Vector-Quantized Zero-Delay Deep Autoencoders for the Compression of Electrical Stimulation Patterns of Cochlear Implants using STOI

1st Reemt Hinrichs
Institut für Informationsverarbeitung
L3S Research Center
Leibniz University Hannover
Hannover, Germany
hinrichs@tnt.uni-hannover.de

2nd Felix Ortmann
Institut für Informationsverarbeitung
Leibniz University Hannover
Hannover, Germany

3rd Jörn Ostermann
Institut für Informationsverarbeitung
Leibniz University Hannover
Hannover, Germany

***Abstract*—Cochlear implants (CIs) are battery-powered, surgically implanted hearing-aids capable of restoring a sense of hearing in people suffering from moderate to profound hearing loss. Wireless transmission of audio from or to signal processors of cochlear implants can be used to improve speech understanding and localization of CI users. Data compression algorithms can be used to conserve battery power in this wireless transmission. However, very low latency is a strict requirement, limiting severly the available source coding algorithms. Previously, instead of coding the audio, coding of the electrical stimulation patterns of CIs was proposed to optimize the trade-off between bit-rate, latency and quality. In this work, a zero-delay deep autoencoder (DAE) for the coding of the electrical stimulation patters of CIs is proposed. Combining for the first time bayesian optimization with numerical approximated gradients of a nondifferential speech intelligibility measure for CIs, the short-time intelligibility measure (STOI), an optimized DAE architecture was found and trained that achieved equal or superior speech understanding at zero delay, outperforming well-known audio codecs. The DAE achieved reference vocoder STOI scores at 13.5 kbit/s compared to 33.6 kbit/s for Opus and 24.5 kbit/s for AMR-WB.**

***Index Terms*—cochlear implants, autoencoder, hyperparameter optimization, Kiefer-Wolfowitz**

I. Introduction

Cochlear implants (CIs) are surgically implanted hearing-aids capable of restoring a sense of hearing in people suffering from moderate to profound hearing loss. While good speech understanding is achieved in high speech-to-background noise environments, more challenging environments as encountered in common social situations like a restaurant setting still pose a problem [1]. Research focuses mostly on these difficult environments and attempts to improve speech understanding. A good review of techniques and algorithms applied can be found in [2]. Beamformers, remote microphones and binaural sound coding strategies [2], [3] are among the techniques used to improve speech understanding and/or localization of CI users. All of these techniques require a wireless transmission of audio to the signal processor of a CI, or between two signal processors. To save power or bandwidth in this wireless transmission, signal compression or coding is commonly applied to reduce the bitrate of the audio signal before transmission. This coding usually introduces an additional delay in the processing chain and thus has to be kept as small as possible, as hearing aid users cannot tolerate delays above the range of $5 - 10$ ms without affecting their speech perception [4]. Due to this delay constraint, the selection of source coding algorithms is severly limited.

For this purpose we proposed [5]–[7] to code and transmit the electrical stimulation pattern or excitation patterns generated by the sound coding strategy of the CI. We proposed [5] a combination of differential pulse-code modulation (DPCM) and arithmetic coding to code the current magnitudes and the band selection of the electrical stimulation patterns generated by the advanced combinational encoder (ACE) sound coding strategy. Using this approach we achieved [6] lower bitrates and zero latency at equal or better speech understanding than state-of-the-art audio codecs. Autoencoders are artificial neural networks specifically designed to learn compressed representations of given input data or signals and have found application in a wide range of tasks [8]–[10]. Habibian et al. [8] proposed rate-distortion autoencoders for video compression achieving

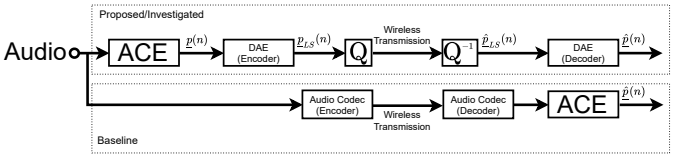

Fig. 1. Two methods to wireless transmission of audio for cochlear implants (CIs). Conventionally, the audio signal would be encoded by an audio codec, transmitted to the signal processor of the CIs where the audio is decoded by the same audio codec. In the investigated approach, the audio signal is first processed by the sound coding strategy of the CI, in our case the advanced combinational encoder (ACE), and then compressed and decompressed before and after transmission by a deep autoencoder (DAE).

