

# FROM HALLUCINATION TO ARTICULATION: LANGUAGE MODEL-DRIVEN LOSSES FOR ULTRA LOW-BITRATE NEURAL SPEECH CODING

Jayeon Yi and Minje Kim

University of Illinois Urbana-Champaign, Siebel School of Computing and Data Science, USA 61801

## ABSTRACT

“Phoneme Hallucinations (PH)” commonly occur in low-bitrate DNN-based codecs. It is the generative decoder’s attempt to synthesize plausible outputs from excessively compressed tokens missing some semantic information. In this work, we propose language model-driven losses (LM loss) and show they may alleviate PHs better than a semantic distillation (SD) objective in very-low-bitrate settings. The proposed LM losses build upon language models pretrained to associate speech with text. When ground-truth transcripts are unavailable, we propose to modify a popular automatic speech recognition (ASR) model, Whisper, to compare the decoded utterance against the ASR-inferred transcriptions of the input speech. Else, we propose to use the timed-text regularizer (TTR) to compare WavLM representations of the decoded utterance against BERT representations of the ground-truth transcriptions. We test and compare LM losses against an SD objective, using a reference codec whose three-stage training regimen was designed after several popular codecs. Subjective and objective evaluations conclude that LM losses may provide stronger guidance to extract semantic information from self-supervised speech representations, boosting human-perceived semantic adherence while preserving overall output quality. Demo samples, code, and checkpoints are available online <sup>1</sup>.

*Index Terms*— Speech codec, language model, loss function

## 1. INTRODUCTION

Deep neural networks (DNNs) have emerged as a viable model for low-bitrate speech codecs. While traditional codecs such as MELPe [1] or Codec2 [2] are known to enable intelligible transmissions at sub-1 kbps rates at the cost of reduced sound quality, many DNN-based codecs have successfully reached such ultra-low bitrates with improved intelligibility and perceptual sound quality [3, 4, 5, 6, 7].

Hence, the training objectives for DNN-based speech codecs are to learn to convert speech to a compact (i.e., low-bitrate) discrete representation and then decode it back to audio with the least possible difference from the input. More specifically, there are two major categories of codecs based on the training procedures and input features, as commonly classified [6, 8, 9]. First, *acoustic* codecs rely more on reconstruction losses defined within a short period of signals [3, 4, 5]. Conversely, *semantic* codecs aim to preserve semantic information during the coding process, often by learning from foundation models trained with longer-term self-supervision objectives, such as HuBERT [10] or WavLM [11], which are thought to contain

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [26ZC1100, Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World].

<sup>1</sup><https://minjekim.com/research-projects/lm-loss#icassp2026>

rich semantic information in their output. Many successful state-of-the-art codecs [6, 7, 12, 13, 14] are of this category.

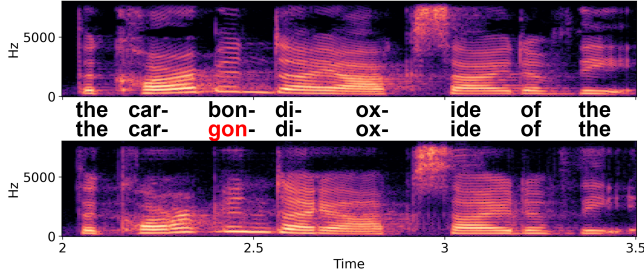
In this paper, we focus on a specific type of coding artifact, “phoneme hallucination (PH),” which results from generative decoders’ attempts to synthesize plausible outputs from excessively compressed codes that may lack some essential semantic information. As a remedy, many codecs propose to adapt semantic distillation (SD) procedures [9, 13, 15], or use self-supervised representations as inputs [6, 7, 16, 17]. Indeed, semantic losses seem to emphasize linguistic accuracy in the reconstruction.

