The Delta-Phase Spectrum with Application to Voice Activity Detection and Speaker Recognition

Iain McCowan *Member IEEE,* David Dean *Member IEEE,* Mitchell McLaren *Student Member IEEE,* Robert Vogt *Member IEEE,* Sridha Sridharan *Senior Member IEEE*

***Abstract*—For several reasons, the Fourier phase domain is less favoured than the magnitude domain in signal processing and modelling of speech. To correctly analyse the phase, several factors must be considered and compensated, including the effect of the step size, windowing function and other processing parameters. Building on a review of these factors, this paper investigates a spectral representation based on the Instantaneous Frequency Deviation, but in which the step size between processing frames is used in calculating phase changes, rather than the traditional single sample interval. Reflecting these longer intervals, the term Delta-Phase Spectrum is used to distinguish this from instantaneous derivatives. Experiments show that mel-frequency cepstral coefficients features derived from the Delta-Phase Spectrum (termed Mel-Frequency Delta-Phase features) can produce broadly similar performance to equivalent magnitude domain features for both voice activity detection and speaker recognition tasks. Further, it is shown that the fusion of the magnitude and phase representations yields performance benefits over either in isolation.**

***Index Terms*—phase, instantaneous frequency, speech analysis, voice activity detection, speaker recognition.**

I. Introduction

Most speech analysis focuses on features derived from the signal's magnitude spectrum, with the phase spectrum discarded. This is due both to mathematical difficulties analysing phase as a function, as well as psychoacoustic and signal processing experimental results that have rarely shown the phase to provide any empirical benefit over magnitude-only features. While well motivated, however, this still effectively discards half of the information present in the original signal. While this discarded information may mostly be redundant in low noise conditions, when the noise energy becomes comparable to the signal energy, sources of discriminative information that may prove complementary to the magnitude spectrum are desirable.

Many efforts to improve the robustness and discriminative ability of speech features have focussed on the importance of encoding temporal information in the feature extraction process, such as RASTA filtering of spectral trajectories [1], temporal pattern (TRAPS) classifiers [2], and the modulation spectrum [3], [4], [5]. Given the Fourier phase domain encodes relative timing information between different spectral components, interest in its use has increased in recent years. Different approaches have included estimating phase changes from an interference model [6], using the phase of the signal autocorrelation at different lags [7], measuring relative phase difference between frequencies [8], and deriving features based on the group delay [9] and instantaneous frequency [10], [11]. A recent review of the use of phase information in speech processing, however, indicates that broadly effective phase-domain features remain elusive [12].

The main difficulty associated with extracting speech features from the phase spectrum is the ambiguity that exists between angles separated by multiples of 2π radians. While the principal phase spectrum can be obtained by choosing the phase angle to lie between $\pm\pi$, this choice is arbitrary and results in regular discontinuities from circular wrapping of values considered over time or frequency. Phase unwrapping may be performed to restore a continuous phase spectrum for analysis, but consistent unwrapping is difficult, relying on different heuristics in practice [12], [13], [14], [15]. This difficulty in obtaining the phase as a continuous function causes both analytical problems as well as modeling difficulties due to inconsistencies in the representation across frames.

This paper commences with a discussion of practical issues that must be considered in analysing phase domain information from the short-time Fourier transform (STFT). While works can be found in the speech signal processing literature that discuss individual issues to varying degrees, the literature on these details is sparse. A first contribution of this paper is therefore to provide a tutorial introduction and brief literature review of practicalities in dealing with short-time Fourier phase in discrete-time frame-based processing algorithms. In particular, compensation must be made for the inter-frame time step and the effect of the windowing function before the phase spectrum can be meaningfully analysed. Another issue concerns the lack of a common temporal origin when making comparisons of the phase spectrum over different sequences, such as when developing statistical models of speech. Finally, there is a need to select processing parameters, such as frame size and window function, that are appropriate for analysing phase information, rather than naively applying parameters that work well for the magnitude domain.

Following this review, this paper investigates the use of phase changes between analysis frames as a representation that can be consistently analysed both within and across sequences. As a temporal difference in phase values, this representation

The authors are with Speech and Audio Research Laboratory, Queensland University of Technology, 2 George St. Brisbane, QLD, Australia. Phone: +61 7 3138 1608, Fax: +61 7 3138 1516

Mitchell McLaren is also with Centre for Language and Speech Technology, Radboud University, PO Box 9102, 6500HC Nijmegen, The Netherlands.

Email: iain@ieee.org, ddean@ieee.org, m.mclaren@let.ru.nl, r.vogt@qut.edu.au, s.sridharan@qut.edu.au

is equivalent to the Instantaneous Frequency Deviation (IFD) spectrum [11], [16]; however rather than estimating the instantaneous derivative using successive samples, the phase delta is analysed over a larger time delta corresponding to the inter-frame step size. As there is no intention to measure the formal derivative (that is, the limit as the time interval approaches zero or equivalently single sample difference in its digital approximation), in order to distinguish from the Instantaneous Frequency Deviation as commonly computed and used, the term Delta-Phase Spectrum is used to describe the quantity used in this article.

The desire to analyse phase differences over frame step intervals is motivated by the success of approaches to model the speech signal as the birth and death of individual sinusoidal components, each lasting several short-term frames [17], [18], [19], [20]. Measuring the simple difference in phase in narrow frequency bins across step-sized intervals may effectively capture information about timing and transitions across the spectrum and between speech units, such as phonemes, syllables and words, potentially leading to useful features for detecting voice activity in noise, or distinguishing voices. In order to demonstrate use of the Delta-Phase Spectrum in practice, therefore, these two applications are investigated. First, a simple Gaussian Mixture Model (GMM) based Voice Activity Detection (VAD) system is evaluated in different noise conditions in Section V. Mel-frequency cepstral coefficient features derived from the Delta-Phase Spectrum, termed Mel-Frequency Delta-Phase (MFDP) features, are compared to and combined with standard Mel-Frequency Cepstral Coefficient (MFCC) features derived from the magnitude spectrum [21]. Similarly, the effectiveness of MFDP features is evaluated for application to speaker recognition in Section VI.

II. A Review of Implementation Practicalities for the Phase Domain

A. Short-Time Fourier Analysis of Speech Signals

While short-time Fourier analysis of speech is a well-known technique, in established use for over 30 years [22], [23], [24], a brief review is provided here as the basis for the subsequent discussion.

The short-time Discrete Fourier Transform (DFT) is defined as:

$$X_m(k) = \sum_{n=-\infty}^{\infty} w(n - mD)x(n)e^{-j\omega_k n} \qquad (1)$$

where m is the frame index, $w(n)$ is a causal window of length T (i.e., zero-valued outside the range $0 \leq n \leq T-1$), D is the number of samples between successive analysis frames (the step size, with $D \leq T$), and $\omega_k = \frac{2\pi k}{L}$, where L is the number of analysis frequencies being considered in the DFT (with $L \geq T$).

B. Selection of Processing Parameters

The above equation can be interpreted as shifting a short-time window function $w(n)$ through progressive D-sized delays over the signal $x(n)$, to obtain successive T-length frames for analysis. Implementation therefore depends on appropriate choice of the window function, the step size (that is, the frame rate) and the frame length.

The window function $w(n)$ is necessary to impose a finite extent on the signal being analysed. Important considerations include minimising spectral leakage through effective tapering (enforcing periodicity in the window length) and the ability to resolve frequency components (effective window bandwidth and sidelobe level). Different windowing functions are analysed in [25]. While windows with smooth tapering, such as the Hamming window, are commonly used in analysing the magnitude spectrum over short-time segments, some studies have shown that the rectangular window is more appropriate when analysing the phase domain [26], [27], [12].

Speech is generally considered to be approximately piecewise-stationary over a period of approximately 20 milliseconds; however it is noted that some sounds are stationary over longer or shorter durations, or may be non-stationary. The choice of analysis frame (that is, window) length T is a trade-off between desired frequency resolution and temporal resolution - a longer frame gives better frequency resolution, however possibly at the expense of blurring out more rapid speech events. A window duration of between 16-32 ms is often used in analysis of the speech magnitude spectrum as an effective balance between these practical considerations.