This work was supported by the DFG Cluster of Excellence EXC 1077/1 Hearing4all and funded by the German Research Foundation (DFG) - Project number: 381895691.

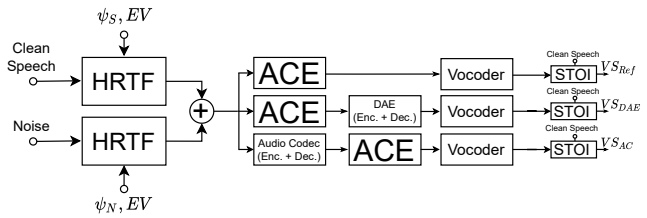

Fig. 2. Block diagramm showing the computation of the VSTOI scores. The clean speech files and the noise are first processed by head-related transfer functions (HRTF) using individual azimuths ψ_S, ψ_N and an identical environment EV. The processed speech and noise are subsequently combined, yielding the signal $x_{N+S}(n)$, and processed further. The reference VSTOI score VS_{Ref} is obtained by first applying the advanced combinational encoder (ACE) and subsequent comparison of the the vocoded signal and the clean speech signal using STOI. Similar for the VSTOI scores of the DAE, VS_{DAE}, and the VSTOI scores of the audio codecs, VS_{AC}. Vocoder settings were identical to [6] to allow comparison to listening test results.

close to state-of-the-art performance. Min et al. [9] used a deep autoencoder for the compression of speech outperforming the MELPe speech codec at very low bitrates of below 2.4 kbit/s. Wand and Saniee [10] used convolutional autoencoders for the compression of ultrasonic data.

In this work, a vector-quantized deep autoencoder (VQ-DAE) is presented for the compression of the electrical stimulation patterns generated by the ACE sound coding strategy. The VQ-DAE was first optimized using a weighted mean square error followed by numerically approximated gradient descent using the nondifferential short-time objective intelligibility (STOI) metric. The entire training sequence as well as the DAE structure was optimized using bayesian optimization. In Section II the CI sound coding strategy, the used dataset as well as the hyperparameter optimization and evaluation used in this work are described. Afterwards, in Section III, the proposed DAE is compared to common audio codecs with respect bitrate and the corresponding intelligibility of their coded speech. Section IV discusses the results and the manuscript is concluded in Section V with a summary of this work.

II. Methods and Materials

A. Advanced Combination Encoder

The advanced combinational encoder (ACE) sound coding strategy is a common sound coding strategy for CIs and described in detail in [3]. The audio captured by the microphone of the CI is split into M subbands using a discrete fourier transform filterbank. For each subband $i \in \{1, 2, \ldots, M\}$, the envelope $a_i(n) \geq 0$ is extracted resulting in the set ENV $:= \{a_1(n), \ldots, a_M(n)\}$, where n is discrete time or the frame number. Then the band-selection is performed: $N < M$ subbands $a_i(n)$ with the largest envelopes are selected, resulting in the set $A := \{a_{i_1}(n), \ldots, a_{i_N}(n)\} \subset$ ENV. For future reference we define the set of selected bands $Sel := \{i_1, \ldots, i_N\}$ and its complement $Sel^c = \{1, \ldots, M\} \setminus Sel$ whose dependency of n was left out for clarity. Then, the loudness growth function (LGF) is applied to each $a \in A$.

TABLE I

Speech and noise azimuths, signal-to-noise ratios (SNRs), noise types and acoustic scenarios considered in this work. A:B:C denotes the set $\{A, A+B, A+2B, \ldots, C\}$. BFR, sample 187511 from freesound.com, is restaurant noise, CCITT is speech-shaped noise.