However, semantic codecs such as TAAE [12], FocalCodec [6], and SemantiCodec [7] do exhibit PHs at low bitrates ( $< 0.4$  kbps), as made evident in listening examples shared online by the respective authors. On the other hand, it has been shown possible to revert self-supervised speech representations to speech [18]. This implies that better distillation objectives may exist, with the potential to benefit existing and new architectures alike.

Meanwhile, automatic speech recognition (ASR) models, such as Whisper [19], are trained to directly estimate subwords given a speech utterance, where the model’s ability in learning the long-term linguistic context is critical to its success. Similarly, the timed-text regularizer (TTR) [20] directly maps WavLM [11] representations to those learned by the corresponding language model, BERT [21], to introduce linguistic contexts to speech enhancement results. Since these models were explicitly trained to associate speech representations with text representations, they may be a richer source for ultra-low bitrate speech codes to learn the semantics from.

Some previous works in speech coding do use such linguistic context to enhance token quality. For instance, LLM-Codec [22] proposes to fill and freeze its codebooks with large language model (LLM) embeddings for thousands of carefully selected words, along with an L2 loss enforcing similarity between the VQ outputs and LLM embeddings. However, the bitrate is not as competitive as other semantic codecs. As another example, XY-tokenizer [14] distills from an LLM to map speech tokens to LLM input so that the LLM would output a transcription for the speech. In this way, the learned tokens embed semantic information transferred from the LLM. Again, XY-tokenizer operates in a much higher bitrate (1 kbps). In summary, while LLMs see some recent usage as a part of distillation objectives for codecs, proposed architectures in the literature introduce additional bits to host the semantic information.

Meanwhile, there are also non-speech-codec works, e.g., [23], that attaches adapters [24] to a speech enhancement model to distill from an LLM. However, adapters induce additional inference overhead, while also blocking the loss flow. In contrast, [20] uses their TTR model as a regularization loss, inducing zero inference overhead while still improving their speech separation model. However, the work does not examine whether their scheme improves the linguistic conformance of output speech. Neither of these approaches has been evaluated for training codecs, either.



**Fig. 1.** (Top) The input speech. (Bottom) The 187.5 bps reference codec with no LM losses exhibits phoneme hallucinations (PHs).

In this work, we formulate and test *language model-driven losses* (LM loss), demonstrating how they can enhance a neural speech codec. We first show that PHs can still manifest in a simple semantic codec trained on widely used training procedures. Then, we formulate two families of LM losses. While these do leverage pretrained models like some of the previous works, our proposed methods do not levy any architectural constraints or require any additional finetuning once pretrained; they can be applied to any model that outputs speech. We show that many tested variants of LM losses result in suppressed PHs. Finally, we compare the strengths and drawbacks of each approach with objective metrics, including word error rate (WER), as well as a MUSHRA-like subjective test and an intelligibility test.