Most studies on the relevance of phase domain information in the speech processing literature have however concluded that longer analysis frames are required than typically used in magnitude domain analysis. For instance, perception tests have repeatedly shown that intelligibility of phase-only stimuli improves over magnitude-only stimuli as the window duration extends from 100-1000 ms, while the converse is true for shorter frames [28], [12], [29]. According to a range of studies on English speech, phonemes typically vary from 50 to 200 milliseconds, syllables from 50 to 500 milliseconds, and words from 80 to 850 milliseconds [30], [3]. Further, it has been shown that inter-word pauses during conversational speech can vary from 100 to 1000 milliseconds [30]. Work modelling the speech signal as the birth and death of individual sinusoidal components, each lasting several short-term frames, suggests that significant phase variations over time and frequency may be produced by these underlying speech units [17], [18].

In selecting the step-size D, consideration must first be given to the bandwidth of the window function and the analysis frame length. For instance, for a Hamming window of length T, the step-size should be less than or equal to $T/4$ to avoid aliasing [23]. This places an upper bound on the step-size, or equivalently a lower bound on the frame rate (although it is common practice to implement a step size of T/2). It is of course necessary that the frame rate be high enough to capture the temporal dynamics in the signal. Following the Nyquist sampling principle, the assumption that speech is quasi-stationary over 20 ms segments motivates new frames being taken at 10 ms intervals, that is at 100 frames per second. This frame rate is commonly used in speech analysis applications. While the preceding paragraph motivated longer analysis frames for the phase domain, a similar 100 Hz frame rate (step size) is still motivated in order to effectively sample these variations in the speech signal.

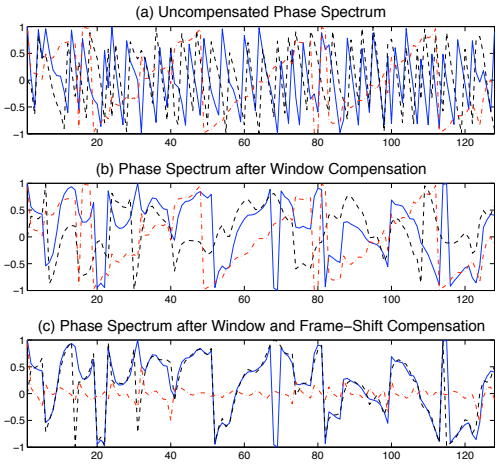


Fig. 1. Plot of the Phase Spectrum for successive input audio frames for (a) uncompensated case, (b) when compensation has been made for the analysis window, and (c) when both window and frame-shift compensation using Equation (3) have been applied to the DFT. The solid blue line shows the current frame, the dashed black line shows the previous frame, and the dot-dashed red line shows the phase difference between these (calculated using the phase of the quotient, to avoid wrapping effects). A window and FFT length of $T = L = 128$ and a step size of $D = 8$ were used on a signal at $F_s = 16$ *kHz*.

C. Compensation for the Analysis Window

It is first necessary to understand the effect of the windowing function on the phase. A description of this effect and a compensation method can be found in [18, Section 9.3.3], summarised here for convenience.

Because the window function is commonly symmetric about its mid-point, and this is aligned with the mid-point of the current frame in practical implementation of the analysis procedure, it has a linear Fourier transform phase of $\omega_k T/2$. A simple way to compensate for its effect is to implement a circular shift of the windowed signal in the time-domain prior to the Fourier transform. Specifically: take the m^{th} input frame $x_m(n)$ of length T, apply the window function $w(n)$, zero-pad as necessary to the FFT length L, then circularly shift the frame by $T/2$ samples such that the latter half of the frame occupies the range $0 \leq n \leq (T/2)$ and the earlier half occupies the range $L-(T/2) \leq n \leq L-1$. Following this, the Fourier transform can be taken and the inter-frame time step compensated as above. An alternative implementation may instead compensate the phase modulation in the frequency domain.

The effect of this window compensation on a single analysis frame is shown in Figure 1(a)-(b).

D. Compensation for Inter-frame Time Step

To implement the short-time DFT in practice, typical speech signal analysis multiplies the window function by T samples from the signal to obtain an analysis frame. A length L DFT is then taken to obtain the spectrum for analysis. Subsequent analysis frames are obtained by shifting the input signal by D samples - that is, discarding the first D samples from the T-length buffer, and appending D new samples at the end.

This procedure in effect implements:

$$\tilde{X}_m(k) = \sum_{n=-\infty}^{\infty} w(n)x(n+mD)e^{-j\omega_k n}, \qquad (2)$$

sometimes referred to as the Running Short-time Fourier Transform (RSTFT) [31]. That is, it is the signal that is effectively being shifted (through progressive advancements) past the fixed window, rather than the window being shifted over the signal. This distinction is important when considering the absolute time origin for each analysis frame. In (1), the temporal origin of each signal frame remains the origin of the original signal $x(n)$. In (2), however, the absolute position of the frame within the original sequence is discarded, by redefining the temporal origin of the frame as mD.

This has no implications for applications that only consider the magnitude spectrum. Further, when the short-term Fourier analysis is being done prior to re-synthesis of the signal, such as when using the overlap-add method to implement frequency-domain filtering, this difference is eventually compensated for by time-shifting the synthesised frame to its correct position in the output.

In applications that seek to analyse the phase spectrum across multiple frames, however, the above distinction has important consequences. Direct inter-frame comparisons of phase values calculated in this way are invalid, due to the changing reference. By accounting for the effect of the step size, however, it is possible to restore a common reference point to the phase values for every frame, allowing meaningful analysis and modelling of the phase information over time.

If we therefore wish to compare the phase spectrum between frames, it is straightforward to see that (1) is related to (2) by

$$X_m(k) = \tilde{X}_m(k)e^{-j\omega_k mD} \qquad (3)$$

Returning to the typical procedure for obtaining the spectrum of successive analysis frames, applying (3) following the DFT compensates the phase spectrum to restore a common reference. The effect of correctly compensating for both the window function and the frame time step is shown in Figure 1(c).

III. The Delta-Phase Spectrum

Two problems still remain with the phase spectrum following the above compensation: the lack of common temporal reference between different sequences, and possible ambiguities arising from phase wrapping. Absolute values of the phase spectrum have little meaning without a common reference point: phase values are by nature relative. While phase values within a single sequence can be compared once they have been compensated back to a virtual zero reference time, comparison of phase values across sequences is problematic due to the arbitrary start-point for the windowing. When developing statistical models of the behaviour of the phase spectrum, therefore, it is necessary to somehow restore some common reference to the values.

A. Review of Spectral Representations based on Instantaneous Frequency

One means of achieving a consistent phase-domain quantity for analysis and modelling is to calculate the temporal derivative, commonly referred to as the Instantaneous Frequency (IF). A number of works in the literature have investigated spectrographic representations which plot the magnitude spectrum with the bin location on the time and frequency axes reassigned according to the instantaneous frequency and group delay [32], [31], [33]. A technical and historical review relating these different approaches is presented in [32], including algorithms for practical digital estimation of the IF using the STFT. One common method for computing the IF without explicit differentiation is to first calculate the IF Deviation as the imaginary part of the ratio between two STFT's, one calculated using the standard window and one calculated by replacing the window with its derivative [31], [34], [33]. The IF can then calculated by compensating this for the centre frequency of each component. Following [35], this is referred to as the Auger and Flandrin method in [32].

Recently a spectrographic representation that is instead based directly on the Instantaneous Frequency Deviation was proposed in [11]. In that work, the IF was first calculated as the phase difference between two successive STFT's calculated with a single sample increment, following [36] and referred to as the finite difference approximation for IF in [32]. Adapting [11], [31] to the notation from the preceding section, let us redefine the short-term Fourier transform in terms of the starting sample q of the frame m (ie $q = mD$), rather than implying this from the frame index m:

$$\tilde{X}(k,q) = \sum_{n=-\infty}^{\infty} w(n)x(n+q)e^{-j\omega_k n}, \tag{4}$$

The Instantaneous Frequency can then be calculated as [11], [32]:

$$\mathbf{v}(k,q) = \arg\left[\tilde{X}(k,q)\tilde{X}^*(k,q-1)\right], \tag{5}$$

where $(\cdot)^*$ indicates the complex conjugate. The Instantaneous Frequency Deviation can be calculated as [16], [31], [11]:

$$\begin{aligned} \boldsymbol{\psi}(k,q) &= \mathbf{v}(k,q) - \omega_k & (6)\\ &= \arg\left[\tilde{X}(k,q)\tilde{X}^*(k,q-1)e^{-j\omega_k}\right], & (7)\end{aligned}$$

