omega)$ is an even function, coefficients of the sine terms are zero.

For a minimum-phase signal, the group delay function can be computed in terms of the cepstral coefficients of the log-magnitude spectrum, as given in [Yegnanarayana et al. (1984)],

$$\begin{aligned} \log|X(\omega)| &= \frac{c_0}{2} + \sum_{k=1}^{\infty} c_k \cos(\omega\, k) \\ \tau(\omega) &= \sum_{k=1}^{\infty} k\, c_k \cos(\omega\, k) \end{aligned} \tag{12}$$

where, $\tau(\omega)$ is the minimum-phase group delay function and a_k are the cepstral coefficients. From Equation 12, it can be observed that the group delay function can be obtained from the log-magnitude response by ignoring the dc term, and by multiplying each coefficient with k. Applying this observation to

Equation 11, we get the group delay function as:

$$\tau_{Y^n}(\omega) \approx \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \tag{13}$$

This expression shows that the group delay function is inversely proportional to the noise power ($\sigma^2(\omega)$) in regions where noise power is greater than the signal power.

3.2 High signal power spectral regions ($P_Y^s(\omega)$)

Now consider frequencies ω such that $P_X(\omega) \gg \sigma^2(\omega)$. Starting with Equation 9, and following the steps similar to those in the previous subsection:

$$\ln\left(P_Y^s(\omega)\right) \approx \ln\left(P_X(\omega)\right) + \frac{\sigma^2(\omega)}{P_X(\omega)} \tag{14}$$

Since $P_X(\omega)$ is non-zero, continuous, and periodic in ω, $\frac{1}{P_X(\omega)}$ is also periodic and continuous. Consequently, $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ can be expanded using Fourier series, giving

$$\ln\left(P_Y^s(\omega)\right) \approx \frac{d_0 + \sigma^2(\omega)\, e_0}{2} + \sum_{k=1}^{\infty} \left(d_k + \sigma^2(\omega)\, e_k\right) \cos(\omega\, k)$$

Using Equation 12, and following the steps in the previous case, we obtain the expression for the group delay function as,

$$\tau_{Y^s}(\omega) \approx \sum_{k=1}^{\infty} k\, (d_k + \sigma^2(\omega)\, e_k) \cos(\omega\, k) \tag{15}$$

where d_k and e_k are the Fourier series coefficients of $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ respectively. It is satisfying to observe that if $\sigma^2(\omega)$ is negligible, the group delay function can be expressed solely in terms of the log-magnitude spectrum.

3.3 Signal power ≈ noise power regions ($P_Y^e(\omega)$)

For frequencies ω such that $P_X(\omega) \approx \sigma^2(\omega)$, we again start with Equation 9, and follow steps similar to those in the previous subsections, except in this case we do not need the Taylor series expansion:

$$
\begin{aligned}
P_Y^e(\omega) &\approx 2P_X(\omega) \\
\ln\left(P_Y^e(\omega)\right) &\approx \ln 2 + \ln\left(P_X(\omega)\right)
\end{aligned} \tag{16}
$$

Expanding $\ln\left(P_X(\omega)\right)$ as a Fourier series, since it is a periodic, continuous, function of ω with a period 2π, the group delay function can be computed as,

$$
\tau_{Y^e}(\omega) \approx \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \tag{17}
$$

where d_k are the Fourier series coefficients of $\ln(P_X(\omega))$.

3.4 Behaviour of minimum-phase group delay functions in noise

From Equations 13, 15, and 17, the estimated group delay functions are summarised respectively for the three cases:

$$
\tau(\omega) \approx \begin{cases}
\frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) & \text{for}\, \omega\ :\ P_X(\omega) \ll \sigma^2(\omega), \\
\sum_{k=1}^{\infty} k\, (d_k + \sigma^2(\omega)\, e_k) \cos(\omega\, k) & \text{for}\, \omega\ :\ P_X(\omega) \gg \sigma^2(\omega), \\
\sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) & \text{for}\, \omega\ :\ P_X(\omega) \approx \sigma^2(\omega)
\end{cases} \tag{18}
$$

From Equation 18, we note that the group delay function of a minimum-phase signal (which is corrupted with additive noise), is *inversely* proportional to the noise power at frequencies corresponding to high noise regions. On the other hand, for low noise regions, the group delay function becomes *directly* proportional to the signal power. In other words, the group delay function tends to follow the magnitude spectrum of the signal, rather than that of the noise. This indicates that the group delay function of a minimum-phase signal preserves the peaks and valleys of the magnitude spectrum well in the presence of additive noise. The group delay representation is thus robust and information like formant peaks are well retained.

