A Hybrid VOX System Using Emulated Hardware Behaviors

Eduardo Gonzalez
Ingram School of Engineering
Texas State University
San Marcos, TX 78666, USA
email: eg1196@txstate.edu

Stan McClellan
Ingram School of Engineering
Texas State University
San Marcos, TX 78666, USA
email: stan.mcclellan@txstate.edu

***Abstract*—This paper analyzes two well-known but complementary speech detection algorithms, and combines them to create a robust, low complexity method of speech detection. Software emulation of behaviors important in venerable hardware-based voice-operated switches is key to hybrid system performance. We test the hybrid system in the context of amateur radio, where speech and in-band data is accurately detected in real-time, even in the presence of significant noise.**

***Keywords*-silence detection, voice activity detection, VAD, VOX, voice-activated squelch, voice-activated switch**

I. INTRODUCTION

Speech detection plays an important role in applications where communication may be intermittent, or hands-free operation is desirable. Examples of this class of applications include emergency radio services, amateur radio, and communications for infrastructure maintenance and development. These environments require monitoring of communications channels for the presence of speech, which places a psychological strain on operators who must listen to constant noise and interference. Often, voice-operated switch systems are used to detect the presence of speech on a channel, and automatically "gate" the signal to an audio amplifier. Automated speech detection can effectively relieve operator strain and mute the speaker/receiver until active speech is present in the incoming transmission.

This paper analyzes two complementary approaches to speech detection, compares their operating characteristics, and presents a combination of elements to produce a hybrid, easily implemented and robust speech detection system. We focus on use of this approach in amateur radio systems and explore the performance and requirements of an automated squelch for convenient, hands-free operation. Conventional terminology among amateur radio operators uses the term "VOX" for a voice-activated switch, or voice-operated squelching unit. Thus, we refer to this system as a "VOX" in the remainder of this paper.

In Section II, we describe a venerable but popular hardware driven approach with some operational features which are very attractive for the user community. In Section III, we examine a software-driven approach which is similar to well-known pitch detection schemes, but optimized for low computational complexity. In Section IV, we describe the characteristics of a hybrid system which derives operational features from both of the preceding architectures. In Section V we evaluate the three complementary approaches and present performance comparisons, and Section VI concludes with observations about the examined systems and their application in real-time systems.

II. HARDWARE DRIVEN APPROACH

In the 1970's, Motorola engineers developed a transistor circuit for hardware-based voice detection [1]. This circuit, which we refer to as the "MICOM" implementation, had very good characteristics for speech detection in noisy analog transmissions, and variants of this system were popular in the amateur radio community. Such variants include the Smart Squelch, popularized in *73 Magazine* [2] and an implementation by the Jet Propulsion Laboratories Amateur Radio Club [3] for retransmission of NASA Select Audio over the JPL voice/packet repeater network in Southern California.

The MICOM circuit 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 the speaker.

MICOM-like circuits exploit the syllabic rate of human speech (3 syllables per second) and include a detector for short-term frequency modulation which is characteristic of voiced speech. The main components of MICOM implementations include a high gain amplifier, a trigger circuit to produce constant width pulses, a 3.25 Hz lowpass filter, comparators and timing circuitry to create hysteresis on the output "voicing" signal.

Motivated by the popularity and continued use of the MICOM VOX architecture [3] we performed an in-depth analysis of of this circuit to understand its behavior and model its features in a software simulation. First, we analyzed the MICOM circuit by hand and modeled it using a SPICE variant (MultiSim [4]) to accurately decompose its functional components. Then, we duplicated these functional

 ISBN: 978-1-61208-193-9

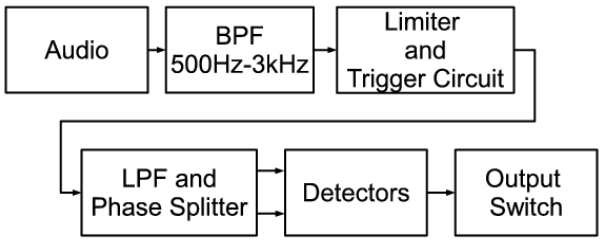

Figure 1. High level block diagram of MICOM algorithm.

components using a simulation package (Simulink [5]) to model the subsystems using signal processing algorithms.

