Speech Separation using Neural Audio Codecs with Embedding Loss

Jia Qi Yip*† Chin Yuen Kwok* Bin Ma† and Eng Siong Chng*
* Nanyang Technological University, Singapore
† Alibaba Group
E-mail: jiaqi006@e.ntu.edu.sg

***Abstract*—Neural audio codecs have revolutionized audio processing by enabling speech tasks to be performed on highly compressed representations. Recent work has shown that speech separation can be achieved within these compressed domains, offering faster training and reduced inference costs. However, current approaches still rely on waveform-based loss functions, necessitating unnecessary decoding steps during training. We propose a novel embedding loss for neural audio codec-based speech separation that operates directly on compressed audio representations, eliminating the need for decoding during training. To validate our approach, we conduct comprehensive evaluations using both objective metrics and perceptual assessment techniques, including intrusive and non-intrusive methods. Our results demonstrate that embedding loss can be used to train codec-based speech separation models with a 2x improvement in training speed and computational cost while achieving better DNSMOS and STOI performance on the WSJ0-2mix dataset across 3 different pre-trained codecs.**

I. Introduction

Speech separation, also known as the cocktail party problem [1], is the task of isolating individual speakers from overlapping audio, has seen remarkable progress in recent years [2]. While current models achieve impressive signal-to-noise ratios on clean datasets [3] [4], the field is now advancing towards more generalized capabilities in diverse environments. This shift towards broader applicability necessitates even larger training datasets [5], a requirement that is currently hindered by the substantial computational costs associated with training speech separation models.

The computational intensity of speech separation stems from two primary factors. First, traditional models typically employ minimal time-compression in their encoders, often using simple 1D convolutions that achieve only 8x downsampling in the time dimension [6]. Recent approaches have revisited hybrid frequency and time domain methods [7] [8], incorporating greater downsampling to enhance computational efficiency without sacrificing performance.

Second, the loss function and evaluation metrics used in speech separation are inherently expensive. The Permutation Invariant Training (PIT) [9] Loss, essential for handling the speaker order ambiguity, requires computing $O(N!)$ combinations of outputs and ground truth, where N is the number of speakers in the mixture. Furthermore, the commonly used Scale-Invariant Signal to Distortion Ratio (SI-SDR) [10] is a waveform comparison loss function, which is computationally demanding for the long sequences typical in audio waveforms.

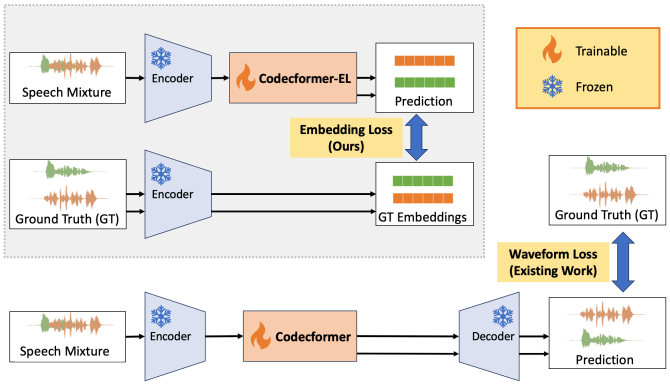

Fig. 1. Comparison between our proposed embedding loss (Top Left) with the conventional waveform loss used in a previous codec-based speech separation approach (Bottom). The key advantage of the embedding loss training method is that the decoder of the neural audio codec is not necessary for training, significantly reducing the computational cost during training.

Neural Audio Codecs [11] [12] [13], self-supervised models trained on vast amounts of general audio data, offer a promising solution to these challenges. These codecs, structured as autoencoders with a quantizer [14] between the encoder and decoder, can achieve time-domain compression ratios of approximately 100x, depending on the specific codec architecture. Due to the strong representations learned by neural audio codecs, they have been found to have useful applications beyond compression. They have been utilized in various downstream tasks, including automatic speech recognition [15] [16], text-to-speech [17] [18], speaker verification [19], and singing voice synthesis [20]. Recently, speech separation has also been added to this list of applications [21].

