MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Factorized RVQ-GAN For Disentangled Speech Tokenization

Khurana, Sameer; Klement, Dominik; Laurent, Antoine; Bobos, Dominik; Novosad, Juraj; Gazdik, Peter; Zhang, Ellen; Huang, Zilli; Hussein, Amir; Marxer, Ricard; Masuyama, Yoshiki; Aihara, Ryo; Hori, Chiori; Germain, François G; Wichern, Gordon; Le Roux, Jonathan

TR2025-123 August 20, 2025

Abstract

We propose Hierarchical Audio Codec (HAC), a unified neural speech codec that factorizes its bottleneck into three linguistic levels—acoustic, phonetic, and lexical—within a single model. HAC leverages two knowledge distillation objectives: one from a pre-trained speech encoder (HuBERT) for phoneme- level structure, and another from a text-based encoder (LaBSE) for lexical cues. Experiments on English and multilingual data show that HAC's factorized bottleneck yields disentangled token sets: one aligns with phonemes, while another captures word-level semantics. Quantitative evaluations confirm that HAC tokens preserve naturalness and provide interpretable linguistic information, outperforming single-level baselines in both disentanglement and reconstruction quality. These findings underscore HAC's potential as a unified discrete speech representation, bridging acoustic detail and lexical meaning for downstream speech generation and understanding tasks.

Interspeech 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Factorized RVQ-GAN For Disentangled Speech Tokenization

Sameer Khurana[1⋆], *Dominik Klement*[2⋆], *Antoine Laurent*[3⋆], *Dominik Bobos*[♠], *Juraj Novosad*[♠], *Peter Gazdik*[♠], *Ellen Zhang*[♢], *Zili Huang*[♡], *Amir Hussein*[♡], *Ricard Marxer*[♣], *Yoshiki Masuyama*[1], *Ryo Aihara*[1], *Chiori Hori*[1], *François G. Germain*[1], *Gordon Wichern*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Speech@FIT, Brno University of Technology, Czech Republic [3]LIUM University, France

sameerkhurana10@gmail.com, xkleme15@vutbr.cz, leroux@merl.com

Abstract

We propose Hierarchical Audio Codec (HAC), a unified neural speech codec that factorizes its bottleneck into three linguistic levels—acoustic, phonetic, and lexical—within a single model. HAC leverages two knowledge distillation objectives: one from a pre-trained speech encoder (HuBERT) for phoneme-level structure, and another from a text-based encoder (LaBSE) for lexical cues. Experiments on English and multilingual data show that HAC's factorized bottleneck yields disentangled token sets: one aligns with phonemes, while another captures word-level semantics. Quantitative evaluations confirm that HAC tokens preserve naturalness and provide interpretable linguistic information, outperforming single-level baselines in both disentanglement and reconstruction quality. These findings underscore HAC's potential as a unified discrete speech representation, bridging acoustic detail and lexical meaning for downstream speech generation and understanding tasks.

Index Terms: RVQ, GAN, Audio Codec, Speech Tokenization

1. Introduction

Neural speech codecs (NSCs) are a family of neural network architectures that convert speech signals into discrete token representations [1–4]. These discrete tokens can then be leveraged in various downstream tasks, ranging from spoken language modeling [5] and speech-to-speech translation [6] to text-to-speech synthesis [7] and speech understanding in large language models. Broadly, NSCs can be classified into phonetic (P-NSC) and acoustic (A-NSC) approaches.

A P-NSC follows a two-stage pipeline: (1) a pre-trained transformer encoder (e.g., HuBERT [8]), trained via self-supervised learning (SSL), outputs contextual acoustic frame embeddings; and (2) those embeddings, selected from the layer that performs best on a phoneme recognition task, are quantized using k-means vector quantization (VQ) to produce discrete tokens [9]. Because these tokens align closely with underlying phoneme labels, P-NSCs excel at tasks requiring high-level linguistic structure [6, 10]. However, speech generated from P-NSCs tends to sound robotic and lacks speaker diversity [5].

By contrast, A-NSCs aim for high-fidelity speech reconstruction through a low-bitrate compression model, often based on the residual VQ-generative adversarial network (RVQ-GAN) framework. An encoder maps input speech to acoustic frame embeddings which are then quantized into discrete token sequences across multiple VQ layers. A decoder reconstructs the speech from these tokens, preserving detailed acoustic nuances, resulting in more natural-sounding speech with broad speaker variation. Nonetheless, because A-NSCs focus on fine-grained acoustic details, they often lack coherent linguistic and grammatical structure. Prominent examples of A-NSCs include SoundStream [1], Encodec [2], and Descript Audio Codec [3].