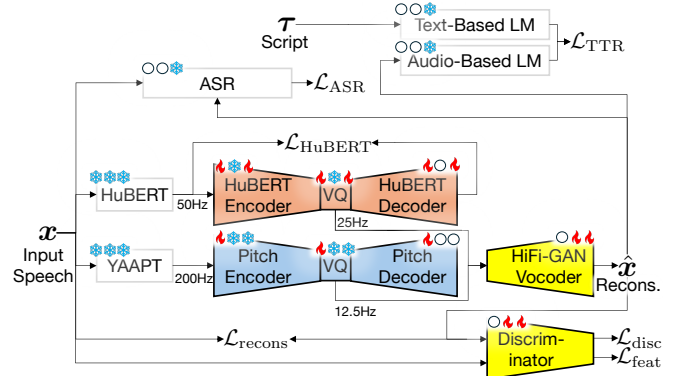
## 2. LM LOSSES FOR CODECS

### 2.1. The Phoneme Hallucination (PH) Phenomenon

Although a past work finds  $\sim 100$ bps to be a minimum bitrate for speech communication [25], many codecs often operate with a higher bitrate due to their inefficiency and the need for sending extensive acoustic information on top of the semantic contents. In this work, we focus on a specific type of artifact a modern neural codec can generate, namely *phoneme hallucination* (PH), where the codec reconstructs incorrect phonemes. It is a failure mode related more to the semantic aspect of the utterance than acoustics, as a PH can sound clean. We observe PHs when an aggressive reduction of bitrate is coupled with generative decoding, as in very-low-bitrate codecs like [6, 7, 12]. For example, the encoder’s temporal decimation could be too much, to the level that the frame rate of the code is too low to represent a phoneme. Having too few codewords in a codebook could also lead to misrepresentation: instead of producing acoustic noise, a generative decoder can try to come up with a plausible phoneme different from the original one. We call it hallucination because the synthesized phoneme could still sound clean, meaning we need an advanced semantic loss function that measures this discrepancy to tame the behavior of the generative decoder, while potentially injecting more information to the low-bitrate code.

### 2.2. LM Losses

The proposed LM losses distill linguistic information from pretrained LMs to explicitly match speech and text. They can be applied to any DNN-based speech codec with no architectural modifications or additional inference overhead. Generally, an LM loss takes the decoded utterance as input and measures its linguistic feasibility by comparing the two semantics extracted from the decoded



**Fig. 2.** Architecture and training of our reference codec. Our three-stage training procedure emulates common codec-training setups.  $\color{red}\bullet$  denotes the modules updated in the given stage, while  $\color{blue}\bullet$  represents frozen ones.  $\circ$  does not participate in. For example,  $\color{red}\bullet\color{blue}\bullet\circ$  means that the module is updated in the first stage, while being frozen for the second stage, and then being idle in the third stage. In the third stage, either a LM loss ( $\mathcal{L}_{ASR}, \mathcal{L}_{TTR}$ ) or the SD loss ( $\mathcal{L}_{HuBERT}$ ) is employed in combination with the others.

utterance and a corresponding script. We propose two loss functions to cover both cases: with and without ground-truth transcripts.

#### 2.2.1. The automatic speech recognition (ASR) loss

The ASR loss leverages ASR models pretrained on a subword-wise autoregressive transcription task as in [19]. Let an ASR model  $\mathcal{S}$  be trained to predict the next subword  $w_{i+1}$  in the learned token space,

$$w_{i+1} \approx \hat{w}_{i+1} \leftarrow \mathcal{S}(x, \widehat{W}_{<i+1}), \quad (1)$$

given a speech utterance  $x$  and all the previously predicted subword tokens  $\widehat{W}_{<i+1} := [\hat{w}_1, \hat{w}_2, \dots, \hat{w}_i]$ .

We repurpose this autoregressive loss function to measure the fitness of the codec’s decoded signal  $\hat{x}$ . To this end, we first predict the token sequence  $\widehat{W}$  by using the clean speech  $x$  as the input audio as in (1). Then, we predict another token sequence  $\widetilde{W}$  by using the decoded signal  $\hat{x}$  and the token sequence  $\widehat{W}$  predicted from the clean speech  $x$ ,

$$\mathcal{L}_{ASR}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{CE}(\mathcal{S}(\hat{x}, \widehat{W}_{<i}) || \hat{w}_i), \quad (2)$$

where  $\mathcal{L}_{CE}$  is the subword level prediction loss defined over tokens, e.g., cross entropy.

The main advantages of the proposed ASR loss are that it can leverage the internal LM that the ASR model already learned from the pairs of the clean utterance  $x$  and its corresponding tokens  $W$ . In addition, the loss defined in the token space does not need a ground-truth script, unleashing the possibility of using any clean speech corpora for codec training. While one can easily modify this loss to use a ground-truth script, we found in preliminary experiments that this may cause significant training instabilities.

