RADE: A Neural Codec for Transmitting Speech over HF Radio Channels

*David Rowe**, *Jean-Marc Valin*†

*Supported by a grant from Amateur Radio Digital Communications
†Xiph.Org Foundation

***Abstract*—Speech compression is commonly used to send voice over radio channels in applications such as mobile telephony and two-way push-to-talk (PTT) radio. In classical systems, the speech codec is combined with forward error correction, modulation and radio hardware. In this paper we describe an autoencoder that replaces many of the traditional signal processing elements with a neural network. The encoder takes a vocoder feature set (short term spectrum, pitch, voicing), and produces discrete time, but continuously valued quadrature amplitude modulation (QAM) symbols. We use orthogonal frequency domain multiplexing (OFDM) to send and receive these symbols over high frequency (HF) radio channels. The decoder converts received QAM symbols to vocoder features suitable for synthesis. The autoencoder has been trained to be robust to additive Gaussian noise and multipath channel impairments while simultaneously maintaining a Peak To Average Power Ratio (PAPR) of less than 1 dB. Over simulated and real world HF radio channels we have achieved output speech intelligibility that clearly surpasses existing analog and digital radio systems over a range of SNRs.**

1. INTRODUCTION

High-frequency (HF) push-to-talk (PTT) radio has the benefit of operating without infrastructure over ranges of several thousands of km. Applications include humanitarian, remote area, emergency and government communication when access to cellular and satellite systems cannot be guaranteed. The HF radio signal propagates from the transmitter to the receiver via reflection from upper layers of the atmosphere. Typically the signal is reflected multiple times by various sub-layers thus multiple, time shifted versions of the signal arrive and are summed at the receiver (multipath propagation). Single side band (SSB) – an analog communications system that was invented in 1915 [1] – entered widespread use in the 1950's [2] and remains the de facto standard on HF due to its power and bandwidth efficiency, and robustness to multipath propagation. The voice quality of analog and digital HF speech systems remains low compared to modern cellular and Internet telephony services and has not seen significant improvement in 70 years. Key requirements for HF voice services are voice quality, narrow RF bandwidth, low SNR operation, robustness to multipath channels, and latency of less than 200 ms to support PTT speech.

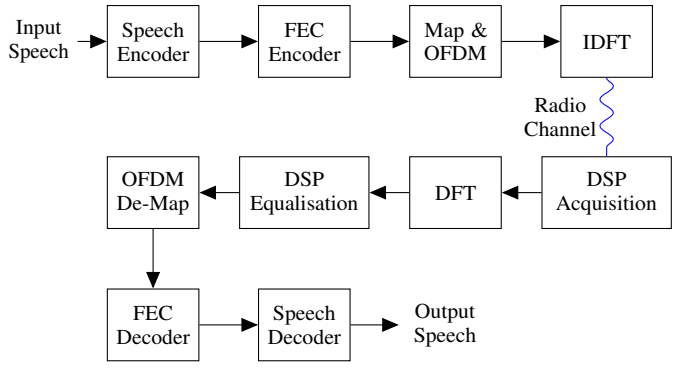

Fig. 1: Classical DSP speech over radio system employing OFDM modulation.

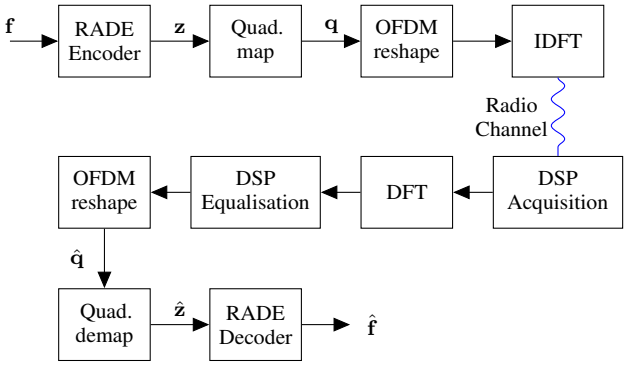

Fig. 2: RADE system, which employs ML combined with OFDM and classical DSP synchronisation. The features $\mathbf{f}$ are obtained from input speech using a feature extractor, and output speech is synthesised from $\hat{\mathbf{f}}$ using the FARGAN vocoder. The features $\mathbf{f}$ are updated at 100 Hz, the latent vector $\mathbf{z}$ at 25 Hz, and the sample rate over the channel is $F_s = 8000$ Hz.

