Vector-Quantized Feedback Recurrent Autoencoders for the Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay

Reemt Hinrichs
Institut für Informationsverarbeitung
Leibniz Universität Hannover, Germany
Email: hinrichs@tnt.uni-hannover.de

Julian Bilsky
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany

Jörn Ostermann
Institut für Informationsverarbeitung
Leibniz Universität Hannover
Hannover, Germany

***Abstract*—Cochlear Implants (CIs) are surgically implanted hearing devices that allow to restore a sense of hearing in people suffering from moderate to profound hearing loss. Modern CIs offer wireless streaming of audio to the signal processor of the CI to improve speech understanding in complex acoustic environments. To conserve energy in this wireless streaming, proprietary source coding of the stimulation patterns of CIs was proposed, achieving state-of-the-art results with respect to bitrate, latency and intelligibility of the coded stimulation patterns. This work investigates vector-quantized feedback recurrent autoencoders (VQ FRAE) to improve source coding of the stimulation patterns of CIs. The VQ FRAE is optimized with respect to the non-differentiable STOI using simultaneous perturbation stochastic approximation. With this approach, a state-of-the-art bitrate of 4.69 kbit/s was achieved, while maintaining zero latency and little to no degredation of intelligibility. The FRAE outperforms audio codecs like Opus with respect to bitrate, intelligibility and latency.**

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 social situations still pose a problem [1]. Wireless streaming of audio as required for, e.g., beamformers, remote microphones [2] or binaural sound coding strategies [3] is among the techniques applied to improve speech understanding in these challenging environments. To save power or bandwidth in this wireless transmission, signal compression is commonly applied to reduce the bitrate of the audio signal before transmission. This coding usually introduces an additional delay and should be kept as small as possible, as speech perception of hearing aid users can be affected by delays above the range of $5-10$ ms [4]. Due to this delay constraint, the selection of source coding algorithms is severly limited. For this purpose, we proposed [5], [6], [7], [8] to code and transmit the electrical stimulation patterns generated by the sound coding strategy of the CI. The signal flow of this approach is depicted in Fig. 1. Initially, we proposed the Electrocodec [5], [6], 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 sound coding strategy. To further reduce the required bitrate while maintaining zero latency, we investigated vector-quantized autoencoders (VQ AEC), whose hyperparameters were optimized using bayesian optimization [8]. We were able to optimize the AEC with respect to the short-time objective intelligibility measure (STOI) [9] used to assess the intelligibility of stimulation patterns. That way, we were able to reduce the bitrate by almost 50 % compared to the Electrocodec. However, the proposed AEC optimization is sub-optimal due to not making use of time-dependencies present in the stimulation patterns, and the structure consisting of AEC and VQ was not optimized together. Recently, feedback recurrent autoencoders (FRAEs) have been proposed [10], [11] for the compression of sequential data. FRAEs feed previous, decoded frames back to the input layers of the encoder and decoder, to allow to make use of redundancies between frames. The number of frames fed back to the input layers is called recurrent dimension. We use FRAEs alongside VQ of the latent space to compress the stimulation patterns of CIs at zero delay, and compare them to regular autoencoders without recurrency (AEC). To optimize the overall structure including the VQ, we use simultaneous perturbation stochastic approximation, a numerical technique for the approximation of gradients [12], that way achieving considerable compression gains. The result is an automatic approach for the development of near-optimal delay-free compressors for the stimulation patterns of CIs. Finally, we introduce a regularization scheme of the vocoder STOI scores used during training, to achieve a compression performance less variant across signal-to-noise ratios (SNRs).

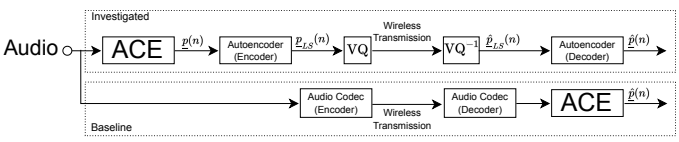


Fig. 1: Two methods to wireless transmission of audio for CIs. Conventionally, the input audio would be encoded by an audio codec, transmitted to the signal processor of the CIs, where the audio is subsequently decoded. In the investigated approach, the audio signal is first processed by the sound coding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

978-1-5386-5541-2/18/$31.00 ©2018 IEEE

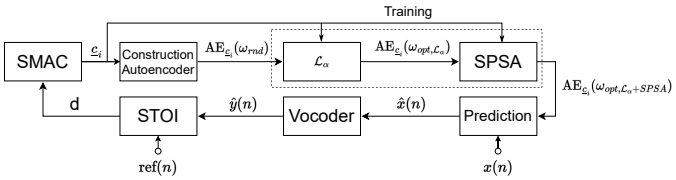


Fig. 2: The hyperparameters $\underline{c}_i$ are used to construct an autoencoder (AE) which is subsequently trained using gradient descent and simultaneous perturbation stochastic approximation (SPSA). Then, the stimulation pattern $x(n)$ of a single speech signal is compressed and decompressed, reconstructed 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 d is then returned to SMAC to assess the quality of the hyperparameters $\underline{c}_i$.

