ARL-TR-9865 ● JAN 2024

Low-Bitrate Speech Compression with a Glottal Pulse Autoencoder

by Michael S Lee

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9865 ● JAN 2024

Low-Bitrate Speech Compression with a Glottal Pulse Autoencoder

Michael S Lee
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
January 2024	Technical Report	10/01/2022	8/01/2023

4. TITLE AND SUBTITLE
Low-Bitrate Speech Compression with a Glottal Pulse Autoencoder

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Michael S Lee

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-NA
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9865

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID ID: Michael S Lee, 0000-0002-0419-6069

14. ABSTRACT
In this report, a convolutional neural network with a custom voicing layer is used to compress English speech to a rate of 1500 bits per second. The model consists of an autoencoder that converts speech input into a quantized latent space and then decodes with voice-like glottal sound and noise layers multiplied by a formant mask. The input and output of the model are magnitude short-time Fourier transform spectrograms. Each 32-ms frame of the input is approximately mapped to an element of the bottleneck layer and quantized to 48 bits via 4 additive layers of 12-bit vector codebooks.

15. SUBJECT TERMS
deep learning; autoencoder; speech compression; formant; glottal pulse; short-time Fourier transform; residual vector quantization; Network, Cyber, and Computational Sciences; Military Information Sciences

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	25

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Michael S Lee	(410) 278-5888

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Acknowledgments

The author would like to thank Dr John Hyatt for helpful discussions. Supercomputing time was provided by the US Army Combat Capabilities Development Command Army Research Laboratory Supercomputing Resource Center.

1. Introduction

Speech compression algorithms enable low bitrate digital voice communication systems, which are useful for applications such as amateur radio and mobile phones. While fully uncompressed 48-kHz, 24-bit audio would require a transmission rate of nearly 1 million bits per second (bps), modern high-quality speech encoders range between 13 to 128 kilobits per second (kbps). Speech compression methods have achieved intelligibility at rates as low as 450 bps (Rowe 2011; Siahkoohi et al. 2022; Jenrungrot et al. 2023). In addition, narrowband codecs have been enhanced to wideband quality via deep learning models (Klein et al. 2018).

1.1 Prior Work

Among the first breakthroughs, popular open-source speech compression protocol CODEC2 uses a sum of sinusoidal harmonics model (Rowe 1997) to achieve bit rates between 450 and 3200 bps. The open-source nature of CODEC2 and the marked speech intelligibility of its lower bitrate models has made this construct a favorite of digital amateur radio enthusiasts who often work with very low bandwidth transmissions (Rowe 2011).

In the last few years, several machine learning (ML)-based speech and audio compression models have been developed furthering the state of the art. For example, the linear predictive codec network (LPCNet) model yields a bit rate of 1.6 kbps using a combination of linear prediction techniques and recurrent neural networks (NNs) (Valin and Skoglund 2019). A notable breakthrough is Soundstream (Zeghidour et al. 2021), which transforms a 1-s time-series (TS) audio input into 75 frames that are encoded via residual convolutional neural networks (CNNs) with dilation layers. Each frame in the bottleneck is quantized to 48 bits, yielding a 3-kbps compression rate. Building on Soundstream, Encodec (Défossez et al. 2022) has achieved a variable bit-rate audio encoder, which roughly matches the quality of our presented model at 1.5 kbps. Even further, Siahkoohi et al. (2022) and Jenrungrot et al. (2023) demonstrate intelligible speech at 600 and 900 bps using a Transformer plus a Soundstream-like CNN architecture. Meanwhile, Lyra (Google Open Source Blog 2022) achieves 3 kbps compression with a different strategy using a wave recurrent neural network (WaveRNN) (Kalchbrenner et al. 2018; Jayashankar et al. 2022) trained from a Karhunen–Loève transform (KLT) compressed spectrogram.

1.2 This Work

In this work, we present an NN model that encodes English speech at 1.5 kbps. The method uses a voice model analogous to linear predictive coding whereby a buzzer-type sound forms the carrier for voiced (mostly vowels) speech and white noise is used for the unvoiced phonemes (i.e., most consonant sounds). A formant "mask" is then multiplied in to modulate both components. In addition, the fundamental frequency and voicing amount are deduced from the latent space generated by the encoder. Like Soundstream, we quantize the bottleneck layer with residual vector quantization. Traditional vector quantization (VQ) assigns continuous-valued vectors to a finite set of codebook vectors that can be discretely indexed by integers. As an extension, residual vector quantization (RVQ) performs a finer granularization of the continuous space via a hierarchy of quantized vectors. RVQ avoids generating a massive amount of codebook vectors by instead building additive combinations of codebooks at different magnitudes. The original work introduced a VQ-based variational autoencoder, which demonstrates how to train a VQ layer using pass-through backpropagation (Van Den Oord and Vinyals 2017). Another work developing the method audio language model (AudioLM) demonstrates that RVQ can be trained just as easily (Borsos et al. 2023).

