BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 10/10 pp106–114
Volume 6, Number 1, March 2023
DOI: 10.26599/BDMA.2022.9020025

Ultra-Short Wave Communication Squelch Algorithm Based on Deep Neural Network

Yuanxin Xiang, Yi Lv*, Wenqiang Lei, and Jiancheng Lv

Abstract: The squelch problem of ultra-short wave communication under non-stationary noise and low Signal-to-Noise Ratio (SNR) in a complex electromagnetic environment is still challenging. To alleviate the problem, we proposed a squelch algorithm for ultra-short wave communication based on a deep neural network and the traditional energy decision method. The proposed algorithm first predicts the speech existence probability using a three-layer Gated Recurrent Unit (GRU) with the speech banding spectrum as the feature. Then it gets the final squelch result by combining the strength of the signal energy and the speech existence probability. Multiple simulations and experiments are done to verify the robustness and effectiveness of the proposed algorithm. We simulate the algorithm in three situations: the typical Amplitude Modulation (AM) and Frequency Modulation (FM) in the ultra-short wave communication under different SNR environments, the non-stationary burst-like noise environments, and the real received signal of the ultra-short wave radio. The experimental results show that the proposed algorithm performs better than the traditional squelch methods in all the simulations and experiments. In particular, the false alarm rate of the proposed squelch algorithm for non-stationary burst-like noise is significantly lower than that of traditional squelch methods.

Key words: squelch; Gated Recurrent Unit (GRU); ultra-short wave communication

1 Introduction

Ultra-short wave communication is widely used in the aviation field[1], it is the most fundamental and important function in the field of military aviation and civil aviation. The speech quality of ultra-short wave communication directly affects the efficiency of command release and air traffic control. But the ultra-short wave signals are disturbed by various noises under the increasingly complex electromagnetic environment and increasing communication distance. Besides, the factors like the energy of channel noise, the sensitivity of the receiving antenna, and the Doppler motion all largely affect the communications of ultra-short wave signals and the quality of the final speech. Therefore, the voice after demodulation contains lots of background noises which will affect the comfort of long-term monitoring and even the intelligibility of the voice.

Voice squelch[2] is an effective technique that can reduce the influence of noise on the quality of communication voice. Its main purpose is to automatically turn off the voice channel when the signal demodulated by the ultra-short wave receiver is noise or has an extremely low Signal-to-Noise Ratio (SNR). It can prevent the user from hearing noises and help achieve a comfortable communication environment.

The voice squelch in ultra-short wave communication is different from the Voice Activity Detection[3] (VAD) in speech signal processing in the following two aspects:

First, they face different noises with different characteristics. For example, VAD faces noise with a

• Yuanxin Xiang, Wenqiang Lei, and Jiancheng Lv are with the College of Computer Science, Sichuan University, Chengdu 610000, China. E-mail: yuanxinxiang@scu.edu.cn; wenqianglei@gmail.com; lvjiancheng@scu.edu.cn.
• Yi Lv is with the Sichuan Research Institute, Shanghai Jiao Tong University, Chengdu 610000, China. E-mail: lvyi_ivy@qq.com.
* To whom correspondence should be addressed.
Manuscript received: 2022-07-13; accepted: 2022-07-28

relatively stable energy level. But the noise is non-stable in ultra-short wave communication since the existence of the automatic gain control circuit in the ultra-short wave receiver. The noise is weak when the received ultra-short wave signal is strong, or the noise is strong when the received ultra-short wave signal is weak. This makes it impossible to calculate the speech existence probability through the adaptive noise estimation method based on statistical models[4].

Second, their application purposes are different. The VAD algorithm is usually applied to the preprocessing of speech recognition, which aims to quickly and precisely distinguish between words using the recognition of speech and non-speech segments. Meanwhile the voice squelch algorithm in ultra-short wave communication aims to robustly avoid the noise affecting the comfort of hearing during long-term monitoring. The difference between the two application purposes leads to a big difference in the final hangover strategy.

Traditional squelch algorithms are based on the energy of carrier and audio[5]. They first determine the existence of a signal using the statistical characteristics of a carrier, and then determine whether the signal is a speech signal through the energy-based audio algorithm. However, the traditional squelch algorithms do not work well under complex or poor electromagnetic environments. In these environments, the traditional squelch algorithms have to face problems, such as long-distance communication, weak signal reception, low SNR, low carrier energy, and high noise energy. In these cases, the noise energy can frequently or even continuously break the energy threshold of the traditional squelch algorithms and affect auditory comfort.