II. Fundamentals

A. Advanced Combination Encoder

The advanced combinational encoder (ACE) sound coding strategy is a common sound coding strategy for CIs [3]. The input audio signal of the CI is split into M subbands by a discrete fourier transform. 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, and $N < M$ 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), which maps from the acoustic to the electric domain, is applied to each $a \in A$. This results in the signal $\underline{p}_n := (LGF(a_1(n)), \ldots, LGF(a_M(n)))^T$. The input signal of all investigated models is $\underline{p}_n$. All CI parameters, whose description can be found in [8], were set to default values. The channel stimulation rate, which is the frequency at which the $\underline{p}_n$ are generated, was set to 900 pulses per second.

B. Datasets

To create realistic noisy speech signals, the TIMIT speech corpus [13] was processed using behind-the-ear head related transfer functions (HRTF) from [14]. 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 $5°$. An azimuth of $-90°$ corresponds to a source located to the left, $+90°$ correspond to a source located to the right, and $0°$ correspond to the front of the listener. Source distance was 80 cm. Each speech recording of the training and test data of TIMIT was processed using random signal-to-noise ratios (SNRs), speech and noise azimuths, acoustic environments and noise type from a list of values given in Tab. I. As noise, we used Comité Consultatif International Téléphonique et Télégraphique (CCITT) [15] noise, bus noise, office noise and restaurant noise. CCITT noise is speech-shaped noise often used in clinical research. The HRTF-processed speech files were then processed by ACE to generate the corresponding stimulation patterns used by the autoencoders.

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 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

TABLE II: Mean VSTOI scores $\overline{VSTOI}$ of the investigated feedback recurrent autoencoders (FRAE) and non-recurrent autoencoders (AEC) on the train set without quantization. ΔVSTOI as defined in Section II-G. The reference mean VSTOI score was 0.6643.

FRAE	$\overline{VSTOI}$	ΔVSTOI	AEC	$\overline{VSTOI}$	ΔVSTOI
FRAE-L5-H2-R1	0.67329	0.00895	AEC-L3-H2	0.66260	-0.00174
FRAE-L5-H2-R2	0.67762	0.01328	AEC-L3-H3	0.64235	-0.02199
FRAE-L5-H2-R3	0.66925	0.00491	AEC-L4-H2	0.67106	0.00672
FRAE-L5-H2-R4	**0.68955**	**0.02522**	AEC-L4-H3	0.66440	0.00006
FRAE-L5-H3-R1	0.68511	0.02077	AEC-L5-H2	0.67952	0.01518
FRAE-L5-H3-R2	0.68688	0.02254	AEC-L5-H3	**0.68001**	**0.01567**
FRAE-L5-H3-R3	0.67851	0.01417			
FRAE-L5-H3-R4	0.68546	0.02112			

C. Loss Function and Pre- and Postprocessing

Due to the N of M band-selection performed by ACE, the distortion of the stimulation patterns 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 loss $\mathcal{L}_\alpha$ defined as

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

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 a weighting factor optimized with SMAC. $\sigma(x)$ is the rectified linear unit, whose usage was motivated by the pre- and postprocessing applied. In the pre-processing, any subband not selected at time n was set to a negative value to distinguish 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 is considered not selected in the post-processing, if $\hat{p}_i(n) < 0$. If $p_i(n) < 0$, no distortion occurs.

TABLE III: Mean VSTOI scores of the best performing models on the test set before and after optimizing the autoencoders including the quantizers using SPSA. VSTOI scores surpassing the mean reference VSTOI scores are highlighted using bold font. The mean reference VSTOI score was 0.6343.