Consider the classical DSP speech over radio system in Fig. 1. Speech samples from a microphone are compressed to a low bit rate using a speech encoder. Forward error correction (FEC) adds redundant bits to protect the sensitive payload speech bits from channel errors. The bitstream is then converted to a signal suitable for transmission over the radio channel using a modulator (e.g. a sequence of QAM symbols mapped to OFDM sub-carriers). The received radio signal is demodulated to a bit stream; the FEC decoder attempts to correct any bit errors, and the speech decoder converts the signal back into a sampled speech signal for replay over a loudspeaker. Classical DSP digital voice systems have the following drawbacks: separating the speech coding and channel protection (FEC) leads to inefficiencies; the use of largely linear DSP means the inability to exploit non-linear dependencies when compressing voice; difficulty in handling multipath channels; long latencies due to use of interleavers/FEC to overcome multipath fading; and low quality speech from low bit rate classical DSP vocoders. They exhibit an abrupt threshold SNR where the system ceases to work, and speech quality does not gracefully scale with available channel SNR (although step changes are possible with mode switching).

This paper proposes RADE (Fig. 2), a RADio autoEncoder [3] designed to efficiently transmit speech over HF radio channels inspired by the RDO-VAE structure from DRED [4].

The use of ML combines the operations of quantisation, channel encoding and modulation, and allows joint training of the entire end-to-end system to minimise distortion when subjected to the HF channel impairments [5], [6]. The powerful non-linear transforms and prediction available in ML allow us to more efficiently model the speech signal and time based evolution of the channel to improve robustness. We employ the FARGAN vocoder [7] for high quality neural speech synthesis, however our work is applicable to any neural and even classical vocoders with a similar feature set.

Our contributions in this paper are:

1) An autoencoder that combines the classical DSP functions of quantisation, channel coding, and modulation to generate discrete time but continuously valued (analog) QAM symbols directly from vocoder features. Unlike classical approaches there is no intermediate bit stream, and the QAM symbols (Fig. 3) emerge from the training process rather than being members of a well defined, discrete constellation.
2) A training procedure that minimises the end-to-end distortion of vocoder features in the presence of additive Gaussian noise and frequency-selective fading, while simultaneously generating an OFDM waveform with low Peak To Average Power Ratio (PAPR).

In Section 2 we describe how we have combined a neural encoder and decoder with OFDM to develop the RADE system. Training over HF channels at low PAPR is discussed in Section 3. For testing we have adopted an automatic speech recognition (ASR) based approach over simulated HF channels which we describe in Section 4, along with an over the air demonstration.

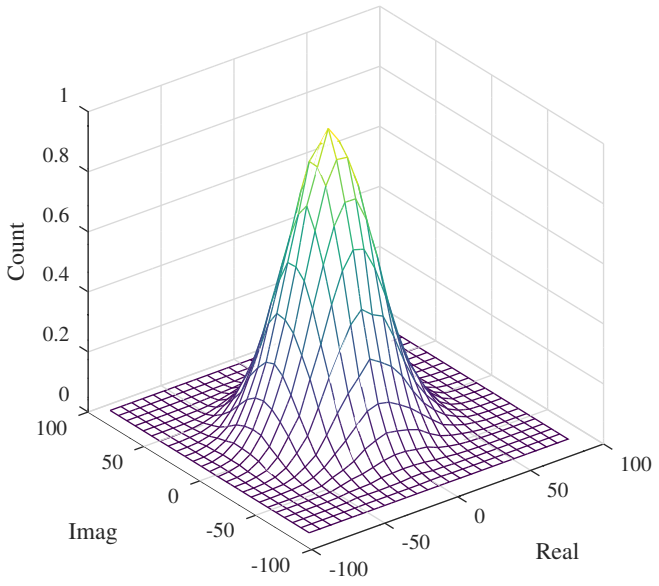

Fig. 3: 2D histogram of complex RADE encoder QAM symbols with all elements of $\mathbf{z}$ superimposed for a 60 second sample containing multiple speakers. The bin count has been scaled to a maximum of 1. Compared to digital QAM constellations, the RADE constellation looks like noise. Plotting individual elements of $\mathbf{z}$ produces a similar histogram (no obvious structure).