In recent years, deep learning based speech processing methods have achieved good performance in non-stationary noise environments, prompting considerable research and applications in conference systems, smart speakers, True Wireless Stereo (TWS), and other fields[6–13]. Deep learning based speech processing methods are mainly divided into two categories: spectral mapping and mask estimation. The spectral mapping method uses the frequency spectrum as the input feature and the training target, and uses the Feedforward Neural Network (FNN) model as the regression model to achieve the mapping from the logarithmic power spectrum of the noisy speech to the logarithmic power spectrum of the target speech[14–17]. The mask estimation method takes different acoustic features as the input, uses neural network models to estimate different masks, processes the spectrum of a noisy speech using masks, and then recovers the noisy speech signal to the time domain[18–20]. The FNN model has a poor generalization ability to speech data from different speakers, but this problem is significantly alleviated by treating the speech enhancement problem as a sequence-to-sequence mapping problem using the Long Short-Term Memory (LSTM) layer[21]. Thus, the related data-driven methods are more suitable for the detection of non-stationary burst-like noise in ultra-short wave communication. However the generalization ability and robustness of a single neural model architecture are relatively limited, so it cannot fully satisfy the requirements of high robustness in squelch.

Since neither the traditional squelch algorithm nor the deep learning based squelch algorithm can solve the squelch problem for ultra-short wave communication alone, we propose a new squelch algorithm that combines the advantages of a deep neural network and the traditional energy decision method. We build a three-layer Gated Recurrent Unit (GRU)[22] and take the speech banding spectrum as a feature to get the speech existence probability. Then the final squelch result is got by comprehensively considering the speech existence probability and the decision result of the traditional energy-based algorithm.

2 Analysis of the Principle of Ultra-Short Wave Voice Reception

Figure 1 shows the schematic block diagram of the current typical ultra-short wave communication in the aviation field. The ultra-short wave signal is first received by the antenna. Next the energy probability of the Radio Frequency (RF) signal is controlled by the analog Automatic Gain Control (AGC) in the ultra-short wave

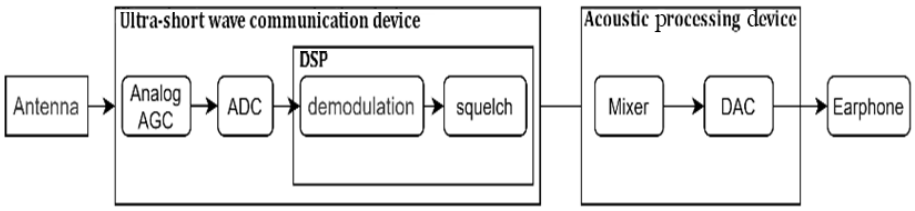

Fig. 1 Block diagram of the principle of ultra-short wave voice reception.

communication device. Then the signal is converted to a digital signal through the Analog-to-Digital Conversion (ADC). Lastly, it is demodulated in the Digital Signal Processor (DSP) and performed the squelch processing. When the ultra-short wave communication device judges that there is speech in the signal, the signal will be packeted and sent to the audio processing system at a certain time interval. After the volume control and the mixing operations are performed inside the acoustic processing device, the signal is converted to an analog signal through the Digital-to-Analog Conversion (DAC). Finally, the audio is outputed to the headphones.

The current noise suppression methods are relatively mature for the noise in the Frequency Modulation (FM) and Amplitude Modulation (AM) of the ultra-short wave communication system, however, they have problems distinguishing between useful signal and noise after the analog AGC circuit amplifies all passing signals. The noise in the FM and AM of the ultra-short wave communication system is relatively single and regular, in which the noise in AM is similar to white noise, and the noise in FM is colored noise. These noises have corresponding mature noise suppression methods. Nonetheless, the analog AGC circuit makes the noise hard to be distinguished from useful signals by amplifying all the passing signals without distinguishing between useful signals and noise, the noise is week in the speech segments, but it is amplified and strong in the non-speech segments.

Figure 2 shows typical waveform data of the ultra-short wave audio received by the earphone under the current system. As shown in the marked section, if the squelch judgment is wrong, it will cause strong noise in the non-speech segment. This kind of noise has the characteristics of strong suddenness, high energy, non-stationary, etc., which is the key problem to be solved by squelch algorithms.

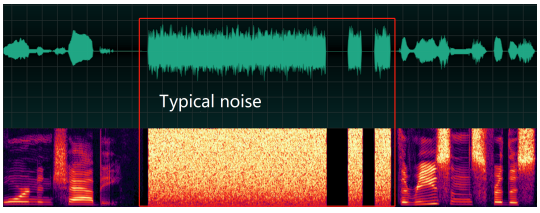

Fig. 2 Typical noise type of squelch.

3 Principle of the Proposed Algorithm

3.1 System block diagram

Figure 3 shows the implementation block diagram of the proposed algorithm, which consists of four parts: feature extraction, Deep Neural Network (DNN) model, energy calculation, and comprehensive decision.

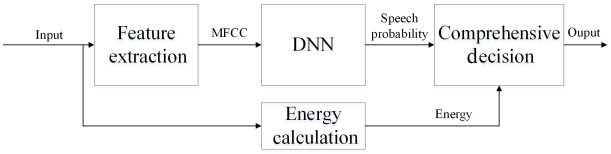

Fig. 3 System block diagram of the proposed algorithm, where MFCC refers to the mel-frequency cepstral coefficients feature extraction method, which is a leading approach for speech feature extraction[23].