In effect, we modeled the hardware implementation to extract performance measurements. This enabled a common reference to compare subsequent speech detection algorithms. Following subsections describe this process and illustrate the performance of the MICOM system.

A. MICOM Subsystems

To baseline MICOM performance, both the MultiSim and Simulink simulations used an 8kHz audio file which was manipulated through the stages shown in Figure 1. Each of the stages play an important role as described below:

- Band-Pass Filter (0.5 - 3 kHz): Removes non-voice-band energy.
- Limiter (85dB amplifier): Amplifies the signal so that non zero samples are saturated at the extrema. The effect of this function is a zero crossing detector for positive going excursions.
- Trigger Circuit (0.33ms pulses): Triggered by the amplified and limited voice band signal to create a steady stream of pulses that have uniform width, one per zero-crossing.
- Low-Pass Filter (3.25 Hz): Extracts the syllabic envelope from the pulse stream, estimating energy $<$ 3 Hz.
- Phase Splitter: The first output of the phase splitter removes the DC component from the LPF output, and the second output inverts the resulting signal. This separates the original output of the LPF into a "top phase" and "bottom phase" for the detector.
- Detector: Creates a detection event if either of the phase voltages is above a manually-set threshold. This threshold must be set by the user each time a different channel is selected (ex. carrier frequency in amateur radio) or if the noise floor of the channel changes. In the analog implementation, a potentiometer provided decent control. However, in software, the tuning of this threshold becomes difficult.
- Output Switch: Incorporates a timing capacitor that creates a one second holdover from a single detection event. This is done in order to remove "dropouts" in the middle of active speech. The output switch also incorporates hysteresis by lowering the threshold whenever a detection event occurs. This, like the holdover capacitor, is intended to reduce false negatives.

B. MICOM Problems

Although the MICOM circuit was robust and simple to implement in an analog system, some subtleties of modeling analog phenomena make it less stable and more difficult to implement directly in a discrete time system. Certain components such as DC removal, which are simply capacitors in an analog circuit, become complicated in a discrete environment. Further, slight usability issues revolve around the threshold setting, which is sensitive and has small tolerance. Issues also arise whenever modulated data is transmitted on the channel, or when noise changes slowly producing localized energy $<$ 3 Hz in the detection circuit.

Although much of the MICOM VOX functionality may have been supplanted by modern signal processing techniques, many of the MICOM operational characteristics are powerful and attractive to the user community. Thus, we attempt to model and emulate selected features in a discrete fashion.

III. Software Driven Approach

Robust speech detection systems often incorporate separate detection or classification of voiced and unvoiced speech. Many approaches to detection of voiced, unvoiced, and silence segments have been described in the literature, including for example: pitch detection [6], spectral characterizations [7], [8], and distance measures or statistical tests applied to harmonic and/or nonparametric models [9], [10].

However, in some classes of systems, detection of voiced segments is performed by subtracting estimated noise power from the output of a comb filter at the dominant frequency of the voiced speech. This result is compared to a threshold that determines whether speech is present. This type of "discriminate and threshold" system is functional, but presents a heavy computing load.

An approach to reducing the compute burden, which we refer to as the "Harris Algorithm," provides an approximation of the voiced detector through a single lag autocorrelation process [11]. This method has been used by Harris Corp. to provide dynamic channel routing and activation for ADPCM (Adaptive Differential Pulse-Code Modulation) channel encoding.

The Harris Algorithm has several useful features for robust speech detection. However, in a complete implementation it may be lacking key features which are provided very effectively by aspects of the MICOM system.

A. Harris Subsystems

The Harris algorithm was designed in the 1990's to meet the demand for a functional and simple voice detector [11]. For the purpose of this paper we summarize the general

 ISBN: 978-1-61208-193-9

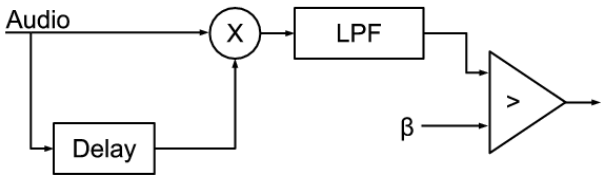

Figure 2. High level diagram showing the Harris algorithm components.