2. RADE DESIGN

Rather than operating directly on the time-domain signal like recent end-to-end neural codecs [8], RADE uses classical acoustic features similar to those used in LPCNet [9]. We use 20-dimensional feature vectors $\mathbf{f}$ that consist of 18 Bark-scale cepstral coefficients, the pitch period, and a voicing parameter. Using classical features not only reduces complexity, but makes it easy to replace the vocoder without breaking compatibility. Moreover, we show in Section 4 that our choice of features is not a limiting factor for our application.

The RADE encoder and decoder form an autoencoder that is trained to minimise the reconstruction loss $\mathcal{L}(\mathbf{f}, \hat{\mathbf{f}})$ between the input features $\mathbf{f}$ and the decoded features $\hat{\mathbf{f}}$, as defined in Eq. (12) of [4]. The feature vectors $\mathbf{f}_n$ generated every 10 ms are concatenated and passed to the RADE encoder every 40 ms (frame rate of 25 Hz). They are transformed to a $d = 80$ dimensional vector $\mathbf{z}$ by a stack of 1D convolutional (conv) and gated recurrent units (GRU), arranged in a DenseNet-like [10] topology. This encoder was derived from RDO-VAE [4], with the quantisation steps deleted. For 40 ms time-step $n = 0, 4, 8, ...$, each stage can be represented by:

$$\mathbf{z}_{i+1} = [\mathbf{z}_i, \mathrm{conv_i}([\mathbf{z}_i,\ \mathrm{GRU_i}(\mathbf{z}_i)])], \quad i = 1..6 \tag{1}$$

where $\mathbf{z}_1 = \mathrm{dense}([\mathbf{f}_n, ..., \mathbf{f}_{n+3}])$, and the encoder output $\mathbf{z} = \mathrm{dense}(\mathbf{z}_6)$. The decoder has a similar, but symmetrical design, but includes an additional gated linear unit (GLU) in each stage.

$$\mathbf{y}_{i+1} = [\mathbf{y}_i, \mathrm{conv_i}([\mathbf{y}_i,\ \mathrm{GLU_i}(\mathrm{GRU_i}(\mathbf{y}_i))])], \quad i = 1..6 \tag{2}$$

where $\mathbf{y}_1 = \mathrm{dense}(\hat{\mathbf{z}})$ and the output feature vectors $[\hat{\mathbf{f}}_n, ..., \hat{\mathbf{f}}_{n+3}] = \mathrm{dense}(\mathbf{y_6})$.

For bandwidth efficient transmission over the channel the elements of $\mathbf{z}$ are mapped to $d/2 = 40$ complex QAM symbols $\mathbf{q}$. Compared to classical digital modulation, the elements of $\mathbf{q}$ can be viewed as continuously valued (analog) QAM symbols.

For transmission over the HF multipath channel we employ OFDM with pilots symbols. We reshape the serial stream of QAM symbols at rate R_q as N_c parallel sub-carriers, each running at a symbol rate of $R_s = R_q/N_c$ symbols/s, where R_s is chosen based on delay spread considerations. We have chosen $N_c = 30$ and $R_s = 50$ Hz.

This hybrid ML-DSP design has several benefits:

1) OFDM performs equalisation using a single complex multiply of each symbol which allows us to efficiently represent the multipath channel fading in the ML frame-rate processing.
2) The $F_s = 8000$ Hz sample rate processing is performed efficiently in classical DSP, with the ML processing at a much slower frame rate $R_f = 25$ Hz.
3) We can perform acquisition, synchronisation, and sample rate conversion using well known DSP techniques, further simplifying the ML processing.

