TECHNICAL APPENDIX DM-CODEC: DISTILLING MULTIMODAL REPRESENTATIONS FOR SPEECH TOKENIZATION

Anonymous authors
Paper under double-blind review

A RESOURCES

We provide the code for training DM-Codec, trained model checkpoint, and Dockerfile for a reproducible code environment. The links are shared anonymously for the double-blind review process. We will publicly share all resources after the completion of the review timeline.

- **DM-Codec codebase**: Codebase
- **Trained model checkpoints for inference**: Model-checkpoints
- **Dockerfile for reproducible environment**: Docker

B MODEL COMPONENTS

Encoder Decoder. The encoder-decoder architecture in DM-Codec is based on SEANet (Tagliasacchi et al., 2020), leveraging the successful design employed in recent speech tokenization models (Zhang et al., 2024a; Défossez et al., 2022; Zeghidour et al., 2021). The architecture is designed to efficiently process and reconstruct speech signals while maintaining high fidelity. The Encoder $\mathbf{E}$ consists of a 1D convolution layer with C channels and a kernel size of 7, followed by B residual convolutional blocks. Each block contains a strided convolutional downsampling layer with kernel size K (where $K = 2S$, and S represents the stride), paired with a residual unit. The residual unit comprises two convolutional layers with a kernel size of 3 and a skip connection, while the number of channels is doubled at each downsampling stage. This is followed by a two-layer BiLSTM and a final 1D convolutional layer with D output channels and a kernel size of 7. The Decoder $\mathbf{D}$ mirrors the encoder's structure but replaces BiLSTM with LSTM, strided convolutions with transposed convolutions, and employs reversed strides for up-sampling. The final audio output is reconstructed from $\mathbf{D}$. For the experiments, we use the following configuration: C = 32, B = 4, and S = (2, 4, 5, 8).

Residual Vector Quantizers. The Residual Vector Quantizer (RVQ) plays a central role in our tokenization process, quantizing the encoder's outputs. Our implementation is inspired by the training procedures described in Encodec (Défossez et al., 2022) and SpeechTokenizer (Zhang et al., 2024a). The RVQ projects input vectors to the most similar entry in a codebook, and the residual is calculated and processed in subsequent quantization steps, each utilizing a different codebook. The codebook entries are updated using an *exponential moving average* (EMA) with a *decay rate* of 0.99 for the matched item, while unmatched entries are replaced by candidates from the current batch. To ensure proper gradient flow during training, we employ a *straight-through estimator*. A *commitment loss* is also computed and added to the total training loss to promote stability. In our experiments, we utilize a codebook size of 1024 and 8 quantization levels.

Discriminators. We incorporate a trio of discriminators to enhance the quality and realism of the generated speech: the Multi-Scale Discriminator (MSD), the Multi-Period Discriminator (MPD), and the Multi-Scale Short-Time Fourier Transform (MS-STFT) discriminator. The MS-STFT discriminator follows the implementation outlined in (Défossez et al., 2022), operating on the real and imaginary components of multi-scale complex-valued STFTs. It begins with a 2D convolutional layer, followed by 2D convolutions with increasing dilation rates in the time dimension (1, 2, and 4) and a stride of 2 across the frequency axis in each sub-network. A final 2D convolution with a kernel size of 3 × 3 and a stride of (1, 1) is applied to produce the prediction. The MSD and MPD discriminators follow the architectures introduced in (Kong et al., 2020), with adjustments to the channel numbers to align the parameter count more closely with the MS-STFT discriminator. This ensemble of discriminators works in concert to provide comprehensive feedback on various aspects of the generated speech, contributing to the overall quality and naturalness of the output.

C RELATED WORK

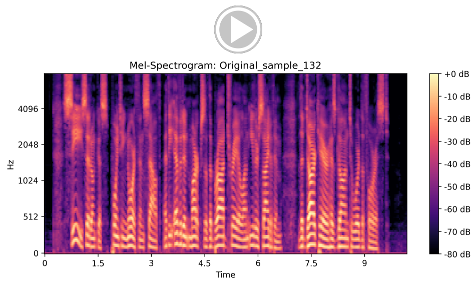
Tokenization Techniques in Speech. Tokenization in speech processing can be broadly categorized into two main approaches: (i) speech encoder-based and (ii) language-based. In the speech encoder-based tokenization approach, a pretrained speech encoder serves as a teacher model, providing semantically rich audio representations. These representations are then used to guide the training model, either through an alignment network (Messica & Adi, 2024) or by optimizing specific losses (Zhang et al., 2024a; Liu et al., 2024). Language-based tokenization approach involves processing audio through a speech encoder to obtain discrete representations or using the corresponding text to feed into a language model. The representations from the language model are then utilized either to learn a tokenizer for speech (Turetzky & Adi, 2024) or to reconstruct speech (Hassid et al., 2024; Zhang et al., 2024b; Wang et al., 2024). Besides, (Zhang et al., 2024b) proposed SpeechLM where two discrete tokenizers were introduced and learned in an unsupervised way and converted the speech and text in a shared discrete space.