#### 2.2.2. The timed-text regularizer (TTR) loss

Another approach to imposing more semantics to codec training is to directly use time-aligned scripts as the target, as proposed in [20] for

source separation. In the context of our neural speech codec (see Fig. 2), the timed-text regularization (TTR) first processes the decoded speech  $\hat{x}$  using the audio-based language model (LM) whose output embeddings are compared with the ones from the text-based LM, assuming their subword-level alignment.

More specifically, the LMs involved in the definition of the TTR loss are pretrained from a clean utterance and its script, so their output embeddings are guaranteed to be similar for clean speech. For the  $i$ -th subword and its corresponding segment of the speech signal  $\mathbf{x}^{(i)}$ , i.e.,  $\mathbf{x} = [\mathbf{x}^{(1)}; \dots; \mathbf{x}^{(N)}]$ , WavLM [11], a self-supervised speech model, transforms the variable-length subword audio signal  $\mathbf{x}^{(i)}$  into a series of embedding vectors  $\mathbf{X}^{(i)} = [\mathbf{X}_{:,1}^{(i)}, \dots, \mathbf{X}_{:,m}^{(i)}]$ . Since the number of embeddings  $m$  varies by the length of the spoken subword, a transformer-based *summarizer* module turns them into a single representative embedding vector:  $\bar{\mathbf{S}}_{:,i} \leftarrow \mathcal{P}_{\text{Sum.}}(\mathbf{X}^{(i)})$ . After constructing  $\bar{\mathbf{S}}$  with  $N$  total embeddings, we impose the inter-subword relations into this audio-based representation via another transformer-based *aggregator* module, which transforms the entire sequence  $\bar{\mathbf{S}}$  into a self-attended version  $\bar{\bar{\mathbf{S}}}$ , i.e.,  $\bar{\bar{\mathbf{S}}} \leftarrow \mathcal{P}_{\text{Agg.}}(\bar{\mathbf{S}})$ .

Meanwhile, the ground-truth script  $\tau$  is processed by the text-based LM, BERT [21], to produce the subword-level embedding sequence  $\mathbf{T}$ . Since both BERT and WavLM models are pretrained and frozen, the summarizer and aggregator layers can be seen as a trainable projection module that converts the audio embeddings into the space defined by  $\mathbf{T}$ . Hence, we define the TTR loss function:

$$\mathcal{L}_{\text{TTR}}(\bar{\bar{\mathbf{S}}}\|\mathbf{T}) = \frac{1}{N} \sum_{i=1}^N \left( 1 - \frac{\bar{\bar{\mathbf{S}}}_{:,i} \cdot \mathbf{T}_{:,i}}{\|\bar{\bar{\mathbf{S}}}_{:,i}\| \|\mathbf{T}_{:,i}\|} \right) + \frac{2}{N(N+1)} \sum_{1 \leq i < j \leq N} \|\bar{\bar{\mathbf{S}}}_{:,i} \cdot \bar{\bar{\mathbf{S}}}_{:,j} - \mathbf{T}_{:,i} \cdot \mathbf{T}_{:,j}\|^2. \quad (3)$$

The first term calculates the cosine similarity between each pair of  $\bar{\bar{\mathbf{S}}}_{:,i}$  and  $\mathbf{T}_{:,i}$ , while the second term scores the similarity of internal pairwise relations between  $\bar{\bar{\mathbf{S}}}$  and  $\mathbf{T}$  vectors.

While BERT and WavLM models are frozen,  $\mathcal{P}_{\text{Sum.}}$  and  $\mathcal{P}_{\text{Agg.}}$  are updated to minimize this loss for the clean signals. Once this pretraining is done, the frozen text-based and audio-based LMs are used to calculate the same TTR loss, but by taking the decoded signal  $\hat{x}$  rather than  $\mathbf{x}$  to update the modules in the codec.

