Do Coarser Units Benefit Cluster Prediction-Based Speech Pre-Training?

Ali Elkahky, Wei-Ning Hsu, Paden Tomasello, Tu Anh Nguyen, Robin Algayres, Yossi Adi, Jade Copet, Emmanuel Dupoux, Abdelrahman Mohamed

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096788

DO COARSER UNITS BENEFIT CLUSTER PREDICTION-BASED SPEECH PRE-TRAINING?

Ali Elkahky, Wei-Ning Hsu, Paden Tomasello, Tu-Anh Nguyen, Robin Algayres, Yossi Adi
Jade Copet, Emmanuel Dupoux, Abdelrahman Mohamed

Meta Inc.

ABSTRACT

The research community has produced many successful self-supervised speech representation learning methods over the past few years. Discrete units have been utilized in various self-supervised learning frameworks, such as VQ-VAE [1], wav2vec 2.0 [2], HuBERT [3], and Wav2Seq [4]. This paper studies the impact of altering the granularity and improving the quality of these discrete acoustic units for pre-training encoder-only and encoder-decoder models. We systematically study the current proposals of using Byte-Pair Encoding (BPE) and new extensions that use cluster smoothing and Brown clustering. The quality of learned units is studied intrinsically using zero speech metrics and on the downstream speech recognition (ASR) task. Our results suggest that longer-range units are helpful for encoder-decoder pre-training; however, encoder-only masked-prediction models cannot yet benefit from self-supervised word-like targets.

Index Terms— self-supervision, representation learning, unit discovery

1. INTRODUCTION

Self-supervised speech representation learning has dramatically improved over the past few years [5], showing impact for multiple downstream tasks [6] in (ultra) low-resource conditions. During their training, many approaches benefited from auxiliary discrete acoustic units derived from latent continuous representations to facilitate learning, e.g., VQ-VAE [1], Wav2vec 2.0 [2], HuBERT [3]. Although they were proposed for just model pre-training with no downstream use, these discrete units ,in later works, used as pseudo-language for Textless speech and dialogue generation [7, 8], speech compression [9] and translation [10]. Given the crucial role of discrete acoustic units during self-supervsied learning, there have been multiple proposals for refining such units; for example [3] interleaved high-level feature learning with updating the learned units, [11] added a duration constraint during k-means inference to control the granularity of inferred units, [4] applied temporal smoothing of learned representation, and [12] applied byte-pair encoding to discrete units.

Another open area of investigation arises from the dependence of most speech representation methods on encoder-only training, which mainly captures acoustic and phonetic information. How can the linguistic content and broader syntactic constraints be integrated to impact learned representations? Would the discovered pseudo-language capture syntactic and semantic information, demonstrated in probing tasks and generated audio content? A recent study [13] showed that feeding word boundary information to a speech representation model leads to better semantics modeling of the spoken audio. An Encoder-decoder self-supervised model would fill this gap where the decoder module would model the regularities of the pseudo-language units, even fixing some of their associated labeling noise [4, 12].

This work systematically evaluates different proposals to improve representation learning, the discovered acoustic units, and the downstream model performance. We use the HuBERT self-supervised approach [3] as our test bed in this paper, both with encoder-only and encoder-decoder pre-training setups. We use some of the zero-speech metrics [14, 15] to describe the quality of discovered discrete acoustic units quantitively. The paper is organized as follows: first, we introduce the zero-speech metrics we used, then dive into different proposals for improving the learned discrete tokens. Our experimental section presents apple-to-apple comparisons between different discrete units on the downstream ASR task. We also study the interplay between the granularity of the discovered units and adopting an encoder-decoder pre-training approach.

2. RELATED WORK

Self-supervised approaches made remarkable progress in Computer Vision (CV) and Natural Language Processing (NLP). NLP models such as BERT [16], and RoBERTa [17] used a Masked Language Model (MLM) loss on top of a Transformer encoder [18]. CV approaches like MoCo [19] and SimCLR [20] relied on contrastive losses for learning representation. Ideas from those models made their way to speech representation learning. Contrastive Predictive Coding (CPC) [21] and Wav2Vec [22] contrasted near-by frames from further away ones. Wav2vec 2.0 [2] used a bidirectional transformer encoder and defined a contrastive loss between the discrete representation of masked segments in the current position and other masked segments in the same utterance. The Hidden Unit BERT (HuBERT) model [3] discretizes the input audio first, then apply the MLM to predict the audio tokens given masked continuous input representations. The WavLM model [23] extends HuBERT by mixing speakers and different types of noise to the input while extending the model to denoise inputs in addition to the masked prediction. For a full discussion of self-supervised approaches for speech processing, we refer the reader to this recent review article [5].

Learned discrete acoustic units were developed both for serving representation learning approaches and for their own sake as pseudo-language to work with in subsequent speech and audio applications. In [11], Dynamic Programming (DP) is employed to constrain unit durations in k-means inference. Using an order of magnitude more discrete units, compared to the HuBERT model, [24] showed that the masked prediction loss would still yield competitive performance with randomly assigned and fixed clusters. Acoustic piece [25] proposed to learn longer-range units by applying the SentencePiece algorithm [26] on top of learned HuBERT units.

Seq2seq pretraining was introduced for speech data following its success with text input with the BART model [27]. [4] extended the encoder-only HuBERT model and showed a solid performance

for speech translation and other NLP tasks. Similarly, [12] proposed a seq2seq extension with a dual loss on both the encoder and the decoder sides during pretraining, fine-tuning, and inference.

3. ACOUSTIC UNITS IMPROVEMENTS AND EVALUATION METRICS

This section discusses different ideas for discrete acoustic unit improvements and unit evaluation metrics derived from zero-speech and unsupervised clustering literature.

