Phase-Processing For Voice Activity Detection: A Statistical Approach

Johannes Stahl, Pejman Mowlaee, and Josef Kulmer
Signal Processing and Speech Communication Laboratory
Department of Electrical Engineering
Graz University of Technology
Graz, Austria
Email: {johannes.stahl,pejman.mowlaee,kulmer}@tugraz.at

***Abstract*—Conventional voice activity detectors (VAD) mostly rely on the magnitude of the complex valued DFT spectral coefficients. In this paper, the circular variance of the Discrete Fourier transform (DFT) coefficients is investigated in terms of its ability to represent speech activity in noise. To this end we model the circular variance as a random variable with different underlying distributions for the speech and the noise class. Based on this, we derive a binary hypothesis test relying only on the circular variance estimated from the noisy speech. The experimental results show a reasonable VAD performance justifying that amplitude-independent information can characterize speech in a convenient way.**

***Index Terms*—Voice activity detection, phase spectrum, circular variance, speech enhancement.**

I. Introduction

For robust speech applications, detection of speech presence is of high importance as an initial processing step. Voice activity detectors are an indispensable component in reliable speech communication systems as they avoid unnecessary processing of non-speech frames. Thus, a voice activity detector (VAD) is often employed as a front-end for various speech processing applications including automatic speech recognition, speaker recognition and speech coding.

Various representations and speech features have been utilized for VAD, including: energy and zero crossing rate [1], Mel-frequency cepstral coefficients (MFCCs) [2], the squared STFT magnitude [3], long-term spectral envelope [4], long-term signal variability (LTSV) [5], perceptual spectral flux [6], long-term temporal information and harmonic-structure [7], and the generalized auto-regressive conditional heteroscedasticity (GARCH) filter to model speech in time domain [8]. Further studies reported fusing multiple features using machine learning techniques, such as deep belief network [9], support vector machine [10] and minimum error classifier [11].

This paper aims to solve the voice activity detection problem in the STFT domain. Similar to many other speech processing applications the spectral phase has been neglected for voice activity detection in the last decades, VAD methods formulated in the spectral domain focused on information carried by the spectral magnitude of speech (e.g. [3]). The reason behind this is the circumstance that the instantaneous phase spectrum does not reveal any intuitive, directly accessible information about the underlying speech signal. In order to circumvent the analysis of the instantaneous phase directly several phase-derived features such as the delta-phase spectrum [12], the base-band phase difference [13], phase distortion deviation [14], group delay and modified group delay [15] have been proposed in order to characterize speech in various applications. Two methods that exploit the complex nature of DFT coefficients are the approaches presented in [12] and [16]. Wisdom et al. [16] propose a method relying on the complex domain to solve VAD. They employ the degree of impropriety (DOI) combined with a generalized likelihood ratio test (GLRT), reporting a successful discrimination between the speech plus noise and noise-only classes. Their proposed features took into account the second-order statistics of the complex data, namely the impropriety of a complex sub-band. As speech shows a higher degree of impropriety than noise it can be classified by means of this feature. Alternatively, the modulation spectrum information modeled by temporal phase changes was used for VAD in speaker recognition [12]. This approach, based on the delta-phase spectrum could successfully employ a phase-derived feature for VAD.

In the last years the discipline of phase-aware speech processing has been an emerging field. For example, some recent studies reported that phase information contributes to push the limited performance of existing solutions [17]–[19]. In this regard, we propose the circular variance of a complex DFT coefficient as a possible amplitude-independent feature for VAD. The classification is achieved by a binary hypothesis test framework. Our experiments show that the proposed VAD performs comparable to magnitude-only approaches highlighting the importance of phase information in the context of speech processing.

The rest of this paper is organized as follows; In Section II we present the underlying signal model and the circular variance as the proposed feature as well as its statistical properties. Section III explains the classification procedure itself and the evaluation of the proposed method is presented in Section IV. Finally, Section V concludes on the work.