For audio signals demodulated from the carrier, we use the frame processing method. Considering the short-term stability of the audio signal, we use 20 ms as the duration of each frame, each frame has an offset of 10 ms and a 50% overlap, and the signal is truncated by the Hanning window[24].

Each frame is processed in two branches: one is the traditional energy-based decision method, the other is the proposed deep neural network based decision method. For the first branch, we calculate the root mean square of each frame as their energy level. For another branch, we first convert the signal into frequency domain using the Fast Fourier Transformation (FFT), second, we extract the features using the sub-band processing method. Then we feed the extracted features into the proposed deep neural network to obtain the speech existence probability of the frame. Finally, the energy level and the speech existence probability of each frame are fed to the comprehensive decision module to get the final squelch result for each frame and decide whether the frame is valid or should be turned off.

The details of each module are described in the following summary.

3.2 Feature extraction

Referring to the related speech recognition processing, we adapt the 39 dimensions MFCC features as the input characteristic of each frame signal (including 12 dimensions MFCC, 12 dimensions first difference MFCC, 12 dimensions second difference MFCC, 1 frame energy, 1 first difference frame energy, and 1 second difference frame energy). The MFCC is a cepstrum parameter extracted in the Mel scale frequency domain, and the Mel scale describes the nonlinear characteristics of the human ear frequency. The selected specific features and processing flow are shown in Fig. 4.

3.2.1 Deep neural network

(1) Model architecture

We choose GRU as the core network since we consider

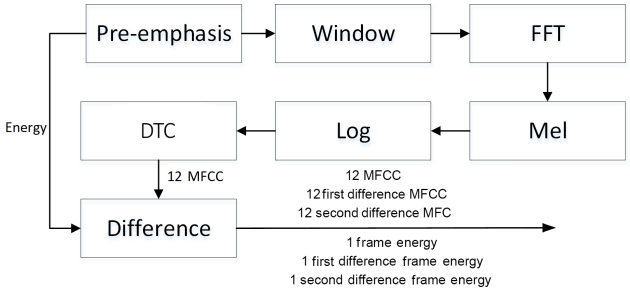

Fig. 4 Selected features. DCT refers to discrete cosine transform, which is a widely used transformation technique in signal processing and data compression.

the following factors: First, the audio signal is a time-series signal, and at the same time it is a quasi-stationary process, which has a strong correlation in the time domain. Second, the GRU has a better achievability of subsequent avionics embedded systems.

The model architecture is shown in Fig. 5. A 3-layer neural network architecture is built. The input layer and the output layer are standard fully connected networks, and the middle layer is composed of a GRU with a width of 32 as the core model for distinguishing between speech and noise.

(2) Training data

The training data are built by mixing noise signals and pure speech signals according to different rules. The noise data include the real noise from different ultra-short wave communication devices, and the demodulated baseband signals are simulated by software under FM, AM communication, and various SNR environments. Besides, to improve the generalization of the algorithm, the noises from other audio environments are added as training data (a total of 104 kinds of noises, including

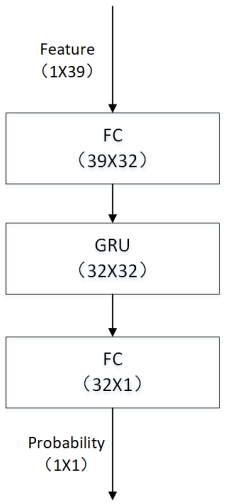

Fig. 5 Architecture of the proposed deep neural network model.

common colored noise, wind noise, and electromagnetic environment noise). For the pure speech data, we use the open-source TIMIT speech database[25,26].

Then the training data are generated by superimposing the noise data and the pure speech data according to different SNRs. The SNR is randomly distributed from 20 dB to –5 dB. In addition, for the characteristics of the ultra-short wave communication that the noise is strong when the signal is weak and the noise is weak when the signal is strong, a large number of strong non-stationary burst-like noise data are added to the training corpus.

(3) Objective function

As the final output is the speech existence probability, the standard cross entropy loss function is chosen as the objective function,

$$L = -\sum_{i=1}^{n} y_i \ln(y_i') \tag{1}$$

where y_i is the true label and y_i' is the Softmax probability for the i-th class.

3.2.2 Comprehensive decision

The process of the comprehensive decision is shown in Fig. 6. The squelch opening threshold is set by a combination of the energy level and speech existence probability. We focus on the case of weak speech and strong noise. The high-energy frames are mainly evaluated by the speech existence probability calculated by the neural network, and they will be continuously classified as speech frames only when their speech existence probability is higher than a high threshold, which process avoids the influence of long-term noise on speech.