4 Representations of group delay functions

Although the previous section showed the robustness of group delay functions in additive noise, it was done under the assumption of the signal being minimum-phase. Furthermore, to gainfully employ group delay functions for a speech processing task, features need to be extracted. This section endeavours to summarise literature relevant in handling both issues.

In general, the group delay function can be effectively utilised only if the signal under consideration is a minimum-phase signal [Hegde et al. (2007b)].

A signal is defined as a minimum phase signal only if both itself and its inverse are causal and stable. The group delay function becomes unstable if the roots of the transfer function are on the unit circle. Since speech is produced by a stable system, the poles of the system are well within the unit circle. The possible presence of zeros near the unit circle lead to an unstable group delay function. For speech signals, the zeros are introduced by pitch, nasals and the short-time analysis window function. Further, zeros are also introduced by the glottal return phase (the time interval between the most negative value of the glottal flow derivative and glottal closure [Plumpe et al. (1999)] when the glottal flow is modelled as a linear filter [Doval et al. (2003)]. The presence of zeros makes the denominator term in Equation 3 vanish, leading to an ill-behaved group delay function. Thus, the analysis presented in the previous section applies to minimum-phase signals.

To compute the group delay function for non-minimum phase signals, two approaches can be used:

4.1 Minimum phase group delay function (MINGD)

In this method, a minimum-phase equivalent is derived from the non minimum-phase signal using the root cepstrum [Murthy and Yegnanarayana (1991)]. In [Nagarajan et al. (2003)], it was shown that for an all-pole, non minimum-phase signal, the causal portion of the inverse Fourier transform of the magnitude spectrum is a minimum-phase signal. The group delay function of this signal is then utilised. The MINGD of a signal is thus derived from its minimum-phase component.

4.2 Modified group delay function (MODGD)

In this method, the zeros near the unit circle in Equation 3 are suppressed by smoothing the magnitude spectrum $X(w)$ [Murthy and Gadde (2003); Hegde et al. (2007b)]. The resulting group delay function is well behaved. The MODGD $\tau_m(\omega)$ is defined as

$$\tau_m(\omega) = (\frac{\tau_s(\omega)}{|\tau_s(\omega)|})(|\tau_s(\omega)|)^{\alpha} \tag{19}$$

where

$$\tau_s(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_I(\omega)\hat{X}_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{20}$$

and $|S(\omega)|$ is a smoothed version of $|X(\omega)|$, while the parameters α and γ are

introduced to control the dynamic range. The length of the cepstral smoothing window is controlled by the parameter lifter_ω.

The MINGD of a signal is extracted from the magnitude spectrum of the signal, whereas, the MODGD can extracted directly from the signal. Studies such as [Hegde et al. (2007b)] have shown the have shown the equivalence between MINGD and MODGD in a least squares sense, when the signal is not minimum phase. Furthermore, it is also demonstrated in [Hegde et al. (2007b)] the application of the MODGD for tasks such as syllable recognition, language identification and speaker recognition. Given these reasons, we validate the robustness property of group delay functions on the MODGD by experimental evaluation.

4.3 Converting the MODGD into features

To convert the MODGD into meaningful features, a decorrelation operation is performed using homomorphic processing. This is achieved by applying a discrete cosine transform (DCT) on the MODGD. Thus, the MODGD is converted into cepstral features and the resulting feature representation is called the modified group delay feature (MODGDF.) The decorrelation implied in the MODGDF features allow the use of diagonal covariances for modelling the feature distribution.

The extraction of MODGDF features from the signal are now described as follows [Hegde et al. (2007b)]:

(1) Compute the DFT of the speech signal $x[n]$ as $X[k]$.
(2) Next, the DFT of the signal $n\,x[n]$ is computed as $\hat{X}[k]$.
(3) Compute the cepstrally smoothed spectra of $X[k]$ and denote it as $S[k]$. The parameter lifter_ω is used to control the length of the window in the cepstral domain.
(4) Compute the MODGD as:

$$\tau_m[k] = \left(\frac{\tau[k]}{|\tau[k]|}\right)(|\tau[k]|)^\alpha \tag{21}$$

where $\tau[k] = \frac{X_R[k]\hat{X}_R[k]+X_I[k]\hat{X}_I[k]}{|S[k]|^{2\gamma}}$ where the parameters α and γ are used to control the dynamic range of the MODGD.
(5) Compute the MODGDF features by taking the DCT:

$$c[n] = \sum_{k=0}^{k=N_f} \tau_m[k]\cos(n(2k+1)\pi/N_f) \quad 0 \le n < N_c \tag{22}$$

where N_f is the DFT size and N_c are the number of cepstral coefficients.

