A VAD/VOX Algorithm for Amateur Radio Applications

William B. D. Forfang*, Eduardo Gonzalez†, Stan McClellan‡, and Vishu Viswanathan‡

* Freescale Semiconductor Inc., Austin TX, USA will.forfang@freescale.com
† FlexRadio Systems, Austin TX, USA ed.gonzalez@flexradio.com
‡ Ingram School of Engineering, Texas State University, San Marcos TX, USA
stan.mcclellan@txstate.edu, vishu.viswanathan@txstate.edu

***Abstract*—In amateur radio applications, voice activity detection (VAD) algorithms enable hands-free, voice-operated transmissions (VOX). In this paper, we first review a recent hybrid VAD algorithm, which was developed by combining features from two legacy speech detection algorithms long used in amateur radio applications. We then propose a novel VAD algorithm whose operating principles are not restricted to those of legacy approaches. The new method employs two key features. The first feature, called sub-band variance ratio, is the ratio of energies calculated over a low-frequency region and over the rest of the spectrum of the input audio signal. The second feature, called temporal formant density, is a running N-frame sum of the number of low-bandwidth formants over a low-frequency region. Both features are shown to yield low values for non-speech segments and relatively high values for speech segments. A two-state decision logic that uses these two features is employed to make frame-by-frame VAD decisions, which are then used in the VOX function for amateur radio transmissions. The proposed new method is compared against the hybrid method using both a simple objective measure involving comparisons against manually derived true VAD data and a subjective pairwise comparison listening test, over audio signal data from amateur radio transmissions at various signal-to-noise ratios. The results from these comparison tests show that the new method provides a better overall performance than the hybrid method. In summary, a new VAD/VOX algorithm for amateur radio applications is proposed that offers performance benefits over existing methods.**

***Keywords*—voice activity detection; VAD; voice-activated switch; voice-activated transmission; VOX.**

I. Introduction

As technology has progressed, speech processing algorithms have found ubiquitous deployment within consumer electronics. In many of these algorithms, Voice Activity Detection (VAD) plays an important role in increasing overall performance and robustness to noise. Amateur radios, digital hearing aid devices, speech recognition software, etc. are common examples of speech processing applications that employ VAD [1]–[3]. Precise discrimination between speech and non-speech allows, for example, an algorithm to capture, characterize, and update an accurate noise profile for adaptive noise cancellation [4]. The integrity of silence compression in discontinuous transmission schemes also relies upon the accuracy of VAD algorithms. Speech coding, speaker recognition, and speech enhancement are all examples of VAD applications [4]–[10].

VAD schemes with basic energy level detection can provide satisfactory performance at high signal-to-noise ratios (SNRs), but often perform poorly in noisy conditions. More robust VAD methods have been developed, which consider statistical features beyond average energy such as long-term spectral divergence [3] or multiple-observation likelihood ratio tests [11]. In 2012, Gonzalez and McClellan [1] published a VAD scheme that performs well in noisy environments while maintaining low computational complexity. Their algorithm specifically targets voice-activated transmission (VOX) applications.

A VOX is an amateur radio application that allows hands-free switching between the operating modes of a transceiver. A radio transceiver with a push-to-talk operating scheme requires a physical 'transmit' button to be pressed and held for the duration of the transmission, whereas a VOX automatically activates 'transmit' mode upon detection of an operator's voice. It then disables 'transmit' mode after observing a sufficient interval of non-speech.

In designing the VAD algorithm, Gonzalez and McClellan emulated a well-known legacy hardware approach to VOX, and then rectified its deficiencies by incorporating ideas from a complementary digital approach. The resulting hybrid algorithm was then tested in the context of amateur radio transmissions and was found to exhibit a higher robustness to noise than its legacy constituents, without an increase in computational cost.

Motivated by the success of the hybrid algorithm, this research investigates the design of a new VOX/VAD algorithm whose operating principles are not restricted to those of legacy devices. Instead, features for speech detection in the new algorithm are extracted from linear prediction coefficients and spectral sub-band energy analysis. Both algorithms target amateur radio applications, and are therefore compared to one another using audio stimuli captured from amateur radio transmissions. Quantitative evaluations of their relative performance were performed across multiple SNRs with both objective and subjective methods. For the remainder of this discussion, the algorithm by Gonzalez and McClellan will be referred to as the "hybrid algorithm," and the new algorithm will be referred to simply as the "new algorithm." Performance comparisons show that the new algorithm has better overall

performance than the hybrid algorithm.

It is worth emphasizing that our goal in this work has been to develop a VAD algorithm specifically for use with VOX transmission for amateur radio application. In this application, a speech transmission decision is made only after a frame has been declared as speech, and the decision to stop speech transmission is made only after a sufficient number of frames have been declared as non-speech. Therefore, for effective VOX transmission, it is not necessary that VAD decisions are made accurately every single frame. Because of the inherent delay in switching from transmit to no-transmit decision mode, occasional frame VAD errors are not significant and do not impact the effectiveness of the VOX transmission system. For this reason, the new algorithm is not compared to industry standard VAD implementations.

In Section II, the theoretical backgrounds behind the hybrid algorithm and the new algorithm are discussed. The legacy VAD approaches used in designing the hybrid algorithm are explained. The operating principles behind the new algorithm are described mathematically and shown graphically. In Section III, two methods used to compare the algorithms' performances are explained. One method is comprised of an objective performance measurement, and the other method is a subjective listening test. The results of the comparisons are given and discussed. The new algorithm was found to achieve equal or better performance under the majority of tested conditions. In Section IV, areas of further research are presented.

II. Theory of Operation

In a VAD paradigm an input audio signal can be generalized to fall into one of three categories at any given instant: noise (Eq. (1)), noise and voiced speech (Eq. (2)), or noise and unvoiced speech (Eq. (3)).

$$y(n) = q(n), \tag{1}$$

$$y(n) = q(n) + v(n), \tag{2}$$

$$y(n) = q(n) + u(n), \tag{3}$$

where $q(n)$ represents noise, $v(n)$ is voiced speech, and $u(n)$ is unvoiced speech. Distinguishing audio signals best modeled by Eq. (1) from those characterized by Eq. (2) or Eq. (3) is the goal of VAD.

In a general sense, speech is a statistically non-stationary signal. That is, its statistical description changes with time. However, a finite number of speech samples observed over a sufficiently short time period will exhibit wide-sense stationary statistical behavior [12]. To exploit such behavior for speech processing, an observed frame of audio samples must be short enough for meaningful statistical analysis but long enough to capture vocal features of interest. A detailed description of the two VAD/VOX algorithms is presented in the following subsections.