II. Proposed phase-based feature

A. Signal Model and Notations

Let $Y(k,\ell) = |Y(k,\ell)|e^{j\vartheta(k,\ell)}$ be the noisy DFT coefficient at frequency bin k and frame index ℓ with $|Y(k,\ell)|$ and $\vartheta(k,\ell)$

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

as the spectral amplitude and phase. Similarly, $S(k,\ell)$ and $D(k,\ell)$ are the DFT coefficients of the clean speech and the noise signal, respectively. The voice activity detection is formulated as a classification of frames where speech is present (hypothesis H_1) or absent (hypothesis H_0)

$$H_0 \quad : \quad Y(k,\ell) = D(k,\ell) \tag{1}$$

$$H_1 \quad : \quad Y(k,\ell) = S(k,\ell) + D(k,\ell) \tag{2}$$

In the following, we describe the proposed phase-derived feature called circular variance that is used for VAD later on.

B. Circular Variance

Let $x(k,\ell)$ denote the circular variance of a random variable with realization

$$z(k,\ell) = e^{j\tilde{\vartheta}(k,\ell)} = u(k,\ell) + jv(k,\ell), \tag{3}$$

where $\tilde{\vartheta}(k,\ell)$ denotes the unwrapped phase derived from the wrapped noisy phase $\vartheta(k,\ell)$ by using e.g. [20].

The circular variance is estimated by taking into account the absolute value of the sample mean $\overline{z}(k,\ell)$ of $z(k,\ell)$ [21]

$$x(k,\ell) \quad = \quad 1 - \overline{R}(k,\ell), \tag{4}$$

$$\overline{R}(k,\ell) \quad = \quad \Big| \underbrace{\frac{1}{L} \sum_{\ell'=\ell-\frac{L}{2}}^{\ell+\frac{L}{2}-1} z(k,\ell')}_{\overline{z}(k,\ell)} \Big|, \tag{5}$$

where $\overline{R}(k,\ell)$ denotes the mean resultant length. It follows that for the circular variance we have $x(k,\ell) \in [0,1]$.

In contrast to the assumption that the phase of the speech DFT-coefficients is uniformly distributed, we argue that the DFT phase is concentrated around a mean value, following a von Mises distribution[1] [22]. By employing the von Mises distribution, high concentration of the phase around a mean value can be modeled as well as the uniform distribution which can be considered as a special case of the von Mises distribution with zero concentration at all, representing the maximum uncertainty in phase. We consider voiced phonemes as a sum of sinusoids. The individual sinusoids' phases in the first frame are denoted as the initial phase values. The phase values of the successive frames are mainly determined by the frame-shift and the initial phase since the sinusoidal parameters do not change abruptly, resulting in a low circular variance $x(k,\ell)$ as illustrated in Figure 1.

From these considerations it follows that low circular variance regions reveal the presence of voiced speech while noise-like components yield a higher circular variance. This motivates us to employ the circular variance as an indicator of speech activity. Furthermore, in previous studies the circular variance has been reported useful for single-channel speech enhancement [24], [25]. In order to support these claims, Figure 1 illustrates the speech structure revealed by the circular variance similar to the spectrogram. Especially the harmonic characteristics of speech are nicely represented by low circular variance.

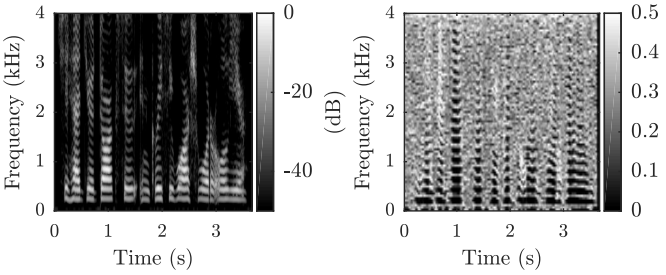

Fig. 1. (Left) Magnitude spectrogram and (Right) circular variance shown for the utterance *"She had your dark suit in greasy wash water all year."* by female speaker from TIMIT [23]. The harmonic structure is revealed by (left) the spectrogram (right) circular variance regions. The circular variance is close to zero in the case of speech presence (justified by the spectrogram).

The proposed VAD works in two stages: first a voice activity decision is made at DFT-bin level. Then, in the second stage the DFT-bins decisions are taken into account to make a frame-level VAD decision. As the circular variance is assumed to be close to zero for speech-present regions and close to one for speech absent regions, the bin-level decision is achieved by a binary hypothesis test based on the estimated circular variance from the noisy observation calculated in (4).