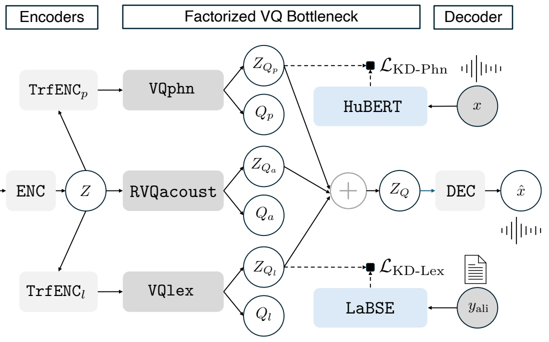

Figure 1: *Diagram of our proposed Hierarchical Audio Codec (HAC). HAC encodes the input speech signal x into a multi-level set of disentangled discrete tokens (lexical Q_l, phonetic Q_p, and acoustic Q_a) capturing distinct aspects of the audio.*

Recent work, such as SpeechTokenizer [11], addresses the limitations of standalone P-NSCs and A-NSCs by combining phonetic and acoustic tokens within a single framework. However, these two-level solutions omit an important lexical representation, which captures word-level or subword-level semantic and syntactic information. In contrast, our proposed Hierarchical Audio Codec (HAC) introduces this third level of abstraction, lexical, alongside phonetic and acoustic tokens, enabling it to jointly model higher-level linguistic structure, mid-level phonetic details, and fine-grained acoustic nuances. Learning this extra lexical layer is particularly valuable for downstream tasks such as spoken language modeling, speech-to-speech translation, and voice-enabled question answering, where a richer, word-oriented representation can substantially improve semantic coherence and contextual accuracy. By disentangling the token space across these three levels, HAC combines the strengths of existing two-level models with a deeper linguistic understanding, all within a unified architecture that eliminates the need for external token merging.

2. Hierarchical Audio Codec (HAC)

HAC, illustrated in Fig. 1, consists of a down-sampling CNN encoder (`ENC`), two transformer encoders (`TrfENC`$_{p,l}$), a factorized bottleneck consisting of three VQ modules (`VQphn`, `RVQacoust`, and `VQlex`), and an up-sampling CNN decoder (`DEC`). HAC is trained using tuples (x, y_{ali}), where $x \in \mathbb{R}^T$ is a speech utterance and y_{ali} is its force-aligned text transcript.

[⋆]Major Contributors. [♠]Phonexia, Czech Republic. [♢]Massachusetts Institute of Technology, USA. [♡]CLSP, Johns Hopkins University, USA. [♣]Univ. de Toulon, Aix Marseille Univ., CNRS, LIS, Toulon, France.

HAC maps x to $\hat{x}$, a reconstruction of x through the following steps:

$$\begin{aligned}
Z &= \mathtt{ENC}(x), \\
Z_{Q_a}, Q_a &= \mathtt{RVQacoust}(Z), \\
Z_{Q_p}, Q_p &= \mathtt{VQphn}(\mathtt{TrfENC}_p(Z)), \\
Z_{Q_l}, Q_l &= \mathtt{VQlex}(\mathtt{TrfENC}_l(Z)), \\
Z_Q &= Z_{Q_p} + Z_{Q_a} + Z_{Q_l}, \\
\hat{x} &= \mathtt{DEC}(Z_Q),
\end{aligned}$$

where $Z,\ Z_{Q_a},\ Z_{Q_p},\ Z_{Q_l} \in \mathbb{R}^{F\times D}$, and Z_{Q_a}, Z_{Q_p}, and Z_{Q_l} are the codebook entries corresponding to token sets $Q_a \in \{1,\ldots,A\}^{F\times N}$, $Q_p \in \{1,\ldots,P\}^F$, and $Q_l \in \{1,\ldots,L\}^F$, respectively. Here, F is the number of acoustic frames, and D is the frame embedding dimension. To maximize codebook utilization, the VQ layers follow the low-dimensional code lookup process described in [3]. `RVQacoust` module performs residual vector quantization (RVQ) and consists of N VQ layers, each with a codebook size of A. `VQphn` and `VQlex` consist of a single VQ layer with codebook sizes P and L, respectively.