The detailed processes are described as follows: when a frame is inputted, if the previous frame is a non-speech frame, and if the frame energy is greater than the high energy threshold ZT1, it is determined to be a speech frame, and the previously pending frame is set to a speech frame, too. Otherwise, if the frame energy is lower than the high energy threshold ZT1, it is determined as a pending frame if the frame energy is greater than the low energy threshold ZT2 or the speech existence probability is greater than the low probability threshold P1. When a frame is inputted, if the previous frame is a speech frame, it will be continuously set as a speech frame if the frame energy is greater than the low energy threshold ZT2 and the speech existence probability is greater than the high probability threshold P2. If the previous frame is a speech frame but the consecutive N frames do not satisfy the above

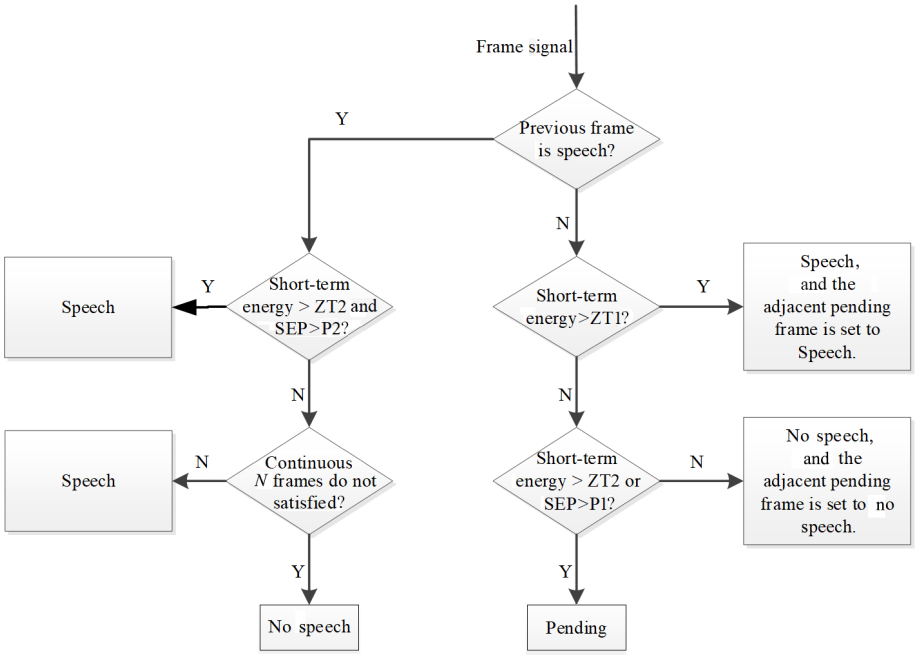

Fig. 6 Process of the comprehensive decision.

requirements, according to the hangover strategy, the current frame is determined to be a non-speech frame and then the squelch process is activated.

4 Simulation

The simulations and experiments of the proposed algorithm are introduced in this section. We first simulate the demodulation of baseband signals with different SNRs in the typical AM and FM communication modes in ultra-short wave communication, and confirm the effectiveness and performance of the proposed algorithm. We then simulate the non-stationary noise and verify the effectiveness and adaptability of the proposed algorithm. At last, we verify the proposed algorithm on the real received signals of ultra-short wave devices to confirm its generalization ability and practicability.

4.1 Simulation under AM communication

We use MATLAB to simulate and generate the baseband signal processed by modulation, channel noise simulation, and demodulation in AM communication mode.

The detailed simulation flow chart is shown in Fig. 7. We first modulate the pure speech signal, then add noise to the modulated signal according to the set SIgnal to Noise And Distortion (SINAD) ratio and the additive white Gaussian noise channel model, next use the demodulation algorithm to obtain the baseband signal. Afterward we use different squelch algorithms to process the baseband signal to get the squelch results. The results are compared to evaluate the effectiveness of the proposed algorithm.

The noise in AM communication has the following characteristics: it is mainly additive noise, and the noise of the baseband signal is in the form of white noise, directly relate to the SINAD, and negatively correlated with the SNR.

Simulations are performed on signals with SNR of 20 dB, 15 dB, 10 dB, and 5 dB. The final sampling rate for the baseband audio signal is 8000 Hz, the frame

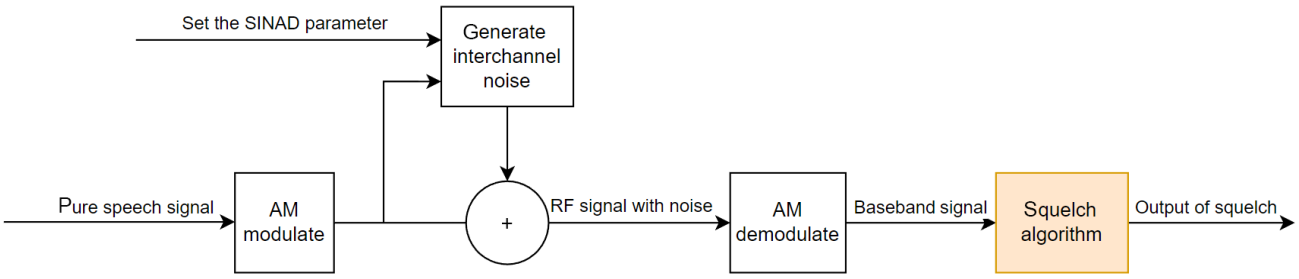

Fig. 7 Schematic diagram of the simulation method for AM communication.