The quality of our English speech model appears to meet or exceed some of the currently available speech encoders at 1.5 kbps. However, we were not able to test against two new ultra-low bitrate encoders (Siahkoohi et al. 2022; Jenrungrot et al. 2023), which we expect may have better quality-to-bitrate performances.

Furthermore, many of the new breed speech encoders have been developed with computational efficiency in mind. With a similar perceptual quality as our model at 1.5 kbps, Encodec, by Meta research (Défossez et al. 2022) also provides exceptional computational efficiency at inference time—where the model can be run in real-time on a CPU. Some of the salient features of Encodec include training with time- and time-frequency based loss functions and encoding/decoding purely in the time domain. In addition, Encodec training incorporates an adversarial model to improve the realism of the generated output. Nonetheless, our model is a good example of what can be achieved with relatively simple and low-parameter NNs.

2. Theory and Methods

2.1 Human Speech Production

Speech is the acoustic result of a biophysical system with multiple components, such as the mouth, vocal cords, and throat. As such, the complexity of spoken audio

is physically bounded by frequency range, volume, and timbre variation. This implies that the subspace dimension of all audio signals corresponding to speech should be relatively small compared to all possible audio waveforms. Therefore, a significant amount of data compression should be possible, in theory.

For example, voiced speech (i.e., all vowel sounds and some consonants like "r"), can be modeled accurately as a periodic glottal pulse shaped by a multi-peaked bandpass filter to simulate formants. The fundamental frequency of the periodic pulse in different peoples' voices ranges from 75 (deep male voice) to 500 Hz (child). Unvoiced speech, such as most consonants, can be modeled as transient white noise also conditioned with various frequency filters. White noise contains a large amount of entropy. However, from a receiver/intelligibility point of view, various clips of white noise should be indistinguishable from each other and thus offer no additional semantic meaning.

2.2 Estimated Lower-Bound Bitrate for Encoding English Speech

One can envision a speech compression algorithm where the input audio is converted directly to text via automated speech recognition (ASR), the text is transmitted, and then the text is converted back to speech at the receiver end with text-to-speech. In this limit, only about 28 bps of information would need to be transmitted (for English, at least). We derived this estimate from the fact that people can speak up to about 170 English words per minute or 2.83 words per second, and without any context to predict the next word, English words have an entropy of around 9.8 bits (Grignetti 1964). However, to yield faithful spoken audio, we need to add non-semantic features such as speaker identity, inflections, mood, and rate, leading to another approximately 50 bps (33 bits, e.g., to identify a unique human speaker, and ~17 more bits to describe every possible way they could say the same text). Thus, accurate speech encoding should theoretically require less than 100 bps. Currently, most speech compression algorithms lose intelligibility below 3000 bps, with the best algorithms reaching approximately 500 bps. This discrepancy suggests there is still considerable research in speech compression that could be done to approach theoretical bit rates.

2.3 Spectrogram vs. TS Processing

Speech data can be input into NNs as a 1-D vector/TS or a 2-D matrix/short-time Fourier transform spectrogram (STFT). The TS format requires presumably lower overall computational cost as a final phase reconstruction step (i.e., via the iterative Griffin–Lim algorithm [Griffin and Lim 1984]) is not required for synthesizing the 1-D audio signal. However, for generating high-quality speech, the phase

component is plausibly a confounding variable with random-like features and, therefore, is not required. Hence, a magnitude STFT is often suitable.

2.4 Proposed Speech Compression Model

As seen in Fig. 1, our proposed speech compression model is a combination of an encoder into a quantized latent space, followed by a decoder and phase reconstruction. The encoder first converts an input audio waveform into a magnitude spectrogram before entering the NN. The decoding component uses the resultant latent space to simultaneously deduce the fundamental frequency, voiced/unvoiced balance, and formant mask of each block. In this work, we only considered a deterministic autoencoder. Certainly, a variational autoencoder could be used instead. In addition, a discriminator model could be trained to detect synthetic versus actual sound samples and challenge the autoencoder to produce more realistic output audio samples (i.e., a generative adversarial network [GAN] model).