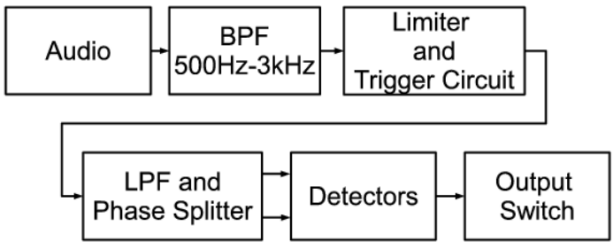

Figure 1. A high-level diagram of the MICOM algorithm, reprinted with permission from [1].

A. Hybrid Algorithm

The hybrid algorithm's operational theory proceeds from a combination of techniques found in two legacy VAD schemes. The first scheme is a hardware-driven approach developed by Motorola in the 1970s [13]. This circuit, which we refer to as the "MICOM" implementation, was popular with amateur radio enthusiasts since it provided a simple and easily implemented speech detection subsystem. The MICOM VOX continuously monitors a specified channel, suppressing non-speech noise in the idle channel while allowing detected speech signals to activate transmission. MICOM-like circuits exploit the syllabic rate of human speech (about 3 syllables per second) and include a detector for short-term frequency modulation, which is characteristic of voiced speech. The main components of the MICOM implementation include a high gain amplifier, a trigger circuit to produce constant width pulses, a 3.25 Hz low-pass filter, comparators, and timing circuitry to create hysteresis on the output "voicing" signal.

To implement MICOM features into the hybrid algorithm, a SPICE variant (Multisim [14]) was used to analyze and accurately decompose MICOM's functional components. These functional components were then duplicated using a simulation package (Simulink [15]) to model the subsystems via signal processing algorithms. Fig. 1 depicts a high-level block diagram of the MICOM algorithm.

Detailed descriptions of the individual blocks in Fig. 1 are given in [1], [13]. Briefly, the bandpass filter BPF extracts the voiceband part of the input audio signal; the Limiter and Trigger Circuit amplifies all non-zero samples to extreme saturation levels as a means of zero-crossing detection and generates a steady stream of uniform-width pulses, one per zero-crossing; the LPF and Phase Splitter block uses a 3.25 Hz low-pass filter to extract the syllabic rate envelope and a phase splitter to separate the signal into "top phase" and "bottom phase" voltages; the Detectors declare a detection event if either of these phase voltages is above a manually-set threshold; and finally, the Output Switch causes a single detection event to lead to a one-second holdover, using a timing capacitor, thereby avoiding "drop-outs" in the middle of active speech.

The second technique that influenced the design of the hy-

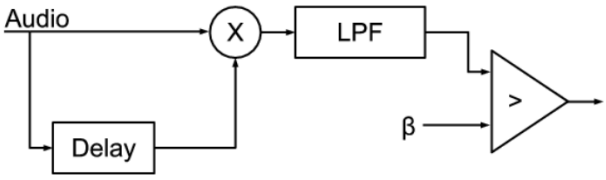

Figure 2. A high-level diagram of the Harris algorithm, reprinted with permission from [1].

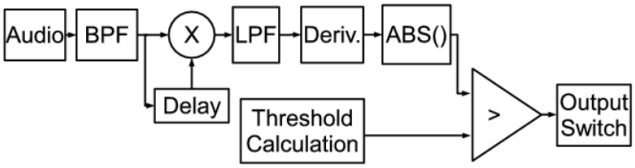

Figure 3. A high-level diagram of the hybrid algorithm, reprinted with permission from [1].

brid algorithm is a software-driven single-lag-autocorrelation process published by Harris Corp. in 1991 [16], which we refer to as the "Harris" algorithm. The Harris algorithm has several useful features for robust speech detection. However, in a complete implementation it may be lacking key features that are provided very effectively by aspects of the MICOM system. A block diagram of the Harris system is shown in Fig. 2.

The system incorporates a fixed delay and a multiply operation, which essentially computes a running autocorrelation at a single predetermined lag, according to Eq. (4).

$$ACF(l) = \sum_{n} X_n \bar{X}_{n-l} \tag{4}$$

The output from this delay and multiply operation is fed into a simple low-pass filter implemented as an accumulator. The resulting low-frequency component of the running autocorrelation is then compared to a threshold β to determine the presence of speech. The effect of the Harris approach is to detect strong, stable correlations around the predetermined lag value, which is related to pitch frequency.

Although the MICOM and Harris systems have advantages, they both also have shortcomings. The MICOM circuit is robust and simple to implement in an analog system, but some subtleties of modeling analog phenomena make it less stable and more difficult to implement directly in a discrete time system. The Harris algorithm performs well in detecting the onset of speech, but is inconsistent during active speech segments. The detector output has many false negatives (namely, non-speech) within active speech, and the resulting audio is choppy and incomprehensible. When the threshold is lowered to prevent these drop-outs, the same results occur during silence intervals since the noise creates a high enough output to repeatedly trigger a detect event.

The idea of detecting strong correlations around a predetermined lag value used in the Harris approach is valuable, but by itself it does not provide a reliable system. The hybrid implementation described here uses aspects of the MICOM system to address these problems.

A high-level diagram of the hybrid algorithm is shown in Fig. 3.

Each of the hybrid algorithm blocks is explained below:

- Bandpass Filter (300-700 Hz): The BPF provides the same function as the BPF in the MICOM circuit but the voiceband is decreased so that processing is done on more selective data.
- Delay and Multiply: Extracts short term periodicities in filtered audio. The chosen delay of 50 samples at a sampling frequency of 8000 Hz provides smooth operation and sufficient sensitivity.
- MICOM Low-pass Filter: Instead of using a simple accumulator, the 3.25 Hz low-pass filter from the MICOM circuit is used to extract syllabic rate information from the delay and multiply block. This filter also provides a much sharper cut-off, eliminating unwanted frequency components that interfere with the estimation in the Harris algorithm.
- Derivative and Absolute Value: The derivative converts the slowly changing output of the LPF into a more defined and faster changing waveform, which increases the tolerance and sensitivity of the threshold. Since the output of the LPF contains information about the changes in syllabic rate, like the phase splitter subsection of the MICOM circuit, both positive and negative deviations are important. The absolute value allows a single threshold to consider both deviations.
- Threshold Calculation: Removes the need for manual setting of the threshold value. To accomplish this, whenever speech is not detected, the energy of the noise is continuously calculated and the baseline threshold is established according to this changing energy level. This allows detection in varying noise floors.
- Modified MICOM Output Switch: Forces a holdover period following a detection event. Instead of using a 1.0s holdover (as in the MICOM circuit) the hybrid algorithm uses a 0.25s holdover, which minimizes drop-outs during active speech without overly extending the detection period.