Discrete Speech Representation. There are two well-known methods for discrete speech representation: semantic tokens and acoustic tokens. Semantic tokens are derived through self-supervised learning (SSL) techniques for speech (Baevski et al., 2019; Hsu et al., 2021; Chung et al., 2021) and capture abstract, high-level features that relate to general, symbolic aspects of speech, while omitting details related to speaker identity and acoustic characteristics. In contrast, acoustic tokens are obtained using neural audio codecs (Zeghidour et al., 2021; Défossez et al., 2022; Yang et al., 2023) and focus on delivering precise reconstructions of acoustic features. However, recent models (Turetzky & Adi, 2024; Liu et al., 2024; Shi et al., 2024) have shown that speech models based on self-supervised learning (SSL) are effective at extracting acoustic representations where LMs be employed to refine these models further, enhancing their ability to extract more nuanced semantic representations.

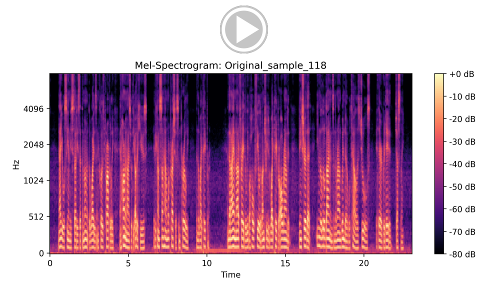
Textual Language Models in Speech. Research on speech models, including works by (Nguyen et al., 2023), (Borsos et al., 2023), and (Kharitonov et al., 2022), has focused on utilizing raw audio to extract prosodic features, identify speaker characteristics, and generate audio without depending on textual features or supervision from textual LMs. In contrast, many newer methods have started using audio encoders to transform audio signals into discrete tokens, which can be processed by textual LMs. *TWIST* method introduced by (Hassid et al., 2024) initializes the weights of the SpeechLM using a pre-trained text LM, showing that this combination significantly improves performance. Similarly, the *SELM* model developed by (Wang et al., 2024) leverages GPT (Radford, 2018; Radford et al., 2019) as its foundation due to its enhanced parallel processing capabilities and capacity. However, text-based LLMs such as GPT-3 (Brown, 2020) and Llama (Touvron et al., 2023) are essential for speech modeling. Once discrete audio representations are obtained, these large text models are trained to align with or enhance the original text embedding space, as explored in studies by (Zhang et al., 2023), (Fathullah et al., 2023), (Shu et al., 2023), and (Rubenstein et al., 2023). This trend of integrating textual LMs into speech modeling has become increasingly popular in recent research.

D RECONSTRUCTED SPEECH COMPARISON

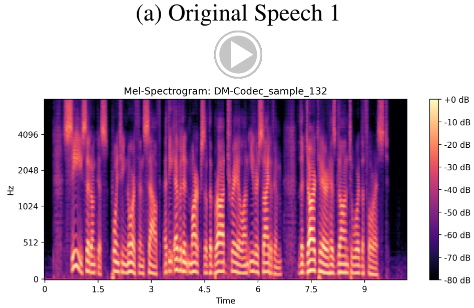
We plot the Mel-Spectrogram of the original speech and the reconstructed speech from DM-Codec and compare them with the reconstructed speech of EnCodec, SpeechTokenizer, and FACodec. Fine-grained differences may not be readily apparent in the Mel-Spectrogram visually; therefore, we encourage readers to click on the respective play button in Figure 1 for a hyperlink to the playable audio file.