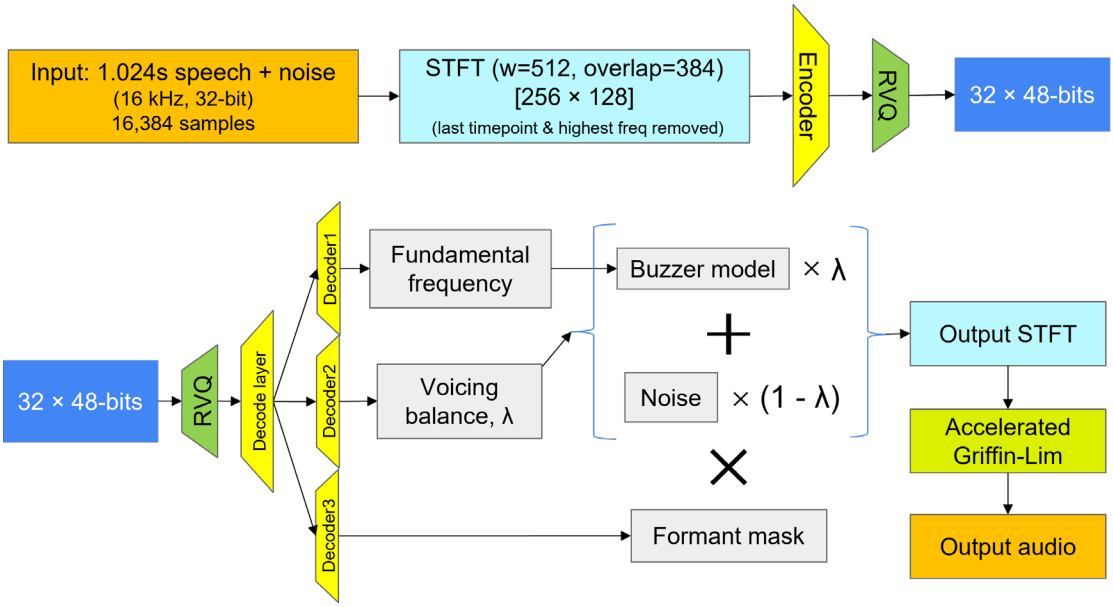

Fig. 1 Proposed speech compression model

2.4.1 Frame Size

Typical speech compression and communication protocols define frames (sometimes overlapping) of a certain number of time samples, n_f. In this work, we split 1.024-s blocks of audio into 32 discrete frames, corresponding to $n_f = 512$ samples and a duration of 32 ms. Other methods in the literature have frame durations that range from 20 to 50 ms. For comparison, at a normal speaking pace,

phonemes in the English language can range from 50 to 300 ms in duration depending on various factors (Crystal and House 1988).

2.4.2 Encoder Layers

The optimal structure of an NN model to encode speech spectrograms is an unsolved problem. Therefore, we have used intuition and trial and error to yield the following topology, as seen in Fig. 2. The number of kernels in the downsampling convolutional layers are 16, 32, 48, 64, 80, 96, 112, 128, 32; the first two downsampling layers use a 5 × 5 kernel with 2 × 2 striding, and the next six layers use 5 × 1 with 2 × 1 striding. The last layer has a kernel size of 1 × 5 with a 1 × 1 striding, which encodes the continuous latent space and permits mixing of adjacent phoneme fragments.