Label	Speech Azi. [°]	Noise Azi. [°]	SNR [dB]	Noise	Scenario
Train	-90:15:90	-90:15:90	-5:5:20 30 50	BFR, Bus, CCITT, Office	Anechoic, Office
Test	-90:5:90	-90:5:90	-2.5:2.5:10 20 40	BFR, Bus, CCITT, Office	Anechoic, Office, Cafeteria

The LGF is the logarithmic mapping between the acoustic and the electric domain with $p_i := p(a_i) := \text{LGF}(a_i) \in [0, 1]$ for $a_i \geq \text{s}_{base}$ and no output is generated for $a < \text{s}_{base}$. s_{base} is the so called base level which represents the threshold of hearing. It is individually determined for each CI user by an audiologist. In this manuscript, the default settings of the research implementation of the ACE sound coding strategy were used which set $N = 8$, $M = 22$ and $\text{s}_{base} = 4/255$. The channel stimulation rate was set to 900 pulses per second (pps). The input signal of the proposed DAE, called a frame, was $(\text{p}(a_1(n)), \text{p}(a_2(n)), \ldots, \text{p}(a_M(n)))^T$, i.e. a $M \times 1$ vector. Thus the DAE compresses across frequency without lookahead and achieves zero delay. The number of bits per frame multiplied by the channel stimulation rate yields the bitrate of the DAE.

B. Datasets

To create realistic noisy speech signals, the TIMIT speech corpus [11] was processed by the behind-the-ear head related transfer functions (HRTF) from [12]. These HRTFs allow to simulate speech in noise scenarios, where the azimuth of each source can be independently varied with respect to its incident azimuth in the range of $\pm 90°$ in steps of $5°$ except for certain acoustic environments. An azimuth of $-90°$ corresponds to a source located to the left of the listener and $+90°$ corresponds to a source located to the right of the listener, $0°$ corresponding to the front of the listener.

The right ear signal of the HRTF-processed speech files was used in all cases and source distance was 80 cm. Each speech recording of the training and test data of TIMIT was processed using signal-to-noise ratios (SNRs), speech and noise azimuths, acoustic environments and noise type from a list of values given in Tab. III. For each category (SNR,

TABLE II

Bounds and default values used in the hyperparameter optimization together with the optimized values. Log denotes logarithmic sampling. Optimized were the number of neurons per layer of the DAE encoder (decoder was symmetric), α as in in Eq. 1, the learning rate lr and the parameters A and c explained in Sec. II-F.

Parameter	Default	Lower Bound	Upper Bound	Log	Optimized
#Neurons (L1)	16	16	30	X	30
#Neurons (L2)	8	6	16	X	14
α	0.5	0.1	0.9	X	0.46834
lr	0.001	0.0001	0.1	✓	0.0016
A	100000	10000	200000	✓	11337
c	0.001	0.0001	0.1	✓	0.0129

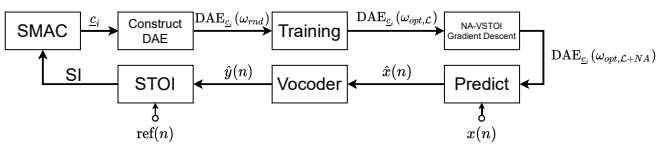

Fig. 3. Optimization loop of the deep autoencoder (DAE) structure. The hyperparameters $\underline{c}_i$ are used to construct the DAE which is subsequently trained using regular gradient descent and numerically approximated VSTOI (NA-VSTOI) gradient descent. Then, the stimluation pattern $x(n)$ of a single speech signal is compressed and decompressed, reconstructed using a vocoder yielding the waveform $\hat{y}(n)$ and compared to the reference, noise-free audio waveform $ref(n)$ by STOI. The resulting speech intelligibility score (SI) is then returned to SMAC to assess the quality of the hyperparameters $\underline{c}_i$.

noise type, ...) and each speech files, values were selected and applied randomly. While combinations of conditions like bus noise in a cafeteria environment are certainly less realistic than others, these still give important information about the robustness and generalization capabilities of the DAE. The range of values for the SNR and other categories for the test set was chosen such that it allowed to assess the impact of out-of-group values, e.g. the impact of a speech azimuth not used in training like $5°$. As noise, we used Comité Consultatif International Téléphonique et Télégraphique (CCITT) noise, bus noise, office noise and restaurant noise. CCITT noise is speech-shaped noise often used in clinical research. Random segments of the noise signals were taken as they were considerably longer than the speech signals to avoid repeating the same segment and thus biasing the training and test data.

For each noise recording, which were considerably longer than the speech recordings, random segments were used to avoid reusing the same segment for every speech file. After HRTF-processing, each audio file was peak normalized. The DAE, after hyperparameter optimization, was trained using a hand-selected subset of the HRTF-processed train data of TIMIT to reduce the training duration. This subset, refered to as the train set, consisted of 100 files covering 50 male and 50 female speakers, all SNR, azimuths, environments and noise types considered in the HRTF-processed training data.