Model\Codebook Size [bit]	6		7		8		10	
	Before	After	Before	After	Before	After	Before	After
AEC-L5-H2					0.5776	0.6336	0.6051	**0.6395**
AEC-L5-H3					0.5627	0.6191	0.5941	**0.6349**
FRAE-L5-H3-R2	0.5770	0.6244	0.5914	**0.6394**	0.5988	**0.6372**	0.6190	**0.6517**
FRAE-L5-H2-R4	0.5363	0.6221			0.5797	0.6241	0.6126	**0.6465**
FRAE-L5-H3-R4	0.6087	**0.6486**	0.6227	**0.6536**	0.6255	**0.6527**	**0.6394**	**0.6555**

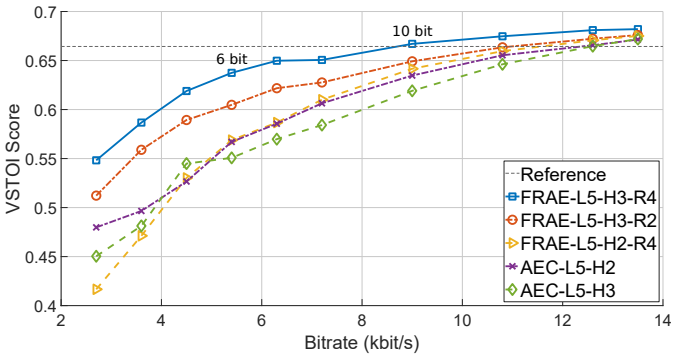


Fig. 3: VSTOI scores across bitrate of the best performing FRAE and AEC models using fixed length codes on the train set before optimizing the models including the quantizers with the SPSA.

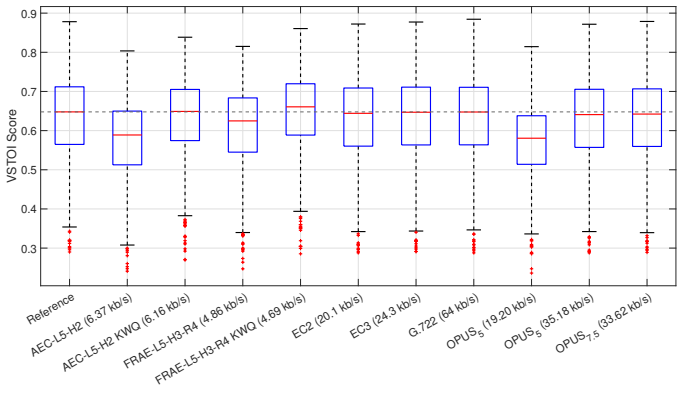


Fig. 4: Comparison of the best performing autoencoders with the Electrocodec, Opus and the G.722 audio codec on the test set. Results after optimizing the entire model including the quantizer with the SPSA are denoted by KWQ. For Opus, the algorithmic latency is specified as an index in milliseconds.

However, for $\hat{p}_i(n) \geq 0$ the subband is incorrectly selected and a distortion value needs to be assigned.

D. Short-Time Objective Intelligibility Measure

We use the well known STOI [9] to objectively assess the intelligibility of stimulation patterns generated by ACE. For its application, the stimulation patterns are resynthesized using a sine vocoder, and the resulting waveform is then compared to the original, clean reference audio input signal. The resulting intelligibility score $d \in [0, 1]$ is called vocoder STOI (VSTOI) score [6], with $d = 1$ denoting the best possible intelligibility, i.e. no difference to the reference, and $d = 0$ denoting the worst possible intelligibility. Even minor VSTOI score differences, starting at about 0.01, can represent a measurable difference in intelligibility [6].

E. Numerical Approximation of Gradients

For the optimization of the AEs including the quantizers with respect to STOI, which, like the quantizers, is nondifferentiable, simultaneous perturbation stochastic approximation (SPSA) [12], [16] was used. The update equation of the SPSA for all parameters $\underline{\omega}$ of the AEs and the quantizers is

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

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 parameters. 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 AE with the weights $\underline{\omega}$. The parameters A and c were obtained through hyperparameter optimization.

TABLE IV: Bitrates in kbit/s of the best performing models on the test set after optimizing the autoencoders including the quantizers using the SPSA and huffman coding.

Model \Bit	6	7	8	10
FRAE-L5-H3-R2	4.80 kbit/s	5.56 kbit/s	6.37 kbit/s	7.96 kbit/s
FRAE-L5-H2-R4	4.73 kbit/s		6.33 kbit/s	7.96 kbit/s
FRAE-L5-H3-R4	4.69 kbit/s	5.54 kbit/s	6.49 kbit/s	8.11 kbit/s
AEC-L5-H2			6.16 kbit/s	7.82 kbit/s
AEC-L5-H3			6.44 kbit/s	8.12 kbit/s

F. Hyperparameter Optimization of the DAE

The hyperparameters of the autoencoders and SPSA were optimized using Bayesian optimization, implemented through sequential model-based algorithm configuration (SMAC) [17]. The steps involved in the hyperparameter optimization are depicted in Fig. 2. The hyperparameters $\underline{c}_i$ tested by SMAC were used to construct the autoencoder, which was then trained with the adam solver on a single stimulation pattern for 500 epochs using the loss function according to Eq. 1, yielding approximately optimal weights. Then, 100 epochs of SPSA training was performed. The VSTOI score achieved on the stimulation pattern was then returned to SMAC.