### 3. EXPERIMENTAL SETUP

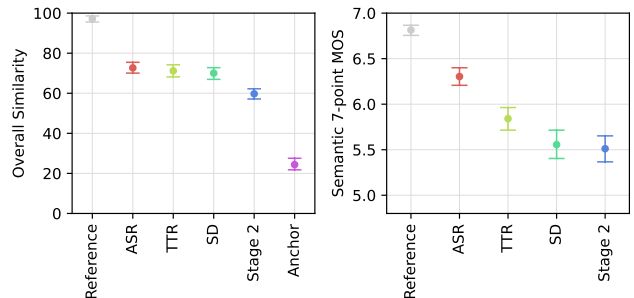
To test if LM losses can enhance neural speech codecs that already include semantic quality considerations, we design experimental procedures with common semantic codecs in mind.

#### 3.1. The Reference Codec

Our reference codec in Fig. 2 is a slight modification of a previous codec [17] that trained a HiFi-GAN vocoder [26] to decode from quantized pitch and HuBERT [10] features. Starting from this codec, we add a HuBERT encoder that decimates the feature rate by a factor of 2. The new encoder is comprised of a `conv1d` to reduce input dimensions from 768 to 128, followed by a ResBlock that shares architecture with the pitch encoder ResBlocks in [17]. Then, a HuBERT VQ codebook quantizes the encoder outputs. This codebook has a size of 32 or 64. Thus, the bitrate of the codec is 187.5 bps or 212.5 bps. We further expanded pitch codebook size to 32, keeping the bitrate same while enhancing acoustic reconstruction quality.

Rate (bps)	LM loss	Semantic		Acoustic	
		WER(%)↓		PESQ↑	WARPQ↑
		Whisper	wav2vec2.0		normalized
187.5	ASR	<b>1.45</b>	<b>4.56</b>	1.35	0.289
	TTR	2.34	7.13	1.39	0.293
	SD	3.33	11.2	<b>1.42</b>	<b>0.295</b>
	S2	3.04	8.82	1.35	0.283
212.5	ASR	<b>1.23</b>	<b>3.63</b>	1.37	.289
	TTR	1.53	5.25	1.44	.293
	SD	2.11	7.04	<b>1.46</b>	<b>.295</b>
	S2	2.09	6.34	1.36	.289
$\infty$	-	0.95	1.74	4.64	1.00

**Table 1.** Objective metrics for all codecs. ASR, TTR, and SD denote codecs from stage 3, respectively trained using the proposed LM losses ( $\mathcal{L}_{\text{ASR}}$ ,  $\mathcal{L}_{\text{TTR}}$ ) and SD loss ( $\mathcal{L}_{\text{HuBERT}}$ ). S2 denotes the stage-2 codec trained on frozen encoder and VQ. Rate of  $\infty$  denotes unencoded utterances.



**Fig. 3.** Overall similarity (left) and the semantic 7-point MOS (right) subjective evaluations; mean and 95% confidence intervals. Codecs trained with LM losses show significantly better semantic performance compared to the others.

Overall, this codec faithfully emulates some common patterns in the literature: self-supervised models as an “encoder” constituent [16, 6, 7, 13], an additional “compressor” to further decimate self-supervised representations [6, 8, 13], and separation of acoustic and semantic “branches” [7, 14].

#### 3.2. Dataset and Linguistics-Preserving Batching

To further simplify the acoustic part of the codec, we chose to train on the single-speaker LJSpeech dataset, which could be seen as a specialized module in the personalized neural speech codec as proposed in [27]. Combined with the architectural choices and the set of evaluations, this lays out a minimal yet sufficiently fertile framework for evaluating both semantic and acoustic effects of LM losses.

To allow codecs to learn longer semantic contexts, we split utterances by their source texts. “LJ021” to “024” utterances comprise the test split, while “025” to “027” comprises the validation split. We also concatenated utterances to train on longer (30-45s) segments from the original texts, where each segment starts with a unique sentence. We used a batch size of 1. The pitch and HuBERT embeddings for each utterance were respectively derived using YAAPT [28] and HuBERT-base-1s960h [10]. Finally, the text-speech alignments required by TTR were derived with the Montreal Forced Aligner [29].