C. Short-Time Objective Intelligibility Measure

The short-time objective intelligibility measure (STOI) [13] is a common algorithm to assess the intelligibility of speech signals which has found application in CI research [14]. There it is used to assess the intelligibility of speech in noise signals processed by the sound coding strategy of a CI [6].

For its application the electrical stimulation pattern corresponding to a given speech signal are resynthesized to waveforms using a vocoder and compared to the unprocessed, speech signal without noise using STOI yielding a value/score in the range of 0 to 1, with 1 being the best and 0 being the worst intelligibility score. In the context of CIs, STOI is also called vocoder STOI (VSTOI). While a precise mapping from VSTOI scores to word recognition scores is data dependent, speech understanding generally increases monotonically with increasing VSTOI scores allowing to compare algorithms

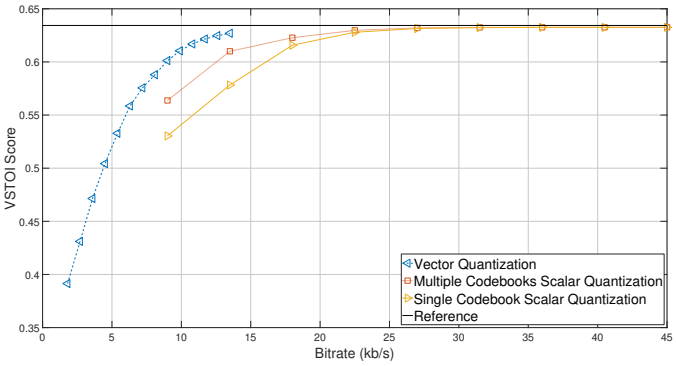

Fig. 4. VSTOI scores across bitrate for the vector quantized deep autoencoder as well as the scalar quantized deep autoencoder on the test set. Latent dimension was five. Multiple codebooks scalar quantization refers to separately optimized quantizers for each dimension in contrast to single codebook scalar quantization which used the same quantizer in each dimension. The latter was introduced as a proxy for the differences in statistics of the latent dimensions.

relatively. Computation of the reference VSTOI score as well as the VSTOI scores of the DAE and the audio codecs is depicted in Fig. 2.

D. Baseline Audio Codecs

To compare our approach to conventional audio coding, three well-known audio codecs were used as baseline. These were the Adaptive Multi-Rate Wideband (AMR-WB), the Opus and the G.722 audio codecs. AMR-WB uses algebraic code excited linear prediction to compress speech and has an algorithmic latency of 25 ms. It can code at several bitrates ranging from 6.6 kbit/s to 23.85 kbit/s. Opus can code at almost any bitrate between 6 kbit/s and 520 kbit/s and at algorithmic latencies between 5 ms and 60 ms. Opus' constraint variable bitrate flag had to be set to achieve the bitrates at the latencies investigated. G.722 is a low delay speech codec using predictive subband coding operating at an algorithmic latency of 1.3 ms. Finally, the Electrocodec [5], [6] was also used as a reference, labeled EC2 and EC3, each number indicating the number of bits per subband DPCM. Unlike the audio codecs it does not code the audio signals directly but compresses the electrical stimulation patterns generated from it by ACE. It has an algorithmic latency of 0 ms. The AMR-WB was selected to investigate the performance state-of-the-art codec without considering the latency constraint. Opus was included as a widely used state-of-the-art codec that satisfies the latency constraint. The G.722 was included as it is actually used in wireless streaming for cochlear implants and because it was expected to yield reference speech understanding allowing to validate VSTOI further. FFMPEG was used to apply the audio codecs except for Opus for which the opus-tools 1.3 were used. For a correct evaluation using STOI, the algorithmic latency of the audio codecs has to be known. While for Opus the official coding software automatically removes any delay, this is not the case for FFMPEG. Therefore, for the G.722 and AMR-WB, the lag maximizing the crosscorrelation between