G. Regularization

Both, FRAE and AEC, improved the VSTOI scores at low SNRs, while showing lower performance at high SNRs. The cause likely is denoising learned during training. To achieve a more balanced compression performance across SNRs, we propose a novel regularization scheme. Let $VSTOI_{Coded}$ be the VSTOI score assigned to a given decoded signal of an AE, and $VSTOI_{Ref}$ be the VSTOI score of the reference stimulation patterns, i.e., the stimulation patterns without any coding applied, then we define $\Delta VSTOI := VSTOI_{Coded} - VSTOI_{Ref}$. To decrease the nominal benefit of improving $VSTOI_{Coded}$ beyond $VSTOI_{Ref}$, we introduce the modified VSTOI score $VSTOI^L_{Coded,mod}$ defined as

$$VSTOI^L_{Coded,mod} := \begin{cases} VSTOI_{Ref} + \frac{1}{L} tanh(L \cdot \Delta VSTOI), A \\ VSTOI_{Coded}, otherwise \end{cases} \tag{3}$$

with $L > 1$ and the condition $A := VSTOI_{Coded} \geq VSTOI_{ref}$. The tanh function allows to smoothly transition from the linear to the nonlinear section, yielding a maximum $\Delta VSTOI$ score of $\frac{1}{L}$.

III. Experiments and Results

First, the hyperparameters of several FRAE and AEC configurations were optimized as described in II-F. Then, all FRAE and AEC models were optimized, without vector quantization, for 7000 iterations with respect to their average VSTOI scores on the train set using the SPSA. Afterwards, the codebooks of the VQ were trained using kmeans and the latent vectors generated by the models on the train set. Codebook sizes between 3 bit to 15 bit were investigated. Then, the best performing models, now together with the quantizers, were optimized for another 7000 iterations using the SPSA and the model performance on the test set was evaluated for selected bitrates, now using huffman coding to minimize the bitrate. Table II shows the mean VSTOI scores across the train set of all configurations investigated, where, e.g., FRAE-L5-H3-R2 denotes the FRAE with a latent dimension of five, three hidden layers of the encoder and a recurrent dimension of two. More hidden layers and a higher recurrency dimension tended to yield better results. Fig. 3 depicts the mean VSTOI scores across bitrate for the best performing models on the train set, before optimizing the models including the VQ by the SPSA. The FRAE-L5-H3-R4 outperformed all other models, despite performing less favorably before quantization, surpassing the reference mean VSTOI scores at 9 kbit/s. Mean VSTOI scores on the test set before and after optimizing the entire structure including the quantizers using the SPSA are given in Table III. Corresponding bitrates, achieved using huffman coding, are summarized in Table IV. A considerable improvement of up to about 0.09 in mean VSTOI score was achieved through optimization of the AEs plus quantizers. The FRAE-L5-H3-R4, achieving a mean VSTOI score of 0.6486 on the test set, surpassed the mean reference VSTOI score of 0.6343 at 4.69 kbit/s. Fig. 4 shows box plots of the VSTOI scores achieved on the test set for the best performing FRAE, using 6 bit VQ, and AEC, using 8 bit VQ, before and after optimizing the entire structure including the quantizers. Additionally, VSTOI scores are depicted for the Electrocodec with 2 bit and 3 bit per subband, labeled EC2 and EC3, respectively, Opus, and the G.722. Opus was used with an algorithmic latency of 5 ms and 7.5 ms. The FRAE-L5-H3-R4, after optimization including the VQ, considerably outperformed all tested audio codecs with respect to VSTOI scores and bitrate. Fig. 4 reveals a minor decrease of performance for high VSTOI scores, which correspond to higher SNRs. In contrast, Fig. 5, which depicts the same evaluation for the subset of the test set with an SNR of $\leq$ 5 dB, reveals a considerable improvement of the VSTOI scores by the FRAE-L5-H3-R4 KWQ at low SNRs. However, a minor bitrate increase of about 1 kbit/s was observed for the AEs, caused by an increased information content of the stimulation patterns. Fig. 6 shows the impact of the regularization according to Eq. 3 on the $\Delta VSTOI$ scores of the FRAE-L5-H2-R4, here without quantization. Due to regularization, extreme values occur less frequently and performance increases at high SNRs at the cost of performance at low SNRs. Regularizing for further epochs or from the beginning can yield further improvements.