Having observed that the IF tracks its harmonic frequency more accurately as the corresponding spectral magnitude increases (ie, IF deviation is inversely proportional to magnitude), the Instantaneous Frequency Deviation Spectrum $\boldsymbol{\alpha}$ was then defined as [11]:

$$\boldsymbol{\alpha}(k,q) = |\boldsymbol{\psi}(k,q)|^{-1} \tag{8}$$

B. Delta-Phase Spectrum

Instead of analysing the instantaneous phase derivatives over single sample intervals, this paper proposes a related representation based on the phase difference between successive frames separated by a step-size time interval. In a similar manner to the Instantaneous Frequency Deviation above, it can be simply calculated as the phase of the ratio of successive complex spectral values:

$$\Delta\phi_m(k) = \arg\left(\frac{X_m(k)}{X_{m-1}(k)}\right) \tag{9}$$

where the use of X from Equation 1 rather than $\tilde{X}$ reflects the fact that the spectrum has been compensated for the inter-frame time step and analysis window to implement the Fourier Transform with a fixed time basis (as described in Section II). Because the phase modulation introduced by the analysis window will be the same for all frames, and will thus be cancelled out during the division, it can be seen that the Delta-Phase Spectrum may simply be implemented as:

$$\begin{aligned}\Delta\phi_m(k) &= \arg\left[\frac{\tilde{X}_m(k)e^{-j\omega_k mD}}{\tilde{X}_{m-1}(k)e^{-j\omega_k(m-1)D}}\right]\\ &= \arg\left[\tilde{X}_m(k)\tilde{X}^*_{m-1}(k)e^{-j\omega_k D}\right] \qquad (10)\end{aligned}$$

where $\tilde{X}$ is the uncompensated short-time Fourier spectrum.

To facilitate direct comparison, the Instantaneous Frequency Deviation from the preceding section may be restated with the frame index m explicit as:

$$\boldsymbol{\psi}_m(k) = \arg\left[\tilde{X}(k,mD)\tilde{X}^*(k,mD-1)e^{-j\omega_k}\right], \tag{11}$$

while the Delta-Phase Spectrum implements:

$$\Delta\phi_m(k) = \arg\left[\tilde{X}(k,mD)\tilde{X}^*(k,mD-D)e^{-j\omega_k D}\right], \tag{12}$$

These expressions are equivalent in the limiting case of $D = 1$ (i.e., a single sample step), however the latter is more general by using the processing parameter D for the time change interval. While the mathematical difference is minor, different information is being captured: rather than estimating the phase derivative at the start time of each individual frame, the simple change in phase between frames is measured. Because the "instantaneous" derivative is not measured, to avoid confusion with the Instantaneous Frequency as commonly calculated and used, the less constrained term Delta Phase is adopted in this article. By adopting a distinct term, it is intended to encourage new interpretation and insights by moving away from the constraint of single sample intervals. The term Delta-Phase draws a clear analogy with delta coefficents commonly used in speech feature vectors derived from the magnitude spectrum [37]. A further computational difference is that two STFTs per frame must be calculated to obtain a spectrogram or derive features from the IF Deviation Spectrum [11] in a standard sliding window procedure, while by reusing the previous frame's FFT in calculating the phase change the Delta Phase Spectrum requires only one.

Rather than calculating phase differences over time, as above, a similar approach in [8] effectively enforced a common reference using a particular frequency bin in the Fourier transform. Such a method however requires an arbitrary frequency bin to be selected (chosen to be $\pi/4$ in [8]), which may or may not provide a robust reference depending on the vocal characteristics of the speaker and the spectral characteristics of the noise.

Finally, it is noted that as well as providing a consistent basis for comparison over different times and sequences, a representation based on the change in phase over a given time allows the issue of phase wrapping to be controlled, as discussed in Section IV.

C. Mel-Frequency Delta-Phase (MFDP) Features

In order to model the speech signal, it is often necessary to extract a pertinent set of features from the raw spectral representation. This section presents one such feature set that may be derived from the Delta Phase Spectrum. The intention is to be illustrative rather than optimal in any sense: clearly other feature representations are possible.

The Mel-Frequency Cepstral Coefficients (MFCC) have proven to be an effective choice of speech features derived from the magnitude spectrum [21]. The MFCC features are formed by first extracting filter bank energies using a bank of band-pass filters on the absolute magnitude spectrum. The filter bank design is inspired by the critical band filtering of the human auditory system [38]. Cepstral coefficients are then derived from these by taking the logarithm of filter bank energies and performing a Discrete Cosine Transform (DCT). The cepstral processing implements a homomorphic transformation, effectively mapping convolutive effects in the original time domain into additive effects in the cepstral domain [39], [40].

This paper proposes extracting Mel-Frequency Delta-Phase Cepstral Coefficients (MFDP) by performing the same operations on the absolute delta-phase spectrum $|\Delta\phi_m(k)|$ from (9), rather than the magnitude spectrum. For these features, the absolute operator is used to measure the amount of change in the phase within each frequency bin without concern for the polarity of this change. It can be seen that the logarithm following filter bank analysis is not strictly motivated for the same reason in the phase domain as in the magnitude domain, as the phase angle is effectively already in the log domain. The logarithm on the filter bank output does however have a second effect in practice: being akin to the application of a soft maximum operation, it effectively emphasises the peak values within each frequency band. In order to avoid smoothing out peaks from the delta-phase spectrum following the filter bank analysis, the logarithm is therefore maintained in the proposed MFDP feature extraction.

An important practical consideration in developing phase domain features is selection of the parameters for the short-time Fourier analysis. Following the rationale presented in Section II-B, in extracting the MFDP features in this paper, a rectangular window function is used on frames of 256 ms duration at a rate of 100 frames per second (i.e., a 10 ms step size). Note that such a frame length corresponds to the analysis interval typically used in other works modelling temporal dynamics of the speech signal [1], [2], [3], [4], [5]. Further motivation for a longer analysis window is the desire to detect phase changes in individual harmonic components in the signal, and thus measure FFT bins that are as narrow as is practical.

IV. Effect of Time Interval on Observed Phase Change

As shown in the previous section, the Delta-Phase extends on the Instantaneous Frequency Deviation by removing the constraint of being a strict instantaneous phase derivative and instead capturing coarser phase changes over longer step-sized intervals. The main impact of increasing the time interval is to broaden the distribution of phase change that may be observed. Two conflicting effects of this are to improve the ability to detect sudden changes in the phase while also introducing the possibility of phase wrapping. This section considers the influence of the step size and FFT length parameters in this context, and presents histograms and spectrograms produced on a sample speech sequence using different parameters. As the closest work from the literature, particular comparison is made between the settings used for the IF Deviation Spectrum [11] and those used for the experiments in the current article.

A. Ability to Detect Sudden Phase Changes

As the phase measurement from the FFT is in some sense an average measure over the frame duration T (here and in the following, assume an FFT length of $L = T$ is used), it becomes increasingly difficult to detect material event-based changes in the phase as D decreases. Consider the case when a substantial change in the signal has occurred in the new D samples from one frame to the next due to some underlying physical event, such as a new speech unit being produced. Following a sinusoidal analysis model for speech [17], [19], [20], this may give rise to the birth or death of a sinusoidal component, causing a rapid shift in the phase of a particular frequency component, rather than a slowly varying modulation in its IF.

The ability of such a rapid phase shift to influence the phase change as measured using two FFT frames depends on the ratio of D to T, as well as the windowing function. For the IF Deviation Spectrum [11], only one of these T samples changes between the two FFTs used to calculate the phase derivative for each frame, and then a D sized step is taken before measuring this again. For example, with $T = 512$, this means that 99.8% of the two frames being compared in each IFD measurement are the same samples, and so any sudden phase change that occurs over a small range of samples in the physical signal will undergo a significant averaging effect, hampering the ability to detect such phase discontinuities. Any sudden phase change will be further smoothed according to the tapering of the window function in the time domain.

The ability to measure sudden phase transitions in the physical signal using phase differences between two FFT's may therefore be expected to improve as the proportion of new samples between the two FFT's (that is, the ratio of D to T) increases, motivating the use of the more general D-sample time delta used in calculating the Delta-Phase Spectrum in the current article.