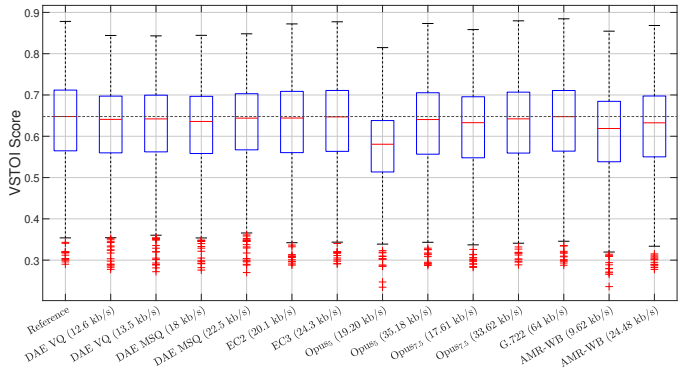

Fig. 5. VSTOI scores of all investigated codecs as well as the reference condition across the entire test set. Mean bitrates across the entire test set are given in parenthesis. Opus' latency is given in milliseconds as subscript. MSQ denotes multicodebook scalar quantization. VQ denotes vector quantization.

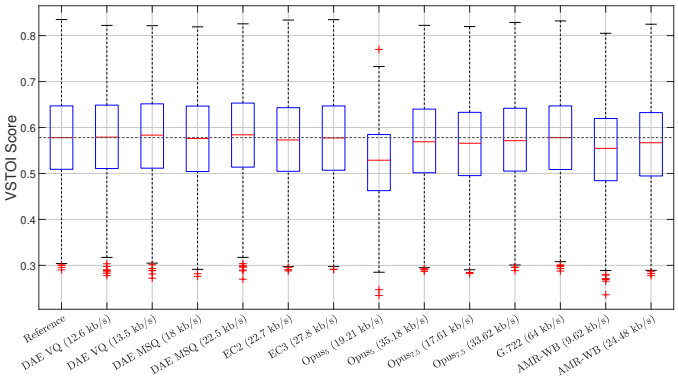

Fig. 6. VSTOI scores of all investigated codecs as well as the reference condition across the subset of the test set with a signal-to-noise ratio ≤ 5 dB. Median performance of the vector quantized deep autoencoder (VQ-DAE) was found to be slightly superior to the reference condition.

the uncoded and the coded audio was chosen as algorithmic latency.

E. Loss Function and Pre- and Postprocessing

Due to the N of M band-selection performed by ACE, the distortion of the excitation pattern is split into two parts: the distortion of the subband envelopes and the distortion of the band-selection. This was taken into account through a weighted mean-square error defined as

$$\frac{1}{M}((1-\alpha)\underbrace{\sum_{i\in Sel}(p_i-\hat{p}_i)^2}_{Envelopes}+\alpha\underbrace{\sum_{i\in Sel^c}\sigma(\hat{p}_i)}_{Band-Selection}), \qquad (1)$$

where p_i is the target value in subband i, $\hat{p}_i$ is the reconstructed value in subband i, M and Sel are as given in II-A. $\alpha \in (0,1)$ is the weighting factor. $\sigma(x)$ is the rectified linear unit (Relu). The Relu function was motivated by the pre- and postprocessing applied. In the pre-processing, any subband not selected at a time n was set to a negative value to distinguishing it from the output range of the LGF, i.e. we have $p_i(n) < 0$ if subband i is not selected. Therefore, a subband i at time n after reconstruction was considered not selected in the post-processing if $\hat{p}_i(n) < 0$. If $p_i(n) < 0$, then no error occurs and the band-selection is not distorted and as such the distortion should be zero independent of the precise value of $\hat{p}_i(n)$. However, for $\hat{p}_i(n) \geq 0$ the subband is incorrectly considered selected and a distortion value needs to be assignd.

F. Numerical Approximation of the Gradient of STOI

While the loss according to Eq. 1 allowed to improve speech intelligibility as measured by STOI, it still did not allow to achieve reference speech intelligibility. Numerical approximation techniques were employed to allow direct optimization of the DAE using STOI. For this purpose, the Kiefer-Wolfowitz algorithm with two-sided randomized differences (KW) [15] was applied to approximate the gradient of STOI. The update equation of the algorithm for the weights $\underline{\omega}$ of the DAE was

$$\underline{\omega}_{k+1} = \underline{\omega}_k + a_k \frac{(y^+_{k+1} - y^-_{k+1})}{c_k}\Delta_k, \qquad (2)$$