### 3.3. LM loss configurations and TTR modules pretraining

We use the pretrained `Whisper-tiny` [19] model as the ASR model  $\mathcal{S}$ . For TTR, we follow [20] to use `BERT-base-uncased` [21] and `WavLM-base` [11]. The summarizer and aggregator are identical four-layer transformer encoder models with a dimension of 768 and a feedforward dimension of 1024. They are jointly pretrained on the `LibriSpeech-960h` dataset using similar splitting and batching procedures. The Adam optimizer was used with a learning rate of  $1 \times 10^{-4}$ ,  $(\beta_1, \beta_2) = (0.8, 0.99)$ , and with an exponential decay scheduler with  $\lambda = 0.999$ . We use the best-performing checkpoint in terms of validation loss after training for 1M steps.

### 3.4. Training the Codec

Many codecs have semantically initialized codebooks [16, 17, 22], often leveraging a pretrained autoencoder to further compress self-supervised representations [6, 8]. Meanwhile, also common is to attach SD objectives to codebooks in settings where all codec weights are jointly trained [9, 13, 15]. To account for both approaches, we adopt a three-stage training approach shown in Fig. 2.

In the first stage, the *HuBERT* codec and the *pitch* codec are trained to reconstruct respective inputs acquired from the frozen HuBERT model and the deterministic YAAPT module. In addition, the VQ codebook is trained with a commitment loss, defined as the MSE loss between VQ inputs and outputs. Following [17], exponential moving average updates are used for all VQ codebook entries, while embeddings unused in each batch are randomly reinitialized.

In the second stage, the modified HiFi-GAN vocoder is trained to reconstruct speech while the encoders and VQs are frozen. The training objective is a combination of log-Mel spectrogram L1 loss, adversarial loss, and feature matching loss. The latter two are derived using multi-scale and multi-period discriminators, following [26]. The stage-two reference codec is our baseline.

Finally, in the third stage, the codec is finetuned on three different settings. The HuBERT codec is unfrozen in the third stage, while the pitch codec stays frozen. The training objective is the sum of the second-stage training objective, commitment loss, and one of  $\mathcal{L}_{ASR}$ ,  $\mathcal{L}_{TTR}$ , or a semantic distillation loss ( $\mathcal{L}_{HuBERT}$ ) defined as the MSE between the input and output of the HuBERT codec. The latter loss is reused from the first stage to simulate past works that use SD objectives on one or more codebooks [9, 15]. For the HuBERT VQ codebook, the first-stage objectives and update schemes are reused.

For all stages, the AdamW optimizer was used with a learning rate of  $2 \times 10^{-4}$ , weight decay of 0.01, and  $(\beta_1, \beta_2) = (0.8, 0.99)$ . Learning rate was reduced every epoch by a factor of  $\lambda = 0.999$ . Validation was run every 1k steps, and training was stopped if validation metrics did not improve for 100k steps.

### 3.5. Evaluation

We conduct two different subjective tests to evaluate the codecs’ perceptual quality as well as the intelligibility improvement we claim to achieve using the LM losses. First, a MUSHRA-like [30] subjective study was conducted to gauge the overall similarity of each codec’s inputs and outputs. We used the decoded utterances at 187.5bps for four codec variants: one from the second stage (as explained in Section 3.1), and three from the third stage. 11 audio experts were asked to score 10 sets of utterances on how similar they felt the decoded utterances are compared to the reference. A 3.5kHz LPF low-anchor was used. Second, a mean opinion score (MOS) subjective study evaluates how well the decoded utterances comply to the LJSpeech transcriptions. 18 English speakers were asked to score 15 sets of

utterances on how well the utterances match the text, from 1 (“no matching words”) to 7 (“perfectly matches the text”). No low anchor was used. We employed webMUSHRA [31] for both studies, randomly sampling utterances from the test set and normalizing to -24 LUFS. WARPQ [32], PESQ [33], and WER were also used for acoustic quality and intelligibility. WARPQ and PESQ do not accurately reflect audio quality if temporal alignment lacks, but may be useful for comparisons. For WER, `Whisper-large-v3` and `wav2vec2.0` [34] were both used, lest `Whisper-tiny` in the ASR loss cause unjustly lower WER on the same family of models.