B. Phase Wrapping Considerations

Contrasting with the above motivation for increasing the interval for calculating phase change is the possibility of introducing phase wrapping. To understand this, let us commence by considering the IF Deviation. IF Deviation measures how the phase-derivative changes relative to the centre frequency of a given FFT bin. Following the filter-bank analogy of the FFT, and neglecting for the moment leakage across bins due to the non-ideal window response, the possible change in the IF relative to the bin centre frequency (ω_k rad/s) is necessarily limited by the bin width. For a frame of length T and taking an FFT length of T, the frequency bandwidth of an ideal individual FFT bin is $\Delta\omega = \frac{2\pi F_s}{T}$ rad/s. If the instantaneous frequency lies outside the range $\omega_k \pm \frac{\pi F_s}{T}$, the component would instead occur predominantly in a neighbouring FFT bin. Over a given time interval, say D samples, the phase change that may be observed in a given frequency bin at the end of the interval is therefore limited to $\pm\frac{\pi D}{T}$ with respect to the phase observed at the start of the interval. For the IF Deviation Spectrum in [11], $F_s = 16000$, $D = 1$ and $T = 512$ were used. In this case, the limiting phase change is $\approx \pm 0.002\pi$ radians for the IF to be within $\pm$ 15.6 Hz of the FFT bin centre frequency, and thus still fall within that bin. Phase wrapping therefore will not occur for the IF Deviation, and this will continue to be the case for the Delta-Phase as long as the interval $D < T$. Considering the parameters used in experiments in the current article, $F_s = 16000$, $D = 160$ and $T = 4096$, the limiting phase change within a bin is $\approx \pm 0.04\pi$ radians.

In practice, due to windowing, the above simplified analysis will not strictly hold: the windowing main lobe width and sidelobes mean that a given component will have some influence over a range of frequency bins. Extending the above analysis, it may be seen that some phase wrapping can theoretically occur for a given frequency component in distant frequency bins that are more than $B = T/D$ bins away from the local bin for that component. The potential influence of such wrapping will depend therefore on the windowing sidelobe level at this frequency bin shift, as well as the strength of the the local frequency component in those bins. For a given frequency component, as long as $D < T$ phase wrapping will only occur as noise in distant FFT bins, progressively affecting less bins and attenuated by the sidelobe level of the windowing as D/T decreases. IF Deviation represents the limiting case of $D = 1$, in which phase wrapping will not occur within the T FFT bins, although general sidelobe leakage may still introduce noise to the measured phase change in each bin.

C. Empirical Analysis

From the above analysis it is apparent that the ratio of the step size to frame length provides a design parameter controlling the observed distribution of delta phase values, playing off the ability to detect sudden phase shifts with the possibility of introducing noise from phase wrapping in distant FFT bins.

1) Phase Change Histograms: To corroborate this analysis, Figure 2 plots the histogram of delta phase values obtained from a sample speech sequence for six different values of D/T with T fixed at 512 samples. The delta phase is approximately uniformly distributed over $\pm\pi$ at $D/T = 1$, and as this ratio decreases the values become more normally distributed with decreasing variance. Case (a) shows the distribution for D=1, as used to calculate the IF Deviation Spectrum in [11]. The settings used in the experiments in this paper ($D/T = 0.039$ using a longer frame of $T = 4096$ samples) corresponds to a setting between cases (b) and (c). This setting was chosen to improve the ability to detect sudden phase shifts while minimising the ability of phase wrapping to significantly affect the measurements.

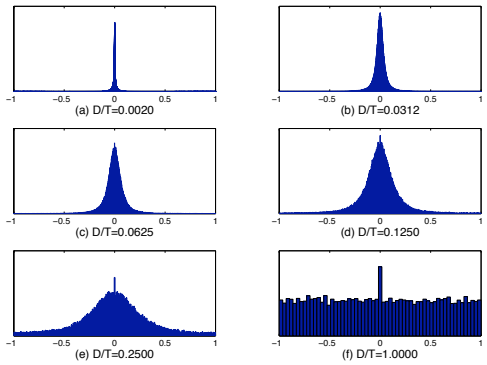


Fig. 2. Histogram of Delta-Phase values (normalised by π radians) for a sample speech sequence (Fs=16 kHz) corrupted by 10dB noise for varying values of D/T, with L=512 in each case. The speech segment consists of a male utterance of *"The decking is quarter-inch mahogany marine plywood"* from the TIMIT database [41]

2) Spectrographic Comparison with IF Deviation: Figures 3-5 demonstrate the effect of different processing parameters on the Delta-Phase Spectrum. In each case the original signal, standard magnitude spectrum and the IF Deviation spectrum [11] are shown for comparison. To facilitate interpretation in terms of phase change, a minor difference is that the absolute IF Deviation is used directly here, rather than its reciprocal as proposed in [11].

For a direct comparison with the IF Deviation spectrum presented in the literature, Figure 3 uses processing parameters taken from the example in [11]. In this case, a Chebyshev 50dB window of length $T = 512$ (32 ms) and a $D = 64$ step size (4 ms) is used. These setting are well suited for calculating the magnitude spectrum, as shown in Figure 3(b). For the IF Deviation spectrum in Figure 3(c), a single sample interval is used to calculate the deviation, while for the Delta-Phase spectrum in Figure 3(d) the step size $D = 64$ is used. It is apparent that the magnitude and phase representations are correlated, with regions of high magnitude often corresponding to regions where there is little change in phase, and vice-versa. As might be expected following the histogram analysis in the preceding section (in effect, $D/T = 0.1250$ is used here), the spectrogram for the Delta-Phase shows a distribution of values with greater variance than the IF Deviation. This appears as a more noisy spectrographic representation that makes the finer structures of the speech less evident for the Delta-Phase than

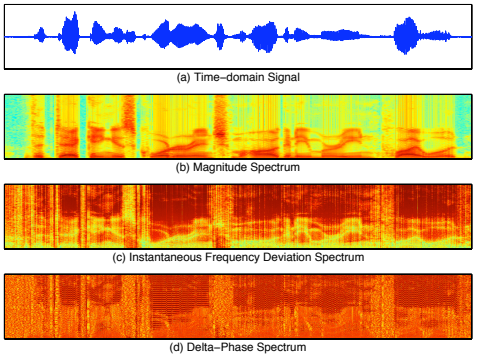


Fig. 3. Sample audio sequence in the (a) Time-domain and spectrographic representations of the (b) Magnitude spectrum, (c) Instantaneous Frequency Deviation spectrum and (d) the Delta-Phase spectrum, using parameters following [11] to facilitate comparison (Fs=16000, T=512, D=64, Chebyshev 50 dB window). For (b)-(d) the y-axis shows increasing FFT bin index (i.e., increasing frequency) and the x-axis shows increasing frame index. The speech segment consists of a male utterance of *"The decking is quarter-inch mahogany marine plywood"* from the TIMIT database [41]

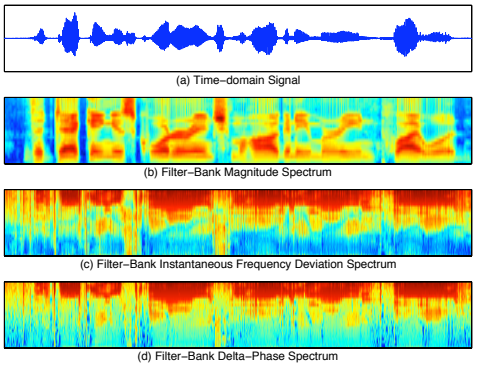


Fig. 4. Sample audio sequence in the (a) Time-domain and Mel-scaled Filter-bank spectrographic representations of the (b) Magnitude spectrum, (c) Instantaneous Frequency Deviation spectrum and (d) the Delta-Phase spectrum, using parameters following [11] to facilitate comparison (Fs=16000, T=512, D=64, Chebyshev 50 dB window). For (b)-(d) the y-axis shows increasing filter-bank index (i.e., increasing frequency) and the x-axis shows increasing frame index.

the IF Deviation.

Figure 4 shows the same sequence using the same processing parameters as Figure 3, however the output of 24 Mel-scaled filter-banks are shown in place of the raw FFT bins. This figure serves simply to illustrate that despite the finer differences between the IF Deviation and Delta-Phase spectrum in Figure 3, when considering Mel-scaled filter-bank outputs as used in extracting features, these differences are less evident.