A disadvantage of OFDM is high peak to average power ratio (PAPR), which for a given transmitter peak power reduces the available SNR at the receiver. We have largely overcome this issue via training, as explained in Section 3.

The OFDM frame is shown in Fig. 4. Pilot symbols are periodically inserted into each OFDM carrier. After the IDFT stage, a cyclic prefix is inserted to guard against inter-symbol interference. To achieve an efficient ratio of pilots to data symbols, we place the QAM symbols from three consecutive $\mathbf{z}$ vectors (120 complex QAM symbols) in each OFDM frame, leading to an OFDM frame duration (and algorithmic delay) of 120 ms.

At the receiver the pilots are used to estimate the time-varying phase of the channel (equalisation), and for initial acquisition (coarse frequency, frame sync) of the received signal. Phase equalisation also allows small frequency offsets (± 2 Hz) to be handled and frequency drift tracked. Neural networks are sensitive to magnitude scaling, so we also use the pilot symbols for coarse magnitude equalisation (gain control).

The insertion of pilot symbols and the cyclic prefix consume carrier power that would otherwise be available for payload symbols, and require the symbol rate and hence overall RF bandwidth to be increased to maintain the payload symbol rate. The overheads for RADE total 4 dB, including a 2-dB difference from ideal performance in the least squares estimation algorithm used for phase equalisation. The

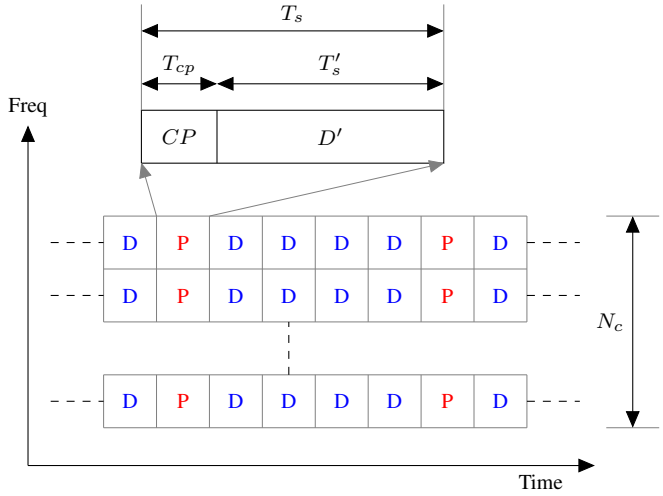

Fig. 4: OFDM modem frame, P denotes pilot symbols, D payload symbols. In each frame we have $N_s = 4$ payload symbols, and $N_c = 30$ carriers. Each symbol is comprised of a $T_{cp} = 0.004$ second Cyclic Prefix and $T'_s = 0.020$ second symbol D'.

resulting waveform has a RF bandwidth of approximately 1500 Hz, with 500 Hz due to overheads.

3. TRAINING

Fig. 5 illustrates the configuration used for training. We use a mixed sample rate design, with most of the signal processing occurring at the subcarrier rate R_s, and selected portions at the sample rate F_s. Only the RADE encoder and decoder have trainable parameters. The bottleneck is defined as:

$$\text{ctanh(x)} = \tanh(|x|)e^{j\arg[x]} \tag{3}$$

which simulates the saturation of a transmitter power amplifier by compressing the magnitude but retaining the phase. We reasoned that a bottleneck applied to the magnitude of the complex time domain signal would encourage the network to maximise the RMS power (and hence minimse the PAPR) given the channel noise and peak power constraint of the bottleneck.

To train we need to apply the $\text{ctanh}(x)$ bottleneck in the time domain, and simultaneously apply a multipath channel model. Applying the multipath model in the time domain introduces phase rotations and inter-symbol interference (ISI), which would then require equalisation and removal inside the training loop. While possible this would require significantly slow down the training process, and require training at the higher rate F_s sample rate ($F_s/R_s = 160$ in our design).