The recently proposed Codecformer [21], a neural audio codec-based speech separation model, has demonstrated the potential of integrating neural audio codecs with speech separation models to reduce computational requirements. However, as shown in Figure 1, this approach still relies on waveform-comparison losses, leaving room for further optimization.

In this paper, we propose a novel embedding loss for speech separation using neural audio codecs as shown in Figure 1. Our approach operates directly on the encoder representations of the codec, eliminating the need for expensive waveform comparisons during training. While PIT is still necessary, the shorter sequence length of the embeddings substantially decreases the computational cost of loss calculation.

Our contributions are threefold:

- We demonstrate that neural audio codec-based speech separation models can be trained from a variety of codecs using only embedding level loss, resulting in 2x improvement in training speech and computational cost with better performance on perceptual metrics, despite lower objective scores.
- While Codecformer [21] only supported the Descript Audio Codec (DAC) [13] and the PESQ [22] perceptual quality metric, we expand support to include EnCodec and SoundStream, as well as include additional perceptual quality metrics, STOI [23] and DNSMOS [24].
- We investigate the impact of pre-training data quantity and diversity on separation performance, comparing separation using the DAC model trained on open source datasets, LibriTTS, AMUSE, and the original DAC [13] dataset.

II. Methodology

In this work we propose Codecformer-EL, based on the recently proposed Codecformer [21], which modifies the training method for Codecformer by employing embedding loss as shown in Figure 1. While Codecformer utilizes both the encoder and decoder of the neural audio codec and performs waveform comparison loss, Codecformer-EL only requires the encoder. The loss computed is mean squared error (MSE) loss wrapped within the PIT [9] algorithm.

As shown in Figure 1, we perform speech separation by training a separator model, Codecformer [21], that makes use of a frozen encoder and decoder from a pre-trained neural audio codec. During inference and evaluation, the decoder is used to generate the necessary waveform for comparison.

A. Embedding Loss Function

Our proposed Embedding loss function can be written as follows:

$$e = \text{Encoder}(s_{gt})$$
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i - e_i)^2 \quad (1)$$

where s_{gt} is the ground truth speech and $\hat{e}$ is the separated embeddings produced by the separator in Codecformer-EL and n is the number of samples in each batch. This loss function is only used during training. Since the encoder used here to produce the groundtruth embeddings, e, is frozen, the embeddings can either be generated during training to save memory, or can be pre-computed before training to save on compute.

The choice of mean squared error loss in this study is inspired by early speech separation work that used this loss over spectrograms. However, because the neural audio codecs are learned representations, the embeddings obtained by the encoder are not spectrograms. Nevertheless, these embeddings are representations of audio and can be decoded into audio. Thus they may be thought of as spectrogram-like, which makes MSE loss more appropriate as opposed to Cosine Similarity loss used for semantic representations.

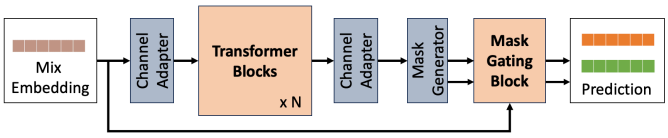

Fig. 2. Overview of the Codecformer model based on [21]. The bulk of the model consists of N transformer blocks along with channel adapter layers, a mask generator and the mask gating block.

B. Neural Audio Codecs

To verify that the embedding loss proposed in this paper is applicable to neural audio codec-based speech separation more generally, we make use of three popular neural audio codecs in this study, SoundStream [11], EnCodec [12] and Descript Audio Codec (DAC) [13] as described in Table I. We also compare the differences in performance across different training datasets, LibriTTS, AMUSE, and the original DAC dataset for the DAC model.