The processing parameters used in Figures 3-4 are however not appropriate for the motivations of the Delta-Phase proposed in the present article, which is to detect significant event-based shifts in the phase of sinusoidal components. The short frame length leads to wider FFT bins than those desired to focus on individual harmonic components, and the

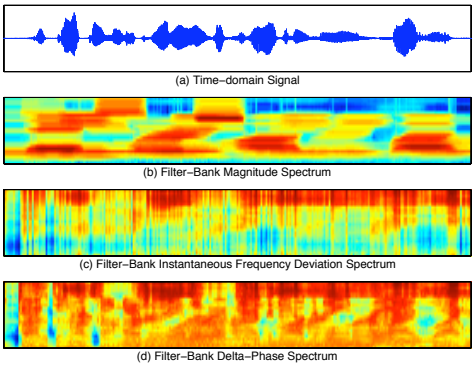


Fig. 5. Sample audio sequence in the (a) Time-domain and Mel-scaled Filter-bank spectrographic representations of the (b) Magnitude spectrum, (c) Instantaneous Frequency Deviation spectrum and (d) the Delta-Phase spectrum, using parameters as used in subsequent experiments in the current article (Fs=16000, T=4096, D=160, Rectangular window). For (b)-(d) the y-axis shows increasing filter-bank index (i.e., increasing frequency) and the x-axis shows increasing frame index.

use of a Chebyshev window sacrifices some ability to detect discontinuous event-based transitions in the signal (albeit, while offering lower sidelobe levels). Figure 5 shows the same sequence using the processing parameters used in the experiments in the current article; that is, a rectangular window of length $T = 4096$ (256 ms) and a $D = 160$ step size (10 ms). As in Figure 4, the output of 24 Mel-scaled filter-banks are shown. It is apparent from Figure 5(b) that the longer analysis window is not an appropriate choice for the magnitude domain. The IF Deviation in Figure 5(c) also shows little information with these processing parameters, as might be expected given that only 1 sample of 4096 has changed in the two FFTs being used to calculate the deviation (that is, 0.024% of the frame). The larger frame size examined here means that in practice averaging affects will hinder the ability to detect any phase change over a single sample interval, particularly at low frequencies.

In contrast, Figure 5(d) confirms that by using a longer frame and increasing the ratio of D to T, the Delta-Phase Spectrum is capturing regions of both high and low phase change in the signal over both time and frequency. These patterns reveal interesting structure in the underlying signal that appear complementary to the information traditionally extracted from the magnitude spectrum, as shown in Figure 4(b). For the Delta-Phase Spectrum in Figure 5(d), 160 of the samples are changing from frame to frame (that is, 3.9% of the frame). This step size (10ms) also has the benefit of matching that commonly used in magnitude domain feature extraction, facilitating fusion of magnitude (MFCC) and phase (MFDP) domain systems in the following VAD and speaker recognition experiments.

V. Application to Voice Activity Detection (VAD)

In order to validate the proposed Delta-Phase Spectrum and Mel-Frequency Delta Phase features derived from it, a first set of experiments was conducted applying the features for

TABLE I
DATABASE NOISE TYPES AND SCENARIOS. THESE FORMAL PARTITIONS OF THE DATABASE ARE REFERRED TO IN THE TEXT USING ITALICISED LABELS, SUCH AS *Street-City*

Type	Scenario 1	Scenario 2
Street	City	Suburb
Car	Windows Up	Windows Down

a simple voice activity, or speech/non-speech, detection task. Note that the goal of these experiments is simply to validate the proposed phase representation, rather than to achieve state-of-the-art VAD performance.

A. Database

In order to evaluate the proposed MFDP speech features for the purposes of voice activity detection, a database of 240 hours of noisy speech over 9600 individual files was constructed through a combination of clean speech and real-world noise recordings. A comprehensive description of the database is available in [42], with relevant details summarised here.

In order to construct the voice activity detection database, two real-world recordings of at least 30 minutes of typical background noise were made in each of 4 scenarios, covering two broad noise types, *Car* and *Street*, as shown in Table I. The two recordings for each scenario were captured at similar times on separate days to ensure adequate temporal difference in the environments. In addition to the noise recording itself, 6 swept-sine sweeps were recorded in the *Car* scenarios in order to allow the reverberant response to be estimated, such that speech may be inserted as if it were captured in that scenario.

For each of the two recordings in each scenario, 200 noisy speech sequences, equally split between lengths of 60 and 120 seconds, were constucted for each of 6 signal-to-noise ratios (SNRs), being -10 dB, -5 dB, 0 dB, 5 dB, 10 dB and 15 dB. These noisy speech sequences were constructed by extracting a random section of the noise recording of the appropriate length and adding clean speech sequences chosen randomly from the TIMIT speech database [41] at the desired SNR. For the *Car* scenarios, the clean speech sequences were first transformed to match the estimated reverberant response of the noise recording. In order to ensure that speech energy is consistent between files, the SNR mixing was performed by adding the inserted speech sequences with an active speech level of -26 dBov (dB overload, following [43]), after first scaling the background noise to match the desired SNR. As the database sequences were constructed, the ground-truth timing for speech events is known precisely for evaluation of VAD algorithms.

B. System Description

A Gaussian Mixture Model (GMM) based speech detection system was used to evaluate the MFDP speech features in comparison to standard MFCC speech features. GMM-based systems using MFCC features have been shown to provide a robust baseline solution across a range of speech classification problems, including speech/non-speech detection [44], [45]. By extracting features from sub-band energies and learning statistical models over training examples, a GMM-MFCC system provides a higher performance baseline than more traditional VAD systems based on thresholding features such as broadband frame energy.

The MFCC and MFDP features used in the experiments were 13-dimensional cepstral coefficients, including the zero'th coefficient, and with first-order regression coefficients appended (i.e., traditional "delta" features such as in [37]), making a 26-dimension feature vector in each case. The MFCC features were calculated using a standard 25 ms Hamming window, with a 10 ms step size (that is, a rate of 100 fps), while the MFDP features used a 256 ms rectangular window, also using a 10 ms step size, following the rationale presented in previous sections. While larger step sizes could be examined for the MFDP papers, to facilitate comparison and fusion with MFCCs, the 10 ms step size was maintained in all experiments.

These speech detection experiments were operated under the assumption that the broad SNR of the target environment is known, but the specific scenario, or type of noise, is not known. To this end the six noise levels were divided into three groups covering two SNRs each, designated as the high (-10 dB, -5 dB), medium (0 dB and 5 dB), and low (10 dB and 15 dB) broad noise levels.

To train speech detection modules under the operating assumption provided, speech and non-speech GMMs were trained based on the known ground truth on one set of scenarios across both noise types, and for each of the three broad noise levels. The other set of scenarios for each of the three broad noise levels was then used to calculate speech scores by taking the difference between the log-likelihoods given for each feature vector by the speech and non-speech GMMs. To give an example for sake of clarity, the low noise *Street-City* data was tested on models trained using low noise data from *Street-Suburb* and *Car-Windows Down*.

The speech scores obtained in this way were then smoothed by a 1-second median filter centred on each feature vector to attenuate short-term variation in favour of the longer term. The MFCC+MFDP results indicate multi-stream fusion of the MFCC and MFDP, in which the log-likelihoods of each stream were combined using addition (equally weighted) prior to the smoothing median filter.

Speech and non-speech segmentation decisions were made by comparing the smoothed speech scores to a threshold. This threshold was estimated by minimising the half total error rate, calculated as the average of the miss and false-alarm rates, on a held-out tuning data set. These tuning scores were calculated similarly to the test scores, but were calculated on the same set as the GMM parameter estimation, to ensure the final testing set is unseen to both the GMM training and threshold tuning. To produce unbiased results for each noise type, the complete results were generated using 2-fold training and testing, split according to the scenario numbers indicated in Table I.

C. Results

Results from the voice activity detection experiments are presented in Tables II and III for the *Car* and *Street* noise

TABLE II
VAD RESULTS FOR CAR NOISE CONDITION

SNR	Features	FAR	MR	HTER
10 to 15 dB	MFCC	2.3%	1.3%	1.8%
	MFDP	3.4%	1.3%	2.3%
	MFCC+MFDP	2.6%	1.0%	1.8%
0 to 5 dB	MFCC	2.6%	2.7%	2.6%
	MFDP	4.6%	1.6%	3.1%
	MFCC+MFDP	3.5%	1.1%	2.3%
-10 to -5 dB	MFCC	3.8%	8.9%	6.4%
	MFDP	7%	8.7%	7.8%
	MFCC+MFDP	7.4%	2.1%	4.7%

TABLE III
VAD RESULTS FOR STREET NOISE CONDITION

SNR	Features	FAR	MR	HTER
10 to 15 dB	MFCC	2.4%	1.7%	2.0%
	MFDP	3.4%	1.5%	2.5%
	MFCC+MFDP	2.5%	1.3%	1.9%
0 to 5 dB	MFCC	3.0%	6.6%	4.8%
	MFDP	4.2%	4.1%	4.2%
	MFCC+MFDP	3.4%	2.7%	3%
-10 to -5 dB	MFCC	6.6%	23%	14.8%
	MFDP	5.1%	17.8%	11.5%
	MFCC+MFDP	8.6%	8.9%	8.8%

types, respectively. Results are presented in terms of percentage False Alarm Rate (FAR), Miss Rate (MR, equivalent to False Rejection Rate) and Half-Total Error Rate (HTER). These demonstrate performance at a particular operating point, selected to optimise HTER on the training data as explained above. To show performance across a range of operating points, the Detection Error Trade-off (DET) plot is shown in Figure 6 [46].