5 Experimental evaluation

In earlier sections, we analysed the effect of additive noise on the MINGD and how the MODGD is an approximation of the MINGD. The previous section described the process of converting the MODGD into cepstral features. In this section, we corroborate the analysis presented in section 3 with experimental results using the MODGDF.

5.1 Voice activity detection

A voice activity detection (VAD) algorithm distinguishes speech and non-speech regions in a signal. VAD is used as a preprocessing stage in many speech applications, including automatic speech recognition (ASR) and speech coding. An effective VAD subsystem improves performance significantly, and achieves better resource utilisation. When viewed as a pattern recognition problem, VAD has to discriminate two classes, (noisy) speech and non-speech. Non-speech may be silence, or acoustic background noise. In high SNR conditions, most VAD algorithms work well. But in low SNR conditions, VAD performance deteriorates considerably, with many detection errors.

Many VAD algorithms use short term energy [Evangelopoulos and Maragos (2005); Parthasarathi et al. (2007)]. Energy-based algorithms usually do not work well in low SNR conditions. Other approaches to VAD use statistical methods [Sohn et al. (1999); Nemer et al. (2001); Li et al. (2005)], which make use of the assumption that the statistics of noise and speech are different. Recently, to perform VAD non-stationary noise, long-term analysis windows have been utilised [Ghosh et al. (2011)].

Standard VAD algorithms include the G.729-Annex B algorithm [Benyassine et al. (1997)] (recommended by the International Telecommunication Union) and the adaptive multi rate (AMR) VAD 1 and 2 [GSM 06.94 (1999)] (recommended by the European Telecommunications Standards Institute).

5.2 Utilising MODGDF for VAD

In this section, we develop a VAD algorithm which utilises MODGDF features for representing the speech signal. The proposed MODGDF-based VAD algorithm is compared to a baseline algorithm as well as standard VAD algorithms.

The use of Gaussian mixture models (GMMs) is popular for modelling generic

probability distributions. In its application to VAD, the underlying principle is that the distributions of speech and non-speech are sufficiently separate in the MODGDF space. In the training phase, MODGDF features extracted from speech and non-speech are used to build GMMs for each class, using framewise labelled data. In the testing phase, each frame of the test utterance is classified as either speech or non-speech. This method is called MODGD-VAD.

The MODGDF features were extracted as given in section 4.3. The speech signal was pre-emphasised, after using a frame size of 25 ms and a frame shift of 10 ms and applying a Hamming window. As specified in [Hegde et al. (2007b)] the MODGDF parameters α and γ were set to 0.4 and 0.9 respectively, and the parameter lifter_ω was set to 8. 13-dimensional MODGDF features were extracted. Cepstral mean subtraction is performed to remove channel effects. Two 64-mixture GMMs were built for speech and non-speech respectively, each initialised using k-means clustering, followed by five iterations of the expectation maximisation (EM) algorithm. Diagonal covariance matrices were utilised.

The performance of MODGD-VAD is compared to a baseline GMM-based algorithm using the Mel frequency cepstra (MFCC). This method is called MFCC-VAD. MFCC features were computed as described in [Murthy et al. (1999)]. A bank of 40 filters, which were spaced between 0 Hz to 4000 Hz are used to compute filterbank energies by integration. The shape of the filters was set by setting the shape constant T_r to 0.0 and the filter warping constant F_ω was set to 0.4. 13-dimensional MFCC coefficients were extracted, and cepstral mean subtraction is performed. 64-mixture GMMs were built for speech and non-speech, initialised by k-means clustering, followed by five EM iterations. As in the case for MODGDF, diagonal covariances were used.

5.3 Reference VAD systems

In addition to the baseline MFCC-VAD, the performance of the MODGD-VAD algorithm is compared to two standard VAD algorithms. We provide a brief description of them below.