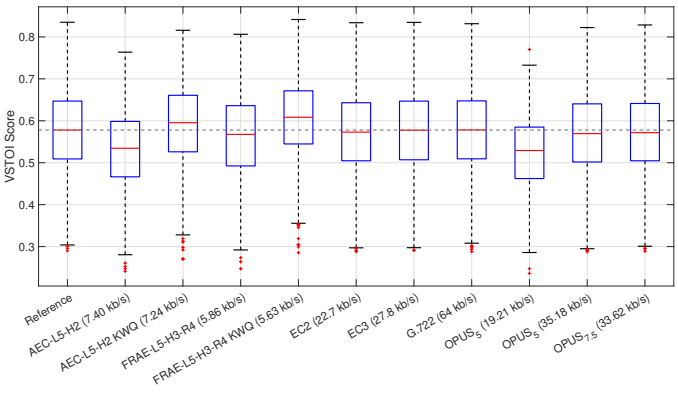


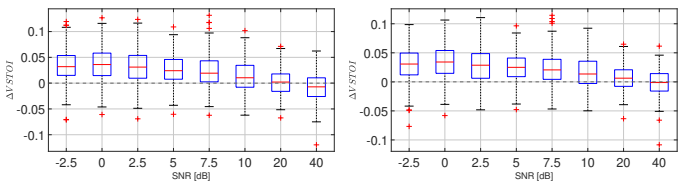
Fig. 5: Comparison of the best performing autoencoders with several audio codec on the subset of the test set with an SNR of $\leq$ 5 dB. Results after optimizing the entire model including the quantizer with the SPSA are denoted by KWQ.

IV. Discussion

The most important result of this work is a method to automatically train near-optimal zero-delay compressors of the stimulation patterns of cochlear implants. While we tested the approach using ACE only, it should be applicable out of the box for any sound coding strategy used in cochlear implants. Larger recurrent dimensions can likely boost the performance further. The achieved bitrates are in the ballpark of Meta's EnCodec [18], however, our approach of coding the stimulation patterns using the FRAE achieves zero latency at lower complexity, unlike 13+ ms for the EnCodec. The results [18], which are based on regular audio signals, suggest possible further gains based on stimulation patterns, which should contain less information than regular audio signals. The SPSA proved to be highly useful yet again, allowing to consistently improve the coding performance of the nondifferentiable autoencoder + quantizer structure, and might be a viable alternative to other common techniques [19], [20] to train nondifferentiable algorithms.

V. Conclusion

This work investigates vector-quantized feedback recurrent autoencoders (VQ FRAE) for the compression of the stimulation patterns of cochlear implants. The nondifferentiable VQ FRAE was optimized with respect to an objective intelligibility measure using simultaneous perturbation stochastic approximation. Considerable coding gains were achieved by optimizing the entire structure, achieving undegraded intelligibility of the stimulation patterns at 4.69 kbit/s and zero latency, outperforming state-of-the-art codecs. A proposed regularization schemes allows to achieve a more balanced coding performance.



(a) No regulation (b) Regulated, $L = 40$

Fig. 6: $\Delta VSTOI$ scores, without quantization, of the FRAE-L5-H2-R4 across SNRs (a) without and (b) with regulation.

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] R. Hinrichs, F. Ortmann, and J. Ostermann, "Vector-quantized zero-delay deep autoencoders for the compression of electrical stimulation patterns of cochlear implants using stoi," in *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES 2022)*, 2022, pp. 159–164.

[9] 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.

[10] Y. Yang, G. Sautière, J. Ryu, and T. Cohen, "Feedback recurrent autoencoder," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3347–3351, 2019.

[11] A. Goliński, R. Pourreza, Y. Yang, G. Sautière, and T. Cohen, "Feedback recurrent autoencoder for video compression," in *Asian Conference on Computer Vision*, 2020.

[12] J. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 3, pp. 817–823, 1998.

[13] 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

[14] 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.

[15] International Telecommuncation Union, "ITU Recommendation G.227," 1993, last access 10.09.2019. [Online]. Available: https://www.itu.int/rec/T-REC-G.227-198811-I/en

[16] 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.

[17] 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

[18] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[19] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. Van Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," 2017. [Online]. Available: https://arxiv.org/abs/1704.00648

[20] S. Uhlich, L. Mauch, K. Yoshiyama, F. Cardinaux, J. A. Garcia, S. Tiedemann, T. Kemp, and A. Nakamura, "Differentiable quantization of deep neural networks," *arXiv preprint arXiv:1905.11452*, vol. 2, no. 8, 2019.