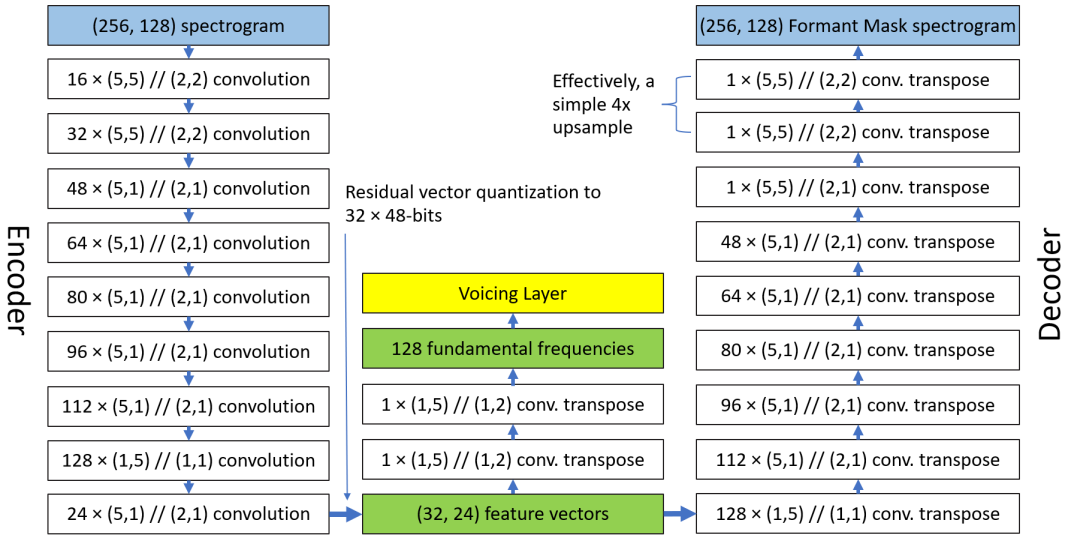

Fig. 2 Specification of most of the layers used in the proposed model

In this work, the encoder outputs a latent space matrix of 32 × 24, where 32 is the number of frames and 24 is an arbitrary number of single-precision floating point channels. The selection of 24 channels is a compromise between competing factors. As the latent space dimension, N, is increased, it is generally easier for the autoencoder to reproduce the input accurately. However, vector quantization can benefit from a smaller latent space, where various vertices of an N-dimensional polytope can be represented. The number of useful vertices associated with a codebook could easily reach 2^N, if we assumed, for example, an N-dimensional hypercube. Thus, a compromise must be made between continuous latent space expressivity and discretization via VQ.

2.4.3 Residual Vector Quantization Layers

Our speech compression model uses residual VQ to convert the 32 × 24 matrix of single-precision (i.e., "float32") numbers into 1536 bits representing 1.024 s of audio. Each frame of 24 values is converted into 48 bits. With traditional VQ, the number of codebook vectors $= 2^b$, where b is the number of desired bits. Predictably, this can lead to unwieldy memory storage for $b >> 10$. Instead, the trick is to build a hierarchy of additive codebooks, each one of size b_i, where $b = \sum_i^{Nc} b_i$, to greatly conserve codebook memory (Adiban et al. 2022). Each codebook is trained on the residual error of the sum of the previous codebooks.

In this work, 4 additive codebooks are used, where each codebook is 12-bit addressable with 4096 learnable entries of size 4096 × 24. The VQ layers are implemented as defined in Van Den Oord and Vinyals (2017). The second through fourth VQ layers are each input with the residual error of the sum of the previous VQ layers (Adiban et al. 2022).

2.4.4 Decoder Layers

The first decoding convolutional layer simultaneously feeds to the fundamental frequency estimation layer, voicing balance layer, and formant mask layers: 128 kernels of size 1 × 5 with a striding of 1 × 1. The frequency and voicing decoders use the same two convolutional transpose layers: 112 kernels of size 1 × 5 with a striding of 1 × 2, and 96 kernels of size 1 × 5 with a striding of 1 × 2. Then, individually the frequency and voicing values are defined by a final sigmoidal activated convolutional layer with a kernel size of 1 × 1. The frequency value is linear mapped from the sigmoidal output of [0, 1] to [75, 300] Hz. Note that while a child's voice can easily exceed a fundamental frequency of 300 Hz, the LibriSpeech training data (used in this work) only contains adult speakers; therefore, we ignore child-voice speech compression in this report.

2.4.5 Formant Mask Layer

Besides generating either a voiced "buzz" or unvoiced noise carrier, the human speech production system shapes the relative amplitudes of certain harmonics (a.k.a., formants) to yield an array of recognizable vowel sounds. To simulate this, we devise a lower-resolution formant "mask", *M*, that is multiplied by the voiced (*B* = buzz) and unvoiced (*N* = noise) carriers. This mask is learned from the same autoencoder used to encode the fundamental frequency and voicing switch. For the speech spectrogram of 256 frequency bands × 128 windows, the mask has a native resolution of 32 × 32 and is up-sampled via three consecutive single kernel layers to 256 × 128 prior to multiplication with the carriers. We also apply the function, $f(x) = 0.1x^4$, to emulate the frequency sharpness of vowel formants.

The formant mask is decoded in reverse order of the encoder: transpose convolutions of $N_{kernels} = (112, 96, 80, 64, 48, 1, 1, 1)$ with the first six having sizes of 5 × 1 and strides of 2 × 1, and the last two having a size of 5 × 5 and strides of 2 × 2. The last three convolution layers act like a smooth up-sampler that is faster than the bicubic "Upsampling2D" operation in TensorFlow. The mask is activated with the non-negative Softplus function, followed by the $f(x) = 0.1x^4$ to generate the sharper formant bumps.