These results show similar performance is achieved using the MFCC or MFDP features. In *Car* noise, the MFCC's show a marginal improvement over MFDP's, and this trend is reversed for the *Street* noise. Without seeking to over-interpret these results, it may be that the MFDP's show benefits in less stationary noise environments due to the longer analysis window of 256 ms. Further, the DET plot in Figure 6 shows that the MFCC features perform better in the high Miss Rate region (high False Rejection), while MFDP features exhibit better performance in the high False Alarm region.

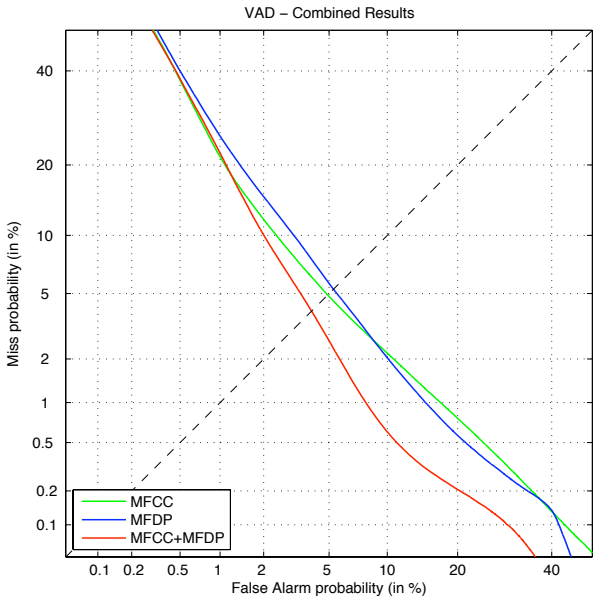


Fig. 6. DET plot of GMM-based VAD results over all noise conditions and levels, using MFCC and MFDP features and their multi-stream fusion.

The two important points to garner from these results are: first, that results using only phase information are comparable to those using magnitude only; second, that in both noise types, and at all noise levels, the multi-stream fusion of the magnitude-domain MFCC's and phase-domain MFDP's yields significant performance benefits over either in isolation, as measured by the HTER. Note that while the FAR increases marginally over MFCC in the fused system, the MR is significantly improved in each case, simply reflecting the fact that the operating point is chosen based on HTER.

VI. APPLICATION TO SPEAKER RECOGNITION

The effectiveness of the proposed MFDP features for voice activity detection was demonstrated in the preceding section, with the fusion results highlighting the complementary information they offer to MFCC features. This section seeks to further validate the proposed phase representation by investigating whether MFDP features are also able to capture speaker discriminative information from the phase domain through their application to the task of speaker recognition. As in the preceding section, the goal of these experiments is to validate the proposed phase representation, rather than to demonstrate state-of-the-art performance.

Speaker recognition is commonly performed using cepstral-based features derived from the magnitude domain. Particularly successful in this research domain are MFCC features. MFCCs provide state-of-the-art speaker recognition performance when used in conjunction with GMMs adapted from a Universal Background Model (UBM) and suitable session variability compensation techniques [47], [48]. Limited research has focussed on the application of phase-related features to speaker recognition due to the belief that the phase component of speech offers little information relative to the magnitude domain [49], [50], [51].

A. Experimental Configuration

The comparison of MFDP and MFCC features will be conducted following the well-known NIST Speaker Recognition Evaluation (SRE) series [52] protocols. Since 1996, NIST have conducted regular evaluations of speaker recognition technology by specifying an evaluation protocol and corresponding corpus predominantly consisting of conversational telephony speech from several hundreds of speakers. The NIST SRE series has driven state-of-the-art in the area of speaker recognition research. For these experiments, the 2006 and 2008 NIST SRE data and protocols were used, specifically the evaluation conditions consisting of 5-minute English-only telephone conversations. This subset of evaluation conditions was selected to allow for a clear analysis of the proposed features without the need to consider additional variability

introduced from microphone, interview or cross-channel trials available in the corpora.

The two feature sets will be examined in the context of a GMM Supervector SVM system. MFCC variants of this system have demonstrated state-of-the-art performance in recent SRE's. The GMM Supervector SVM system [53] combines robust yet straightforward acoustic modelling in the form of mean-adapted high-order Gaussian mixture models (GMM) with more recent discriminative machine learning approaches through Support Vector Machine (SVM) classification.

In this approach, each utterance, in both training and testing, is first used to estimate a mean-adapted GMM through maximum a posteriori (MAP) adaptation from a universal background model (UBM). In this work, gender-dependent, 512-component UBM's are used for this purpose. This form of MAP adaptation has been a well-established approach in speaker recognition for over a decade [47]. The component mean vectors of the adapted GMM are then concatenated together to form a single large vector, known as a *supervector*; the supervector thus provides a convenient, fixed dimension representation of each utterance for use within an SVM classifier.

Speaker SVM training and classification was performed using the GMM mean supervector kernel [53]. This kernel performs a weighted dot-product between the GMM mean supervectors. Support vector machines are discriminative classifiers and thus are trained on both positive and negative examples of a speaker. In the context of NIST evaluations, there is typically only a single (positive) training example of a speaker while a substantial number of impostor (negative) examples are drawn from previous NIST evaluation corpora.

Zero and Test-norm score normalisation was applied to all scores to reduce the statistical variation observed in scores [54]. Both normalisation techniques utilise a large set of impostor speech segments to calculate a set of normalisation statistics. Zero-norm is a speaker-centric technique in which a speaker's scores are scaled by the mean μ_i and standard deviation σ_i, obtained when scoring the impostor cohort against the *speaker* model, such that,

$$\overline{score} = \frac{score - \mu_i}{\sigma_i} \tag{13}$$

Similarly, test-norm calculates μ_i and σ_i by trialling a given *test* segment against a set of speaker models trained using the impostor cohort. In this work, the impostor speech segments were extracted from the NIST 2004 dataset.

The MFDP and MFCC features for these experiments were formed from 12 cepstral coefficients with appended deltas. In contrast to features used in previous sections, the 0th cepstrum was removed from the MFDP features to match the existing MFCC configuration. This was empirically found to provide marginal improvements to MFDP-based speaker recognition. The reader is referred to [55] for more details of the configuration and implementation of the GMM Supervector SVM system used in these experiments.

Two well-established techniques for robust speaker verification were progressively incorporated into the baseline configuration described above in order to observe whether they could offer similar benefits to proposed MFDP's as they do to the magnitude-based features on which they were developed. The first technique, feature-warping [56], applies short-time Gaussianisation to the feature vector stream extracted from an utterance using a sliding window to counteract the adverse effects of channel mismatch and additive noise. A window of 5 seconds is utilised in this work. The second technique, Nuisance Attribute Projection (NAP) [57], aims to reduce the adverse effects of inter-session variation in the SVM kernel space. Inter-session variation, such as differences in channel and background noise, is well known as a major source of error in speaker recognition. NAP addresses this issue by removing the directions of greatest inter-session variation from the supervector space. Based on empirical results, forty directions were removed from the MFCC supervector space and twenty dimensions in the case of the MFDP configuration.

