VOICE ACTIVITY DETECTION USING SUBBAND NONCIRCULARITY

Scott Wisdom[⋆]*, Greg Okopal*[†]*, Les Atlas*[⋆]*, and James Pitton*[†⋆]

[⋆]Department of Electrical Engineering, University of Washington, Seattle, USA
[†]Applied Physics Laboratory, University of Washington, Seattle, USA

ABSTRACT

Many voice activity detection (VAD) systems use the magnitude of complex-valued spectral representations. However, using only the magnitude often does not fully characterize the statistical behavior of the complex values. We present two novel methods for performing VAD on single- and dual-channel audio that do completely account for the second-order statistical behavior of complex data. Our methods exploit the second-order noncircularity (also known as impropriety) of complex subbands of speech and noise. Since speech tends to be more improper than noise, higher impropriety suggests speech activity. Our single-channel method is blind in the sense that it is unsupervised and, unlike many VAD systems, does not rely on non-speech periods for noise parameter estimation. Our methods achieve improved performance over other state-of-the-art magnitude-based VADs on the QUT-NOISE-TIMIT corpus, which indicates that impropriety is a compelling new feature for voice activity detection.

Index Terms— Voice activity detection, spectral impropriety, complex-valued data, second-order statistics

1. INTRODUCTION

Voice activity detection (VAD) consists of classifying short periods of a noisy speech signal as either noise-only or speech-plus-noise. VAD algorithms serve an important role in many speech processing applications, including enhancement, separation, and automatic speech recognition.

A great variety of VAD systems has been proposed. These systems cover a wide range, from unsupervised [1, 2, 3], to semi-supervised [4], to supervised [5, 6, 7]. Various properties of speech signals are used, including long-term spectral envelopes [2], Mel-frequency cepstral coefficients (MFCCs) [5, 3], the magnitude-squared of short-time Fourier transforms (STFTs) [1, 2] and other time-frequency representations, and time-domain features such as zero-crossing rate [8]. For two-channel data, spatial properties have also been exploited [9, 10]. Many VAD systems that use complex spectral representations share a common feature: they only use the magnitude (or magnitude-squared) of the complex values. However, recent work [11, 12, 13] on the statistics of complex data has shown that using only the magnitude-squared statistic of complex random data does not fully characterize the data's second-order statistical behavior. If the complex data is second-order noncircular, or improper, then computing an additional second-order statistic, the complementary covariance, can provide improved estimation algorithms and new blind source separation procedures [14].

All speech processing approaches that use only the magnitude (-squared) of complex speech spectra make an implicit assumption that the complex values are second-order circular, or proper. However, recent work [15, 16] has shown that complex-valued subbands of speech tend to be highly improper, especially in subbands that contain a narrowband harmonic of voiced speech or speech onsets and offsets. Impropriety in the frequency domain is also closely related to the modulation frequency content of signals, which can correspond to syllabic rates of speech [17, 18]. Furthermore, we recently showed promising initial results [19] using impropriety for speech processing. Here, we more extensively explore the usefulness of impropriety for VAD.

In this paper, we propose two VAD systems that account for the impropriety of subbands of noisy speech. We show that our systems are robust to realistic additive noise and mild reverberation conditions, achieving equivalent or superior performance over other state-of-the-art methods that only use the magnitude (-squared) of complex STFTs. Thus, we show that impropriety is a promising new feature for voice activity detection.

This work is funded by ONR contract N00014-12-G-0078, delivery order 0013, and ARO grant number W911NF1210277.

2. BACKGROUND

2.1. Second-order statistics of complex-valued data

There has been a recent interest in improved processing of complex-valued signals [11, 12, 13]. For example, an additional second-order statistic, the complementary (or pseudo-) covariance, of a zero-mean, complex-valued random variable x can be computed: $\widetilde{R}_{xx} = E[x^2]$, where E is the expected value. This statistic is complementary to the conventional Hermitian covariance, $R_{xx} = E\left[|x|^2\right]$. When the complementary covariance is zero, the random variable is said to be proper. When $R_{xx} \neq 0$, a normalized circularity coefficient (CC) can be defined as $k_x = \left|\widetilde{R}_{xx}\right| / R_{xx}$, with the property that $0 \leq k_x \leq 1$. When $k_x = 0$, x is proper, and when $k_x = 1$, x is maximally improper, or rectilinear. Given M samples of x, the sample statistic for k_x is