The performance of the hybrid algorithm was compared against the MICOM system and the Harris algorithm in [1]. The hybrid algorithm was shown to exceed the performances of the Harris algorithm as well as the MICOM approach in both stability and robustness to noise. Fig. 4 shows performances of the Harris, MICOM, and hybrid VOX implementations in a low-noise condition. Although Fig. 4 seems to display a fairly "clean" or lab quality original signal, the signal is actually a speech utterance captured from an amateur radio

transmission, and contains some objectionable, non-speech noise. In the figure, several error conditions are labeled. Note the highly erratic performance of the Harris approach in voiced segments ("A"), but also the ability of the Harris approach to reliably (albeit aggressively) determine non-speech segments ("B"). Also note the inaccurate speech/non-speech decisions of the MICOM approach ("C"). The hybrid approach typically produces accurate voicing indicators with acceptable overhang, and without aggressive penetration into non-speech segments. There are a few exceptions (e.g., a missed onset at "D"). However, this level of performance is quite acceptable for real-time implementation, which avoids clipping, slow-attack, and other behaviors that are objectionable to amateur radio operators. For a more detailed discussion of the comparison among the hybrid algorithm, the Harris algorithm, and the MICOM implementation, the reader is referred to [1]. The performance comparison results presented in [1] show that the hybrid algorithm performs significantly better than the other two algorithms and provides robust and stable speech detection performance in realistic operational conditions.

B. New Algorithm

The operating principle behind the new algorithm relies on the predictability of formant locations during voiced speech. A spectral estimate of audio samples can be analyzed to exploit this property. The frame length must be kept small enough to minimize computational latency, but large enough to resolve the temporal events of interest. A 30 millisecond buffer with a 50% overlap was found to fall appropriately within this envelope of efficiency and speed, yielding a new data set every 15 milliseconds. Two features, or speech indicators, are implemented in the new algorithm. Section II-B1 explains the first feature, which is a ratio of energy levels at predetermined spectral locations. A summary of the second feature, which tracks formant activity, is given in Section II-B2. Finally, the logic that determines speech presence given the available feature data is discussed in Section II-B3.

1) Feature 1, Sub-band Variance Ratio: To analyze the spectral energy, the 30 millisecond frame is multiplied by a Hamming window and a 512 point fast Fourier transform (FFT) is applied. The absolute value of the FFT can be seen in Fig. 5, in which the shaded range (with width W, and distance from the origin d) represents the test area for active speech information.

To test for increased spectral activity in this range, the ratio of the variance within the shaded region to the variance of the remaining spectrum is compared to a threshold value. The derivation of this ratio, Γ_{var} can be seen in Eq. (5) - Eq. (7).

$$\sigma_1^2 = \frac{1}{W-1}\sum_{k=1}^{W}(s_1(k)-\mu_{s1})^2, \tag{5}$$

$$\sigma_2^2 = \frac{1}{N-W-1}\sum_{k=1}^{N-W}(s_2(k)-\mu_{s2})^2, \tag{6}$$

$$\Gamma_{\text{var}} = \frac{\sigma_1^2}{\sigma_2^2}, \tag{7}$$

where $s_1(k)$ is the shaded portion of the spectral estimate; $s_2(k)$ is the spectrum with $s_1(k)$ removed and the remainder concatenated together. μ_{s1} and μ_{s2} are the means of $s_1(k)$ and $s_2(k)$, respectively.

To maximize the selectivity of the Γ_{var} metric, the statistical distance between distributions of Γ_{var} values during speech and during non-speech should be maximized. A script was created, which adjusts the W and d values incrementally and creates distributions of Γ_{var} values for speech and non-speech audio after each adjustment. It then calculates the statistical distances between the distributions, using the well-known Bhattacharyya distance measure, and stores them into a two-dimensional array. By doing this many times, for many combinations of W and d values, a three-dimensional space can be analyzed in which peaks represent W and d combinations that yield high statistical separability of classes. Fig. 6 shows the described three-dimensional space for a range of W and d values obtained from processing test data from amateur radio transmissions with varying SNRs.

Through this analysis, the determined values for the W and d parameters were 781 Hz and 219 Hz, respectively. Fig. 7 displays the Γ_{var} function (bottom graph) and the utterance from which it was extracted (top graph). The utterance was sourced from an amateur radio broadcast and is mostly clean speech. Clearly, Γ_{var} is relatively small for non-speech segments and relatively large during speech segments.

2) Feature 2, Temporal Formant Density: The second feature of interest is also motivated by spectral distribution analysis, but is realized in a different manner. As described by Eq. (8), the coefficients (a_i) of a 10th order forward linear predictor (LPC) are calculated for each windowed frame of time-domain data $y(n)$. The general idea behind Eq. (8) is that a given speech sample at time n can be approximated as the linear combination of the past 10 samples (10th order) weighted by their respective coefficients, a_i.

$$\overline{y} = \sum_{i=1}^{10} a_i y(n-i) \tag{8}$$

The predictor coefficients (a_i) for each frame are computed via autocorrelation and Levinson-Durbin recursion [12]. Once these coefficients are found, they are treated as the coefficients of a polynomial whose roots are then solved for. These complex conjugate pairs of roots are expressed as $r_0 e^{\pm i\phi_0}$, where i is the square root of -1, r_0 is the root magnitude, and ϕ_0 is the root angle. From these roots, Eq. (9) and Eq. (10) are used to estimate formant center frequencies (F) and bandwidths (BW), respectively [12]; f_s denotes the sampling frequency in Eq. (9) and Eq. (10). These estimates are valid for high Q formants [12], which is the case for our formant analysis.

$$F = \frac{f_s}{2\pi}\phi_0 \tag{9}$$

$$BW = \frac{-f_s}{\pi}\ln(r_0) \tag{10}$$

For a more intuitive understanding of the formant estimation process, Fig. 8 graphically summarizes the steps from raw