HAC is trained in an adversarial framework, where the HAC generator is paired with a discriminator as described in [3]. The overall training objective includes: 1) a frequency-domain reconstruction loss to ensure faithful spectral recovery, 2) an adversarial loss to encourage natural-sounding outputs, and 3) codebook learning losses to update the codebook entries. These objectives are described in detail in [3]. To ensure that each token set encodes the intended type of information, we introduce knowledge distillation (KD) losses on the phonetic and lexical bottlenecks: 1) $\mathcal{L}_{\text{KD-Phn}}$ encourages Q_p to represent phoneme-level features, and 2) $\mathcal{L}_{\text{KD-Lex}}$ encourages Q_l to represent word-level (lexical) information. By providing the decoder with phonetic (Z_{Q_p}) and high-level lexical (Z_{Q_l}) information, the acoustic bottleneck `RVQacoust` is free to focus on fine-grained acoustic details, critical for high-fidelity signal reconstruction.

Following [11], we compute the KD losses as follows:

$$\begin{aligned}
&\tilde{Z}_{Q_p} = Z_{Q_p}A_p, \quad \tilde{Z}_{Q_l} = Z_{Q_l}A_l, \\
&\mathcal{L}_{\text{KD-Phn}} = -\frac{1}{D'}\sum_{d=1}^{D'} \log(\sigma(\cos\_\text{sim}(\tilde{Z}_{Q_p}[:,d], Z_{\text{hubert}}[:,d]))), \\
&\mathcal{L}_{\text{KD-Lex}} = -\frac{1}{D''}\sum_{d=1}^{D''} \log(\sigma(\cos\_\text{sim}(\tilde{Z}_{Q_l}[:,d], Z_{\text{labse}}[:,d]))), \\
&Z_{\text{hubert}} = \mathtt{Avg}(\mathtt{HuBERT}(x)), \quad Z_{\text{labse}} = \mathtt{Avg}(\mathtt{LaBSE}(y_{\text{ali}})),
\end{aligned}$$

where D' and D'' are the embedding dimensionalities of HuBERT and LaBSE, respectively, $\cos\_\text{sim}(\cdot)$ is the cosine similarity, $\sigma(\cdot)$ is the sigmoid function, A_p and A_l are projection matrices for dimension matching, and $\mathtt{Avg}(\cdot)$ averages representations over all layers of HuBERT or LaBSE [12].

While we explore other settings in Section 3, our best-performing HAC model has the following hyperparameters: 1) Training input: each speech recording x is 3.8 seconds long and has a 16 kHz sampling rate; 2) Downsampling factor: `ENC` reduces the input time resolution by a factor of 320; 3) Frame embedding dimensionality: $D = 1024$; 4) Phonetic and lexical codebooks: both `VQphn` and `VQlex` use a single VQ layer with codebook size 16,384 (14-bit), and codebook entries are 128-dimensional; and 5) Acoustic codebook: `RVQacoust` has $N = 7$ VQ layers, each with codebook size $A = 1024$, and each codebook entry is 8-dimensional. Each acoustic frame is represented by 9 tokens (the 7 acoustic tokens plus 1 phonetic token and 1 lexical token). The `ENC` and `DEC` follow the same CNN architecture described in [3]. $\mathtt{TrfENC}_{p,l}$ has 4 layers, 8 attention heads, model embedding dimensionality of 768, and feed-forward dimensionality 3072. Layer-normalization is applied to each layer's input, and the encoder has learnable convolutional positional embeddings [13].

HAC is optimized on eight A40 GPUs with a total batch size of 60 seconds. We use the AdamW optimizer with a learning rate 1e-4, $\beta_1 = 0.8$, and $\beta_2 = 0.9$. We train the model for 400K iterations and decay the learning rate at every step with $\gamma = 0.999996$.

3. Experiments

We train English-language models on the LibriSpeech dataset [14], consisting of 960 hours of transcribed English speech, and multilingual models on the VoxPopuli dataset [15], which includes 1.7K hours of transcribed speech across 16 languages. For evaluation, we use forced-aligned test sets from LibriSpeech and Multilingual LibriSpeech (MLS) [16]. The MLS corpus contains eight languages: English (EN), French (FR), German (DE), Italian (IT), Polish (PL), Portuguese (PT), Spanish (ES), and Dutch (NL). We obtain forced alignments using the Montreal Forced Aligner [17]. We train this aligner for languages other than English on a subset of each respective MLS training set.