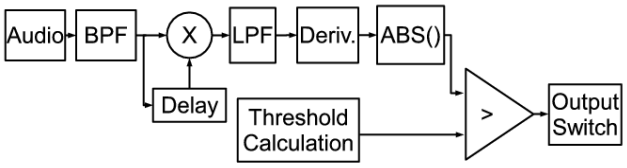

Figure 3. High level diagram showing the hybrid algorithm incorporating MICOM, Harris and new components.

operation of the Harris algorithm and refer the reader to the literature for a complete discussion.

A block diagram of the Harris system is shown in Figure 2. The system incorporates a delay and multiply operation which essentially computes a running autocorrelation at a single pre-determined lag, according to Equation 1. In the equation, l is the fixed lag and $\bar{X}$ is the complex conjugate of X:

$$ACF(l) = \sum_{n} X_n \bar{X}_{n-l} \quad (1)$$

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

B. Harris Problems

The Harris Algorithm performs well in detecting the onset of speech, but is inconsistent during active speech segments. The detect output has many false negatives within active speech, and resulting audio is choppy and incomprehensible. When the threshold is lowered to prevent these dropouts, the same results occur during silence intervals since the noise creates a high enough output to repeatedly trigger a detect event. Furthermore, since the Harris Algorithm relies on the low frequency components of the ACF, the slow spectral rolloff caused by an accumulator (a poor lowpass filter) allows low-frequency components to interfere with the approximation.

The core idea within 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.

IV. Hybrid Approach

In order to achieve a robust hybrid speech detection algorithm, fundamental features of the MICOM circuit and the Harris Algorithm were taken into consideration and then extended. The components that are used from each system are outlined below, as well as the additional modifications made to increase detection speed, reduce false positives, and reduce the need for manual operation of the threshold.

A. Hybrid Inner Workings

Figure 3 presents a high level block diagram of the hybrid system. Each of the hybrid blocks is explained below:

- Band-Pass Filter (300-700 Hz): The BPF provides the same function as the BPF in the MICOM circuit but the voice band is decreased so that processing is done on more selective data.
- Delay and Multiply: Extracts short term periodicities in filtered audio. The delay chosen of 50 samples with a sampling frequency of 8000Hz provides smooth operation and good sensitivity.
- MICOM Low-Pass Filter: Instead of using a simple accumulator, the 3.25Hz lowpass filter from the MICOM circuit is used to extract syllabic rate information from the delay and multiply. This filter also provides a much sharper cut-off, eliminating unwanted frequency components that interfered 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 considers 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 in detection via a counter that resets every time there is a detect event. The output is turned off only when the counter saturates to a holdover value. Instead of using a 1 second holdover (as in the MICOM circuit) the hybrid algorithm uses a 0.25s holdover which results in few dropouts and does not overly extend a detect event.

V. Performance and Comparisons

The hybrid algorithm accurately performs the VOX function in low-noise as well as high-noise conditions. Figure