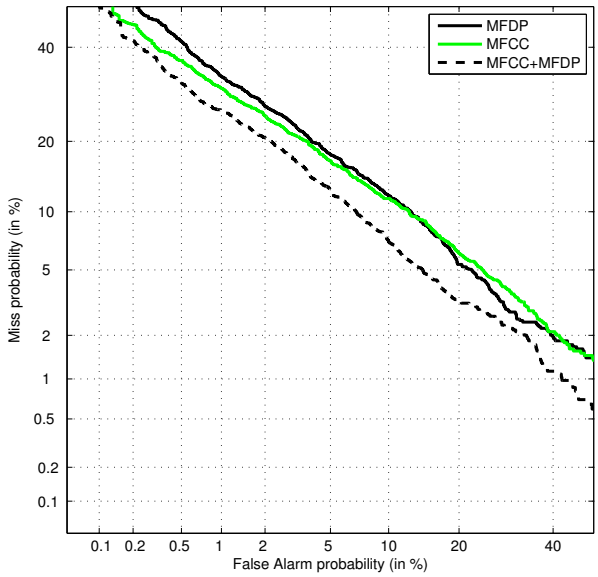


Fig. 7. DET plot of 1-sided, English-only trials from the NIST 2006 speaker recognition evaluation using MFDP and MFCC features.

Evaluation of the feature sets was performed using the English-only trials from the 1-sided training condition of the NIST 2006 and 2008 speaker recognition evaluation (SRE). Classification performance was measured in terms of minimum decision cost function (DCF) and equal error rate (EER), as defined in the NIST SRE protocol [52]. Score-level fusion was implemented using the FoCal toolkit [58] to optimise linear log regression. The fusion weights for the NIST 2006 SRE trials were learned using scores from the 2008 SRE, and similarly, 2008 SRE fusion weights were learned from the 2006 SRE scores. This approach to fusion ensured that the fused weights were not optimistically biased for a given corpus.

B. Results

Figure 7 depicts the DET curves from the English-only trials on the NIST 2006 SRE involving MFDP and MFCC features and their score-level fusion without incorporating feature-warping or NAP. The performance of the proposed

MFDP features demonstrates their effectiveness in capturing speaker discriminative information from the phase domain. It is also clear from the DET plot that the fusion of the features is highly complementary.

Table IV details specific operating performance statistics from trials on the NIST 2006 and 2008 SRE. Several configurations are presented for a thorough analysis of the proposed MFDP features; a Baseline, Feature-Warping, NAP and Feature-Warping+NAP configuration, the last of which amounts to a state-of-the-art configuration developed for MFCC features. The objective of these experiments was to observe whether techniques developed for magnitude-based features were also suited to the proposed MFDP feature set.

The baseline results in Table IV indicate that both MFDP and MFCC features provided broadly comparable performance on the NIST 2006 SRE (These results correspond to the DET curve in Figure 7). On the more challenging NIST 2008 SRE, MFCCs offered a relative gain of 13% and 15% in minimum DCF and EER, respectively, over the proposed MFDP results. Score-level fusion of the baseline MFDP and MFCC configurations resulted in a relative improvement of 22% and 11% in the EER of the 2006 and 2008 corpora, respectively, indicating that the MFDP features offer considerable complimentary information to MFCC's in the baseline configuration.

The introduction of feature-warping to the baseline system provided a relative improvement of 7-17% in MFDP performance statistics across the NIST corpora and a significant relative gain of 50% in the MFCC-based results. Interestingly, MFDP features provided little complimentary information to the feature-warped MFCC's. The large discrepancy in the gains offered by feature-warping may be explained by the relatively large window used during MFDP feature extraction. In using a 256ms window of analysis for MFDP extraction compared to 32ms for the MFCC feature stream, a relatively large correlation between sequential features is expected thereby potentially reducing the effectiveness or necessity of feature-warping. An alternative explanation can be derived from the objective of the feature-warping process. Specifically, MFDP features may be inherently more robust to channel distortions and additive noise than the MFCC feature set. This hypothesis is explored through the application of inter-session variability compensation via NAP.

The application of NAP to the baseline configurations provided significant improvements in excess of 32% (relative) to performance statistics across the NIST corpora. Similarly, the MFCC configuration obtained a 50% relative improvement over baseline results from the application of NAP. As with the baseline results, the fusion of the NAP systems provided a further gain of 23% and 13% EER over the MFCC configuration alone in the NIST 2006 and 2008 SRE, respectively.

The final configuration employing feature-warping and NAP represents a state-of-the-art SVM configuration developed on MFCC features. MFCC-based results in Table IV indicate that NAP provided an average relative gain of 34% in minimum DCF and 27% in EER across the evaluated corpora over the use of feature-warping alone. Comparably, MFDP results obtained an average relative improvement of 26% and 21% in minimum DCF and EER, respectively, from the application of NAP. Interestingly, the optimised number of nuisance directions removed via NAP was only twenty in the case of MFDP features and forty for the MFCC system. Comparable performance gains from NAP suggest that MFDP features exhibit less inter-session variation than MFCCs and allow such variation to be robustly estimated using fewer directions. Such a trait is highly desired of features for speaker verification as inter-session variability continues to be a major cause of classification error. The phase-based MFDP features provided reasonable classification performance on the SRE task with the magnitude-based MFCC features offering relative improvements of more than 50% in the evaluation of both corpora. Score-level fusion of both configurations provided the best performance statistics with relative improvements of up to 5% being obtained over the MFCC configuration in the 2006 SRE. Similar improvements were observed in the 2008 SRE trials through fusion. This demonstrates that the MFDP features extract some speaker specific information from the phase domain that is complementary to magnitude-based features.

TABLE IV
MINIMUM DCF AND EER OBTAINED FROM 1-SIDED, ENGLISH-ONLY NIST 2006 AND 2008 SPEAKER RECOGNITION EVALUATIONS WHEN USING MFCC AND MFDP FEATURE SETS FOR SVM TRAINING AND CLASSIFICATION.

	NIST 2006		**NIST 2008**	
Features	Min. DCF	EER	Min. DCF	EER
Baseline				
MFDP	.0429	11.10%	.0573	15.24%
MFCC	.0400	10.89%	.0498	12.92%
MFCC+MFDP	**.0346**	**8.45%**	**.0465**	**11.45%**
Feature-Warping				
MFDP	.0398	9.37%	.0508	12.69%
MFCC	.0188	4.55%	.0259	6.34%
MFCC+MFDP	**.0179**	**4.54%**	**.0258**	**6.18%**
NAP				
MFDP	.0273	6.46%	.0387	9.81%
MFCC	.0184	4.17%	.0245	6.02%
MFCC+MFDP	**.0180**	**3.20%**	**.0232**	**5.19%**
Feature-Warping + NAP				
MFDP	.0292	6.28%	.0398	9.29%
MFCC	.0130	2.87%	.0190	4.61%
MFCC+MFDP	**.0125**	**2.72%**	**.0185**	**4.37%**

In operating at a comparable level to the MFCC feature set in the baseline configuration, offering robustness to inter-session variation and by providing complementary information to commonly employed MFCC features, the proposed MFDP feature set shows high potential for further application in the field of speaker recognition research. Building on these preliminary experimental results, investigations into SVM kernels tailored to the MFDP feature set and their application to GMM-based classification are likely to better exploit the speaker discriminative information found in MFDP features.

VII. CONCLUSION

This paper has revisited the use of the phase domain in short-time Fourier analysis of the speech signal, highlighting the factors that must be considered and compensated before the phase can be meaningfully analysed. The Delta-Phase Spectrum computed at a frame advance rate of D samples was

proposed as a simple phase domain representation similar that allows consistent comparison over multiple frames and across sequences, while also minimising practical issues associated with phase wrapping. The Delta-Phase extends the Instantaneous Frequency Deviation, removing the constraint of being a strict instantaneous phase derivative and instead capturing coarser changes in the phase structure of the signal from one frame to the next.

Building upon this representation, it was shown that Mel-Frequency Delta Phase features extracted purely from the phase domain could be used to achieve broadly similar performance to the common magnitude-domain Mel-Frequency Cepstral Coefficients for distinguishing speech from noise, and also for distinguishing between voices of different people. Further, it was shown that principled fusion of the magnitude and phase domain information could achieve performance improvements over either in isolation.

There remains much scope for research building upon this work, both in optimising phase domain feature representations and models, and in understanding whether these findings can be applied to automatic speech recognition, which needs to capture shorter-term units than the voice activity and speaker recognition applications considered here.

ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for their valuable comments which has enabled us to improve the quality of the manuscript, as well as Dan Ellis of Columbia University for his help in clarifying latter revisions of this article. The research was supported in part by the Australian Research Council (ARC) Discovery Grant DP0877835.

REFERENCES

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[2] H. Hermansky and S. Sharma, "TRAPS-Classifiers of temporal patterns," in *Fifth International Conference on Spoken Language Processing*. ISCA, 1998.