TABLE I
COMPARISON OF THE DIFFERENT CODECS USED IN THIS WORK

Model	Pre-training Dataset	Embedding Size	Params (M)
SoundStream	AMUSE	512	19.7
EnCodec	AMUSE	512	19.7
DAC	LibriTTS, AMUSE	512	19.7
DAC	Original	1024	74.2

SoundStream [11] pioneered the use of Residual Vector Quantization (RVQ) in neural audio codec development. Its architecture combined a SEANet-based encoder-decoder with an RVQ quantizer. For discrimination, it employed both waveform and spectral discriminators. The model's loss functions included adversarial and feature-matching losses from these discriminators, along with a multi-resolution mel-spectrogram loss.

EnCodec [12] built upon SoundStream's design, substituting the STFT discriminator with a multi-scale version. It introduced a probabilistic approach to discriminator updates and implemented a loss balancing mechanism to automatically adjust loss scales during training.

DAC [13] further refined EnCodec by incorporating several enhancements. These included the adoption of the snake activation function, advanced RVQ featuring factorized and L2-normalized codes, increased use of quantizer dropout, and implementation of an MSMPMB discriminator. These improvements aimed to tackle codebook collapse, the phenomenon where a fraction of the codes are unused.

The LibriTTS dataset [25] is a commonly used for training neural audio codecs. The Audio, Music, and Speech Ensemble (AMUSE) dataset developed for ESPnet-codec [26], is a fused corpus consisting of multiple open source high quality audio data, which also includes the original DAC dataset. Thus, the models trained on the AMUSE dataset represents the models with the largest-scale pre-training.

C. Codecformer-EL

The Codecformer with embedding loss (Codecformer-EL) model proposed in this study makes use of the same basic architecture as Codecformer [21] as shown in Figure 2. The bulk of the model consists of a stack of N transformer blocks followed by a mask gating block that is responsible for modifying the mix embedding received from the encoder. The Channel Adapters are linear layers that adapt the codec embedding size to the required dimensions of the transformer blocks. The mask generator is a linear layer that increase the number of channels to match the number of speakers in the mixture.

One weakness of the original Codecformer [21] it that the model only supported the DAC encoders and decoders. In this work we adapted Codecformer to support EnCodec and SoundStream as well. This required modifying the activation function in the Mask Gating Block from the Snake activation function used in DAC, to the ELU [27] activation function used in EnCodec and SoundStream. By matching the activation function in Codecformer with that of the neural audio codecs that they are embedded in, the generated outputs of the separation model are more likely to adopt a distribution more similar to the distribution expected by the decoder, potentially resulting in better generation by the decoder.

III. Experiments

A. Dataset

Our experiments utilize the widely-used WSJ0-2mix benchmark dataset [28]. This dataset is derived from the WSJ0 corpus, which consists of English read speech from the Wall Street Journal. The dataset comprises speech mixtures created from clean audio recordings. All audio, including both the mixtures and their corresponding ground truth, is sampled at 8kHz and recorded in acoustically controlled settings. The synthetic mixtures are generated using the 'min' condition, where the length of each mixture is cropped to match the duration of the shorter of the two input speeches. The WSJ0-2mix dataset is structured as follows: A training set consisting of 20,000 mixtures (30 hours) from the WSJ0 si_tr_s set, validation set consisting of 5,000 mixtures (10 hours) from the WSJ0 si_dt_05 set and a test set consisting of 3,000 mixtures (5 hours) from the si_et_05 set.

B. Codec Implementation

To ensure a consistent implementation, the model code and pre-trained weights for the codecs used were obtained from the ESPnet-Codec repository [26] except for the DAC Original implementation where the code and weights were obtained from the official release of the original authors[1]. Meanwhile, the training and evaluation of the models was implemented on the speechbrain toolkit [29] with the models from ESPnet handled by a wrapper written for speechbrain. For all neural audio codec models, the 16kHz version of the model is used, which is the lowest sampling rate available across all models. To manage the difference in sampling rate of the dataset and the neural audio codecs, the input audio is resampled using the torchaudio resampling method. Although this means that the codec receives audio that has missing frequencies, it is not expected that this results in significant performance impact.