The final output spectrogram, *D*, is defined as

$$D = aM[vB + (1 - v)N], \tag{1}$$

where $v \in [0,1]$, is the softmax output of the voicing layer governing the balance of voiced and unvoiced speech, *M* is the formant mask, *B* is the custom voicing layer, *N* is the random white noise layer, and arbitrary constant, *a*, is set to 0.1.

2.4.6 Custom Glottal/Buzzer Layer for a Voiced Carrier

The voiced aspect of speech is often modeled as a buzzer sound (i.e., a set of equally spaced harmonics starting from the deduced fundamental frequency of the speaker's voice). The exact progression of harmonic amplitudes can vary from speaker to speaker. A simple summation of equal amplitude harmonics can lead to a "glottal pulse" structure in TS audio space. The frequency of the buzzing sound can be detected from the input speech through a series of encoding convolutional layers.

We define our buzzer sound within a frame (column) of the spectrogram, $B(b)$, as

$$B(\mathrm{b}) = exp\left\{\cos\left[\max\left(\pi, \frac{2\pi f_s b}{f_0 n_{seg}}\right)\right] - 1\right\}, \tag{2}$$

where $b \in [0, 255]$, is the frequency band index, f_0 is the fundamental frequency output from the frequency prediction layer, f_s is the audio sampling rate (= 16 kHz throughout this report), and $n_{seg}$is the number of samples per STFT segment (= 512 for this work). Figure 3 compares an STFT segment of the author vocalizing the "ah" sound with the buzzer model set to a frequency of 112.5 Hz.

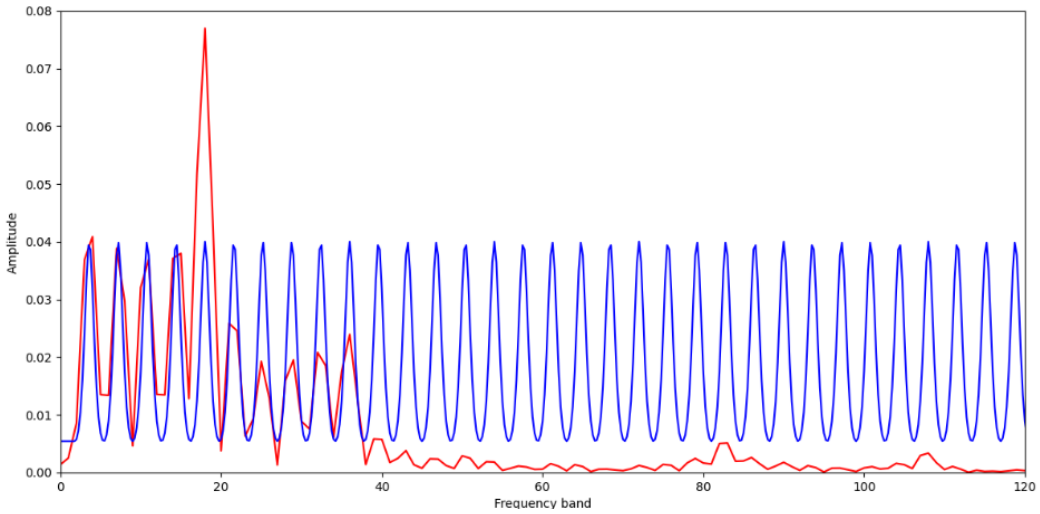

Fig. 3 Spectral comparison of a real "ah" recording (red) and buzzer model set to 112.5 Hz (blue)

Notably, glottal pulses have very sharp harmonics when analyzed over a larger number of time samples/frequency bands. Limiting the Fourier analysis to 256 frequency bands causes significant blurring of the spectrum leading to a rather wide curve for each harmonic.

2.4.7 Custom Noise Layer for an Unvoiced Carrier

The base sound of unvoiced speech can be modeled as white noise, where all frequencies are roughly equal. Randomly generated white noise is transformed into STFT space for use within the decoder. Random white noise can also be used at inference time. Randomness should not affect real intelligibility; however, it could affect the reconstruction loss and objective metrics of speech similarity and intelligibility.

2.4.8 Voicing Balance Parameter

The same first decoding convolutional layer output can also be used as an input branch to classify the speech frames as voiced, unvoiced, or a linear combination of the two via the voicing layer output as v. A sigmoid activation is applied at the output layer to ensure this switch is constrained between 0 (fully unvoiced) and 1 (fully voiced).