## 4. EXPERIMENTAL RESULTS

The MOS results on the right side of Fig. 3 show that models trained with LM losses boost the semantic compliance of decoded utterances with fewer PHs. In the figure and hereafter, ASR, TTR, and SD denote codecs from stage 3, respectively trained using the proposed ASR loss, TTR loss, and  $\mathcal{L}_{HuBERT}$ . S2 denotes the stage-two codec trained on frozen encoder and VQ. A Wilcoxon signed-rank test with Bonferroni correction shows that all differences except S2 and SD are significant ( $p < 0.05$ ), meaning that from an experienced English speaker’s perspective, the ASR loss-aided codec is the best at preserving phonetic and linguistic content, followed by TTR, then S2 and SD. This result is also validated by Table 1, where ASR and TTR exhibits consistently lower WERs on both ASR models.

Meanwhile, the MUSHRA-style similarity scores from the left side of Fig. 3 show that ASR, TTR, and SD show comparable performances, while the differences between this group of three and S2 are quite significant. Clearly, both LM losses and SD objective boost overall quality aspects. In summary, our LM losses do not improve overall quality compared to SD, but they significantly outperform S2 (i.e. sans LM losses) in both tests and SD in the semantic test.

Perhaps most significantly, LM losses seem to widen the breadth of the semantic-acoustic tradeoff [9] than allowed by  $\mathcal{L}_{HuBERT}$ . HuBERT-feature-matching losses like  $\mathcal{L}_{HuBERT}$  are known to well-preserve semantic content [8]. Therefore, the second-stage tokens, trained solely using this signal, are an approximate “upper bound” as to how much semantic content can be preserved using  $\mathcal{L}_{HuBERT}$ , especially since the acoustic losses seem to encourage semantic deviations [9, 14, 15]. However, models trained using LM losses clearly surpass the second-stage codecs in both WER and MOS, suggesting that more semantic context has been learned during the third-stage and was stored in the resulting tokens.

Another reason for the success of LM losses may be that they are end-to-end losses, in contrast to common SD objectives that cannot influence the decoders in any way. Instead of only disentangling away redundant acoustic information in the semantic codes [9], our decoder is tamed directly by the LM losses as well, so that it may use its generative capability in more linguistically-plausible ways.

## 5. CONCLUSION

This work proposed two LM losses and showed that they alleviate PHs in a very-low-bit setting that already includes semantic quality considerations. Both of our proposed ASR and TTR losses seemed to expand the breadth of semantic-acoustic tradeoff than allowed by the SD and acoustic objectives, by boosting the semantic quality of the output while preserving the overall perceived output quality. As the LM losses are end-to-end, they can be flexibly applied to any DNN-based codec that outputs speech. The proposed LM losses may prove to be useful for training very-low-bitrate speech codecs when strong semantic compliance is desired.