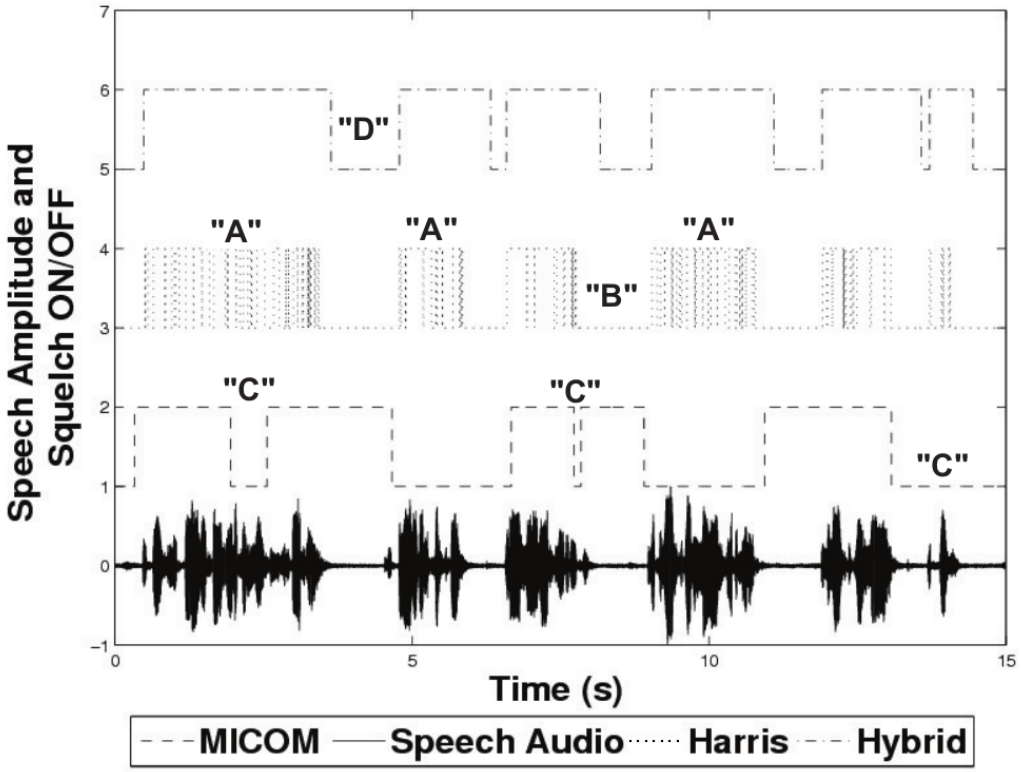

Figure 4. Performance of all three voice detection algorithms in a low noise, natural environment. The utterance was captured from an amateur radio transmission, and contains some non-speech noise. Annotations "A" through "D" indicate detection errors in each algorithm.

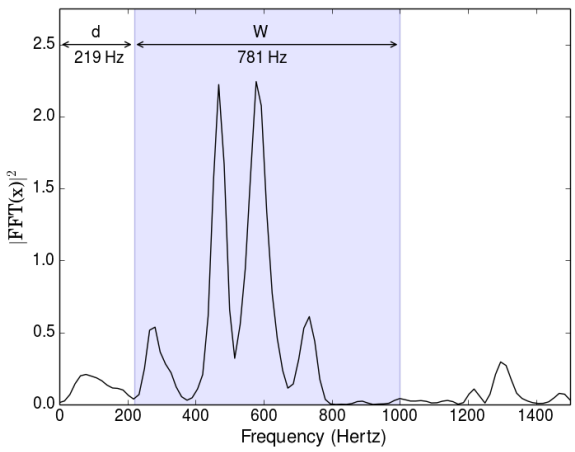

Figure 5. Spectral estimate of a 30 millisecond frame of audio data. The variance of the signal within the gray area with width W, and distance from origin d, is divided by the variance of the remaining spectrum. This quotient, Γ_{var}, signifies the level of spectral activity within the shaded area relative to the remaining spectrum.

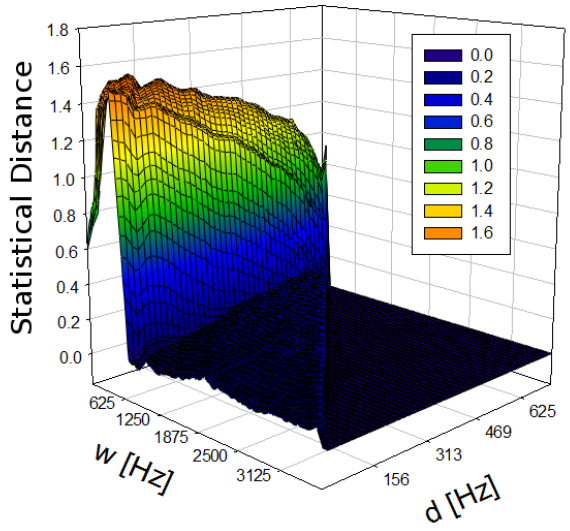

Figure 6. This plot shows the statistical distance between distributions of Γ_{var} values while processing speech audio versus non-speech audio for a large combination of W and d values. The highest peak in this space, which corresponds to a d value of 219 Hz and a W value of 781 Hz, represents a configuration that yields the highest statistical separability between speech/non-speech classes for the given test data.

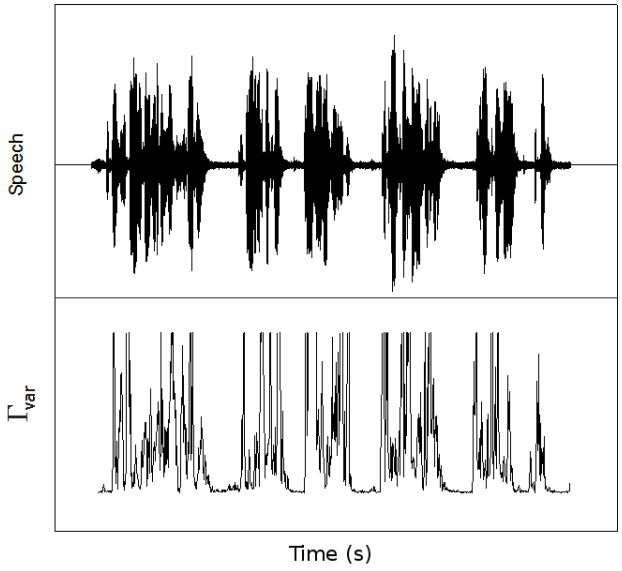

Figure 7. A recorded utterance sourced from an amateur radio transmission (upper) and its resulting Γ_{var} function (lower). The Γ_{var} function is truncated above an arbitrary value along the y-axis for clearer representation. The graph shows that the Γ_{var} feature yields relatively high values during speech segments and relatively low values during segments of non-speech.

audio data to formant estimation via Eq. (8) and Eq. (9). First, the raw audio data for a 30 millisecond frame (Fig. 8(A)) is multiplied by a Hamming window. From the weighted data, the autocorrelation and Levinson-Durbin recursion methods are used to calculate an LPC model for the given frame, as defined by Eq. (8). The LPC coefficients are treated as polynomial coefficients whose roots are then extracted and are plotted in polar form in Fig. 8(B). Eq. (9) is then used to translate root angles to formant center frequencies. The resulting formant estimations are plotted over the LPC spectrum in Fig. 8(C), along with the raw audio spectrum overlaid for comparison.

By comparing each root in Fig. 8(B) (in order of increasing angle) to their respective formant estimates in Fig. 8(C) (in order of increasing frequency) it is clear that roots farther from the origin correspond to estimated formants with smaller bandwidths.