2.4.9 Training Data

The open source LibriSpeech data set of LibriVox English-spoken audiobooks (Panayotov et al. 2015) were used to train our speech compression model. To ensure maximum input variability, random 1.024-s clips are extracted from 20-min segments of 747 different speakers for training, and another 83 speakers for validation.

2.4.10 Phase Reconstruction with Griffin–Lim Algorithm

From a computational efficiency standpoint, spectrogram output from the model is not ideal because it requires at least one additional stage of postprocessing to convert the 2-D spectrogram into TS audio. One common solution is a neural vocoder, which is a distinct ML model that converts a full or compressed spectrogram into a viable TS audio clip. Another method, used in this work, is the fast Griffin–Lim iterative phase reconstruction algorithm (Griffin and Lim 1984; Perraudin et al. 2013).

2.4.11 Neural Network Hyperparameters

We experimented with the number of kernels per convolutional layer but found the final set to be adequate. Future work could explore other architectures such as residual networks (ResNets) (He et al. 2015).

2.5 Objective Metrics of Speech Intelligibility

Several objective metrics have been developed to evaluate the quality of reconstructed speech. Most of these metrics were calibrated against subjective human listening experiments, such as MUSHRA (International Telecommunications Union 2014). In this work, we use three objective metrics:

1) Extended short-time objective intelligibility (E-STOI) (Jensen and Taal 2016),
2) Perceptual Evaluation of Speech Quality (PESQ) (Rix et al. 2001), and
3) WARP-Q, a metric that performs dynamic time warping against mel-frequency cepstral coefficients (MFCCs) (Jassim et al. 2021).

E-STOI scores samples between 0 and 1 (perfect reproduction). PESQ scores clips between –0.5 and 4.5 (perfect); WARP-Q scores were mapped to a mean opinion score (MOS) score between 1 and 5 (perfect).

3. Results

Table 1 shows a comparison of four 1.5-kbps speech codecs on 30-s speech clips from 50 speakers in the LibriSpeech data set (Panayotov et al. 2015) using the E-STOI, PESQ, and WARP-Q speech metrics. Our model slightly outperforms the universal audio encoder (Encodec) with the PESQ metric but underperforms according to the E-STOI metric. With the WARP-Q metric, Encodec seems to struggle, while LPCNet has a slight lead over our model.

Table 1 Speech quality metrics for various 1500-bps compression techniques

Compression method	E-STOI	PESQ	WARP-Q MOS
CODEC2	0.37 (0.05)	1.28 (0.12)	1.60 (0.10)
Encodec	**0.71** (0.07)	1.46 (0.11)	2.36 (0.38)
LPCNet[a]	0.60 (0.04)	1.43 (0.15)	**3.10** (0.42)
This work	0.61 (0.04)	**1.53** (0.14)	3.04 (0.42)

[a] LPCNet has a bit rate of 1600 bps.

Figure 4 illustrates visual differences in the spectrograms between the four codecs on a snippet of speech uttered by a male speaker. CODEC2 at 1.5 kbps clearly lacks some of the finer details in the formant regions. Of the remaining three, our model seems to produce the blurriest results. This is probably because we use the lower-resolution formant mask that is up-sampled by a factor of 4. In addition, our model does not use a discriminator model that could potentially sharpen up the spectrogram. Visually, LPCnet seems to provide the best results for this speech sample. This model is continually updated on their website; therefore, it is highly optimized for speech compared to Encodec. Encodec provides the "grittiest" spectrogram of the top three. This is likely because this codec is optimized for all types of audio signals—including music.

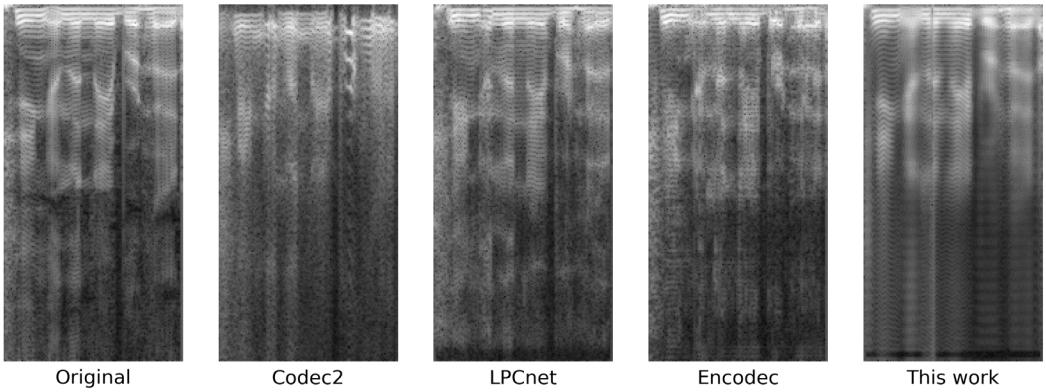

Fig. 4 Comparison of the magnitude spectrograms of a 1.024-s clip with a male speaker using the four 1.5-kbps codecs