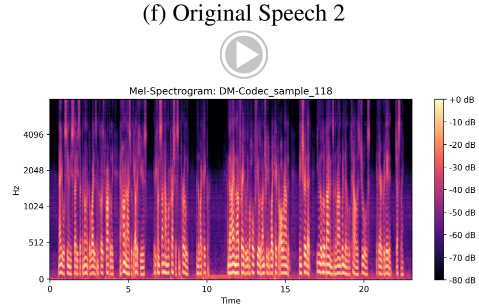
(a) Original Speech 1



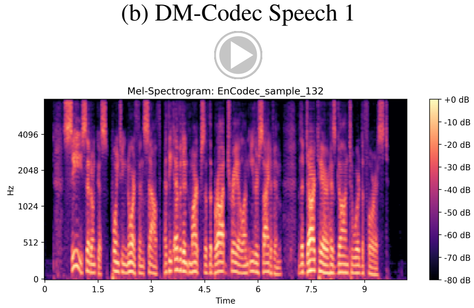
(f) Original Speech 2



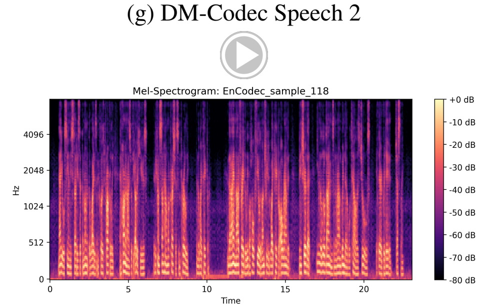
(b) DM-Codec Speech 1



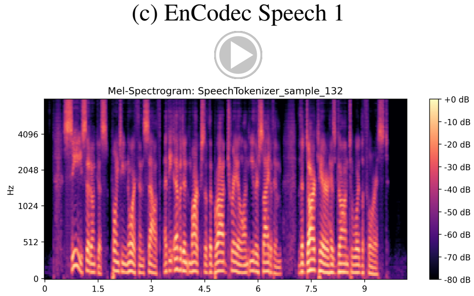
(g) DM-Codec Speech 2



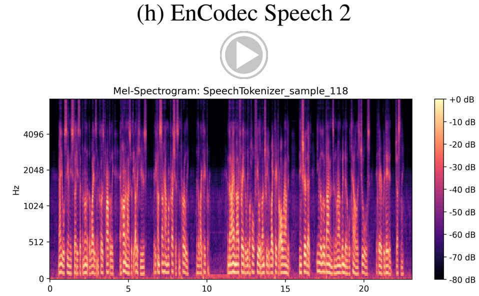
(c) EnCodec Speech 1



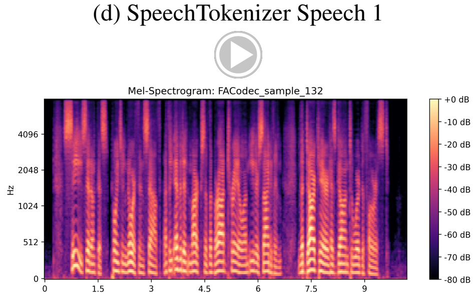
(h) EnCodec Speech 2



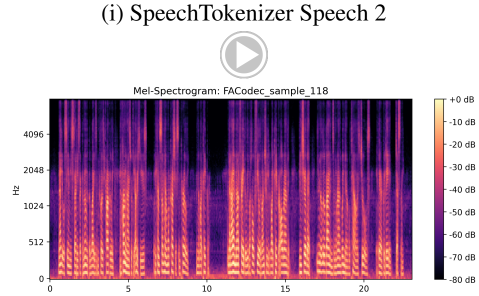
(d) SpeechTokenizer Speech 1



(i) SpeechTokenizer Speech 2



(e) FACodec Speech 1



(j) FACodec Speech 2

Figure 1: Reconstructed speech examples with clickable play buttons above each Mel-spectrogram.

Table 1: Significance Analysis and Comparison of DM-Codec (**D**), EnCodec (**E**), SpeechTokenizer (**S**), and FACodec (**F**). Results reveal DM-Codec consistently achieves significantly better scores. ✓ indicates significantly better, a ★ denotes significant dominance, and a ✗ means no significance in comparison. Avg and Std mean the average and standard deviation of each score.

WER					**WIL**					**ViSQOL**					**STOI**				
DM-Codec																			
Avg	**Std**	**E**	**S**	**F**	**Avg**	**Std**	**E**	**S**	**F**	**Avg**	**Std**	**E**	**S**	**F**	**Avg**	**Std**	**E**	**S**	**F**
0.053	0.113	✓	✓	✓	0.082	0.157	✓	✓	✓	3.258	0.184	★	✓	✓	0.937	0.019	✓	✓	✗
EnCodec																			
Avg	**Std**	**D**	**S**	**F**	**Avg**	**Std**	**D**	**S**	**F**	**Avg**	**Std**	**D**	**S**	**F**	**Avg**	**Std**	**D**	**S**	**F**
0.061	0.131	✗	✗	✗	0.090	0.158	✗	✗	✗	3.078	0.201	✗	✗	✗	0.920	0.017	✗	✗	✗
SpeechTokenizer																			
Avg	**Std**	**E**	**D**	**F**	**Avg**	**Std**	**E**	**D**	**F**	**Avg**	**Std**	**E**	**D**	**F**	**Avg**	**Std**	**E**	**D**	**F**
0.060	0.139	✓	✗	✗	0.089	0.166	✓	✗	✗	3.087	0.190	✓	✗	✗	0.923	0.021	✓	✗	✗
FACodec																			
Avg	**Std**	**E**	**S**	**D**	**Avg**	**Std**	**E**	**S**	**D**	**Avg**	**Std**	**E**	**S**	**D**	**Avg**	**Std**	**E**	**S**	**D**
0.057	0.123	✓	✓	✗	0.086	0.163	✓	✓	✗	3.129	0.250	✓	✓	✗	0.949	0.923	✓	✓	✓