By analyzing the formant locations and their bandwidths, an estimate of the spectral energy distribution can be made for each frame of time-domain data. In Fig. 9 and Fig. 10, two separate audio frames are displayed above scatter-plots of their respective formant estimates.

In these scatter plots, the y-axis represents formant bandwidth and the x-axis represents formant frequency. The shaded rectangles represent a decision space, within which formants will tend to lie when voiced speech is present. Adopting a similar approach to that used for determining the W and d values in the Γ_{var} analysis of Section II-B1, we chose the decision space dimensions to include formants with center frequencies between 100 and 1000 Hz, and bandwidths under 100 Hz. In an LPC spectrum, these formants would manifest as relatively lower frequency, pronounced spectral peaks. Notice that two formants fall within the shaded region during the particular speech frame displayed in Fig. 10 while no formants do so in Fig. 9. The speech frame in Fig. 10 is extracted from a voiced segment of the word 'principle,' recorded over an amateur radio transmission; the non-speech frame in Fig. 9 is background noise extracted from the same transmission.

By monitoring the number of formants within the decision space and computing a running sum of this number over the last N frames (labeled ρ_f below), a *temporal density* of formants is computed. That is, the number of formants that have landed within the decision space over the last N frames is computed to detect the presence of speech. By choosing N = 10, some hysteresis is introduced into ρ_f. In Fig. 11, the ρ_f feature is plotted above the utterance from which it was extracted. Clearly, non-speech segments correspond to values of ρ_f closer to zero while speech segments correspond to ρ_f values closer to 10.

3) Two-State Logic: To combine the above-mentioned features (Γ_{var} and ρ_f) into a speech detection scheme, decision thresholds were chosen statistically. Histograms of feature values during both speech and non-speech were calculated over a variety of utterances and noise levels. Given some overlap in these distributions, two thresholds for Γ_{var} were chosen to indicate a high or low likelihood of speech presence, labeled T_Γ^1 and T_Γ^0, respectively. This creates three possible sample spaces for Γ_{var} values at a given instance: likely speech ($\Gamma_{\text{var}} >= T_\Gamma^1$), likely non-speech ($\Gamma_{\text{var}} <= T_\Gamma^0$), or inconclusive ($T_\Gamma^0 < \Gamma_{\text{var}} < T_\Gamma^1$). Given the higher statistical distances between distributions of ρ_f values during speech and non-speech, only one threshold T_ρ is used.

Finite-state logic is then employed to allow state changes only when both features indicate that the current frame of audio differs from the previous frame of audio. With three possible Γ_{var} interpretations and two possible ρ_f interpretations, six combinations of feature indications can be realized, but only two merit a state change. The state logic is summarized in the diagram of Fig. 12, where **a** represents detected speech from both features, **b** represents detected non-speech from both features. For hands-free VOX switching applications, the transceiver would begin transmissions when the algorithm moves to the "speech" state. It would then end transmissions when the algorithm moves back to the "non-speech" state, and remains there for a sufficient duration.

III. Performance Comparison

A human subject's perceptual evaluation of audio stimuli is the outcome of a complex physiological process. In this process, the quality aspects of the audio are considered along with the subject's expectations, mood, context, etc. When assessing the quality of audio stimuli, the criteria upon which opinions are formed are difficult to characterize, even for the assessor. Forming a mathematically predictive model to forecast such psychoacoustic assessments is therefore difficult. In this research, a simple objective metric was designed to estimate the relative performance of the algorithms. Although

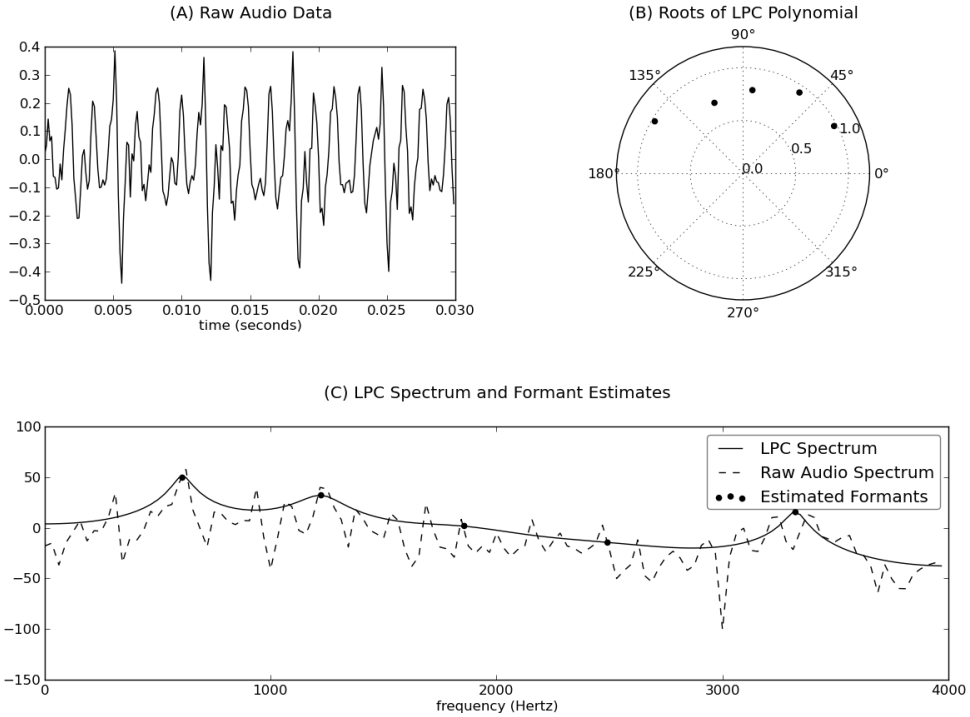

Figure 8. A graphical summary of the mathematical transition from raw audio to LPC formant estimates. A 30 millisecond sample of audio data is plotted in (A). After computing the LPC coefficients from the weighted data, the LPC polynomial roots are extracted and plotted in polar form (B). The formant center frequencies are estimated from the roots with Eq. (9) and plotted over the LPC spectrum in (C). The raw audio spectrum is also displayed in (C).