Below, we summarize the models trained in this work. Each model name encodes its key features, including codebook size, the teacher model for KD, and whether a transformer encoder is employed.

ST-10-HuB-en (Baseline): This is our English (en) baseline, SpeechTokenizer (ST) [11]. It is based on an RVQ-GAN framework trained with a generative objective and the phonetic KD loss ($\mathcal{L}_{\text{KD-Phn}}$) described in Section 2. ST's RVQ bottleneck consists of 9 VQ layers. The model distills from HuBERT-Base by matching the first VQ layer's quantized frame embeddings to the averaged HuBERT-Base embeddings. Each VQ layer has a 10-bit codebook of dimensionality 1024. ST extends Encodec [2] (an earlier RVQ-GAN for audio compression) by adding the Phoneme level KD loss to Encodec's original generative and codebook learning objectives. Like Encodec, ST updates its codebooks via exponential moving average (EMA) and periodically re-initializes them to maximize utilization. To have a fair comparison with other models, we train the baseline using our codebase, instead of using the publicly available checkpoint.

DAC-10-HuB-en: This model adds the phonetic KD loss to DAC [3], an improved version of Encodec that uses low-dimensional code lookups for improved codebook utilization and reconstruction. Like the baseline, it has 9 10-bit VQ layers but has a codebook dimensionality of 8, much lower than the baseline. Since a very low-dimensional codebook is less ideal for KD, we increase the first VQ layer's dimensionality to 128 while keeping the remaining layers at 8 dimensions. Furthermore, unlike ST-10-HuB-en, where the phonetic VQ layer (the KD student) is part of the RVQ module, in DAC-10-HuB-en, we factor out this layer, leading to slightly improved token quality and reconstruction performance. **DAC-14-HuB-en:** This variant extends DAC-10-HuB-en by increasing the phonetic VQ layer's codebook size from 10-bit to 14-bit, enabling a larger vocabulary of phoneme-level tokens. **DAC-14-HuB-T-en:** Building on DAC-14-HuB-en, this model adds a transformer encoder ($\mathtt{TrfENC}_p$ described in Section 2) before the factored-out pho-

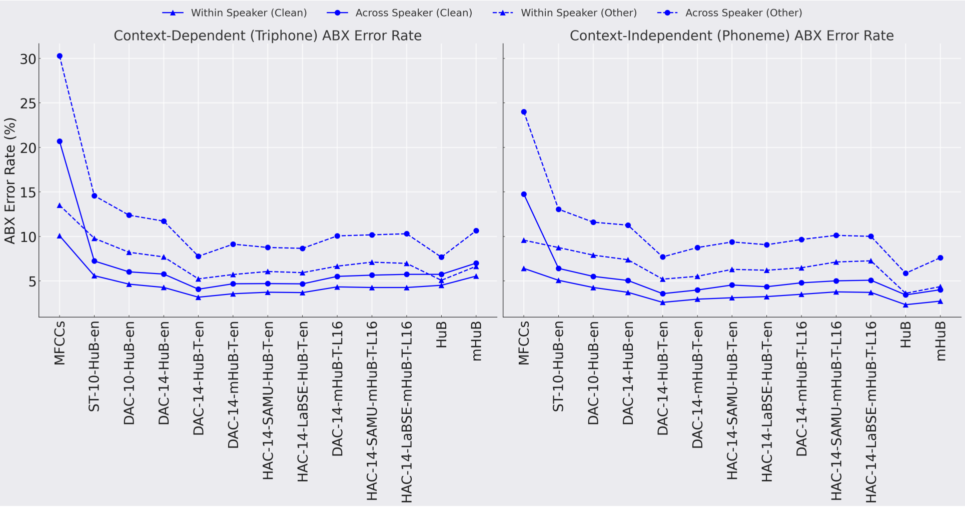

Figure 2: *ABX error rate across different models and scenarios.*

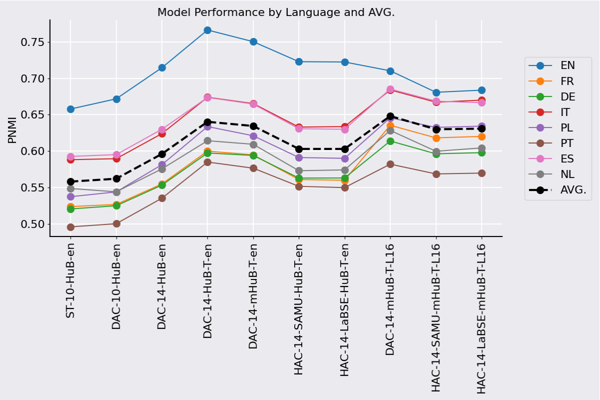

Figure 3: *Phoneme Normalized Mutual Information (PNMI) across different models and languages.*