In Fig. 5, the spectrograms of a clip of speech from a female speaker are shown. CODEC2 at 1.5 kbps does well estimating fundamental frequencies and producing the glottal banding; however, it struggles with some of the frame boundaries. Once again, our model produces the blurriest result with mostly overall correct formant features. Encodec produces some areas where the fundamental frequency prediction is a bit flat (i.e., lacks natural intonation) especially at the higher frequencies. LPCnet, once again, has arguably the best visual similarity to the reference.

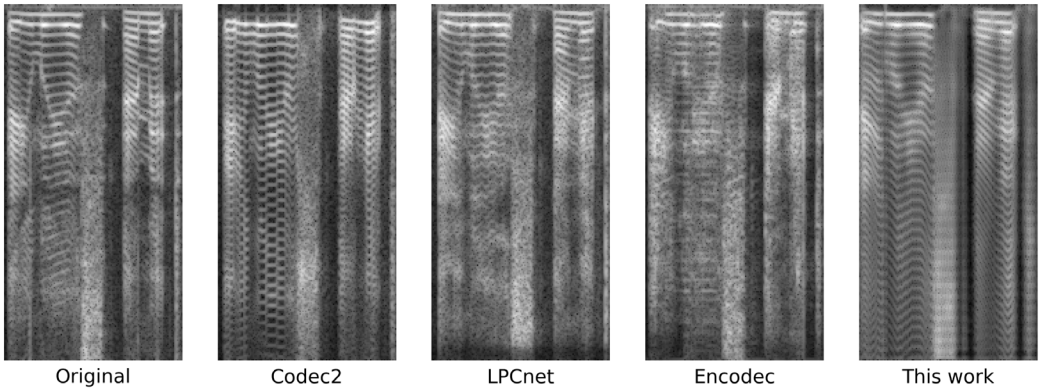

Fig. 5 Comparison of the magnitude spectrograms of a 1.024-s clip with a female speaker using the four 1.5-kbps codecs

4. Discussion and Conclusion

Speech compression has an intriguingly favorable theoretical compression limit of around 100 bps that deep learning is beginning to approach (Siahkoohi et al. 2022; Jenrungrot et al. 2023). The qualities of speech that make it eminently compressible include the fact that speech is physically generated by a common apparatus in humans. It also encodes language, which has been shown to have small entropic limits at normal speaking speeds. Therefore, it is highly likely that future speech compression systems will involve encoding and decoding of the transferred audio into language and speaker-specific parameters.

The model demonstrated in this report outlines how the sound produced by vocal cords (i.e., the glottal pulse can be represented mathematically as a layer within a NN that can be then modulated with a lower-resolution formant mask for optimal feature synthesis). Based on our results and analysis, we expect that language-assisted models such as LMcodec (Jenrungrot et al. 2023) will further drive bit rates down to near theoretical limits while still preserving speech intelligibility.

Future work should consider spoken languages besides English and other patterns present in speech that can be crafted into custom layers. In addition, work should

be done to correctly account for the high entropy but low semantic value of noise in speech at both the loss function and metric levels. In addition, given that speech derives from language, longer-term correlations between phonemes, words, and even phrases should be leveraged to improve compressibility. These strategies and others should lead to compression rates approaching the theoretical bounds of speech entropy.

5. References

Adiban M, Siniscalchi M, Stefanov K, Salvi G. Hierarchical residual learning based vector quantized variational autoencorder for image reconstruction and generation. In 33rd British Machine Vision Conference; 2022.

Borsos Z, Marinier R, Vincent D, Kharitonov E, Pietquin O, Sharifi M, Roblek D, Teboul O, Grangier D, Tagliasacchi M, et al. Audiolm: a language modeling approach to audio generation. IEEE/ACM Trans Audio Speech Language Process. 2023 June 21.