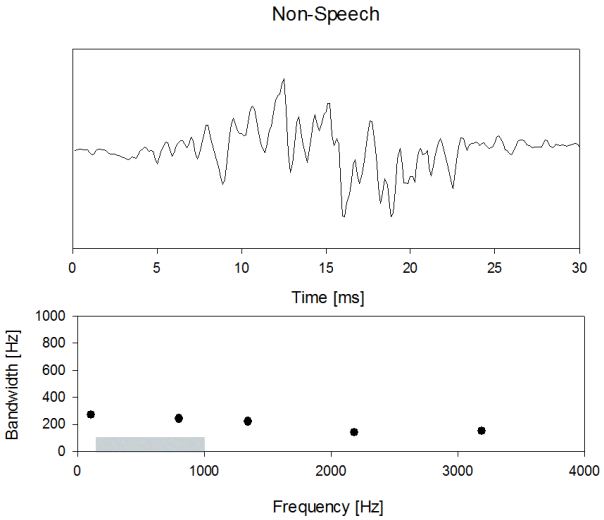

Figure 9. A 30 millisecond frame of non-speech audio (above) and its corresponding formant estimates (below). The y-axis of the lower plot is the estimated formant bandwidth, as defined by Eq. (10), and the x-axis is frequency. The shaded area signifies the decision space for speech-data. Notice that no formants lie within the shaded area.

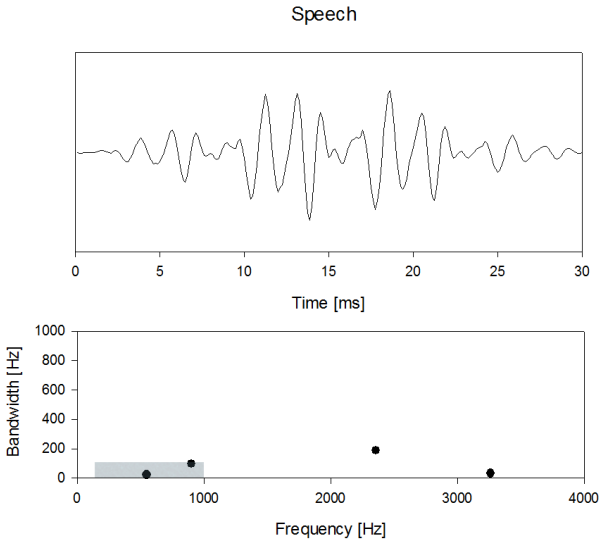

Figure 10. A 30 millisecond frame of voiced speech data (above) and its corresponding formant estimates (below). The y-axis of the lower plot is the estimated formant bandwidth, as defined by Eq. (10), and the x-axis is frequency. Notice two formants fall within the perimeter of the decision space, indicating an increased likelihood of speech presence.

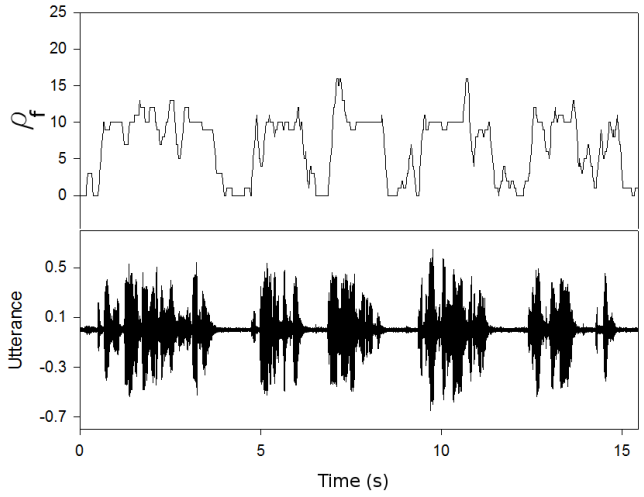

Figure 11. A running temporal formant density (ρ_f) (upper graph) extracted from an amateur radio speech recording (lower graph). It is clear from the graph that higher (ρ_f) values correspond to speech segments and lower (ρ_f) values correspond to non-speech segments.

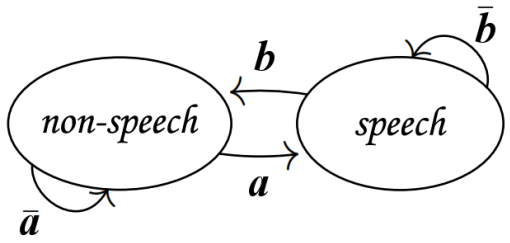

Figure 12. Two-state logic used for making VAD decisions. The audio is deemed either speech or non-speech by the algorithm depending on its previous state and current inputs. Path $\mathbf{a}$ is taken if both features indicate speech presence, path $\mathbf{b}$ is taken if both features indicate non-speech presence. Path symbols $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ refer the logical complements of $\mathbf{a}$ and $\mathbf{b}$.

this metric was not intended to predict subjective assessments, it provided valuable feedback during algorithm design at little cost.

In addition, a subjective evaluation using a paired comparison listening test was administered. Untrained subjects were briefly informed on the purpose and scope of the VAD algorithms, and then presented with a series of audio samples to evaluate. The samples were organized into pairs, and the subjects were asked to choose the preferred audio from each pair. The following sections provide greater detail into these test methods and their results.

A. Objective Comparison

To objectively evaluate a given VAD performance, its resulting activity mask was first compared to a manually derived true marker mask as seen in Fig. 13. In this figure, the audio is plotted under the resulting VAD activity mask. The absolute value of the difference between the VAD mask and the true marker mask is shaded in gray. Taking the integral of the

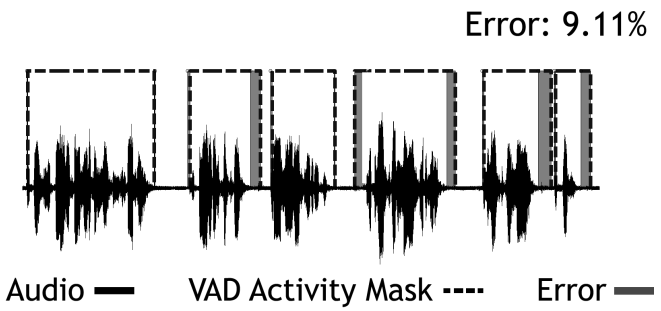

Figure 13. Audio with VAD activity mask overlaid. Gray shaded areas represent errors relative to true markers. The integral of the shaded area divided by the audio sample length gives the error percentage.

shaded areas and normalizing by the audio sample length gives a percentage error relative to the true markers, which was used to gauge the VAD accuracy.

To employ this test metric, the algorithms first processed relatively clean speech captured from radio transmissions and were scored accordingly. Noise was then added systematically to the speech files and the algorithms were graded objectively over a variety of SNRs. Fig. 14 plots the quantitative results from two radio speech transmissions processed at three different SNRs (30 dB, 15 dB, and 0 dB). Admittedly, the database used for evaluation was relatively small but was found to be adequate for our purposes. In Section IV, we suggest test methods that use a larger speech database.

In Fig. 14(a), both algorithms perform similarly. In Fig. 14(b) the hybrid algorithm fails in 'open' mode during the 0 dB test while the new algorithm remains functional. That is, the hybrid algorithm reports detected speech during the entirety of the test. A closer analysis of the algorithms' behavior during this particular test can be seen in Fig. 15.