G.729B VAD: ITU's G.729B VAD [Benyassine et al. (1997)] makes use of spectral distortion, full band and low band energy and zero crossings to make a VAD decision every 10 ms. These parameters are estimated as the difference from the respective running averages. Piecewise linear hyperplanes are used to separate speech and non-speech in the 4-dimensional parameter space, followed by a four stage hangover scheme.

AMR VAD options 1 and 2: ETSI's AMR option 1 and 2 [GSM 06.94 (1999)]are sub-band based methods. AMR 1 uses nine non-uniform sub-bands

and AMR 2 uses sixteen. Both methods estimate sub-band SNRs using a first-order auto-regressive model. VAD decisions are made every 20 ms using an adaptive threshold, followed by a hangover scheme.

5.4 Dataset used for the evaluation

The dataset used for the study consisted of 426 speech utterances (213 female, 213 male), selected from the TIMIT training set (phonetically balanced, read speech [Garofolo (1993)]). The utterances were downsampled to 8kHz, because the software implementations of the G.729B VAD and AMR VAD we used required so. Framewise VAD decisions manually labelled on these utterances formed the reference labels. These utterances were then concatenated in sets of three and silence was inserted between them so that the ratio of speech to non-speech frames was 40% to 60%. This models the average activity in a typical telephone conversation [Beritelli et al. (2001)]. Three different types of noise (babble, pink and white) from the NOISEX-92 database [Varga et al. (1992)] were added to these files, resulting in four datasets, each having an SNR of 0, 5, 10 and 15 dB. It is to be noted that SNR computation is over the whole utterance, including silences.

Six such utterances at 10 dB SNR (comprising of about two minutes each of speech and non-speech) were used as training data. Using framewise speech and non-speech reference labels, the training data was used to estimate parameters the respective GMMs. On the other hand, the total test dataset consists of about thirty hours of noisy data.

5.5 Evaluation

The evaluation data for each noise type consists of four SNR levels. Framewise speech/non-speech outputs of the five VAD methods considered (G.729 B, AMR 1, AMR 2, MFCC-VAD and MODGD-VAD) are compared with the reference labels. The performance evaluation is based on the metrics defined in [Li et al. (2005)], namely:

- P_{cs}: Percentage of correct speech identification,
- P_{cn}: Percentage of correct non-speech identification and
- P_f: Percentage of error (misses and false alarms.)

Table 1. Performance of VAD methods in different noise environments.

Method	G.729B			AMR 1			AMR 2			MFCC-VAD			MODGD-VAD		
SNR	P_{cn}	P_{cs}	P_f	P_{cn}	P_{cs}	P_f	P_{cn}	P_{cs}	P_f	P_{cn}	P_{cs}	P_f	P_{cn}	P_{cs}	P_f
Babble noise environment															
0	65.3	57.3	37.8	47.8	95.7	33.0	44.5	95.6	34.9	88.7	78.1	15.6	95.7	79.3	10.9
5	65.4	68.2	33.4	56.9	97.5	26.8	42.7	99.1	34.6	95.0	83.7	9.5	95.9	88.4	7.1
10	65.3	77.4	29.8	72.2	97.2	17.7	45.4	99.7	32.8	97.3	88.0	6.4	95.0	93.8	5.4
15	65.8	86.4	25.9	84.5	97.1	10.4	57.0	99.5	25.9	96.7	91.4	5.4	93.4	96.6	5.3
Avg.	**65.5**	**72.3**	**31.7**	**65.3**	**96.9**	**22.0**	**47.4**	**98.5**	**32.1**	**94.5**	**85.3**	**9.2**	**95.0**	**89.6**	**7.1**
White noise environment															
0	89.5	54.2	24.5	87.5	67.5	20.4	94.7	89.0	7.5	86.3	76.7	17.5	90.6	85.8	11.4
5	89.4	66.8	19.5	90.2	84.6	11.9	93.6	96.4	5.2	94.5	82.5	10.3	91.9	90.2	10.1
10	89.3	77.4	15.4	92.4	93.4	7.1	93.2	97.9	4.8	97.1	87.9	6.6	92.6	94.1	6.8
15	93.1	87.4	9.1	95.6	95.7	4.3	93.0	98.7	4.6	96.3	92.7	5.16	90.6	97.9	6.4
Avg.	**90.3**	**71.5**	**17.1**	**91.4**	**85.3**	**10.9**	**93.6**	**95.5**	**5.5**	**93.5**	**85.0**	**9.9**	**91.0**	**92.0**	**8.3**
Pink noise environment															
0	89.5	60.5	22.0	89.6	69.2	18.5	95.0	91.4	6.4	77.3	82.4	20.6	91.3	80.4	13.1
5	88.8	71.4	18.1	92.6	86.2	9.9	93.6	97.2	4.9	92.8	86.1	9.9	94.3	85.2	9.4
10	89.3	81.4	13.8	94.7	93.0	5.8	93.4	98.3	4.6	97.0	90.5	5.6	95.1	91.0	6.5
15	88.4	88.6	11.5	95.6	96.6	3.9	92.2	99.2	4.9	93.0	94.6	6.3	94.4	95.0	5.3
Avg.	**89.0**	**75.5**	**16.3**	**93.1**	**86.3**	**9.5**	**93.6**	**96.5**	**5.2**	**90.0**	**88.4**	**10.6**	**93.8**	**87.9**	**8.6**
All	**81.6**	**73.1**	**21.7**	**83.7**	**89.5**	**14.1**	**78.2**	**96.8**	**14.3**	**92.7**	**86.2**	**9.9**	**93.3**	**89.8**	**8.2**