Namely, the binary hypothesis test classifies the observed noisy speech into either of the two classes H_0 (noise only) and H_1 (speech plus noise). To this end we examine the circular variance feature with respect to its distribution for each of the two classes. In order to derive a distribution for the circular variance estimate of noise, we rewrite the sample mean of the complex variable $z(k,\ell)$ as follows

$$\overline{z}(k,\ell) = \overline{u}(k,\ell) + j\overline{v}(k,\ell), \tag{6}$$

with $\overline{u}(k,\ell)$ and $\overline{v}(k,\ell)$ denoting the sample mean values of the real and imaginary parts of $z(k,\ell)$ in (3). The unwrapped phase $\tilde{\vartheta}(k,\ell)$ is assumed to be uniformly distributed for noise dominated regions which implies highly uncorrelated realizations of the random variable $z(k,\ell)$. Using the *central limit theorem* we model the real and imaginary parts as mutually independent, normal distributed random variables $(u,v) \overset{iid}{\sim} \mathcal{N}\big(0,\sigma^2\big)$. If the real and imaginary part of a complex variable $\overline{z}(k,\ell)$ are independent and normal distributed then its absolute value $\overline{R}(k,\ell)$ follows a Rayleigh distribution.

For the speech class we expect a higher correlation among successive samples used to estimate the circular variance, imposing a more heavy-tailed distribution for the circular variance. This indicates that the phase in speech present regions is not uniformly distributed but is rather concentrated around a mean value. This assumption in particular holds for voiced speech whereas for unvoiced speech a noise-like distribution is appropriate. To model the speech, for the voiced portion an Exponential and for the unvoiced portion a Rayleigh distribution is employed. The outcome of these approximations is illustrated on the left panel in Figure 2. It follows that the speech class can only be reliably discriminated from the

[1]The von Mises distribution, also known as Tikhonov distribution, is a circular distribution, parametrized by the mean direction (angle) μ and the concentration parameter κ.

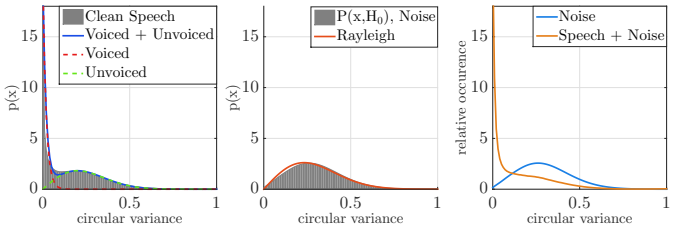

Fig. 2. (Left) empirical circular variance distribution for 50 minutes of clean speech [23] with distribution-fits modeling the voiced (dashed red curve) and unvoiced (green dashed curve) portions of the speech (blue solid curve) (Middle) Rayleigh distribution and empirical distribution for 50 minutes of car noise, window down [26] (Right) empirical circular variance distributions for 50 minutes of noise corrupted speech at 0 dB (car noise, window down).

noise class in the presence of voiced speech. For the sake of simplicity we will drop the indices k and ℓ in the following. Hence, the distributions for the two hypotheses H_0 and H_1 are therefore given by

$$p(x,H_1) = \begin{cases} P\lambda e^{-\lambda x} + (1-P)\frac{2x}{\sigma_1^2} e^{-\frac{x^2}{\sigma_1^2}}, & \text{if } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

$$p(x,H_0) = \begin{cases} \frac{2x}{\sigma_0^2} e^{-\frac{x^2}{\sigma_0^2}}, & \text{if } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where P is the prior probability that speech is voiced and λ is the real valued parameter of the Exponential distribution. The scale parameters of the Rayleigh distributions, σ_1 and σ_0, account for the unvoiced speech together with the noise contribution in equation (7) and the noise circular variance in equation (8). Here, we confine the conventional distributions to the range of the circular variance, i.e. $[0,1]$, resulting in a small truncation error of $< 0.04\%$.

To further justify the selected feature for VAD, Figure 2 shows the empirical distributions as well as the fitted pdfs for the two classes evaluated over 50 minutes of car noise from QUT-NOISE-TIMIT [26]. The Rayleigh distribution accurately approximates noise classes matching the typical assumption of low correlation among successive samples very well. Unvoiced speech follows a more noise-like distribution than voiced speech and, therefore, is less discriminated from the noise class. Since the circular variance can only discriminate voiced speech from noise, we detect voiced frames only. The so-obtained decisions are extended to general VAD by using a longer window when smoothing the raw VAD decisions similar as reported in [27]. Based on this we approximate the pdf of speech with an Exponential distribution, which models the low circular variance regions corresponding to voiced frames very well.