Crystal TH, House AS. The duration of American-English vowels: an overview. Journal of Phonetics. 1988 July 1;16(3):263–284.

Défossez A, Copet J, Synnaeve G, Adi Y. High fidelity neural audio compression. 2022 Oct 24. arXiv:2210.13438.

Google Open Source Blog. Lyra V2 - a better, faster, and more versatile speech codec. Google Open Source Blog; 2022 Sep 30. https://opensource.googleblog.com/2022/09/lyra-v2-a-better-faster-and-more-versatile-speech-codec.html

Griffin D, Lim J. Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing. 1984 Apr;32(2):236–243.

Grignetti MC. A note on the entropy of words in printed English. Information and Control. 1964 Sep 1;7(3):304–306.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 770–778.

International Telecommunications Union. Method for the subjective assessment of intermediate quality levels of coding systems. International Telecommunications Union (US); 2014 July. Report No.: ITU-R BS.1534.

Jassim WA, Skoglund J, Chinen M, Hines A. WARP-Q: quality prediction for generative neural speech codecs. ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021 June 6. p. 401–405.

Jayashankar T, Koehler T, Kalgaonkar K, Xiu Z, Wu J, Lin J, Agrawal P, He Q. Architecture for variable bitrate neural speech codec with configurable computation complexity. ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23. p. 861–865.

Jenrungrot T, Chinen M, Kleijn WB, Skoglund J, Borsos Z, Zeghidour N, Tagliasacchi M. LMCodec: a low bitrate speech codec with causal transformer models. ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 June 4, p. 1–5.

Jensen J, Taal CH. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans Audio Speech Language Process. 2016 Aug 10;24(11):2009–2022.

Kalchbrenner N, Elsen E, Simonyan K, Noury S, Casagrande N, Lockhart E, Stimberg F, Oord A, Dieleman S, Kavukcuoglu K. Efficient neural audio synthesis. International Conference on Machine Learning; 2018 July 3. p. 2410–2419.

Klein WB, Lim FS, Luebs A, Skoglund J, Stimberg F, Wang Q, Walters TC. Wavenet based low rate speech coding. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018 Apr 15. p. 676–680.

Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2015 Apr 19. p. 5206–5210.

Perraudin N, Balazs P, Søndergaard PL. A fast Griffin–Lim algorithm. 2013 IEEE workshop on applications of signal processing to audio and acoustics; 2013 Oct 20. p. 1–4.

Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing; 2001 May 7. Vol. 2, p. 749–752, Proceedings (Cat. No. 01CH37221).

Rowe DG. Techniques for harmonic sinusoidal coding. University of South Australia; 1997 July.

Rowe D. Codec 2-open source speech coding at 2400 bits/s and below. TAPR and ARRL 30th Digital Communications Conference; 2011 Sep. p. 80–84.

Siahkoohi A, Chinen M, Denton T, Kleijn WB, Skoglund J. Ultra-low-bitrate speech coding with pretrained transformers. 2022. arXiv:2207.02262.

Valin JM, Skoglund J. LPCNet: improving neural speech synthesis through linear prediction. ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019 May 12. p. 5891–5895.

Van Den Oord A, Vinyals O. Neural discrete representation learning. Adv Neural Inf Process Syst. 2017;30.

Zeghidour N, Luebs A, Omran A, Skoglund J, Tagliasacchi M. Soundstream: an end-to-end neural audio codec. IEEE/ACM Trans Audio Speech Language Process. 2021 Nov 23;30:495–507.

List of Symbols, Abbreviations, and Acronyms

1-/2-D	one-/two-dimensional
bps	bits per second
ASR	automated speech recognition
AudioLM	audio language model
CNN	convolutional neural network
CPU	central processing unit
E-STOI	extended short-time objective intelligibility
GAN	generative adversarial network
kbps	kilobits per second
KLT	Karhunen–Loève transform
LPCNet	linear predictive codec network
MFCC	mel-frequency cepstral coefficient
ML	machine learning
MOS	mean opinion score
NN	neural network
PESQ	perceptual evaluation of speech quality
ResNet	residual network
RVQ	residual vector quantization
STFT	short-time Fourier transform
TS	time series
VQ	vector quantization
WaveRNN	wave recurrent neural network

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB
10 (PDF)	DEVCOM ARL FCDD RLA A A SWAMI FCDD RLA N M FRAME B RIVERA FCDD RLA NA M LEE E MARK M ZIEMANN B KRACZEK FCDD RLA NB P WANG FCDD RLA NC M VINDIOLA FCDD RLR EI J HYATT