length is 16 ms, and the overlap between frames is 8 ms. We use energy judgment or neural network algorithm to preliminarily judge whether a frame is a speech frame, and then decide the final squelch result through the hangover test. VAD (Sohn)[14] is a traditional and one of the most widely used VAD methods at present. It distinguishes speech and non-speech frames based on the signal energy through the minimum mean square error criterion. Thus, we choose it as a comparison method.

We use the non-speech hit-rate (HR0) and the speech hit-rate (HR1) as the evaluation metrics[27,28] to evaluate the performance of the proposed algorithm and VAD (Sohn) method,

$$\mathrm{HR0} = \frac{N_{0,0}}{N_0^{ref}} \tag{2}$$

$$\mathrm{HR1} = \frac{N_{1,1}}{N_1^{ref}} \tag{3}$$

where N_0^{ref} and N_1^{ref} denote the numbers of non-speech and speech frames in the original audio, respectively. $N_{0,0}$ and $N_{1,1}$ denote the numbers of non-speech and speech frames correctly recognized by the squelch algorithms, respectively.

Figures 8 and 9 show the HR1 and HR0 results under different SNRs, respectively.

The HR1 results show that the proposed deep neural network based squelch algorithm is significantly higher than the traditional energy-based VAD (Sohn) method in terms of the speech hit-rate, especially when SNR is low. The main reason is that the deep neural network based method effectively utilizes the characteristics of the speech spectrum, including energy, spectral characteristics, and the correlation between the previous and subsequent frames, while the energy-based method only utilizes the frame energy or the subband energy and can be easily fooled by the situation that the SNR is low and the noise energy is much higher than the speech energy. Thus, the proposed method is less affected by

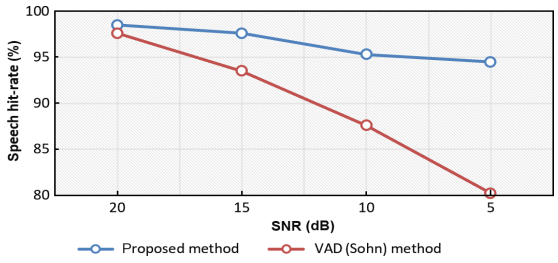

Fig. 8 Simulation results of HR1 on AM under different SNRs.

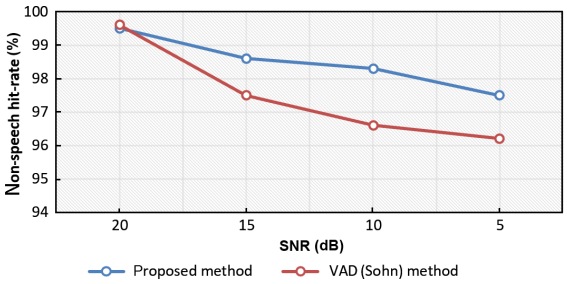

Fig. 9 Simulation results of HR0 on AM under different SNRs.

SNR on the speech hit-rate and it performs much better than the VAD (Sohn) method.

The HR0 results show that the proposed algorithm is slightly better than the traditional VAD (Sohn) method on the non-speech hit-rate. It is because the noise in AM communication is close to white noise and has strong stability, both methods can effectively detect the noise frame.

Figure 10 shows the squelch results of a speech segment under a 10 dB SNR. The red line and green line show the squelch result of the proposed algorithm and VAD (Sohn) method, respectively, and both methods use the same hangover strategy. The proposed algorithm completely detects the entire speech, whereas the VAD (Sohn) method misjudges the low SNR speech segment from the sample number 1.5×10^4 to 1.8×10^4 as noise and erroneously starts the squelch operation.

In summary, in AM communication, the proposed algorithm has a significantly better speech detection accuracy than the traditional energy-based method.