We use the simple equalisation properties of OFDM and perform multipath and AWGN channel simulation in the frequency domain. We assume phase equalisation and ISI removal is performed by the classical DSP stages of the receiver and ignore these these steps during training. This reduces multipath channel simulation to magnitude-only fading applied to each frequency domain QAM symbol via a single real-complex multiplication.

The mixed-rate training system works as follows. The transmit QAM symbols are transformed to the time domain with an inverse DFT, the bottleneck applied, then immediately transformed back to the frequency domain. The real valued multipath model magnitude samples $\mathbf{h}$ are applied to the rate R_s (frequency domain) QAM symbols via a simple multiplication. They are derived from a two path Watterson model [11]:

$$y(t) = x(t)G_1(t) + x(t-d)G_2(t) \tag{4}$$

where $x(t)$ is the time domain signal from the transmitter, and $y(t)$ is the output of the multipath fading model. G_1 and G_2 are two band-limited complex Gaussian signals with *Doppler Spread* bandwidth B_{ds}, and d is the delay spread (path delay) in seconds. Typically, $B_{ds} \approx 1$ Hz, therefore G_1 and G_2 slowly vary in amplitude and phase, modelling reflection of the transmitted signal from separate layers of the ionosphere. The sum of the two terms of (4) causes notches separated by $1/d$ to appear in the simulated channel, with the position and depth of the notches varying as G_1 and G_2 evolve. As $B_{ds} << R_s$ the channel can be considered stationary over the period of one symbol. By taking the z-transform and evaluating at the centre of each carrier frequency ω_c, the elements h_c of $\mathbf{h}$ can be computed as:

$$h_c = |H(e^{j\omega_c})| = |G_1 + e^{-j\omega_c dF_s}G_2| \tag{5}$$

We train with a delay spread $t = 2$ ms and Doppler spreading bandwidth of $B_{ds} = 1$ Hz which we denote a multipath poor (MPP) channel. G_1 and G_2 are sampled and $\mathbf{h}$ updated at rate R_s.

We use the same 205-hour training set as [4], which includes more than 900 speakers in 34 languages and dialects. It is reshaped into 4-second sequences, with the AWGN noise for each sequence chosen at random over a 20 dB range $-3 < E_q/N_0 < 17$ dB to encourage operation at a range of SNRs, where E_q is the energy of each QAM symbol, and N_0 is the noise power per unit bandwidth. Given a symbol magnitude A_q, the total RMS noise summed across the real an imaginary components can be computed as $\sigma = A_q/\sqrt{E_q/N_0}$. Training using this model resulted in signals with a PAPR of less that 1 dB and intelligible speech down to $E_q/N_0 = -3$ dB on AWGN channels.

The system is trained without pilot symbols or cyclic prefix insertion. After training, classical DSP techniques for equalisation and acquisition illustrated in Fig. 2 are wrapped around the core ML to develop the practical, rate F_s speech over HF system described in Section 2.

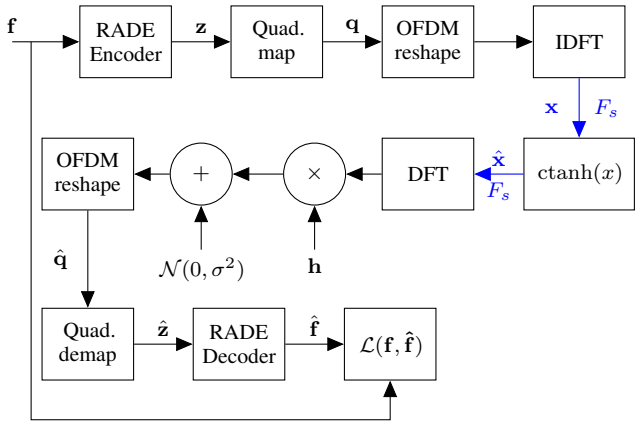

Fig. 5: Configuration used for training. A mixed sample rate model is used for joint PAPR minimisation and optimisation for multipath channels. The rate F_s signals are blue, all other signals are rate R_s. The channel model comprised of $\mathbf{h}$ and $\mathcal{N}(0,\sigma^2)$ is applied only during training.