[3] S. Greenberg and T. Arai, "What are the essential cues for understanding spoken language," *IEICE Transactions on Information and Systems*, vol. E87-D, no. 5, pp. 1059–1070, May 2004.

[4] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.

[5] S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in *Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

[6] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 1.

[7] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings.*, 2003, vol. 2.

[8] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Computer Society, 2009, pp. 4529–4532.

[9] HA Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, 2003, vol. 1.

[10] Y. Wang, J. Hansen, G.K. Allu, and R. Kumaresan, "Average instantaneous frequency (AIF) and average log-envelopes (ALE) for ASR with the AURORA 2 Database," in *Eighth European Conference on Speech Communication and Technology*. ISCA, 2003.

[11] A.P. Stark and K.K. Paliwal, "Speech analysis using instantaneous frequency deviation," *Proceedings Interspeech 2008*, pp. 2602–2605, 2008.

[12] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[13] D.C. Ghiglia and M.D. Pritt, *Two-dimensional phase unwrapping: theory, algorithms, and software*, Wiley New York, 1998.

[14] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 170–177, 1977.

[15] G. Nico and J. Fortuny, "Using the matrix pencil method to solve phase unwrapping," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 886–888, 2003.

[16] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Prentice-hall, 1978.

[17] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[18] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2002.

[19] T.N. Sainath, *Acoustic landmark detection and segmentation using the Mcaulay-Quatieri sinusoidal model*, Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[20] T.N. Sainath and T.J. Hazen, "A sinusoidal model approach to acoustic landmark detection and segmentation for robust segment-based speech recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1.

[21] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[22] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[23] JB Allen and LR Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[24] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 1, pp. 55–69, 1980.

[25] FJ Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[26] N. Reddy and M. Swamy, "Derivative of phase spectrum of truncated autoregressive signals," *IEEE Transactions on Circuits and Systems*, vol. 32, no. 6, pp. 616–618, 1985.

[27] L.D. Alsteris and K.K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," in *Proc. International Conf. Acoustics, Speech, Signal Processing*, 2004, pp. 573–576.

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

[29] MR Schroeder, "Models of hearing," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1332–1350, 1975.

[30] D. O'Shaughnessy, "Timing patterns in fluent and disfluent spontaneous speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995, pp. 600–603.

[31] D. Friedman, "Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1985, vol. 10.

[32] S.A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *The Journal of the Acoustical Society of America*, vol. 119, pp. 360–371, 2006.

[33] T. Abe, T. Kobayashi, and S. Imai, "The IF spectrogram: a new spectral representation," in *Proceedings of ASVA 97*, 2007, pp. 423–430.

[34] F. Charpentier, "Pitch detection using the short-term phase spectrum," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on.*, 1986, vol. 11.

[35] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *Signal Processing, IEEE Transactions on*, vol. 43, no. 5, pp. 1068–1089, 2002.

[36] S. Kay, "A fast and accurate single frequency estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1987 –1990, dec. 1989.

[37] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University*, vol. 1996, 1995.

[38] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, no. 1, pp. 47–65, 1940.

[39] AV Oppenheim and RW Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[40] B.P. Bogert, M.J.R. Healy, and J.W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, 1963, pp. 209–243.

[41] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NTIS order number PB91-100354*, 1993.

[42] David Dean, Sridha Sridharan, Robert Vogt, and Michael Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech*, Makuhari, Japan, September 2010.

[43] I. Rec, "830, subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[44] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2003, vol. 1, pp. 432–435.

[45] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Lecture Notes in Computer Science*, vol. 4625, pp. 509–519, 2008.

[46] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Fifth European Conference on Speech Communication and Technology*. Citeseer, 1997.

[47] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[48] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[49] H.A. Murthy and V.R.R. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003, vol. 1, pp. 68–71.

[50] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[51] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[52] National Institute of Standards and Technology, "The NIST year 2006 speaker recognition evaluation plan," 2006, Available from: http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf.

[53] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2006, vol. 1, pp. 97–100.

[54] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[55] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation techniques for SVM-based speaker recognition," in *Proc. Interspeech*, 2007, pp. 790–793.

[56] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *A Speaker Odyssey, The Speaker Recognition Workshop*, vol. 2001, pp. 213–218, 2001.

[57] A Solomonoff, C Quillen, and W.M. Campbell, "Channel compensation for SVM speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 57–62.

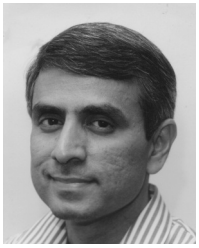
[58] N. Brummer, "FoCal: Tools for fusion and calibration of automatic speaker detection systems," July 2005.

Iain McCowan (M'97) received the B.E. and B.InfoTech. from the Queensland University of Technology (QUT), Brisbane, in 1996. In 2001, he completed his PhD with the Research Concentration in Speech, Audio and Video Technology at QUT, including a period of research at France Telecom R&D. In 2001 he joined the IDIAP Research Institute, Switzerland, progressing to the post of Senior Researcher in 2003. While at IDIAP, he worked on a number of applied research projects in the areas of automatic speech recognition, content-based multimedia retrieval, multimodal event recognition and modeling of human interactions. From 2005-2008 he was with the CSIRO eHealth Research Centre, Brisbane as Project Leader in the area of multimedia content analysis. In 2008 he founded Dev-Audio Pty Ltd to commercialise microphone array technology. He holds an adjunct appointment as Associate Professor at QUT in Brisbane.

David Dean holds Bachelor degrees in Engineering (with Honours) and Information Technology. He completed his PhD programme in 2008 in the area of audio-visual speech technology with his dissertation entitled Synchronous HMMs for Audio-Visual Speech Processing. As a post-doctoral fellow of the Speech, Audio, Image and Video Technology program at the Queensland University of Technology, his research has focused on both acoustic and audio-visual speech processing with a recent focus on speech detection and speaker verification.

Mitchell McLaren received his PhD with the Speech, Audio, Image and Video Technologies (SAIVT) at the Queensland University of Technology (QUT), Brisbane, Australia, in 2010. He received his BCompSysEng also from QUT in 2006. Mitchell has been with the Centre for Language and Speech Technology (CLST) at Radboud University Nijmegen, The Netherlands, since 2010 where he is currently in a post-doctoral role. In 2007, he was a visiting intern within the Laboratoire Informatique D'Avignon in Avignon, France. His PhD research concentrated on speaker verification using support vector machine techniques. Mitchell was awarded the 'Best Student Paper Award' at Interspeech 2008 and the 'IEEE 2009 Spoken Language Processing Student Grant' at ICASSP 2009.

Robert Vogt received his PhD degree in electrical engineering at the Queensland University of Technology (QUT), Brisbane, Australia, in 2006 and a BEng/BInfTech degrees also at QUT in 2002. Robert has been with the Speech, Audio, Image and Video Technologies (SAIVT) group at QUT since 2002 where he is currently a research fellow. His research interests include speaker recognition, speaker diarisation and spoken term detection. During his time with the SAIVT group, Robert has participated in the successful commercialisation of speech research outcomes and helped secure several large grants through competitive funding schemes. In 2008, he was invited to participate in the robust speaker recognition stream at the CSLP Summer Workshop, hosted at Johns Hopkins University, Baltimore, MD.



Professor Sridha Sridharan has a BSc (Electrical Engineering) degree and obtained a MSc (Communication Engineering) degree from the University of Manchester Institute of Science and Technology (UMIST), UK and a PhD degree in the area of Signal Processing from University of New South Wales, Australia. He is a Senior Member of the Institute of Electrical and Electronic Engineers - IEEE (USA). He is currently with the Queensland University of Technology (QUT) where he is a full Professor in the School of Engineering Systems. Professor Sridharan is the Deputy Director of the Information Security Institute and the Leader of the Research Program in Speech, Audio, Image and Video Technologies at QUT. He has published over 300 papers consisting of publications in journals and in refereed international conferences in the areas of Image and Speech technologies during the period 1990- 2010. During this period he has also graduated 24 PhD students as their Principal Supervisor and 15 PhD students as their Associate Supervisor in the areas of Image and Speech technologies. Prof Sridharan has also received a number of research grants from various funding bodies including Commonwealth competitive funding schemes such the Australian Research Council (ARC) and the National Security Science and Technology (NSST) unit. Several of his research outcomes have been commercialised.