5.6 Results and discussion

The experimental performance of the proposed MODGD-VAD and other methods are given in Table 1.

The proposed MODGD-VAD achieves the least average VAD error (8.2%) across all noise types. Also, in babble noise, MODGD-VAD also achieves the least VAD error (7.1%) compared to the other methods. Among the three noise types, it achieves the best non-speech detection (92.0%) in white noise, whereas the best speech detection is achieved in babble noise (95.0%).

The performance of MFCC-VAD is similar, although the algorithm attains a higher average VAD error. At high noise levels (0 dB SNR), MFCC-VAD has a 6% performance reduction in compared MODGD-VAD. Similar to MODGD-VAD, MFCC-VAD achieves the least VAD error and the best speech detection in babble noise (9.2% and 94.5% respectively), whereas the best non-speech detection is achieved in pink noise (88.4.%) In particular, both these supervised methods perform better than the standard methods in babble noise.

G.729B has the highest average VAD error across all environments. The average speech detection rate is similar across various noise types, although the non-speech detection is poor in babble noise. Among the three noise types considered, G.729 obtains the least VAD error and best non-speech detection in pink noise (16.3% and 75.5%), whereas the best speech detection is achieved in white noise (90.3%)

The AMR VAD 2 algorithm achieves the best speech detection performance in pink and white noise (93.6%). In babble noise, it has good speech detection performance, although at the cost of noise detection. The AMR VAD 1 algorithm on the other hand, fares better in babble noise (average VAD error of 22.0% when compared to 32.1% obtained by AMR VAD 2.) Both the AMR algorithms achieve the least VAD error in pink noise (9.5% for AMR 1 and 5.2% for AMR 2), and the best non-speech detection in babble noise (96.9% for AMR 1 and 98.5% for AMR 2.)

As expected, all VAD methods considered show improvement in average error as the SNR increases. In babble noise, MODGD-VAD outperforms all other VAD methods considered. In pink noise and white noise, it performs better than AMR option 1 and is comparable to AMR option 2. Although G.729B achieves the highest average VAD error, it fares better than AMR VAD option 2 in babble noise. In babble noise, MODGD-VAD gives the least VAD error (7.1%), whereas in both white and pink noise AMR VAD option 2 gives the least error (5.5% and 5.2% respectively.)

The supervised VAD methods (MFCC-VAD) and (MODGD-VAD) fare bet-

ter than the unsupervised methods in all conditions. Being non-stationary in nature, babble noise degrades performance of the standard VAD methods considerably. The GMM is able to capture the non-stationarity and separate the speech and non-speech in a better manner.

The standard VAD algorithms implement a hang-over scheme in which the VAD decision of the current frame is based on that of the earlier frames. To reduce the misdetection of speech, the hang-over scheme usually delays the transition from speech to non-speech, at the cost of increased false alarm rate [Sohn et al. (1999)]. It is to be noted that a hang-over scheme was not used in the GMM-based methods.

6 Conclusions

In this paper, we demonstrated the robustness of group delay based representations to additive noise. Analysis of group delay representations in noise showed that the group delay spectrum tends to follow the signal spectrum, rather than the noise spectrum. To support the analysis, a GMM-based voice activity detector was developed using modified group delay based features. The modified group delay features achieved performance comparable to or better than standard VAD methods in various environments. The performance was also compared to an MFCC feature based GMM voice activity detector, which achieved an average error rate slightly more than the MODGD-based system.