## 6. REFERENCES

- [1] NATO, “The 600 bit/s, 1200 bit/s and 2400 bit/s NATO interoperable narrow band voice coder,” *STANAG 4591 C3*, 2008.
- [2] D. Rowe, “Codec 2 [online],” 2011, [https://www.rowetel.com/?page\\_id=452](https://www.rowetel.com/?page_id=452).
- [3] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “SoundStream: An end-to-end neural audio codec,” *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 30, pp. 495–507, Jan 2022.
- [4] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Trans. on Machine Learning Research*, 2023.
- [5] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2023.
- [6] L. Della Libera, F. Paissan, C. Subakan, and M. Ravanelli, “FocalCodec: Low-bitrate speech coding via focal modulation networks,” in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2025.
- [7] H. Liu et al., “SemantiCodec: An ultra low bitrate semantic audio codec for general sound,” *IEEE J. of Selected Topics in Signal Proc.*, vol. 18, no. 8, pp. 1448–1461, Dec. 2024.
- [8] Z. Huang, C. Meng, and T. Ko, “RepCodec: A speech representation codec for speech tokenization,” in *Proc. of the 62nd Annu. Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- [9] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Speech-Tokenizer: Unified speech tokenizer for speech large language models,” in *Proc. of the Int’l Conf. on Learning Representations (ICLR)*, 2024.
- [10] W.-N. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 29, pp. 3451–3460, 2021.
- [11] S. Chen et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. of Selected Topics in Signal Proc.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [12] J. D. Parker et al., “Scaling transformers for low-bitrate high-quality speech coding,” in *Proc. of the Int’l Conf. on Learning Representations (ICLR)*, 2025.
- [13] Z. Ye et al., “Codec does matter: Exploring the semantic shortcoming of codec for audio language model,” in *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, 2025.
- [14] Y. Gong et al., “XY-Tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs,” *arXiv:2506.23325*, 2025.
- [15] A. Défossez et al., “Moshi: a speech-text foundation model for real-time dialogue,” *arXiv:2410.00037*, 2024.
- [16] A. Siahkoobi, M. Chinen, T. Denton, W. Kleijn, and J. Skoglund, “Ultra-low-bitrate speech coding with pretrained transformers,” in *Proc. Interspeech*, 2022.
- [17] A. Polyak et al., “Speech resynthesis from discrete disentangled self-supervised representations,” in *Proc. Interspeech*, 2021.
- [18] H.-S. Choi et al., “Neural analysis and synthesis: Reconstructing speech from self-supervised representations,” in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2021.
- [19] A. Radford et al., “Robust speech recognition via large-scale weak supervision,” in *Proc. of the Int’l Conf. on Learning Representations (ICLR)*, 2023.
- [20] T.-A. Hsieh, H. Choi, and M. Kim, “Multimodal representation loss between timed text and audio for regularized speech separation,” in *Proc. Interspeech*, 2024.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.
- [22] D. Yang et al., “UniAudio 1.5: Large language model-driven audio codec is a few-shot audio task learner,” in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2024.
- [23] K.-H. Hung et al., “Linguistic knowledge transfer learning for speech enhancement,” *arXiv:2503.07078*, 2025.
- [24] N. Houlsby et al., “Parameter-efficient transfer learning for NLP,” in *Proc. of the Int’l Conf. on Machine Learning (ICML)*, 2019.
- [25] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “On the information rate of speech communication,” in *Proc. of the IEEE Int’l Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2017.
- [26] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2020.
- [27] I. Jang, H. Yang, W. Lim, S. Beack, and M. Kim, “Personalized neural speech codec,” in *Proc. of the IEEE Int’l Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2024.
- [28] S. A. Zahorian and H. Hu, “A spectral/temporal method for robust fundamental frequency tracking,” *The J. of the Acoustical Soc. of America*, vol. 123, no. 6, pp. 4559–4571, 2008.
- [29] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” in *Proc. Interspeech*, 2017.
- [30] ITU-R Recommendation BS 1534-3, “Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA),” 2015.
- [31] M. Schoeffler et al., “webMUSHRA — a comprehensive framework for web-based listening tests,” *J. of Open Research Software*, vol. 6, no. 1, pp. 8, 2018.
- [32] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, “WARP-Q: Quality prediction for generative neural speech codecs,” in *Proc. of the IEEE Int’l Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2021.
- [33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. of the IEEE Int’l Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2001.
- [34] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2020.