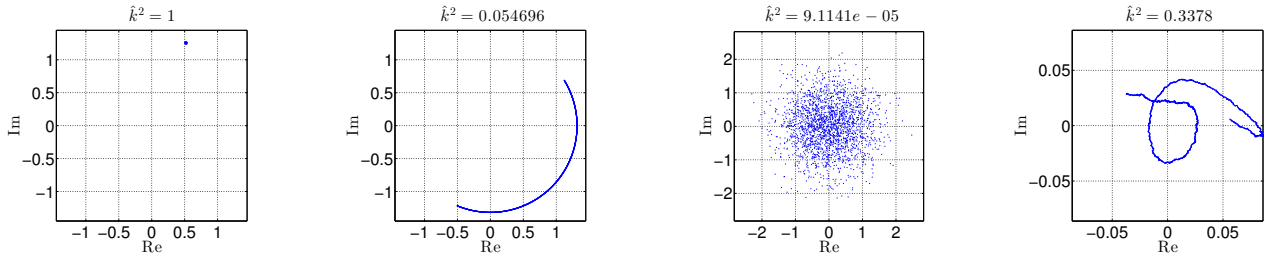
$$\hat{k}_x = \frac{\left|\hat{\widetilde{R}}_{xx}\right|}{\hat{R}_{xx}} = \frac{\left|\frac{1}{M}\sum_{m=0}^{M-1} x(m)x(m)\right|}{\frac{1}{M}\sum_{m=0}^{N-1}|x(m)|^2}. \quad (1)$$

It can be shown that the degree of impropriety (DOI) $\hat{k}_x^2$ is a generalized likelihood ratio test (GLRT) statistic for impropriety [12, Result 3.8].

When the complex data is a random vector, $\mathbf{x} \in \mathbb{C}^N$, the circularity spectrum (CS) characterizes the impropriety of $\mathbf{x}$. The CS is given by the singular values $\{k_i\}_{i\in[1,N]}$ of the coherence matrix $\mathbf{C_{xx}}$, given by [12, Section 3.2]

$$\mathbf{C_{xx}} = \mathbf{R_{xx}}^{-1/2}\widetilde{\mathbf{R}}_{\mathbf{xx}}\mathbf{R_{xx}}^{-T/2} = \mathbf{UKV}^H, \quad (2)$$

where $\widetilde{\mathbf{R}}_{\mathbf{xx}} = E\left[\mathbf{x}\mathbf{x}^T\right]$ and $\mathbf{R_{xx}} = E\left[\mathbf{x}\mathbf{x}^H\right]$. The CS has two compelling properties: it is invariant to nonsingular linear trans-



(a) Tone in subband, $\omega_0 = \omega$. (b) Tone in subband, $\omega_0 \neq \omega$. (c) Demodulated, fullband WGN. (d) Narrowband-filtered WGN.

Fig. 1: Constellations of complex-valued signals in the complex plane.

formations of $\mathbf{x}$, and its elements are the circularity coefficients of the maximally-decorrelated components of $\mathbf{x}$. That is, if $\mathbf{x} = \mathbf{G}\mathbf{y}$, where $\mathbf{G} \in \mathbb{C}^{N \times N}$ is an nonsingular linear transform and the elements of $\mathbf{y}$ are uncorrelated, then $k_i = \mathrm{CC}\{y_i\} = \left|\widetilde{R}_{y_i y_i}\right| / R_{y_i y_i}$. Because of this decorrelation property, the CS is an essential component in complex-valued blind source separation [14].

2.2. Impropriety of complex subbands of noisy speech

In this paper we consider complex-valued subbands of real-valued noisy speech signals. Consider a discrete-time noisy speech signal

$$y(n) = s(n) + v(n), \tag{3}$$

where $s(n)$ is speech and $v(n)$ is additive noise. We consider a complex-valued subband with center frequency $\omega = 2\pi \frac{n_{FFT}}{N_{FFT}}$, $n_{FFT} = 0, ..., N_\omega - 1$ with $N_\omega = \frac{N_{FFT}}{2} + 1$, to be

$$Y(\omega, n) = h(n) * \left(y(n)e^{-j\omega n}\right), \tag{4}$$

where $h(n)$ is a real-valued subband filter. If $h(n)$ is symmetric, (4) can be written as

$$Y^S(\omega, n) = \sum_m h(m - nN_{hop})y(m)e^{-j\omega m}, \tag{5}$$

with $N_{hop} = 1$, which is easily recognized as a maximally-oversampled STFT where $h(n)$ is the analysis window.