We provided a mathematical explanation to the noise robustness of group delay functions. The experiments also demonstrated that features derived from group delay functions separate noisy-speech and non-speech in a reliable manner. Thus, the use of the phase-based features is as appealing as magnitude-based features for this important task.

References

Alsteris, L. D., Paliwal, K. K., 2006. Short-time phase spectrum in speech processing: a review and some experimental results. Digital Signal Process. 17, 578–616.

Benyassine, A., Shlomot, E., Su, H.-Y., 1997. ITU Recommendation G.729 Annex B: A Silence Compression Scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. IEEE Comm. Mag. 35 (9), 64–73.

Beritelli, F., Casale, S., Ruggeri, G., 2001. Performance evaluation and comparison of ITU-T/ETSI voice activity detectors. In: Proc. ICASSP. Vol. 3. pp. 1425–1428.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech Signal Process. 28 (4), 357–366.

Doval, B., d'Alessandro, C., Henrich, N., 2003. The voice source as a causal/anticausal linear filter. In: Proc. ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis.

Evangelopoulos, G., Maragos, P., 2005. Speech event detection using multiband modulation energy. In: Proc. Interspeech. pp. 685–688.

Garofolo, J. S., 1993. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia.

Ghosh, P. K., Tsiartas, A., Narayanan, S., 2011. Robust voice activity detection using long-term signal variability. IEEE Trans. Audio, Speech, Lang. Process. 19 (3), 600–613.

GSM 06.94, 1999. Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channels.

Hegde, R. M., Murthy, H. A., Gadde, V. R., 2007a. Significance of joint features derived from the modified group delay function in speech processing. EURASIP Jnl. Audio, Speech, Music Process.

Hegde, R. M., Murthy, H. A., Gadde, V. R. R., 2007b. Significance of the Modified Group Delay Feature in Speech Recognition. IEEE Trans. Audio, Speech Lang. Process. 15 (1), 190–202.

Hermansky, H., 1990. Perceptual linear predictive (plp) analysis of speech. Jnl. Acoust. Soc. America 87 (4), 1738–1752.

Li, K., Swamy, M. N. S., Ahmad, M. O., 2005. An improved voice activity detection using higher order statistics. IEEE Trans. Speech Audio Process. 13 (5), 965–974.

Murthy, H., Beaufays, F., Heck, L., Weintraub, M., 1999. Robust text-independent speaker identification over telephone channels. IEEE Trans. Speech Audio Process. 7 (5), 554–568.

Murthy, H. A., Gadde, V. R. R., 2003. The modified group delay function and its application to phoneme recognition. In: Proc. ICASSP. Vol. 1. pp. 68–71.

Murthy, H. A., Yegnanarayana, B., 1991. Formant extraction from minimum phase group delay function. Speech Comm. 10 (3), 209–221.

Nagarajan, T., Prasad, V. K., Murthy, H. A., 2003. Minimum phase signal derived from root cepstrum. IEE Electronics Lett. 39 (12), 941– 942.

Nemer, E., Goubran, R., Mahmoud, S., 2001. Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans. Speech Audio Process. 9 (3), 217–231.

Oppenheim, A. V., Schafer, R. W., 2000. Discrete-time Signal Processing. Prentice-Hall.

Paliwal, K. K., Wojcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53, 465–494.

Parthasarathi, S. H. K., Rajan, P., Murthy, H. A., 2007. Voice activity detection using group delay processing on buffered short term energy. In: Proc.

National Conference Comm.

Plumpe, M. D., Quatieri, T. F., Reynolds, D. A., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. 7 (5), 569–586.

Sohn, J., Kim, N. S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6, 1–3.

Varga, A. P., Steeneken, H. J. M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Tech. Report DRA Speech Research Unit.

Yegnanarayana, B., Murthy, H. A., 1992. Significance of group delay functions in spectrum estimation. IEEE Trans. Signal Process. 40, 2281–2289.

Yegnanarayana, B., Saikia, D. K., Krishnan, T. R., Jun. 1984. Significance of group delay functions in signal reconstruction from spectral magnitude or phase. IEEE Trans. Acoust., Speech, Signal Process., 610–622.

Zhu, D., Paliwal, K., 2004. Product of power spectrum and group delay function for speech recognition. In: Proc. ICASSP. Vol. 1.