netic VQ layer, allowing the VQ layer to capture richer context before quantization. **DAC-14-mHuB-T-en:** This version uses multilingual HuBERT (mHuBERT) [18] as the teacher for phonetic KD loss. It follows the same architecture as DAC-14-HuB-T-en but replaces HuBERT-Base with a multilingual variant. **DAC-14-mHuB-T-L16:** A multilingual DAC model that is trained on both VoxPopuli and LibriSpeech. It retains the 14-bit phonetic VQ layer and uses mHuBERT, matching DAC-14-mHuB-T-en's setup, but extends training across 16 languages.

HAC-14-SAMU-HuB-T-en: This is the Hierarchical Audio Codec (HAC) variant with 9 VQ layers, of which two are factored out of the RVQ module. Both factored-out layers have 14-bit codebooks of dimensionality 128; the remaining layers have 10-bit codebooks of dimension 8. For the lexical-level KD loss ($\mathcal{L}_{\text{KD-Lex}}$), we use SAMU-XLS-R (SAMU) [19] as the teacher. SAMU is a language-agnostic semantic speech encoder obtained by distilling LaBSE into a speech model. SAMU removes the need for forced-aligned transcripts when computing the lexical KD loss. Meanwhile, phonetic KD loss still uses HuBERT-Base. **HAC-14-LaBSE-HuB-T-en:** Identical to HAC-14-SAMU-HuB-T-en, except the lexical KD loss is computed directly from LaBSE embeddings that correspond to force-aligned transcripts rather than from SAMU. **HAC-14-SAMU-mHuB-T-L16:** A multilingual HAC model, extending HAC-14-SAMU-HuB-T-en to 16 languages. It uses mHuBERT for phoneme-level KD and SAMU for lexical-level KD. **HAC-14-LaBSE-mHuB-T-L16:** The multilingual version of HAC-14-LaBSE-HuB-T-en. It uses LaBSE for lexical-level KD and mHuBERT for phoneme-level KD.

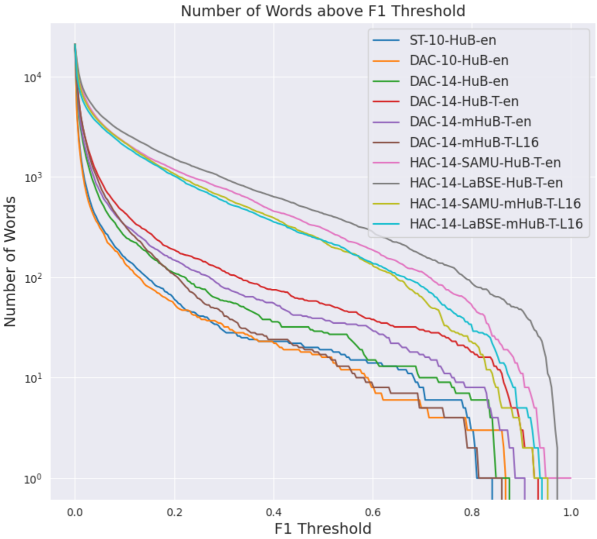

Figure 4: *F1 Scores of VQ tokens when treated as word detectors on the LibriSpeech corpus.*

Figures 2 and 3 evaluate how well the phonetic VQ layer (attached to $\mathcal{L}_{\text{KD-Phn}}$) captures phoneme-level information.