C. Evaluation Methods

In addition to the standard objective metrics used for speech separation, such as SI-SDR, SI-SDRi, SDR, SDRi, we make use of perceptual evaluation metrics to measure the performance of the models. The perceptual evaluation metrics are necessary when performing neural audio codec-based speech separation because distortions introduced by the codecs result in distortions that are heavily penalized by objective metrics but are do not affect perceptual quality.

The perceptual quality metrics used in this study are deep noise suppression mean opinion score (DNSMOS) [24], perceptual evaluation of speech quality (PESQ) [22] and short-time objective intelligibility (STOI) [23]. Among these metrics, DNSMOS is a non-intrusive metric that does not rely on reference speech, while PESQ and STOI are intrusive metrics that do rely on some reference speech. Due to the ambiguity in the output order of the model, the ideal permutation for each of these metrics is computed based on maximizing the SI-SDR over all possible permutations. The permutation of outputs against the ground-truth that produces the highest average SI-SDR is used to compute all of the perceptual quality metrics. This also ensures that a consistent permutation is chosen across all objective and perceptual evaluation metrics.

D. Training Procedure

All models were trained for 20 epochs on a V100 with 32GB RAM. The Adam optimizer was used with an initial learning rate of $1.5e^{-4}$ and the LR scheduler was set to halve the learning rate with a patience of 2 after epoch 5. The models trained using the embedding loss were trained with a batch size of 20, while the those trained using the waveform loss were trained with a batch size of 3 due to memory limitations.

Following the settings of [21], the Codecformer separator consists of 16 transformer blocks with an input embedding size of 256. This is regardless of the embedding size of the neural audio codec used, which is either 1024 in the case of the original DAC and 512 for the ESPnet models, ensuring a fair comparison. The adapter layer in Codecformer was used to map the representations to the correct number of channels.

IV. Results

A. Improvements to Training Speed and Computation

One of the key features of Codecformer-EL is the improvement in training speech and computational efficiently. We measure this in Table IV using Multiple and Accumulate operations (MACs) computed using the PyTorch-OpCounter[2] as well as the number of hours required to train each epoch on the WSJ0-2mix dataset. We can see that Codecformer-EL trained 2.5x

[1] https://github.com/descriptinc/descript-audio-codec

[2] https://github.com/Lyken17/pytorch-OpCounter

TABLE II
COMPARISON OF SEPARATION PERFORMANCE ON DIFFERENT NEURAL AUDIO CODEC MODELS TRAINED ON EMBEDDING AND WAVEFORM LOSSES. ALL CODEC MODELS WERE PRE-TRAINED ON THE AMUSE DATASET

Model	**Loss Type**	**Objective Metrics**				**DNSMOS**				**PESQ**	**STOI**
		SI-SDR	**SI-SDRi**	**SDR**	**SDRi**	**OVRL**	**SIG**	**BAK**	**p808**		
DAC	Embedding	-1.2	-1.3	1.1	0.9	1.80	**2.12**	2.99	**2.58**	1.80	**0.81**
DAC	Waveform	2.8	2.8	4.9	4.7	**1.81**	2.04	**3.38**	2.55	**2.06**	0.79
EnCodec	Embedding	-29.1	-29.1	-8.5	-8.6	**1.89**	**2.20**	**3.18**	**2.65**	**1.92**	**0.80**
EnCodec	Waveform	-9.0	-9.0	-2.6	-2.8	1.24	1.26	2.75	2.29	1.28	0.51
SoundStream	Embedding	-14.2	-14.2	-4.9	-5.00	**1.80**	**2.07**	2.98	**2.63**	**1.94**	**0.83**
SoundStream	Waveform	-2.5	-2.4	1.2	1.0	1.73	1.95	**3.35**	2.45	1.77	0.74