4. EVALUATION AND RESULTS

Informal listening tests on simulated and over the air samples suggest RADE significantly outperforms SSB. Since intelligibility – more than quality – is the primary goal for HF radio, we use Automatic Speech Recognition (ASR) to evaluate the performance of the proposed system. Five hundred samples from the Librispeech dataset were passed through RADE, SSB, and FreeDV 700D simulations at a range of SNRs, then post processed by the Whisper ASR system [12], and the Word Error Rate (WER) measured (Fig. 6). SSB was simulated by band limiting the input speech to 300-2700 Hz, and applying Hilbert compression such that the mean PAPR was around 8 dB. FreeDV 700D is an open-source HF digital voice protocol using an OFDM modem, rate $1/2$ FEC, and the Codec 2 classical vocoder [13]. The MPP channel was simulated using the time domain Watterson model Eq. (4). The Librispeech speech and Watterson model G_1 and G_2 datasets used for the evaluation were not part of the training dataset.

We consider two thresholds (a) Link closure - the point where barely intelligible speech can be sent over the system and (b) Good quality, effortless communication. At the 30% WER level (link closure), the ASR results indicate a 4 dB improvement for RADE over SSB for both AWGN and MPP channels. At the 5% WER level (good quality), the improvement is 13 dB for both channels. FreeDV 700D exhibits low speech quality with the Librispeech dataset, although we note it is in common use on HF by trained operators. Similarly, skilled operators can use SSB down to 0 dB SNR. Note the sharp knee in the FreeDV 700D AWGN curve at -2 dB SNR, common in digital speech system due to the abrupt breakdown of the FEC. RADE and SSB have a more desirable gradual trade off between SNR and speech quality.

These results are based on the SNR at the receiver, and exclude the potential PAPR improvement of 7 dB. For the same peak power transmitter, the RADE waveform would have a mean signal power up to 7 dB higher than SSB at the receiver, leading to an additional 7 dB improvement. This is however dependant on the SSB compressor employed (there is no standard PAPR for SSB), and for the RADE case assumes a transmitter that doesn't degrade the PAPR of the RADE signal.

Results using FARGAN on clean features show that the resulting WER is very close to that of the clean speech. The use of our classical features and the choice of vocoder are thus not a limiting factor in the performance of the proposed system.

4.1. Complexity

The encoder and decoder each require 1 MB of read-only storage (1M weights with 8-bit quantization) and require 32 MMACs for real-time operation. Both easily operate in real time on a laptop using an unoptimised PyTorch implementation. The overall complexity of the complete system is dominated by the FARGAN vocoder's 300 MMACs [7].

4.2. Over the Air Demonstration

To test RADE over real world HF channels licensed Amateur Radio operators from around the world were invited to record a 10 second input sample of their own voice. This was converted to a RADE waveform sample, and concatenated with a Hilbert-compressed version of the same input sample to emulate an analog SSB signal with a 6-8 dB PAPR. Participants then played the concatenated sample through their SSB transmitters over real world HF Radio channels to remote KiwiSDR receivers of their choice (KiwiSDRs are SSB radio receivers that are connected to the public Internet). The received off air signal was processed to obtain a file of SSB and RADE output speech samples, and an estimate of channel SNR. The use of stored files enabled the same Tx signal to be transmitted to different KiwiSDR receivers under different channel conditions, and different transmit power levels. Tests were performed in English, Japanese, Cantonese, and German, over a variety of HF channels of up to 14,000 km (e.g. direct transmission from North America to Australia). Speech samples are available at [14].