Figure 2 compares various models on the ABX discrimination task [20], where three samples (A, B, and X) are presented, A and B differ (e.g., distinct phonemes), and X matches either A or B. The ABX error rate is the proportion of incorrect classifications. We consider two setups: (1) Context-Independent (CI), where A and B are different phonemes in isolation, and (2) Context-Dependent (CD), where A and B are triphones to account for coarticulation. The ABX triples are extracted from the LibriSpeech test sets (clean & other), available in abxLS, an evaluation task in the zero-resource-speech benchmark [21]. All models outperform the reference MFCC-based error rate, with DAC and HAC outperforming the baseline ST-10-HuB-en. Among the DAC variants, DAC-14-HuB-en performs better than DAC-10-HuB-en, likely due to its higher bitrate. Introducing a transformer encoder before the student VQ layer as

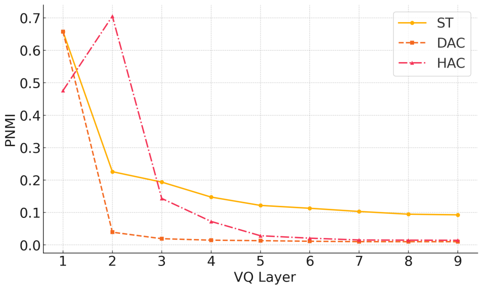

Figure 5: *Layer-wise Phoneme Normalized Mutual Information for different models.*

in DAC-14-HuB-T-en further reduces ABX error. HAC models show slightly higher ABX errors than DAC possibly because juggling both phoneme-level ($\mathcal{L}_{\text{KD-Phn}}$) and lexical-level ($\mathcal{L}_{\text{KD-Lex}}$) losses introduces additional training complexity. We also report ABX error rates using HuBERT (HuB) and mHuBERT (mHuB) embeddings for reference.

Figure 3 shows results for Phoneme Normalized Mutual Information (PNMI) [8], which measures the mutual information between tokens and true phoneme labels, normalized by phoneme label entropy. Values range from 0 (no correspondence) to 1 (perfect alignment). We evaluate PNMI across eight MLS languages. The multilingual DAC-14-mHuB-T-L16 model achieves the highest average PNMI, followed by the HAC multilingual variants; ST-10-HuB-en ranks lowest. Consistent with the ABX results, adding a transformer encoder before the phonetic VQ layer (e.g., DAC-14-HuB-T-en vs. DAC-14-HuB-en) significantly boosts PNMI. English performance is strongest overall, reflecting its greater share of training data compared to other languages.

Figure 4 explores how well the tokens align with word labels. We extract tokens from the `VQlex` layer for HAC models. For DAC and ST models, we use the phonetic VQ layer. Following [22], we compute the F1 score for each (word, token) pair to see if the discovered tokens capture lexical items. Figure 4 plots how many tokens achieve an F1 score above various thresholds. HAC models exhibit a clear advantage, producing substantially more word detectors than their DAC or ST counterparts. HAC-14-LaBSE-HuB-T-en attains the highest number of strong word detectors, leveraging LaBSE-based text embeddings for lexical-level KD. HAC-14-SAMU-HuB-T-en ranks a close second; notably, it does not require text transcripts to compute the KD loss because it uses SAMU, a language-agnostic semantic speech encoder. Multilingual HAC models perform slightly worse, as expected given that evaluation is on English words only.

Figures 5 and 6 provide evidence of how different VQ layers in HAC encode distinct linguistic abstractions.

Figure 5 compares layer-wise PNMI for HAC (HAC-10-SAMU-HuB-T-en), DAC (DAC-10-HuB-T-en), and ST (ST-10-HuB-T-en). VQ layers 1 and 2 in HAC refer to the lexical (attached to $\mathcal{L}_{\text{KD-Lex}}$) and phonetic (attached to $\mathcal{L}_{\text{KD-Phn}}$) VQ layers, respectively. In DAC and ST, VQ layer 1 refers to the phonetic layer. All models show partial disentanglement: phonetic layers yield higher PNMI, whereas other layers do not align with phoneme labels. Notably, HAC's lexical VQ layer (Layer 1) also exhibits some phoneme alignment but remains less phonemically oriented than the phonetic layer. Overall, DAC and HAC achieve stronger disentanglement than ST.

Figure 6 shows correspondence between tokens from different layers of HAC and the word labels. We observe that the layer `VQlex` has significantly more codebook entries that act as word detectors compared to other layers. Some codebook entries of the `VQphn` layer act as word detectors, and virtually no codebook entries from the `RVQacoust` module of HAC act as word detectors.