III. Proposed VAD

A. Bin-Level Processing

To classify a single observed DFT-bin we apply the binary decision rule $x \underset{H_1}{\overset{H_0}{\gtrless}} x_{th}$ which results in the binary hypothesis test

$$H_0 \quad : \quad x > x_{th}, \quad (9)$$

$$H_1 \quad : \quad x < x_{th}, \quad (10)$$

where x_{th} is defined as the threshold of circular variance discriminating between the speech-absent and speech-present class. The observed circular variance is interpreted as a random variable with statistical independent realizations. Throughout our experiments we observed that the circular variance structure of higher order harmonics of the fundamental frequency is likely to be impaired by the additive noise. Therefore, in order to achieve more distinctive characteristics for voice activity detection, the frequency range considered is restricted to the interval $[80, 500]$ Hz. The choice of this interval could be further optimized by considering additional prior information such as an f_0-estimate, which would on the other hand add more complexity to the proposed algorithm.

B. Frame-Level Processing

To achieve a frame-level VAD, the DFT-bin level decisions have to be interpreted accordingly. Frame ℓ is classified with respect to the number of voice-active bins denoted by $n(l)$. We seek for a threshold n_{th} that distinguishes between the two classes based on the number of voice active bins per frame. This can be accomplished by a binomial test, described in the following.

The probability to observe a circular variance that exceeds the threshold x_{th} for the speech-absent case, similar to [28] is given by

$$P_{H_0}\big(x > x_{th}\big) = \int_{x_{th}}^{1} p(x, H_0)\,\mathrm{d}x. \quad (11)$$

Since the realizations of x are statistically independent, the probability of observing more than n_{th} speech active bins in an speech inactive frame can be expressed as follows

$$P\big(n \ge n_{th}\big) = \sum_{n=n_{th}}^{N} \binom{N}{n} \big(1 - P_{H_0}\big(x > x_{th}\big)\big)^n \big(P_{H_0}\big(x > x_{th}\big)\big)^{N-n}, \quad (12)$$

where N is the total number of DFT-bins within the analyzed range. The value obtained in (12) is proportional to the risk of a false alarm and depends on the threshold n_{th}. By thresholding and the choice of P_{th} and we have

$$P\big(n \ge n_{th}\big) \le P_{th}. \quad (13)$$

The threshold n_{th} is the smallest value that satisfies (13). For the proposed method, P_{th} was chosen by means of the cross-validation scheme described in Section IV.

To deal with different non-stationary noise types, the empirical distribution $p(x, H_0)$ is updated after the first 200 voice-inactive frames to adapt the threshold n_{th} accordingly. To this end, it is important to keep the miss rate relatively low at the beginning of the analysis procedure, otherwise voice activity could influence the empirical noise distribution. Thus,

an initial $P_{H_0}(x > x_{th}) = P_{H_0,\text{init}}$ needs to be selected, low enough to keep the risk of such errors small. On the other hand, the parameter $P_{H_0,\text{init}}$ should still allow for the detection of speech activity at the beginning of the analysis. This is why we chose $P_{H_0,\text{init}} = 0.5$.

The parameter x_{th} was set to 0.1 motivated by the intersection point of the empirical distributions. Finally, to cope with the fluctuations in the raw VAD decisions, a moving average filter of $800\,\text{ms}$ is applied, similar to [27].

IV. Evaluation

A. Experiment Setup

The DFT size in the STFT is 256 samples. The frame-shift is 1 sample, in order to avoid phase-unwrapping inaccuracies. In the course of our experiments we found that the sampling rate of the original signal can be reduced up to a certain degree without affecting the performance of the algorithm. Therefore, for the sake of reduced computational complexity, we down-sampled the signal to $2\,\text{kHz}$. The circular variance is estimated by taking into account a $40\,\text{ms}$ ($L = 80$ samples) time span for each frequency bin k.