If a subband contains both speech and noise, it contains two complex-valued components, $S(\omega, n)$ and $V(\omega, n)$. We consider the estimated impropriety of these two components over a short period $n_0 \leq n < n_0 + M$ under the following assumptions:

1. $h(n)$ is a real-valued subband filter of duration N_{win}.
2. $s(n)$ is voiced during the period $n_0 \leq n < n_0 + M$, so it can be approximated within the narrow subband as a complex-valued tone, given by
$$S(\omega, n) \approx Ae^{j(\omega_0 - \omega)n + j\theta}. \tag{6}$$
3. $v(n)$ is zero-mean, real-valued, white Gaussian noise (WGN). Within the subband, $V(\omega, n)$ is complex-valued, narrowband Gaussian noise.

When the center frequency ω of a subband matches the frequency ω_0 of a speech harmonic, $S(\omega, n)$ is maximally improper, because according to (6), $S(\omega, n) \approx Ae^{j\theta}$, a constant complex value. This constant is shown as the single fixed point in panel A of figure 1. However, when the subband is not aligned, i.e. $\omega_0 \neq \omega$, then $S(\omega, n)$ will slowly rotate over time, appearing more and more circular. This is shown in panel B of figure 1.

Demodulating WGN, even real-valued WGN, does not affect its whiteness; an example is shown in panel C of figure 1. Thus, since $v(n)$ is real-valued WGN, $v_\omega(n) = v(n)e^{-j\omega n}$ is complex-valued WGN. When $v_\omega(n)$ is narrowband-filtered, it appears in the complex plane as a slowly wandering trajectory. An example of narrowband-filtered, demodulated WGN is shown in panel D of figure 1.

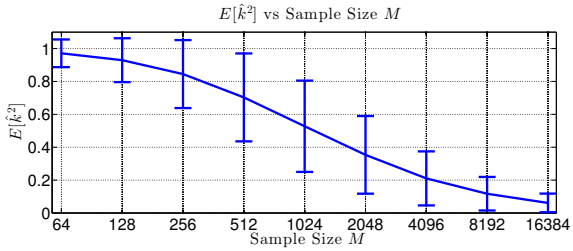


Fig. 2: Results of Monte Carlo experiment (10000 trials) showing the expected sample impropriety of narrowband-filtered, complex-valued WGN versus sample size M.

The sample impropriety of filtered WGN depends on the bandwidth of the filter $h(n)$ and the sample size M. We elect to use a 1024-length Hamming window for $h(n)$. Given this subband filter, figure 2 shows the results of a Monte Carlo experiment measuring the mean and standard deviation of the squared sample circularity coefficient $\hat{k}^2$ from (1) versus sample size M. Notice that $\hat{k}^2$ decreases with increasing M.

3. PROPOSED METHODS

We propose two impropriety-based VADs, one for single-channel and one for dual-channel audio. For both approaches, the complex filterbank in (4) of each channel can be efficiently computed using the STFT (5) with $h(n)$ of duration N_{win}, a window hop of N_{hop}, and a N_{FFT}-length FFT. To correct for the phase rotations induced by the STFT window hops, a phase modification is applied to each subband, which results in a complex-valued filterbank $Y(\omega, n)$ with subbands downsampled by the factor N_{hop}:

$$Y(\omega, n) = \left|Y^S(\omega, n)\right| \exp j\left(\angle Y^S(\omega, n) - n\omega N_{hop}\right), \tag{7}$$

where $\angle Y^S(\omega, n)$ is the unwrapped phase of the STFT $Y^S(\omega, n)$.