The objective metric in Fig. 14 proves suitable for distinguishing between large performance disparities (such as those seen in Fig. 15). However, small differences in VAD performances may go unnoticed in this basic test paradigm. To the human auditory system, these small differences may exhibit high perceptual significance depending upon their context within an utterance. Therefore, the objective metric may be limited in its ability to accurately predict perceptual evaluations. To investigate this possibility, a paired comparison listening test was conducted.

B. Subjective Comparison

Although the objective metric described in Section III-A offers a relative evaluation, its comparison proved inconclusive for all but extreme disparities in algorithm performances. To pursue the possibility of subtle differences in algorithm performances (such as those in Fig. 14(a)) containing overlooked perceptual significance, a paired comparison listening test was designed and administered to 10 untrained participants.

Two utterances captured from radio transmissions were used as test audio for the subjective evaluations. Gaussian

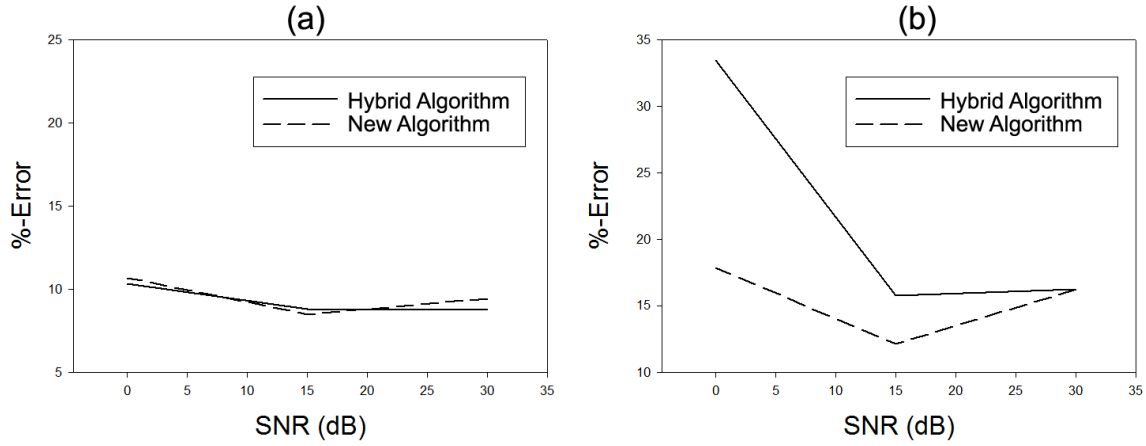

Figure 14. A comparison of algorithm performance swept over SNR. The raw test audio for the comparison in (a) was captured from a radio transmission containing clean speech. In (b) the raw test audio was captured from a radio transmission containing mostly clean speech with occasional hand clapping sounds. Noise was added systematically to both files to create the SNR sweep.

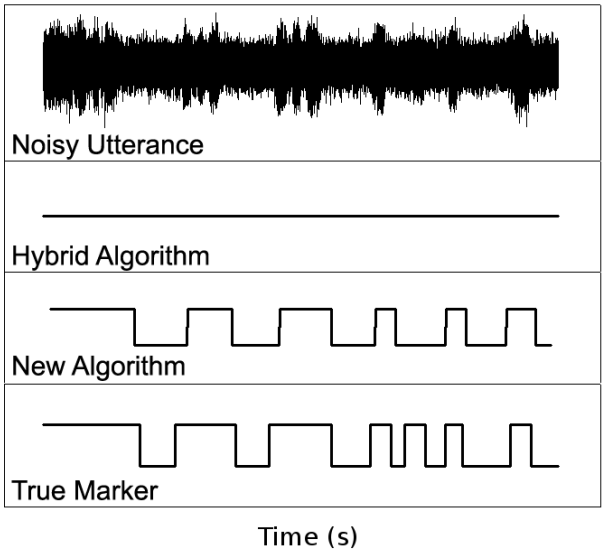

Figure 15. The noisy utterance used during the 0 dB objective measurement from Fig. 14(b) along with the algorithms' resulting VAD activity masks and the utterance's true marker mask. Here, the hybrid algorithm fails 'open,' falsely indicating speech throughout the utterance. The new algorithm maintains overall functionality.

white noise was added to produce three different SNRs per utterance. Signal power measurements for SNR calculations were taken only over the speech sections of the utterances, as determined by manually derived true markers. The test files were processed by the two VAD algorithms, which were configured to zero (mute) the sections they deemed non-speech and preserve the sections deemed speech. Eight pairs of the resulting processed audio files were played to listening test participants in the order presented in Table I.

Rows 6 and 8 in Table I contain pairs of audio samples with significantly different SNRs, which we included for control purposes. Test results from these pairs were used as a means

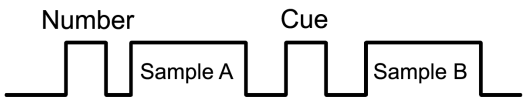

Figure 16. The format for each pair of audio stimuli administered in the paired comparison listening test. The pair number is first announced, followed by the two samples, which are separated by a cue tone for clarity.

to identify potentially unreliable listeners and to exclude their scores from statistical analysis.

The untrained test participants were provided a brief description of the definition and purpose of VAD algorithms before taking the test. An ordinal A/B multiple choice answer-sheet comprised the response format. Each of the 8 pairs of audio stimuli presented in Table I were played to the test participants in the fashion depicted by Fig. 16. The 'number' in Fig. 16 indicates which pair is about to be played; the 'cue' tone serves to separate the samples in each pair.

The mean of the responses was calculated at each tested SNR, and is plotted in Fig. 17. This data indicates that the new algorithm was preferred over the hybrid algorithm at SNRs above 4 dB. The hybrid algorithm was preferred slightly at 0 dB SNR. Thus, the new algorithm provides a better overall performance than the hybrid algorithm.

IV. Future Research Directions

We suggest below a few areas for further research. The first area is to investigate ways of reducing the computational complexity of the proposed new VAD/VOX algorithm. Recall that the temporal formant density feature used in the new algorithm requires solving for the roots of the LPC polynomial and calculating the number of low-bandwidth formants in the designated low-frequency region (see Section II-B2). Solving

TABLE I
LISTENING TEST DETAILS.