 ISBN: 978-1-61208-193-9

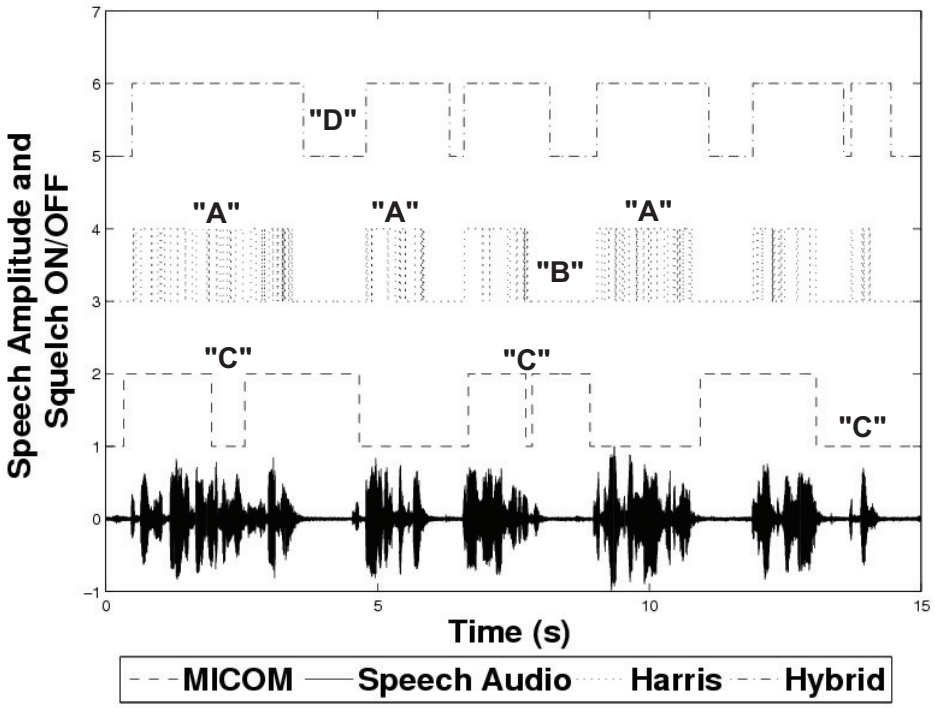

Figure 4. Performance of all three voice detection algorithms in a low noise, natural environment. The utterance was captured from an amateur radio transmission, and contains some non-speech noise. Annotations “A” through “D” indicate detection errors in each algorithm.

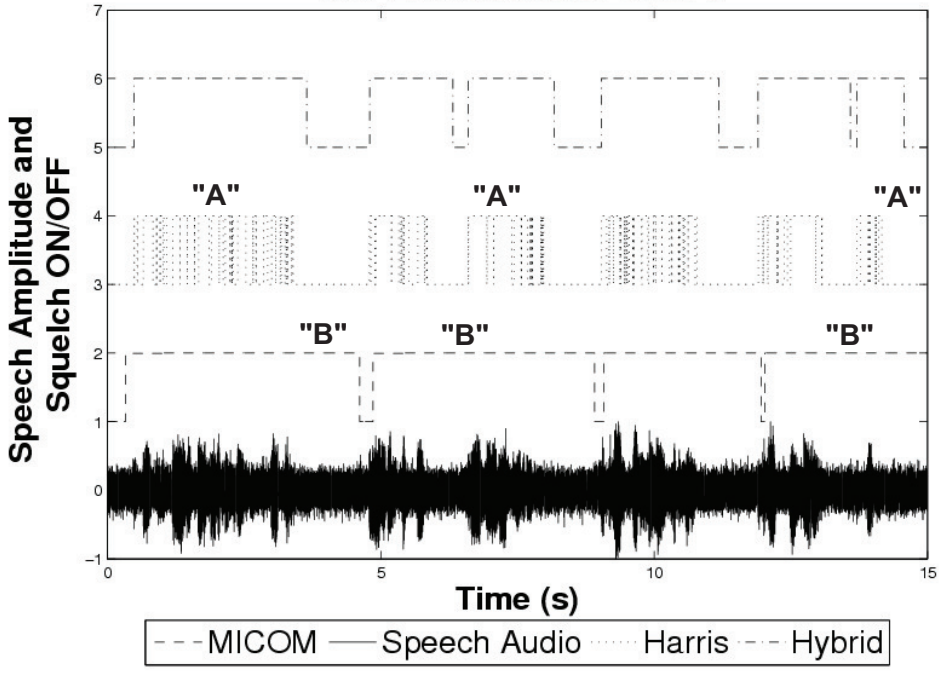

Figure 5. Performance of all three voice detection algorithms in high levels of additive Gaussian noise. In this case, the maximum noise amplitude is half the audio waveform maximum amplitude. Note the erratic performance of the Harris approach in voiced segments (“A”), and the inability of the MICOM approach to discriminate between noise and silence (“B”).

 ISBN: 978-1-61208-193-9

4 shows performance of the Harris, MICOM, and hybrid VOX implementations in low-noise conditions. Although Figure 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 the ability of the Harris approach to reliably (albeit aggressively) determine non-speech segments ("B"). Also note the inaccurate voiced/non-voiced decisions of the MICOM approach ("C"). The hybrid approach typically produces accurate voicing indicators with acceptable overhang, and without aggressive penetration into non-voiced segments. There are a few exceptions (e.g. a missed onset at "D"). However, this style of performance is quite acceptable for real-time implementation, which avoids clipping, slow-attack, and other behaviors which are objectionable to amateur radio operators.

The performance of the Harris, MICOM, and Hybrid approaches in noisy environments is shown in Figure 5, where Gaussian noise was added to a speech signal to simulate poor quality amateur radio channels. The additive noise amplitude is adjusted to be half of the waveform's maximum, or 6dBV down from the signal's peak amplitude. In the figure, several error conditions are labeled.