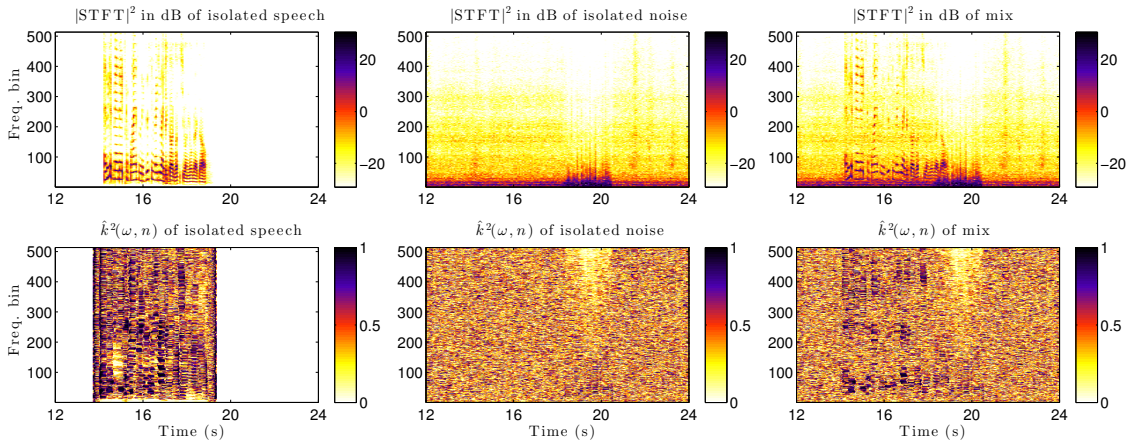


Fig. 3: The high impropriety of speech (lower left, quiet shown as white) tends to dominate over more proper noise (lower middle) in the estimated impropriety of the mixed speech and noise (lower right). Spectrograms are shown in top panels.

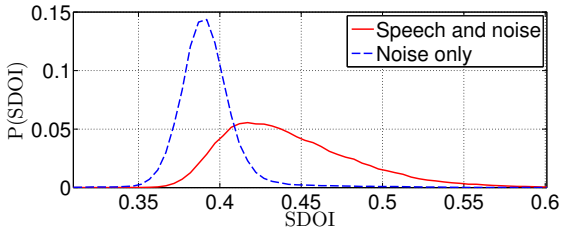


Fig. 4: Empirical distributions of summed degree of impropriety (SDOI) under two hypotheses for 50 minutes of speech in 0 dB SNR car noise (windows down, highway driving).

3.1. Single-channel method

Given a single-channel audio mix $y(n)$ of speech and noise, we use $Y(\omega, n)$ to estimate the instantaneous CC at each time/frequency point (ω, n). We use a sliding window of duration $M^d = M/N_{hop}$ and hop $M^d_{hop} = M_{hop}/N_{hop}$ in each subband to compute the CC estimate using (1):

$$\hat{k}_Y(\omega, n) = \frac{\left|\frac{1}{M^d}\sum_{m=0}^{M^d-1} Y^2(\omega, nM^d_{hop}+m)\right|}{\frac{1}{M^d}\sum_{m=0}^{M^d-1}\left|Y(\omega, nM^d_{hop}+m)\right|^2}. \quad (8)$$

For each time/frequency point, the CC estimate will be between 0 and 1. Recall from §2.1 that the DOI $\hat{k}^2_Y(\omega, n)$ corresponds to a GLRT for the impropriety of $Y(\omega, n)$.

We choose the test statistic for VAD to be the estimated DOIs averaged across frequency, which we will refer to as the "summed degree of impropriety" (SDOI). The SDOI for frame n is given by

$$\text{SDOI}(n) = \frac{1}{N_\omega}\sum_\omega \hat{k}^2_Y(\omega, n), \quad (9)$$

The SDOI will take on values between 0 and 1. To see the effectiveness of the SDOI for discriminating between speech-plus-noise and noise-only, we examine its empirical distribution under the two hypotheses. Using $N_{hop} = 16$ and a sliding window length of $M = 2048$ with hop $M_{hop} = 80$, figure 4 shows the distributions of the SDOI for 50 minutes of audio with $f_s = 8$ kHz, consisting of TIMIT utterances [20] embedded at 0 dB SNR in car noise (window down, highway driving) from the QUT-NOISE [21] database.

To get intuition about why these distributions are different, figure 3 compares the estimated DOIs of speech, noise, and the mix for a short clip of the audio used for figure 4. Notice that time/frequency points where speech is highly improper tend to dominate over more proper noise.

3.2. Two-channel method

For two-channel data, we use the M-length sliding window to estimate the 2×2 spatial covariance matrices $\widetilde{\mathbf{R}}_{\mathbf{YY}}(\omega, n)$ and $\mathbf{R}_{\mathbf{YY}}(\omega, n)$ at each time/frequency point (ω, n). These estimated covariances are used to compute the two-dimensional circularity spectrum $\mathbf{k}(\omega, n)$ using (2).