where $y^{\pm}_{k+1} = f(\underline{\omega}_k \pm c_k\Delta_k)$, $\Delta_k \in \{-1,1\}^N$ a vector of iid noise, $a_k, c_k > 0$ with $a_k, c_k \to 0$. N is the total number of weights. In our work we used $a_k = \frac{a}{(A+k+1)^\gamma}$ with $a = 1$ and $\gamma = 0.602$ as well as $c_k = \frac{c}{(k+1)^\beta}$ with $\beta = 0.101$. $f(\omega)$ returns the VSTOI score achieved using the DAE with the weights $\underline{\omega}$. The parameters A and c were obtained through hyperparameter optimization.

G. Hyperparameter Optimization of the DAE

Bayesian optimization, implemented through sequential model-based algorithm configuration (SMAC) [16], was used to find optimal hyperparameters for the DAE and the KW. A single noisy speech file was coded for a given DAE configuration and evaluated with respect to its VSTOI score. The steps involved in the hyperparameter optimization are depicted in Fig. 3. The hyperparameters $\underline{c}_i$ tested by SMAC were used to construct the DAE which was then trained with the adam solver on a single speech file for 500 epochs using the loss function according to Eq. 1 yielding approximately optimal weights. Then, 100 epochs of numerically approximated VSTOI (NA-VSTOI) gradient descent was performed. The VSTOI score achieved on the speech file was then returned to SMAC. The optimized hyperparameters are summarized in Table II. Total number of layers including the latent space was five, encoder and decoder using the same number of neurons per layer. Swish was always used as activation functions.

H. Training of the DAE

Using the optimized hyperparameters, the DAE was trained using the train set. Again, the DAE was first trained for 500 epochs witha batchsize of 128 to achieve a reasonable starting point for the NA-VSTOI gradient descent. Then, NA-VSTOI gradient descent was performed for 7000 iterations. After about 3600 iterations the DAE with five dimensions

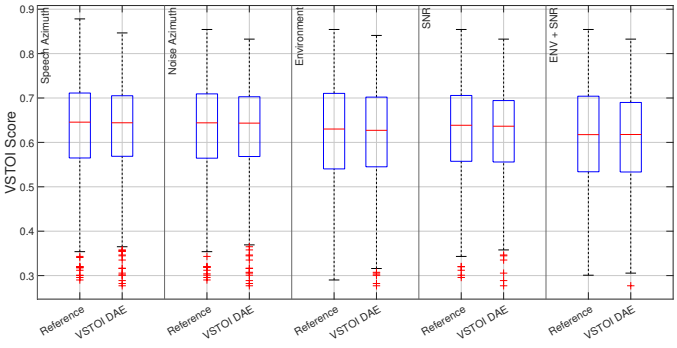

Fig. 7. Out-of-group VSTOI scores for the VQ-DAE (13.5 kbit/s). From left to right: Speech azimuth, noise azimuth, environment (ENV), signal-to-noise ratio (SNR), ENV´+ SNR.

achieved the reference mean VSTOI score on the train set, even reaching superior mean VSTOI scores in successive iterations. In contrast, the DAE with four latent dimensions required more than twice the number of iterations to approach the reference. Subsequently, only a latent space dimension of five was considered.

III. Results

VSTOI score across bitrate for the vector quantized (VQ) and scalar quantized (SQ) DAE on the test set are given in Fig. 4. Single codebook SQ used the same quantizer in each latent dimension while multi codebook SQ used individually trained quantizers in each dimension. While the VQ-DAE achieved the approximate reference mean VSTOI score at 13.5 kbit/s, both the multi codebook and the single codebook SQ-DAE required about 22.5 kbit/s. However, at lower bitrates the multi codebook SQ-DAE performed considerably better. Fig. 5 shows boxplots of the VSTOI scores across the entire test set of the reference condition and the VSTOI scores of the investigated codecs. A dashed line indicating the median of the reference condition was included as orientation. Fig. 6 presents boxplots of the VSTOI scores across the subset files with an SNR ≤ 5 dB. The top whiskers of the reference and the VQ-DAE condition at 13.5 kbit/s differed by 0.035 for the entire test set and 0.013 when considering only SNRs ≤ 5 dB. Median VSTOI scores are summarized in Table III. The performance of the VQ-DAE on the out-of-group files of the test set is shown in Fig. 7. From left to right the columns show the VSTOI scores when only the files of the test set are considered which use speech azimuths, noise azimuths, an environment, SNR and SNR + environment not used in the train set. No significant difference was observed and the VQ-DAE generalizes well to unseen conditions. The largest difference in the medians of about 0.003 was observed when considering only cafeteria samples.