The top trace of Figure 5 shows voicing indicators generated by the hybrid implementation, which accurately track the voicing segments of the original speech, even in the presence of significant additive noise.

The second trace shows voicing indicators generated by the Harris algorithm, which switches erratically between ON/OFF states during voiced segments ("A"), generating numerous false positives and false negatives for voiced and unvoiced speech, as well as inter-word gaps.

The third trace from the top of Figure 5 shows voicing indicators generated by our software emulation of the the MICOM VOX system. In noisy environments, the MICOM system remains in the ON or "voicing" state for the majority of the utterance, and has difficulty discriminating between noise and silence ("B").

Neither the MICOM VOX nor the Harris algorithm are sufficiently robust to generate stable voicing indicators in the presence of mild to moderate additive noise. Furthermore, and not discussed here in detail, the MICOM and Harris approaches are highly susceptible to colored noise, tone bursts, and in-band data.

The hybrid implementation works significantly better than the other two approaches even though the thresholds of the other systems were carefully set to extract maximum performance for the tests typified by Figure 5. In contrast with the other approaches, the hybrid system meshes the MICOM and Harris extremes together and tracks the speech in real-time, with minimal computational burden, and only a small, configurable detection delay.

To complete our analysis, the hybrid algorithm was also tested using several "in-band" data transmissions which are popular in amateur radio [12]. In-band tests included modulation schemes such as WSJT, CW1, PSK31, FSK, Pactor 1&2, and RTTY. Figure 6 provides the combined results of this testing. As shown in the figure, none of these modulation schemes triggered a speech detection event in the hybrid VOX, which would have been indicated by a low-to-high excursion of the voicing indicator. In the figure, the voicing indicator is shown as a dotted line just above each data sequence.

This testing demonstrates the robustness and stability of the hybrid approach in realistic applications and environments. These results are important in amateur radio and infrastructure applications where operators rely on hands-free VOX operation and robust voicing detection in noisy channels.

VI. Conclusion

The results of our comparison of VOX systems has shown that a combination of features from hardware-driven and software-driven approaches provides a robust and low complexity system capable of meeting important application requirements in a variety of environments.

In particular, amateur radio channels with in-band data transmissions and significant noise and non-speech interference are well-served by the hybrid VOX system. The approach described here combining venerable techniques with newer signal processing approaches and emulated hardware behaviors results in a stable, sensitive speech detection algorithm.

Further development and testing will improve the performance of the hybrid implementation in other environments and in different applications. Specifically, work is ongoing to compare the hybrid VOX system to well-known VAD schemes via standardized test frameworks, such as [13].

References

[1] *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.

[2] F. Reid and L. David, "Smart Squelch for SSB," *73 Magazine*, pp. 44–49, Aug. 1982.

[3] J. Tarsala and R. Hammock, "The Jet Propulsion Laboratories 'Smart VOX'," Available: http://www.repeater-builder.com/projects/jpl-vox-sq/ssb-squelch.html, 2005.

[4] *Multisim 11.0*, National Instruments, Jul. 2011.

[5] *Simulink 2011a*, MathWorks Inc., Jan. 2011.