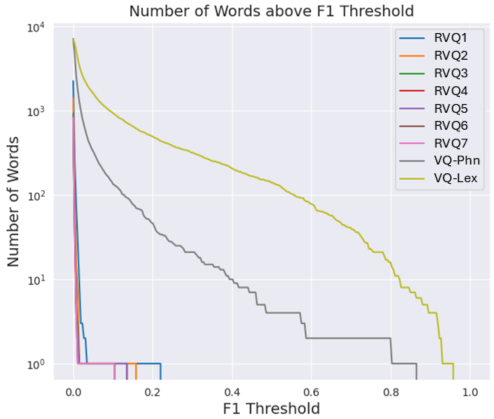

Figure 6: *F1 Scores of VQ tokens from different VQ layers of HAC when treated as word detectors.*

Table 1: *Reconstruction metrics across different models.*

Model	Mel-D ↓	STFT-D ↓	SI-SDR [dB] ↑	ViSQOL ↑
ST	0.64	1.42	6.24	3.89
HAC	0.58	1.37	7.42	4.34
DAC	0.55	1.34	7.82	4.50
DAC (Orig.)	0.51	1.29	8.01	4.55

Finally, Table 1 compares the reconstruction quality of different models on LibriSpeech clean test set using standard reconstruction metrics (see Section 4.4 of [3] for details). Both DAC and HAC outperform the baseline ST and achieve reconstruction quality on par with DAC (Orig.) [3], an RVQ-GAN trained solely on generative losses. For brevity, DAC refers to DAC-10-HuB-T-en, ST to ST-10-HuB-en, and HAC to HAC-10-SAMU-HuB-T-en throughout the table.

4. Conclusions

This paper introduced Hierarchical Audio Codec (HAC), a factorized RVQ-GAN framework that unifies acoustic, phonetic, and lexical token sets within a single model. Through dedicated knowledge distillation losses from speech-focused (HuBERT) and text-based (LaBSE) encoders, HAC learns complementary token groups at different levels of linguistic abstraction. Specifically, the acoustic tokens capture the fine-grained spectral details needed for natural-sounding speech reconstruction, while the phonetic tokens align closely with underlying phoneme sequences, and the lexical tokens detect word-level distinctions. Experiments demonstrate that this disentangled multi-level representation not only preserves high-fidelity audio quality but also offers strong linguistic and semantic interpretability across various languages. Overall, our results highlight the potential of factorized tokenization to bridge the gap between high-fidelity audio compression and linguistically rich speech representations.

Acknowledgements—Dominik Klement was supported by Czech Ministry of Interior project No. VK01020132 ”112”. The work reported here was started at JSALT 2024, and supported by JHU with gifts from Amazon, Facebook, Google, and Microsoft.

5. References

[1] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 495–507, 2021.

[2] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *TMLR*, 2023.

[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proc. NeurIPS*, 2023.

[4] J. Shi, J. Tian, Y. Wu, J.-W. Jung, J. Q. Yip, Y. Masuyama, W. Chen, Y. Wu, Y. Tang, M. Baali *et al.*, "ESPnet-Codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech," in *Proc. SLT*, 2024, pp. 562–569.

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "AudioLM: a language modeling approach to audio generation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[6] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang *et al.*, "Direct speech-to-speech translation with discrete units," in *Proc. ACL*, May 2022, pp. 3327–3339.

[7] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 705–718, 2025.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451 – 3460, 2021.

[9] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe *et al.*, "Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study," in *Proc. ICASSP*, 2024, pp. 11 481–11 485.

[10] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *TACL*, vol. 9, pp. 1336–1354, 2021.

[11] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "SpeechTokenizer: Unified speech tokenizer for speech large language models," in *Proc. ICLR*, 2024.

[12] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proc. ACL*, May 2022, pp. 878–891.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[15] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL*, Aug. 2021, pp. 993–1003.

[16] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech*, 2020, pp. 2757–2761.

[17] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[18] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, "mHuBERT-147: A compact multilingual HuBERT model," *arXiv preprint arXiv:2406.06371*, 2024.

[19] S. Khurana, A. Laurent, and J. Glass, "SAMU-XLSR: Semantically-aligned multimodal utterance-level cross-lingual speech representation," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1493–1504, 2022.

[20] T. Schatz, "Abx-discriminability measures and applications," Ph.D. dissertation, Université Paris 6 (UPMC), Sep. 2016.

[21] E. Dunbar, M. Bernard, N. Hamilakis, T. A. Nguyen, M. De Seyssel, P. Rozé, M. Rivière, E. Kharitonov, and E. Dupoux, "The zero resource speech challenge 2021: Spoken language modelling," *arXiv preprint arXiv:2104.14700*, 2021.

[22] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *Proc. ICLR*, 2020.