TABLE III
COMPARISON OF CODECFORMER SEPARATION PERFORMANCE ON THE DAC MODEL WITH DIFFERENT PRE-TRAINING DATASETS

Pre-training Dataset	**Loss Type**	**Objective Metrics**				**DNSMOS**				**PESQ**	**STOI**
		SI-SDR	**SI-SDRi**	**SDR**	**SDRi**	**OVRL**	**SIG**	**BAK**	**p808**		
AMUSE	Embedding	-1.16	-1.16	1.10	0.94	1.80	2.12	2.99	2.58	1.80	0.81
	Waveform	2.8	2.7	4.9	4.7	1.81	2.04	3.38	2.55	2.06	0.79
Original	Embedding	-21.9	-22.0	-2.4	-2.6	1.57	1.91	2.53	2.38	1.57	0.71
	Waveform	6.7	6.7	7.9	7.7	2.19	2.53	3.40	2.80	2.20	0.85
LibriTTS	Embedding	-3.4	-3.3	-0.1	-0.2	1.89	2.26	2.98	2.62	1.75	0.80
	Waveform	1.0	1.1	3.0	2.9	1.82	2.03	3.43	2.60	1.76	0.75

faster and with 1.9x fewer MACs compared to Codecformer. Additionally, it trained 6.8x faster than Sepformer and requires 97x fewer MACs. The MACs were calculated on 2 seconds of 8kHz audio data and a V100 GPU with 16GB of RAM utilizing the maximum possible batch size was used to calculate the training speed.

TABLE IV
COMPUTATION AND TRAINING SPEED OF CODECFORMER-EL COMPARED AGAINST CODECFORMER AND SEPFORMER

Model	GMACs	Training Time (h/epoch)
Sepformer [2]	77.3	2.7
Codecformer [21]	1.5	1.0
Codecformer-EL (Ours)	**0.8**	**0.4**

B. Comparison between Codecformer-EL and Codecformer

Table II presents the separation performance of three neural audio codecs pre-trained on the AMUSE dataset, comparing our Codecformer-EL (embedding loss) method with the original Codecformer (waveform loss) approach.

A consistent pattern emerges: Codecformer models achieve higher scores on objective metrics (SI-SDR, SI-SDRi, SDR, SDRi), while Codecformer-EL models perform better on perceptual metrics, despite lower objective scores. This discrepancy highlights a known limitation of objective waveform-matching metrics for evaluating resynthesized speech [30].

The neural audio codec's decoder, trained with GAN-based loss, prioritizes generating perceptually natural speech over exact waveform reconstruction. This approach explains the disparity between objective and perceptual metric scores. Across the models, we observe varying degrees of performance difference. DAC models show the smallest gap between methods, with embedding loss outperforming on some perceptual metrics. EnCodec models display the most dramatic difference, with embedding loss achieving the highest perceptual scores despite poor objective metrics. SoundStream models also demonstrate a clear advantage for embedding loss in perceptual metrics.

These results underscore the potential of Codecformer-EL for real-world applications where perceived quality is crucial. They also emphasize the importance of using perceptual quality metrics for evaluating resynthesized speech separation performance. The success of our method across different codec architectures demonstrates its versatility and potential for wide applicability in speech separation tasks.

C. Comparison of speech separation performance across different codec pre-training datasets

In Table III we show the performance of the DAC model on speech separation after pre-training on different sets of data. In this analysis we see that regardless of the pre-training data, the DAC model performs similarly on the AMUSE and LibriTTS dataset, with embedding loss having a similar or better separation performance than waveform loss. However, in the model trained on the Original DAC dataset, the embedding method performs worse than the waveform loss and the Original DAC model with waveform loss performs the best out of all the models. This could potentially be due to the larger embedding size of 1024 compared to 512 in the other models. The embedding loss training method, due to its heavy reliance on the compressed representations, could be penalised more for the mismatch in dimension sizes between the transformer blocks and the DAC encoder. This could be further investigated in future work with ablation studies over different embedding sizes for the transformer in Codecformer-EL.

V. Conclusions

In this paper, we introduced Codecformer-EL, a novel approach to speech separation that leverages neural audio codecs and employs an embedding-level loss function. Our method demonstrates significant improvements in computational efficiency and training speed compared to previous approaches, while maintaining comparable or superior perceptual quality in separated speech. We showed that Codecformer-EL is effective across multiple neural audio codecs and pre-training datasets, highlighting its versatility and potential for widespread application.