 ISBN: 978-1-61208-193-9

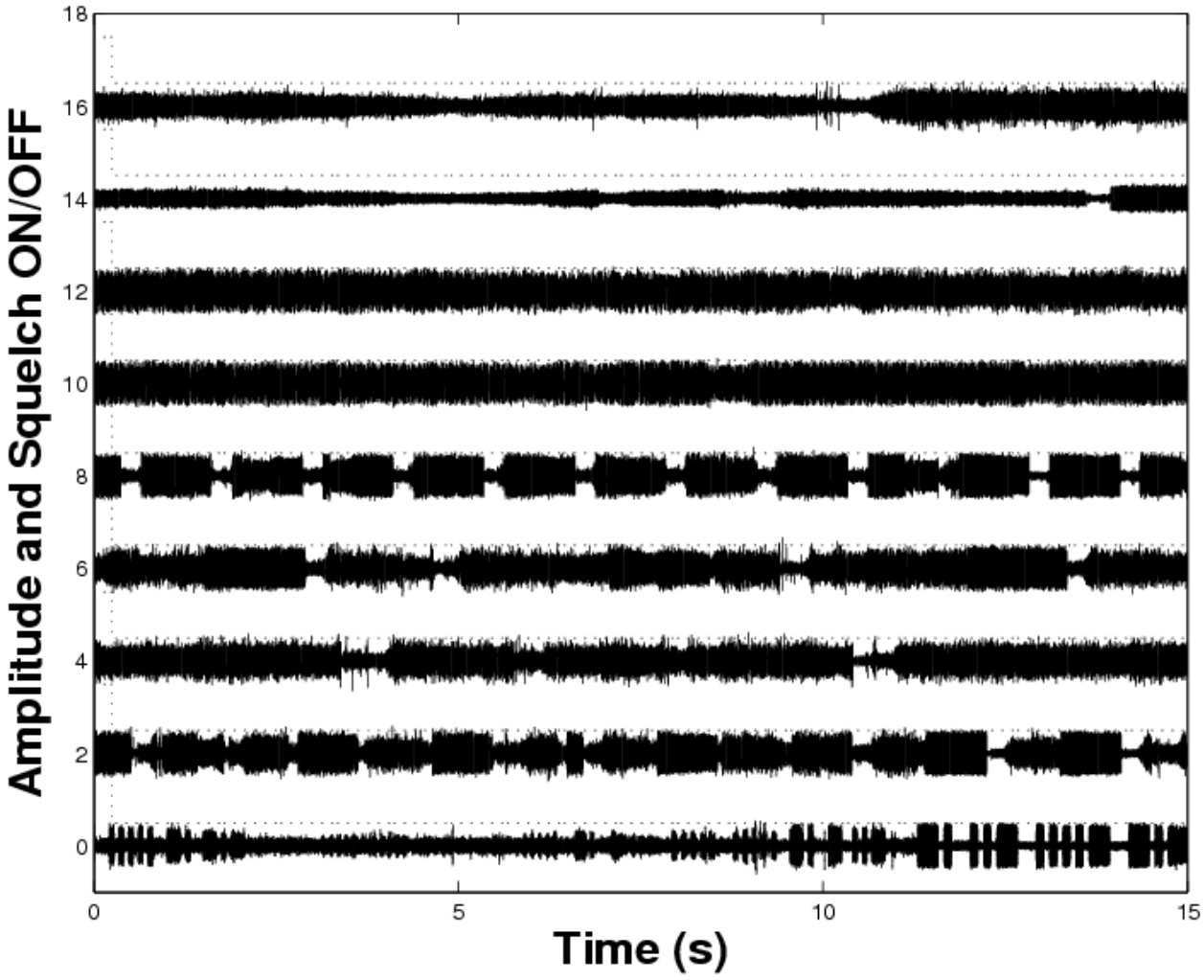

Figure 6. Performance of the hybrid system for common amateur radio data transmissions [12]. Top to bottom: WSJT, RTTY, Strong PSK31, PSK31, PACTOR2, PACTOR1, Noise, FSK, and CW1. The hybrid voicing indicator is shown as a dotted line just above each data sequence.

[6] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, Oct. 1976.

[7] L. Rabiner and M. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 4, pp. 338–343, Aug. 1977.

[8] S. McClellan and J. Gibson, "Variable-rate CELP based on subband flatness," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 2, pp. 120–130, Mar. 1997.

[9] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 14, no. 2, pp. 502–510, Mar. 2006.

[10] B. Cox and L. Timothy, "Nonparametric rank-order statistics applied to robust voiced-unvoiced-silence classification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 5, pp. 550–561, Oct. 1980.

[11] M. Webster, G. Sinclair, and T. Wright, "An efficient, digitally-based, single-lag autocorrelation-derived, voice-operated transmit (VOX) algorithm," in *Military Communications Conference (MILCOM'91)*, vol. 3, Nov. 1991, pp. 1192–1196.

[12] *The ARRL Handbook for Radio Communications*, 89th ed., The American Radio Relay League, Newington, CT, Oct. 2011.

[13] *IEEE Std 269-2010 (Revision of IEEE Std 269-2002): IEEE Standard Methods for Measuring Transmission Performance of Analog and Digital Telephone Sets, Handsets, and Headsets*, IEEE, 2010.

 ISBN: 978-1-61208-193-9