For the evaluation of the proposed VAD method we chose the scheme recommended in [26] together with the QUT-NOISE-TIMIT database specified therein. The database consists of 600 hours of noise-corrupted speech. While the clean speech files were obtained from TIMIT [23], the noise was recorded at 10 different locations in 2 sessions where 2 locations form 1 scenario, resulting in 5 distinct noise scenarios. This allows for a two-fold cross-validation of algorithmic parameters (in our case to tune for P_{th}) between two locations for each noise scenario, providing unbiased test results over the entire corpus. The noise corrupted speech is obtained by randomly selecting clean speech files and mixing them with the noise recordings at various SNRs. The resulting files have a length of 60 and 120 seconds. The amount of speech within a file is set so that 25% of the noisy files have $0\% - 25\%$ of speech, 50% have $25\%-75\%$ and again 25% have $75\%-100\%$ of speech. The start positions of the utterances are selected randomly. In addition to the audio-data, the reference VAD labels for evaluation are provided from [26]. For a detailed description of the QUT-NOISE-TIMIT database we refer to [26].

In our evaluation the following benchmark methods are chosen: Sohn [3] as a standard amplitude-only statistical model based method, the impropriety-based algorithm [16] which takes into account not only the amplitude information of the complex DFT coefficients but also the phase by analyzing its impropriety, and AZR [27] which combines the auto-correlation function (ACF) with the zero crossing rate (ZCR), both revealing the signal periodicity. The cross-validation scheme depicted above is employed to obtain the parameter settings of the benchmark methods. The implementation of Sohn's method utilizes the minimum-statistics noise PSD estimator [29].

Similar to [16], here we quantify the VAD performance in terms of the following evaluation criteria: i) false alarm rate (FAR), ii) miss target rate (MR), and iii) half total error rate (HTER $= \frac{\text{FAR}+\text{MR}}{2}$). It is important to note that the MR and the FAR are strongly influenced by the length of the moving average filter and the threshold P_{th}: the longer the filter and the lower the parameter P_{th}, the higher the FAR gets while the MR decreases.

B. VAD Results

The results shown in Table I are averaged over all noise scenarios. Following [26], we summarize certain SNRs to regions of Low Noise (10 or $15\,\text{dB}$), Medium Noise (0 or $5\,\text{dB}$) and High Noise (-10 or $-5\,\text{dB}$). Additionally, to give more insights into the performance of the particular VAD methods, in Figure 3 we report bar plots, illustrating the HTER performance for each noise scenario and SNR. The following observations are made:

- The AZR method [27] consistently performs best, illustrating the successful fusion of two features: ACF and the ZCR.
- Among the VADs using a single feature, impropriety performs best in terms of HTER for high and medium noise.
- The proposed VAD performs comparable to the amplitude-only approach of Sohn in most scenarios.
- The circular variance, although not being the best performing feature, turns out to be a reliable feature for VAD. It is capable of detecting speech activity in an adverse noisy scenario.

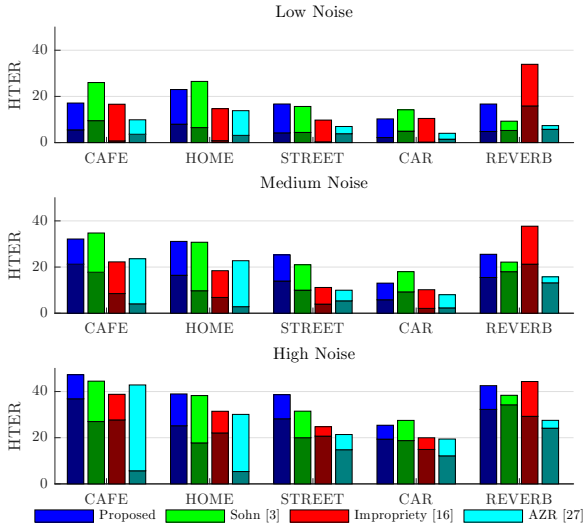

Fig. 3. Individual HTER (%) results for different noise scenarios. The bars are divided into two panels, indicating the MR (%) (darker panel) and FAR (%) (lighter panel).

V. Conclusions

We presented a new voice activity detector (VAD) relying on information extracted from the noisy observation. In the proposed method, a binary hypothesis test framework was derived in the circular variance domain with no requirement of

TABLE I
Overall VAD results averaged across 10 different noise types for different SNR regions.