E Limitations and Broader Impact

Limitations. In this work, we present the effectiveness of our proposed method, DM-Codec, based on the LibriSpeech dataset. Future research could investigate its performance across a variety of datasets and domains. Additionally, exploring the capabilities of DM-Codec in multilingual contexts would be valuable. Another limitation of our work is the absence of experiments with emerging LLMs. Currently, we focus solely on masked language models to derive representations. Further investigation into these decoder-based LLMs' impact on DM-Codec can be studied and addressed.

Broader Impact. The integration of language models in speech processing has traditionally focused on model-specific implementations or specific training objectives. In this work, we propose a novel approach by leveraging language models during the tokenization phase through our model, DM-Codec. By incorporating language-specific representations from the corresponding text, DM-Codec enhances the quality of discrete speech representations. This method bridges the gap between language and speech models, offering a more unified approach to multimodal representation learning. DM-Codec provides a robust framework for generating high-quality audio representations, with potential applications in various domains, including multilingual speech processing, low-resource languages, and other audio-related tasks. Our findings pave the way for more effective and contextually aware speech processing models, contributing to advancements in the broader field of speech and language technologies.

F Significance Analysis and Comparison

To compare the significance of speech reconstruction results between our proposed DM-Codec and baselines EnCodec (Défossez et al., 2022), SpeechTokenizer (Zhang et al., 2024a), and FACodec (Ju et al., 2024), we follow the approach of Dror et al. (2019) and measure the stochastic dominance at $\alpha = 0.05$. We compute the inverse cumulative distribution functions (CDFs) for individual WER, WIL, ViSQOL, and STOI scores obtained for 300 randomly sampled speech from the LibriSpeech test clean subset. Significance was evaluated using the ϵ value, indicating dominance. Scores were categorized as: significantly better when $0.0 < \epsilon \leq 0.5$, significantly dominant when $\epsilon = 0.0$, and not significantly better when $\epsilon > 0.5$.

Table 1 shows the full significance analysis, comparing between DM-Codec (**D**) and the baselines: EnCodec (**E**), SpeechTokenizer (**S**), and FACodec (**F**). The significance of DM-Codec is indicated by it outperforming all baselines across most metrics with better average and standard deviation. Among the baseline, however, FACodec achieves improved results over EnCodec and SpeechTokenizer, whereas SpeechTokenizer surpasses EnCodec in performance.

REFERENCES

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250. IEEE, 2021.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1266. URL https://aclanthology.org/P19-1266.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL https://arxiv.org/abs/2210.13438.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Towards general-purpose speech abilities for large language models using unpaired data. *arXiv preprint arXiv:2311.06753*, 2023.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models, 2024. URL https://arxiv.org/abs/2403.03100.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. Text-free prosody-aware generative spoken language modeling. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8666–8681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.593. URL https://aclanthology.org/2022.acl-long.593.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alexander H Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. Dinosr: Self-distillation and online clustering for self-supervised speech representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Shoval Messica and Yossi Adi. Nast: Noise aware speech tokenization for speech language models. *arXiv preprint arXiv:2406.11037*, 2024.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11: 250–266, 2023.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Jiatong Shi, Xutai Ma, Hirofumi Inaguma, Anna Sun, and Shinji Watanabe. Mmm: Multi-layer multi-residual multi-stream discrete speech representation from self-supervised learning model. *arXiv preprint arXiv:2406.09869*, 2024.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930*, 2023.

Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. SEANet: A Multi-Modal Speech Enhancement Network. In *Proc. Interspeech 2020*, pp. 1126–1130, 2020. doi: 10.21437/Interspeech.2020-1563.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Arnon Turetzky and Yossi Adi. Last: Language model aware speech tokenization, 2024. URL `https://arxiv.org/abs/2409.03701`.

Ziqian Wang, Xinfa Zhu, Zihan Zhang, YuanJun Lv, Ning Jiang, Guoqing Zhao, and Lei Xie. Selm: Speech enhancement using discrete tokens and language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11561–11565. IEEE, 2024.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec, 2023. URL `https://arxiv.org/abs/2305.02765`.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507, nov 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129994. URL `https://doi.org/10.1109/TASLP.2021.3129994`.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models, 2024a. URL `https://arxiv.org/abs/2308.16692`.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, et al. Speechlm: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024b.