To get a maximally discriminative test statistic, we use linear discriminant analysis (LDA) [22, Section 8.6.3] to train two N_{FFT}-length vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ using $\mathbf{k}(\omega, n)$ and ground-truth labels. We will refer to the resulting statistic as "circularity spectrum with linear discriminant analysis" (CS-LDA), and it is given by

$$\text{CS-LDA}(n) = \frac{1}{2}\sum_{i=1}^{2}\left(\frac{1}{N_\omega}\sum_\omega a_i(\omega)k^2_i(\omega, n)\right). \quad (10)$$

4. EXPERIMENTS

To test the performance of VAD using the SDOI (9) and CS-LDA (10), we use the QUT-NOISE-TIMIT corpus [21], which consists of TIMIT utterances [20] embedded in two-channel recordings of five different types of real-world noise, with each noise type recorded at two different locations for two different sessions (total of 20 noise conditions). Two noise types, CAR and REVERB, also include exponentially-swept sines, which allows estimation of two-channel reverberation impulse responses (RIRs) using the technique of Farina [23]. The CAR RIRs are relatively short, while the REVERB RIRs are highly reverberant. The locations of the noise types given by Dean et al. [21] are listed in table 2.

There are six SNRs ranging from -10 dB to 15 dB. We use the `sA` subset of QUT-NOISE-TIMIT, which consists of 6000, 60-second files, for 100 total hours of audio. For each of the six SNRs,

Method	Low noise (10 or 15 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
DSB + Sohn et al.	15.17	21.49	18.33	25.91	23.24	24.58	40.55	**29.77**	35.16
DSB + Ramirez et al.	11.63	13.54	12.58	18.44	21.31	19.87	28.38	36.64	32.51
CS-LDA (2ch)	10.80	18.16	14.48	**14.75**	25.17	19.96	29.26	30.01	**29.63**
SDOI (1ch)	**8.03**	**9.86**	**8.95**	16.48	**13.94**	**15.21**	**29.13**	32.47	30.80

Table 1: Overall VAD results averaged across noise types on the QUT-NOISE-TIMIT corpus.

Noise type	Location 1	Location 2
CAFE	Cafe	Food court
HOME	Kitchen	Living room
STREET	City	Suburb
CAR	Window down	Window up
REVERB	Car park	Pool

Table 2: QUT-NOISE noise types and locations.

there are 50 files for each noise type, location, and session. 25% of the files contain 25% or less of speech, 50% have 25% to 75% speech, and the remainder have 75% or more speech. We use a sampling rate of $f_s = 8$ kHz. For noise types without RIRs provided, the second channel of speech is a copy of the first channel of speech.

We compare to two baseline VAD methods: Sohn et al.'s statistical model-based likelihood ratio test [1] and Ramírez et al.'s long-term spectral divergence (LTSD) [2]. These methods are well-suited for comparison, because they use the magnitude (-squared) of complex STFTs to derive their detection statistics.

To try and ensure a fair comparison to the two-channel CS-LDA method, we apply a delay-and-sum beamformer (DSB) to two-channel data before applying single-channel baselines, which gives an average relative improvement of 4.57% in overall HTER versus the single-channel baselines alone.

Our procedure for evaluating VAD performance is exactly the same as that used by Dean et al. [21] and Ghaemmaghami et al. [8], which is computing the half-total error rate (HTER) at an optimal detection threshold. The HTER is given by the average of the false-alarm rate (FAR) and the miss rate (MR). The FAR and MR are given by

$$\text{FAR} = \frac{\text{\# false positives}}{\text{\# negatives}} \qquad \text{MR} = \frac{\text{\# false negatives}}{\text{\# positives}} \tag{11}$$

To choose the detection threshold for each method—and, in the case of CS-LDA, to train the LDA vectors $\mathbf{a}_i(\omega)$—we follow the same procedure as Ghaemmaghami et al. [8], which uses cross-validation between noise locations to choose detection thresholds for a particular noise type. For example, the detection threshold for STREET-City noise is given by the threshold that minimizes HTER on STREET-Suburb noise, and vice versa. This cross-validation ensures that the methods are robust across different realizations of the same type of noise. Also, the six SNRs are grouped into three subsets (low, medium, and high noise), which evaluates the robustness of the methods to different levels of noise.

Decisions for the methods are made in 10 ms increments. For Sohn et al.'s method, we use the implementation from the VOICEBOX Matlab toolbox with default parameters [24]. For Ramírez et al., we use default parameters from their paper [2]. For SDOI and CS-LDA, we use $N_{win} = N_{FFT} = 1024$, $N_{hop} = 16$, $M = 2048$, and $M_{hop} = 80$. The resulting decision labels are median filtered with a window of 1 second.