IV. Discussion

The DAE was found, in either configuration, to generalize well from the train to the test set and appeared to be robust to changes of the acoustic scenario. The VQ-DAE at 13.5 kbit/s achieved higher VSTOI scores at lower latency than all tested audio codecs except for the G.722. As speech intelligibility presumably increases monotonically with increasing VSTOI scores, it can be concluded that the VQ-DAE matches or surpasses the intelligibilty of Opus at half or less its bitrate and at zero latency. For Opus, we found [6] VSTOI scores to correspond well to the measured speech understanding, but no such investigation exists for the AMR-WB, which was considerably outperformed regarding VSTOI scores and bitrate. It is possible, that VSTOI somewhat misjudges the true intelligibility to CI users for the AMR-WB or that the FFMPEG implementation is suboptimal. Subjectively, the AMR-WB coded files at the highest bitrates sounded like the reference. For the G.722 results were as expected, further supporting the usefulness of VSTOI for algorithm development. In [6] the EC2 achieved reference speech intelligibility, while its VSTOI score was at least 0.01 lower than the reference VSTOI score. While the mapping from VSTOI scores to word recognition scores is certainly data dependent, this should suggests that the VQ-DAE, with a VSTOI score difference of 0.007 to the reference, yields reference speech intelligibility. It is interesting to compare the Electrocodec to the VQ-DAE, as the Electrcodec utilizes time-dependencies for its compression scheme, whereas the VQ-DAE solely relies on frequency-dependencies. While the Electrocodec could still be somewhat improved, a gap of about 7-9 kbit/s is unlikely to be closed in the future. These result might suggest that frequency-dependencies are of greater importance for the coding of the electrical stimulation patterns. In contrast to conventional audio coding, with sampling rates of 16 kHz or more, the stimulation patterns are usually generated at a channel stimulation rate of 500 - 1200 pulses per second. This considerably decreases the autocorrelation in the stimulation patterns, and thus makes predictive coding less effective. The observed improvement in VSTOI scores of the VQ-DAE, when considering SNRs of 5 dB or less, can be explained by learned denoising. While the train set was fairly balanced with respect to SNR, 50 % of its files having an SNR ≥ 15 dB, it is reasonable that the NA-VSTOI gradient descent made the DAE learn a denoising algorithm to improve speech understanding, as less noisy excitation patterns cannot be improved in the same manner. From Fig. 5 a slightly poorer performance of the DAE can be observed for speech files with higher VSTOI scores. These generally correspond to higher SNR files agreeing with the hypothesis put forward. However, CI research is more concerned to improve speech understanding in lower SNR situations, and thus a somewhat poorer performance at higher SNRs is not necessarily a downside. Additionally, training the VQ-DAE on a larger train set of 200 files allowed to considerably improve performance at higher SNRs as well. Interestingly, every time for four and five latent dimensions, hyperparameter optimization yielded a first hidden layer that was wider than the input layer. This might suggest that an additional preprocessing step is necessary before compression can efficiently performed. Initially, 16 and 8 neurons were tested in the encoder and decoder per layer, which did not allow to achieve reference VSTOI scores. In [17] STOI was

TABLE III

MEDIAN VSTOI SCORES OF THE VECTOR-QUANTIZED (VQ) AND SCALAR QUANTIZED (MSQ) DEEP AUTOENCODER (DAE) AND THE OTHER INVESTIGATED CODECS AND THE REFERENCE CONDITION (REF) ACROSS THE ENTIRE TESTSET AND THE SUBSET OF CONDITIONS WITH A SIGNAL TO NOISE RATIO $\leq$ 5 DB. VALUES IN PARENTHESES ARE THE RESPECTIVE BITRATE IN KBIT/S.