4.2 Simulation under FM communication

We do similar simulations and experiments on FM communication to further evaluate the effectiveness of the proposed algorithm. The noise of the baseband signal in FM communication is different from that of the

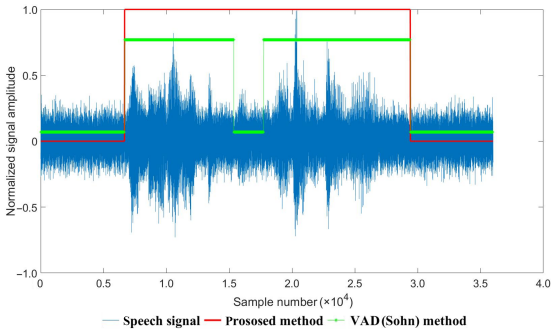

Fig. 10 Squelch results of a speech segment under 10 dB SNR.

baseband signal in AM communication. The phase of the main channel noise in FM communication is disordered, resulting in different noises in the final demodulated baseband signal. When the SNR is low, due to the fast phase change, the final demodulated baseband signal has lots of non-stationary noises which are similar to flicker noise. Figures 11 and 12 compare the HR1 and HR0 results of the two methods under different SNRs in FM communication, respectively.

The HR1 result shows that the proposed algorithm is better than the traditional VAD (Sohn) method on the speech hit-rate. The HR0 result shows that the proposed algorithm is significantly better than the VAD (Sohn) method on the non-speech hit-rate. The main reason is that the traditional method will misjudge the speech frames under the interference of non-stationary noises. The proposed algorithm can alleviate this kind of problem well.

Figure 13 shows the simulation results under 5 dB SNR. The entire speech segment is correctly detected by the proposed algorithm, while the VAD (Sohn) method makes multiple wrong judgments affected by the non-stationary noises.

In summary, the proposed algorithm is also effective under FM communication and performs significantly better than the traditional method.

4.3 Simulation under non-stationary noise situation

In this study, the proposed method is simulated under

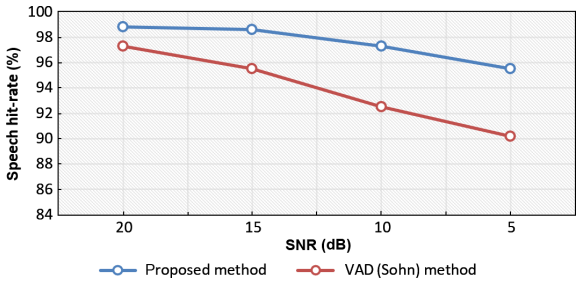

Fig. 11 Simulation results of HR1 on FM under different SNRs.

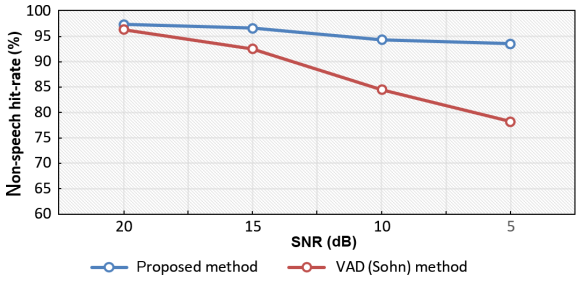

Fig. 12 Simulation results of HR0 on FM under different SNRs.

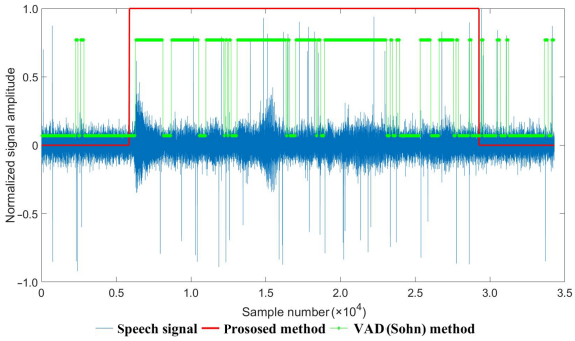

Fig. 13 Simulation results under a 5 dB SNR.

the most typical and bad situation in ultra-short wave communication where the carrier energy is weak and the noise energy is strong. In this kind of situation, the traditional method can hardly work, which will cause long-term noise interference. Figure 14 shows simulation results of our method and the traditional VAD (Sohn) method. The data simulate a long-lasting high-energy noise following a speech. The traditional energy-based squelch algorithm is easy to misjudge the noise as the speech signal under this situation, thus causing the receiver to be in an environment with strong noise for a long time, causing a bad influence on the communication.

While the proposed algorithm not only effectively detects the speech in the previous section, but also recognizes the high-energy noise in the back end after a short convergence, avoiding the long-term interference of the noise. The short convergence is mainly for reducing the missed alarm rate and avoiding missing words. In the process from the signals without speech to the signal with speech, we intentionally allowed our algorithm to focus on the energy judgment, and supplemented by speech probability, allowing short-term convergence, to improve the robustness of our algorithm.