Our findings reveal an important dichotomy between objective and perceptual metrics in evaluating resynthesized speech. While objective metrics sometimes indicated poorer performance, perceptual quality metrics often showed comparable or superior results, underscoring the importance of using appropriate evaluation methods for this type of speech processing task.

Codecformer-EL's improved efficiency paves the way for more scalable speech separation models capable of handling real-world audio challenges, particularly in applications where audio compression is necessary for transmission, such as smartphones offloading computationally intensive tasks to cloud servers [21]. However, it's crucial to acknowledge potential limitations, including the trade-off between compression and information preservation in codec embeddings. Highly compressed embeddings may lead to loss of fine-grained audio details, potentially affecting separation quality in complex acoustic environments [19]. Moreover, Codecformer-EL's performance is inherently tied to the quality and generalizability of the pre-trained neural audio codec, which may lead to suboptimal results when the codec's pre-training data significantly differs from the target separation task.

Future work could explore techniques to mitigate these limitations, such as fine-tuning the codec encoder for specific separation tasks or investigating adaptive compression levels based on input complexity. Additionally, research into the optimal balance between compression and separation quality could further enhance the practical applicability of this approach.

In conclusion, Codecformer-EL represents a significant step forward in efficient, high-quality speech separation, opening new avenues for both research and real-world applications in audio processing and communication technologies

Acknowledgment

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore. We would like to acknowledge Alibaba-NTU Joint Research Institute, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[3] J. Q. Yip, S. Zhao, Y. Ma, *et al.*, "Spgm: Prioritizing local features for enhanced speech separation performance," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[4] S. Zhao, Y. Ma, C. Ni, *et al.*, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[5] J. Pons, X. Liu, S. Pascual, and J. Serrà, "Gass: Generalizing audio source separation with large-scale data," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 546–550.

[6] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "Exploring self-attention mechanisms for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, 2022.

[7] Z. Wang, S. Cornell, S. Choi, Y. Lee, B. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *TASLP*, vol. 31, pp. 3221–3236, 2022.

[8] C. Chen, C.-H. H. Yang, K. Li, Y. Hu, P.-J. Ku, and E. S. Chng, "A neural state-space model approach to efficient speech separation," in *Proc. Interspeech*, 2023.

[9] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. DOI: 10.1109/ICASSP.2017.7952154.

[10] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 626–630.

[11] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

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

[13] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, pp. 4–29, 1984.

[15] T. Wang, L. Zhou, Z. Zhang, *et al.*, "Viola: Unified codec language models for speech recognition, synthesis, and translation," *arXiv:2305.16107*, 2023.

[16] A. Gupta, G. Saon, and B. Kingsbury, "Exploring the limits of decoder-only models trained on public speech recognition corpora," *arXiv:2402.00235*, 2024.

[17] Z. Ju, Y. Wang, K. Shen, *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *ICML*, 2024.

[18] D. Yang, D. Wang, H. Guo, X. Chen, X. Wu, and H. Meng, "Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models," *arXiv:2406.02328*, 2024.

[19] K. C. Puvvada, N. R. Koluguri, K. Dhawan, J. Balam, and B. Ginsburg, "Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 111–12 115.

[20] X. Chang, J. Shi, J. Tian, *et al.*, "The Interspeech 2024 challenge on speech processing using discrete units," in *Proc. Interspeech*, 2024.

[21] J. Q. Yip, S. Zhao, D. Ng, E. S. Chng, and B. Ma, "Towards audio codec-based speech separation," in *Proc. Interspeech*, 2024.

[22] 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. ICASSP*, IEEE, vol. 2, 2001, pp. 749–752.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *TASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[24] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6493–6497.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[26] *Espnet-codec*, https://github.com/espnet/espnet/tree/codec, 2024.

[27] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[28] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 31–35.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106.04624 [eess.AS].

[30] J. Shi, X. Chang, T. Hayashi, Y.-J. Lu, S. Watanabe, and B. Xu, *Discretization and re-synthesis: An alternative method to solve the cocktail party problem*, 2022. arXiv: 2112.09382 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2112.09382.