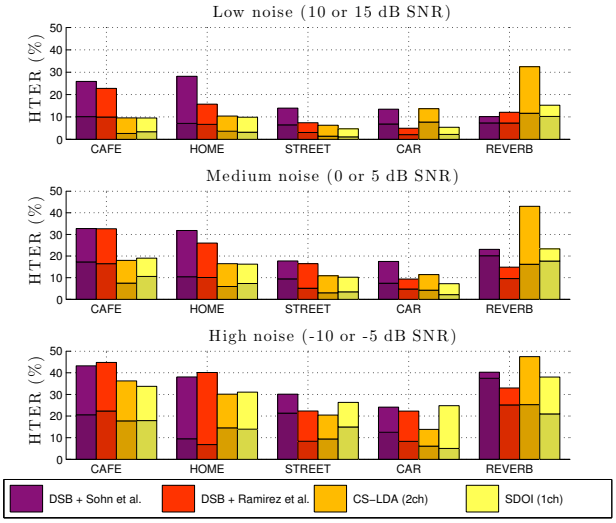


Fig. 5: VAD results on the QUT-NOISE-TIMIT corpus. In each stack, top lighter-shaded bar corresponds to $\%\text{FAR}/2$, and bottom darker-shaded bar corresponds to $\%\text{MR}/2$.

Our results are shown in table 1 and figure 5. Our new VADs achieve equivalent or superior performance to the baseline methods on all noise types except for the REVERB noise type. Interestingly, the unsupervised single-channel SDOI statistic performs better than the supervised two-channel CS-LDA at lower noise levels (0 to 15 dB SNR), while CS-LDA performs better at higher noise levels (−10 and −5 dB SNR). Vulnerability to high amounts of reverberation is expected, since adding many shifted copies of improper signals together in a subband tends to make them look more proper.

MATLAB code for our impropriety-based features and methods is available at `github.com/impropriety`.

5. CONCLUSION

In this paper, we have proposed two new methods for voice activity detection that exploit a previously overlooked property of complex subbands of speech: second-order noncircularity, or impropriety. Our methods work by estimating the instantaneous degree of impropriety at each time/frequency point and combining the results across frequency. We tested our method on a challenging VAD corpus, QUT-NOISE-TIMIT, and our new methods achieved equivalent or superior performance versus conventional baselines on all noise types, except for heavy reverberation. This performance indicates the importance and potential usefulness of impropriety in speech processing. Future work will explore using impropriety-based methods and features for speech enhancement and recognition.

6. REFERENCES

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

[2] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[3] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, Nov. 2011.

[4] F. Germain, D. L. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection.," in *Proc. Interspeech*, 2013, pp. 732–736.

[5] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and support vector machine," in *Proc. Int. Conf. on Speech and Computer*, Moscow, Russia, 2007, vol. 2, pp. 556–561.

[6] Q.-H. Jo, J.-H. Chang, J. W. Shin, and N. S. Kim, "Statistical model-based voice activity detection using support vector machine," *Signal Processing, IET*, vol. 3, no. 3, pp. 205–210, 2009.

[7] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[8] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[9] J. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2007, vol. 4, pp. 385–388.

[10] H.-D. Kim, K. Komatani, T. Ogata, and H. Okuno, "Two-channel-based voice activity detection for humanoid robots in noisy home environments," in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, Pasadena, CA, May 2008, pp. 3495–3501.

[11] D. P. Mandic and V. S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear, and Neural Models*, Wiley, Hoboken, N.J., 2009.

[12] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*, Cambridge University Press, Feb. 2010.

[13] T. Adali and P. Schreier, "Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 112–128, Sept. 2014.

[14] J. Eriksson and V. Koivunen, "Complex random vectors and ICA models: identifiability, uniqueness, and separability," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1017–1029, Mar. 2006.

[15] B. Rivet, L. Girin, and C. Jutten, "Log-rayleigh distribution: A simple and efficient statistical representation of log-spectral coefficients," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 796–802, Mar. 2007.

[16] P. Clark, *Coherent Demodulation of Nonstationary Random Processes*, Ph.D. thesis, University of Washington, 2012.

[17] P. Clark, I. Kirsteins, and L. Atlas, "Existence and estimation of impropriety in real rhythmic signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 3713–3716.

[18] S. Wisdom, L. Atlas, and J. Pitton, "Extending coherence time for analysis of modulated random processes," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[19] G. Okopal, S. Wisdom, and L. Atlas, "Estimating the noncircularity of latent components within complex-valued subband mixtures with applications to speech processing," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2014.

[20] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[21] 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. Interspeech*, Makuhari, Japan, Sept. 2010.

[22] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambridge, MA, Aug. 2012.

[23] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*. 2000, Audio Engineering Society.

[24] M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.