4.4 Validation under real noise environment

To verify the generalization ability of our algorithm,

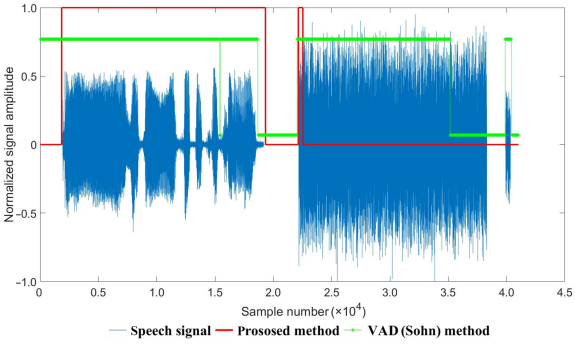

Fig. 14 Simulation results on non-stationary noise.

we simulated the actual noise environment of an ultra-short wave communication device with an SNR of 5 dB, which is the noise data that out of the training library. Figure 15 shows the simulation results, the entire speech segment is still detected, which verifies the effectiveness of our algorithm against the noise out of the training library and ensures that the proposed algorithm has a good generalization ability.

5 Conclusion

We proposed a new squelch algorithm for ultra-short wave communication, it is based on a deep neural network and traditional energy decision to solve the problem that the traditional squelch algorithms for ultra-short wave communication do not work well under non-stationary noise and low SNR in the complex electromagnetic environment. We use the speech banding spectrum as the features and use a 3-layer GRU as the model to judge the existence probability of the voice, and then the integrated squelch result is computed by integrating the existence probability and the energy threshold based decision. For the low energy signal, the final decision is mainly based on the energy threshold and the speech probability is auxiliary. For the high-energy signal, we use speech probability as the main factor and energy judgment as a supplement. At the same time, to reduce the missed alarm rate, the energy level and the speech existence probability are used as the primary basis for detecting the emergence of the speech signal, and the speech existence probability is the main evidence for detecting the disappearance of the speech signal.

The simulation and experimental results on AM and FM communication modes show that the proposed algorithm is significantly better than the traditional VAD (Sohn) method. At the same time, the simulation and experimental results in the non-stationary noise, noise out of the training library, and noise from real radio data show that the proposed algorithm has a strong generalization on the squelch for multiple noises. In addition, the proposed algorithm performs well under non-stationary burst-like noise, making it highly suitable for the squelch of ultra-short wave communication than the traditional methods.

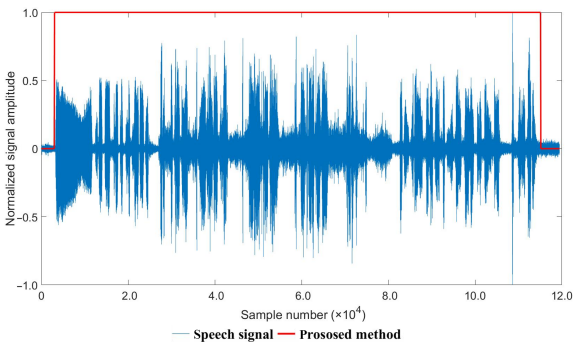

Fig. 15 Simulation results on real noise.

References

[1] Z. Wang, Application and development of civil aviation communication technology, (in Chinese), *China Civil Aviat.*, vol. 9, no. 1, p. 231, 2020.

[2] W. W. Wang, Voice squelch method in civil aviation VHF anti-jamming transceiver, (in Chinese), CN Patent CN112532259A, March 19, 2021.

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

[4] I. Cohen, Relaxed statistical model for speech enhancement and a priori SNR estimation, *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 5, pp. 870–881, 2005.

[5] Y. B. Li, An effective integrated processing method for quieting tone and its application, (in Chinese), *Telecommun. Eng.*, vol. 52, no. 1, pp. 54–57, 2012.

[6] D. L. Wang and J. T. Chen, Supervised speech separation based on deep learning: An overview, *IEEE/ACM Trans. Audio, Speech, Lang. Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] F. G. Liu, Z. W. Zhang, and R. L. Zhou, Automatic modulation recognition based on CNN and GRU, *Tsinghua Science and Technology*, vol. 27, no. 2, pp. 422–431, 2022.

[8] X. D. Tang, J. X. Guo, P. Li, and J. C. Lv, A surgical simulation system for predicting facial soft tissue deformation, *Comput. Visual Media*, vol. 2, no. 2, pp. 163–171, 2016.

[9] X. L. Xu, T. Gao, Y. X. Wang, and X. L. Xuan, Event temporal relation extraction with attention mechanism and graph neural network, *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 79–90, 2022.

[10] B. M. Oloulade, J. L. Gao, J. M. Chen, T. F. Lyu, and R. Al-Sabri, Graph neural architecture search: A survey, *Tsinghua Science and Technology*, vol. 27, no. 4, pp. 692–708, 2022.

[11] C. Y. Hou, J. W. Wu, B. Cao, and J. Fan, A deep-learning prediction model for imbalanced time series data forecasting, *Big Data Mining and Analytics*, vol. 4, no. 4, pp. 266–278, 2021.

[12] S. K. Patnaik, C. N. Babu, and M. Bhave, Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks, *Big Data Mining and Analytics*, vol. 4, no. 4, pp. 279–297, 2021.