	SAMPLE A			SAMPLE B		
	ALGORITHM	SNR	UTTERANCE	ALGORITHM	SNR	UTTERANCE
1	hybrid	30 dB	1	new	30 dB	1
2	new	0 dB	2	hybrid	0 dB	2
3	hybrid	15 dB	1	new	15 dB	2
4	new	15 dB	2	hybrid	15 dB	2
5	hybrid	30 dB	2	new	30 dB	2
6	hybrid	15 dB	1	new	0 dB	1
7	new	0 dB	1	hybrid	0 dB	1
8	new	30 dB	2	hybrid	15 dB	2

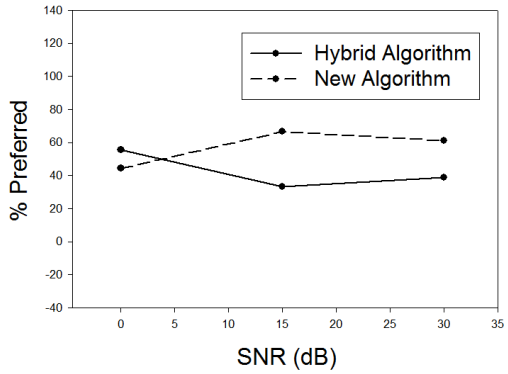

Figure 17. The averaged preference of test participants at different SNRs. The two preference functions sum to 100%. The listening test results reported a preference for the new algorithm when processing stimuli with SNRs over 4 dB. The hybrid algorithm was preferred at 0 dB.

for the polynomial roots is a computationally intensive task. Reference [17] proposes a way of directly calculating the number of formants in a given frequency range without first solving for the roots of the polynomial and shows significant computational savings. This approach may be modified to directly calculate the number of formants with bandwidths smaller than a specified value. The same paper also contains a procedure for calculating the bandwidths of formants in a specified frequency region.

We can also reduce the computational complexity of the sub-band variance ratio feature used in the new algorithm. Since the LPC analysis (used in the temporal formant density feature) provides the zeroth autocorrelation coefficient R(0), by Parseval's theorem, R(0) equals the energy under the entire power spectrum. By computing the power spectrum only over the shaded region in Fig. 5 and by using R(0), we can compute a ratio similar to the one given by Eq. (7). The approximation involves not using the means in Eq. (5) and Eq. (6), yielding an energy ratio rather than a variance ratio. The modest computational savings here arise from not having to compute the spectral values in the frequency range outside the shaded region.

As described in Section II, both sub-band variance ratio and temporal formant density features focus on the low-frequency region and as such are effective in identifying voiced frames. By the same token, the new method may likely misclassify unvoiced speech frames as non-speech. While these errors should not affect the effectiveness of our target VOX transmission application, it may be desirable to minimize misclassification of unvoiced speech as non-speech. To this end, a third feature may be developed by focusing on energy ratio and/or formant density in a designated high-frequency region; VAD decisions are then made using all three features.

Although the focus of this work was on VOX transmission for amateur radio application, it may be of interest to evaluate the effectiveness of the proposed new algorithm by comparing against industry standard VAD algorithms like the ones used in the ITU standard G.729 or 3GPP Adaptive Multi-Rate (AMR) standard [4]. The challenge here is to have access to a speech database with known ground-truth VAD decision data. We have identified one such database that is included in a 1998 TIA standard called TIA/EIA/IS 727 [18]. This database includes 5 male and 5 female clean speech sentences, noise files in four noise environments (which can be used to generate noisy speech files at different signal-to-noise ratios), and ground-truth VAD decision data for the ten clean speech files. Both objective and subjective tests may be performed in comparing the new algorithm against widely used industry standard VAD algorithms.

V. CONCLUSION

Motivated by the success of a hybrid VAD algorithm described in [1], a new algorithm targeting amateur radio applications was developed. Unlike the hybrid algorithm, whose design combines ideas from two well-known methods, the new algorithm was designed without restricting its operating principles to those of legacy approaches. The performance of the new algorithm was compared to the hybrid algorithm, both objectively and subjectively, in the context of amateur radio transmission data. The objective evaluations, which were computed by comparing the algorithm behavior to true VAD markers as described in Section III-A, indicate that the new algorithm achieves equal or higher performance than the hybrid algorithm under the tested noise conditions. The subjective evaluations, which were performed through the listening test described in Section III-B, show that the new algorithm was preferred over the hybrid algorithm by the majority of listeners, particularly at higher SNRs. In Section IV, future research ideas are suggested for (i) reducing the

computational complexity of the new algorithm, (ii) preventing the new algorithm from misclassifying unvoiced speech as non-speech, and (iii) enlarging the performance testing by using an industry standard speech database and by comparing the new algorithm against well-known industry standard VAD algorithms.

REFERENCES

[1] E. Gonzalez and S. McClellan, "A hybrid VOX system using emulated hardware behaviors," in *International Conference on Digital Telecommunications (ICDT)*, April 2012, pp. 105 – 110.

[2] T. V. den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, April 2007, pp. IV–565 – IV–568.

[3] J. Ramírez, J. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, April 2004.

[4] J. Ramírez, M. Górriz, and J. Segura, "Voice activity detection. Fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech, June 2007, pp. 1–22.

[5] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 478–482, July 2000.

[6] J. Séris, C. Gargour, and F. Laville, "VAD INNES: A voice activity detector for noisy industrial environments," in *50th Midwest Symposium on Circuits and Systems*, August 2007, pp. 377–380.

[7] B. V. Ilarsha, "A noise robust speech activity detection algorithm," in *International Symposium on Intelligent Multimedia, Video and Speech Processing*, October 2004, pp. 322–325.

[8] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2024–2038, November 2006.

[9] L. Ye, W. Tong, C. Huijuan, and T. Kun, "Voice activity detection in non-stationary noise," in *IMACS Multiconference on Computer Engineering in Systems Applications*, October 2006, pp. 1573–1575.

[10] D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador, and X. Menendez-Pidal, "A robust speech detection algorithm for speech activated hands-free applications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, March 1999, pp. 2407–2410.

[11] J. M. Górriz, J. Ramírez, J. Segura, and C. G. Puntonet, "An improved MO-LRT VAD based on bispectra Gaussian model," *Electronics Letters*, vol. 41, no. 15, pp. 877 – 879, July 2005.

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[13] *Service Manual for Motorola Micom HF SSB Transceiver*, Motorola, Inc., 1975, part No. 68-81025E95A, The "Constant SINAD" Squelch was used in the Motorola Micom HF SSB Transceiver. The MICOM squelch board part number is TRN6175.

[14] *Multisim 11.0*, National Instruments, July 2011.

[15] *Simulink 2011a*, MathWorks Inc., January 2011.

[16] M. Webster, G. Sinclair, and T. Wright, "An efficient, digitally-based, single-lag autocorrelation-derived, voice-operated transmit (VOX) algorithm," *Military Communications Conference, 1991. MILCOM'*, November 1991.

[17] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129 – 134, April 1993.

[18] *TDMA Cellular/PCS - Radio Interface - Minimum Performance Standards for Discontinuous Transmission Operation of Mobile Stations*, Telecommunications Industry Association, June 1998.