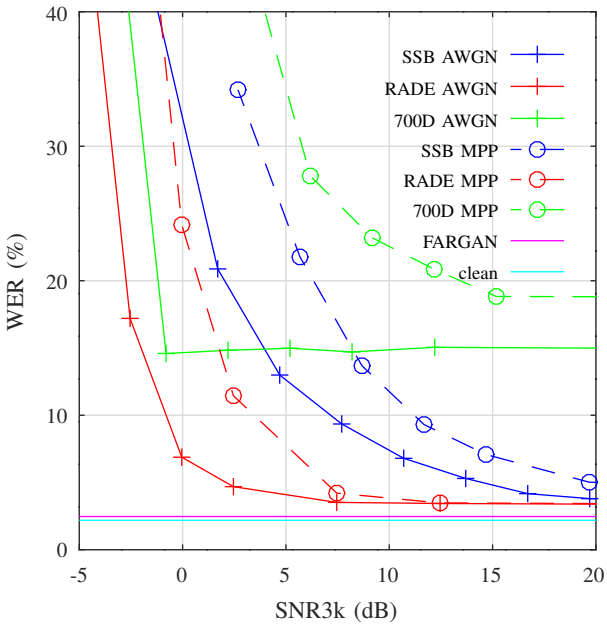

Fig. 6: Word error rate % versus SNR for simulated AWGN and MPP channels, with clean and uncoded FARGAN as controls. SNR for all signals is normalised to a 3000 Hz noise bandwidth.

5. CONCLUSION

We have combined a ML vocoder, ML autoencoder and classical DSP OFDM to build a system capable of sending speech over HF radio channels. It is robust to AWGN and multipath channel impairments, and the transmit signal has a PAPR of less than 1 dB. Unusually for HF speech systems, the audio bandwidth is 8000 Hz, despite requiring just 1500 Hz of RF bandwidth. Our ASR simulation and real world demonstration show performance significantly exceeding that of the analog SSB at the same SNR. Unlike classical DSP HF systems, speech quality improves gradually with channel SNR without any mode switching. Interestingly, our system also shows robustness to channel impairments we did not train for, e.g. impulse noise.

The experimental HF OTA results demonstrate surprisingly good performance on multipath channels where the period of the fading (100s of ms) is large compared to the 40 ms analysis window of the autoencoder. To overcome fading with classical DSP requires an interleaver of several times the fading period (e.g. 1000-2000 ms) which introduces significantly more algorithmic delay than our system for a similar level of robustness.

The OFDM frame design contains three latent vectors $\mathbf{z}$, so introduces an algorithmic delay of 120 ms. This is comparable to other digital PTT radio systems, e.g. P.25 [15].

6. REFERENCES

[1] A. A. Oswald, "Early history of single-sideband transmission," *Proceedings of the IRE*, vol. 44, no. 12, pp. 1676–1679, 1956.

[2] J. F. Honey and D. K. Weaver, "An introduction to single-sideband communications," *Proceedings of the IRE*, vol. 44, no. 12, pp. 1667–1675, 1956.

[3] T. J. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," 2017. [Online]. Available: https://arxiv.org/abs/1702.00832

[4] J.-M. Valin, J. Büthe, A. Mustafa, and M. Klingbeil, "DRED: Deep REDundancy coding of speech using a rate-distortion-optimized variational autoencoder," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[5] Y. Chen, B. Dong, X. Zhang, P. Gao, and S. Li, "A hybrid deep-learning approach for single channel HF-SSB speech enhancement," *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2165–2169, 2021.

[6] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Low-latency deep analog speech transmission using joint source channel coding," *IEEE Journal of Selected Topics in Signal Processing*, 2025.

[7] J.-M. Valin, A. Mustafa, and J. Büthe, "Very low complexity speech synthesis using framewise autoregressive GAN (FARGAN) with pitch prediction," 2024.

[8] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[9] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5891–5895.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[11] "ITU-R F.1487: Testing of HF modems with bandwidths of up to about 12 kHz using ionospheric channel simulators," 2000.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[13] D. Rowe, "Codec 2 Algorithm Description," https://rowetel.com/downloads/codec2_doc.html.

[14] ——, "September 2024 RADE Demonstration," https://freedv.org/davids-freedv-update-september-2024.

[15] T. I. Association *et al.*, "Project 25-DataOverview-NewTechStandards," *ANSI/TIA-102.BAEA-A*.