Dataset\Condition	Ref	VQ-DAE (12.6)	VQ-DAE (13.5)	DAE MSQ (18)	DAE MSQ (22.5)	EC2 (20.1)	EC3 (24.3)	$Opus_{5ms}$ (19.2)	$Opus_{5ms}$ (35.18)	$Opus_{7.5ms}$ (17.61)	$Opus_{7.5ms}$ (33.62)	G.722 (64)	AMR-WB (9.62)	AMR-WB (24.48)
Test Set	0.648	0.641	0.642	0.636	0.644	0.644	0.647	0.581	0.641	0.633	0.642	0.648	0.619	0.633
Test Set ($\leq$ 5dB)	0.578	0.579	0.583	0.576	0.584	0.573	0.577	0.529	0.569	0.566	0.571	0.578	0.555	0.567

used to optimize a neural network as well. However, there STOI was not combined with hyperparameter optimization and had to be approximated, which was not applicable for our work due to the additional CI processing involved. Instead, we used an approximation of the gradient of STOI for training.

V. CONCLUSION

This work investigated vector-quantized deep autoencoders (VQ-DAE) for the compression of the excitation patterns of cochlear implants. At a bitrate of 13.5 kbit/s and zero delay, the VQ-DAE achieved equal or superior speech intelligibility measured through an objective intelligibility measure while reducing the bitrate by up to $50\,\%$ compared to state-of-the-art audio codecs. The VQ-DAE was found to generalize well to unseen acoustic scenarios and was able to slightly improve speech intelligibility in low signal-to-noise ratio conditions.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Waldo Nogueira and Tom Gajecki for their comments and suggestions.

REFERENCES

[1] T. Goehring, M. Keshavarzi, R. P. Carlyon, and B. C. Moore, "Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 705–718, 2019.

[2] F. Henry, M. Glavin, and E. Jones, "Noise reduction in cochlear implant signal processing: A review and recent developments," *IEEE Reviews in Biomedical Engineering*, pp. 1–1, 2021.

[3] T. Gajecki and W. Nogueira, "A synchronized binaural n-of-m sound coding strategy for bilateral cochlear implant users," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[4] M. Stone, B. Moore, K. Meisenbacher, and P. Derleth, "Tolerable hearing aid delays. v. estimation of limits for open canal fittings," *Ear and hearing*, vol. 29, pp. 601–17, 09 2008.

[5] R. Hinrichs, T. Gajecki, J. Ostermann, and W. Nogueira, "Coding of electrical stimulation patterns for binaural sound coding strategies for cochlear implants," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 4168–4172.

[6] R. Hinrichs, T. Gajecki, J. Ostermann, and W. Nogueira, "A subjective and objective evaluation of a codec for the electrical stimulation patterns of cochlear implants," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1324–1337, 2021. [Online]. Available: https://doi.org/10.1121/10.0003571

[7] R. Hinrichs, L. Ehmann, H. Heise, and J. Ostermann, "Lossless compression at zero delay of the electrical stimulation patterns of cochlear implants for wireless streaming of audio using artificial neural networks," in *2022 7th International Conference on Frontiers of Signal Processing (ICFSP)*, 2022, pp. 159–164.

[8] A. Habibian, T. van Rozendaal, J. M. Tomczak, and T. Cohen, "Video compression with rate-distortion autoencoders," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7032–7041, 2019.

[9] G. Min, C. Zhang, X. Zhang, and W. Tan, "Deep vocoder: Low bit rate compression of speech with deep autoencoder," in *2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2019, pp. 372–377.

[10] B. Wang and J. Saniie, "Massive ultrasonic data compression using wavelet packet transformation optimized by convolutional autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2021.

[11] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167639390900107

[12] H. Kayser, S. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 6, 12 2009.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[14] E. H.-H. Huang, C.-M. Wu, and H.-C. Lin, "Combination and comparison of sound coding strategies using cochlear implant simulation with mandarin speech," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2407–2416, 2021.

[15] H. Chen, T. Duncan, and B. Pasik-Duncan, "A kiefer-wolfowitz algorithm with randomized differences," *IEEE Transactions on Automatic Control*, vol. 44, no. 3, pp. 442–453, 1999.

[16] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter, "Smac3: A versatile bayesian optimization package for hyperparameter optimization," *Journal of Machine Learning Research*, vol. 23, no. 54, pp. 1–9, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-0888.html

[17] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5059–5063.