[13] F. Fourati and M. S. Alouini, Artificial intelligence for satellite communication: A review, *Intell. Converged Netw.*, vol. 2, no. 3, pp. 213–243, 2021.

[14] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, An experimental study on speech enhancement based on deep neural networks, *IEEE Signal Proc. Lett.*, vol. 21, no. 1, pp. 65–68,

2014.

[15] J. M. Valin, A hybrid DSP/Deep learning approach to real-time full-band speech enhancement, in *Proc. IEEE 20th Int. Workshop on Multimedia Signal Processing*, Vancouver, Canada, 2018, pp. 1–5.

[16] X. H. Le, H. S. Chen, K. Chen, and J. Lu, DPCRN: Dual-path convolution recurrent network for single channel speech enhancement, in *Proc. 22nd Annu. Conf. Int. Speech Communication Association*, Brno, Czechia, 2021, pp. 2811–2815.

[17] Z. Z. Xu, T. Jiang, C. Li, and J. C. Yu, An attention-augmented fully convolutional neural network for monaural speech enhancement, in *Proc. 12th Int. Symp. Chinese Spoken Language Processing*, Hong Kong, China, 2021, pp. 1–5.

[18] Y. X. Wang and D. L. Wang, A deep neural network for time-domain signal reconstruction, in *Proc. 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, South Brisbane, Australia, 2015, pp. 4390–4394.

[19] Y. H. Wang, W. X. Zhang, Z. Wu, X. X. Kong, Y. B. Wang, and H. X. Zhang, Noise modeling to build training sets for robust speech enhancement, *Appl. Sci.*, vol. 12, no. 4, p. 1905, 2022.

[20] L. M. Zhou, Y. Y. Gao, Z. L. Wang, J. W. Li, and W. B. Zhang, Complex spectral mapping with attention based convolution recurrent neural network for speech enhancement, arXiv preprint arXiv: 2104.05267, 2021.

[21] K. Tan, B. Y. Xu, A. Kumar, E. Nachmani, and Y. Adi, SAGRNN: Self-attentive gated RNN for binaural Speaker separation with interaural cue preservation, *IEEE Signal Proc. Lett.*, vol. 28, pp. 26–30, 2021.

[22] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555, 2014.

[23] A. Hossan, S. Memon, and M. A. Gregory, A novel approach for MFCC feature extraction, in *Proc. 4th Int. Conf. Signal Processing and Communication Systems*, Gold Coast, Australia, 2010, pp. 1–5.

[24] J. O. Smith III, *Spectral Audio Signal Processing*. Stanford, CA, USA: W3K Publishing, 2011.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, https://catalog.ldc.upenn.edu/LDC93s1,2022.

[26] N. Chanchaochai, C. Cieri, J. Debrah, H. W. Ding, Y. Jiang, S. S. Liao, M. Liberman, J. Wright, J. H. Yuan, J. H. Zhan, et al., GlobalTIMIT: Acoustic-phonetic datasets for the world's languages, in *Proc. 19th Annu. Conf. Int. Speech Communication Association*, Hyderabad, India, 2018, pp. 192–196.

[27] 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 Commun.*, vol. 42, nos. 3&4, pp. 271–287, 2004.

[28] D. Ghosh, R. Muralishankar, and S. Gurugopinath, Robust voice activity detection using frequency domain long-term differential entropy, in *Proc. 19th Annu. Conf. Int. Speech Communication Association*, Hyderabad, India, 2018, pp. 1220–1224.

Yuanxin Xiang received the BEng degree from Wuhan University of Technology, China in 2014. He received the MEng degree from National University of Singapore, Singapore in 2018. He is currently a research assistant at Sichuan University. His research interests include deep learning and natural language processing.

Wenqiang Lei received the PhD in computer science from the National University of Singapore, Singapore in 2019. He is currently a professor at the College of Computer Science, Sichuan University, Chengdu, China. His research interests cover natural language processing, particularly dialogue systems and discourse analysis, and conversational recommendations. He has published multiple papers at top conferences like ACL, IJCAI, AAAI, EMNLP, and WSDM. He served as PC member on toptier conferences, including ACL, EMNLP, SIGIR, and AAAI. He is a reviewer for journals like *ASLP* and *TKDE*. He is also the winner of ACM MM 2020 Best Paper Award.

Yi Lv received the BEng degree from Beijing Normal University, China in 2008, and the PhD degree from Institute of Acoustics, Chinese Academy of Sciences, China in 2013. He is now a senior engineer at Shanghai Jiao Tong University. His research interests focus on audio signal processing and avionics.

Jiancheng Lv received the PhD degree in computer science and engineering from the University of Electronic Science and Technology of China, Chengdu, China in 2006. He is currently a professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Prior to that, he was a research fellow at the Department of Electrical and Computer Engineering, National University of Singapore. He is the coauthor of the book *Subspace Learning of Neural Networks*. His research interests include neural networks and machine learning.