Method	Low Noise: 15 or 10 dB SNR			Medium Noise: 0 or 5 dB SNR			High Noise: -10 or -5 dB SNR		
	%FAR	%MR	%HTER	%FAR	%MR	%HTER	%FAR	%MR	%HTER
Sohn [3]	24.4	12.2	18.3	24.8	25.9	25.3	25.0	47.1	36.0
AZR [27]	15.6	6.6	11.1	20.5	12.1	16.3	31.9	25.5	28.7
Impropriety [16]	27	7.2	17.1	22.8	17.1	20.0	17.9	45.9	31.9
Proposed	23.7	9.9	16.8	21.7	29.2	25.4	20.4	56.8	38.6

a noise PSD estimator. The intention about our VAD proposal was to emphasize that there are amplitude-independent characteristics in speech eligible to discriminate it from noise. Our results demonstrated that such a VAD is capable of yielding comparable results to amplitude-only benchmarks.

The current work motivates for further studies on combining the conventional amplitude-only VADs with the phase-based proposal, in order to benefit from the complementary sources of information, especially for unvoiced speech to improve the overall VAD performance.

VI. Acknowledgements

We thank Scott Wisdom for sharing the implementation. This work was supported by the Austrian Science Fund (project number P28070-N33). The work was partially funded by the K-Project ASD in the context of COMET - Competence Centers for Excellent Technologies by BMVIT, BMWFW, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria and Vienna Business Agency. The programme COMET is conducted by the Austrian Research Promotion Agency (FFG)

References

[1] "ITU-T recommendation G.729-Annex B, A silence compression scheme for G. 729 optimized for terminals conforming for recommendtion V.70," 1996.

[2] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Proc. ISCA Interspeech*, Sept. 2007, pp. 556–561.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[4] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *speech communication*, vol. 42, no. 3, pp. 271–287, Apr. 2004.

[5] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 3, pp. 600–613, Mar. 2011.

[6] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, Mar. 2013.

[7] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE J. Sel. Topics in Signal Processing.*, vol. 4, no. 5, pp. 834–844, Oct. 2010.

[8] S. Mousazadeh and I. Cohen, "AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 916–926, May 2011.

[9] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.

[10] J. Wu and X. L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 466–469, Aug. 2011.

[11] J. W. Shin, J-. H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Elsevier Computer Speech and Language*, vol. 24, no. 3, pp. 515 – 530, July 2010.

[12] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The Delta-Phase Spectrum With Application to Voice Activity Detection and Speaker Recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2026–2038, Sept. 2011.

[13] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement.*, Sept. 2012, pp. 1–4.

[14] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP J. on Audio, Speech, and Music Processing*, 2014.

[15] H. A. Hegde, R.M. Murthy and V. R. R Gadde, "Significance of the Modified Group Delay Feature in Speech Recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, Jan. 2007.

[16] S. Wisdom, G. Okopal, L. Atlas, and J. Pitton, "Voice activity detection using subband noncircularity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Apr. 2015, pp. 4505 – 4509.

[17] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[18] T. Gerkmann, M. Krawczyk, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[19] P. Mowlaee, J. Kulmer, F. Mayer, and J. Stahl, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*, John Wiley & Sons, 2016.

[20] T. Drugman and Y. Stylianou, "Fast and accurate phase unwrapping," in *Proc. ISCA Interspeech*, Sep. 2015, pp. 1171–1175.

[21] K. V. Mardia and P. E. Jupp, *Directional statistics*, vol. 494, John Wiley & Sons, 2009.

[22] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May 2015.

[23] 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 CDROM," 1993.

[24] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 8, pp. 1283–1294, Aug. 2015.

[25] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 9, pp. 1521–1532, Sept. 2015.

[26] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. ISCA Interspeech*, Sept. 2010, pp. 3110–3113.

[27] H. Ghaemmaghami, B. Brendan, R. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. ISCA Interspeech*, Sept. 2010, pp. 3118–3121.

[28] C. Breithaupt and R. Martin, "Voice activity detection in the DFT domain based on a parametric noise model," in *Proc. International Workshop on Acoustic Signal Enhancement.*, Sept. 2006.

[29] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Elsevier Signal Processing*, vol. 86, no. 6, pp. 1215 – 1229, 2006.