3.1. Zero Speech Metric

Different metrics have been introduced to evaluate automatic unit discovery and automatic speech segmentation in zero-speech research. Precision, Recall, and F-score have been used for evaluating the quality of detected segmentation boundaries given gold phonemes segmentation. Precision computes how many gold segment boundaries the system got out of all the predicted boundaries. The Recall is how many gold segments the system got out of the segments in the reference. Since Recall can be improved by adding extra random boundaries, [15] proposed the Over-Segmentation (OS) and R-value metrics. OS is defined as the ratio of the number of predicted segments to the number of reference segments minus one:

$$OS = 100 \times \frac{\#Spans_{pred}}{\#Spans_{ref}} - 1 \tag{1}$$

By defining the Recall Error RE as $(1 - Recall) \times 100$, The R-value can be calculated as follows:

$$r_1 = \sqrt{OS^2 + RE^2} \tag{2}$$

$$r_2 = \frac{RE + OS}{\sqrt{2}} \tag{3}$$

$$R = 1 - \frac{|r_1| + |r_2|}{200} \tag{4}$$

The R-value balances over-segmentation and recall, so lowering the R-value amounts to detecting the right boundaries in the reference segmentation without proposing too many arbitrary spans. The F-score aims at a similar balance by combining precision and recall, but the R-value drops quicker when OS is greater than 1 [15].

The above metrics measure the quality of the segmentation boundaries but not the nature and coherence of the assigned discrete units, i.e., labels, for each detected audio span. The V-measure [14] is the harmonic mean of the Homogeneity and Completeness of unsupervised clustering. Homogeneity is based on the conditional entropy (H) of the gold class distribution(C) given the clusters' distribution (K) normalized by the gold class entropy. In the same way, Completeness is based on the conditional entropy of clusters K given classes C normalized by the clusters K entropy. Those measures work on the segment level by aligning predicted segment to the reference segment and take the label of segment with most overlap as with the reference as the assigned cluster for this segment.

$$Homogeneity = \begin{cases} 1, & \text{if } H(C,K) = 0 \\ 1 - \frac{H(C|K)}{H(C)}, & \text{otherwise} \end{cases} \tag{5}$$

$$Completeness = \begin{cases} 1, & \text{if } H(C,K) = 0 \\ 1 - \frac{H(K|C)}{H(K)}, & \text{otherwise} \end{cases} \tag{6}$$

In the experiments, we reason about the learned HuBERT units using these metrics rather than relying solely on the downstream ASR performance. We used the clean validation set of the Librispeech dataset [28] for evaluating these zero-speech metrics. The validation set ground-truth text is converted to phoneme sequences and was aligned to the raw speech using a baseline ASR model. These alignments serve as our reference for metric computation.

3.2. Encoder-Decoder Models

Conducting seq2seq pretraining through adding an autoregressive decoder to the HuBERT model was proposed in [4, 12] as an extension of the standard encoder-only Hubert pretraining. The primary motivation for adding a decoder during pretraining is to capture long-range linguistic relations between acoustic units, leading to better input semantics modeling. Both proposals rely on a baseline HuBERT model to estimate discrete acoustic units for each input utterance. They then deduplicate units to represent an input segment rather than a 20ms frame (the output frame rate of baseline HuBERT model) and use this unit sequence as the target for label-smoothed cross-entropy loss of the decoder. However, these two proposals differ in two points: [4] uses average pooling as in Sec 3.3.2, deduplicates units, then builds a large vocabulary using BPE on the resulted units, which reduces the output sequence length even more. On the other hand, [12] applies a dual loss on the encoder and the decoder. Since this work aims to analyze the impact of units when used as encoder and decoder targets, respectively, we do not consider dual loss pre-training.

3.3. Ways to improve discrete acoustic units

3.3.1. DP Smoothing

After running the k-means algorithm to discover clusters in learned representations, [11] replaces the greedy assignment of cluster IDs to representation frames with a DP search constrained by the inferred segment length. A segment score $w_{seg}(a, b)$ is defined for a segment starts at a position a and ends at position b:

$$w_{seg}(a,b) = min_{e_k \in Clusters\_Center} \sum_{i \in [a,b]} \|x_i - e_k\| \tag{7}$$

where x_i is the continuous representation vector at time i. The duration score $w_{dur}(a, b)$ is defined to penalize shorter segments (i.e $w_{dur}(a,b) \propto \frac{1}{(b-a+1)}$). At inference time, a weighted sum of these two scores is minimized:

$$w(a,b) = w_{seg}(a,b) + \lambda w_{dur}(a,b) \tag{8}$$

The minimum cost segmentation based on equation 8 is found by running a dynamic program over possible segments (a, b) with a cost $\mathcal{O}(N^2)$ where N is the sequence length. This smoothing is expected to impact the over-segmentation metric positively but could worsen recall.

3.3.2. Time Averaging

Pooling continuous speech representations over time before clustering was used in [4] to reduce the sequence target length in seq2seq pretraining when combined with byte-pair encoding (BPE) [26]; however, the impact of the temporal average of features on unit segmentation was not studied. Like DP-smoothing, averaging features over time would improve the detection precision of phonetic segments and reduce over-segmentation. Although time-averaging

n features is easier to tune and understand, DP-smoothing offers a less aggressive alternative with its real-valued weighting between segment and duration scores.

3.3.3. Layer Averaging

The original Hubert paper [3] used a single latent layer of representations to train discrete codebooks; however, multiple layers of representations were used in [29] to represent the distillation targets of a HuBERT model. In the experiments, we examine an instance-normalized average of multiple layers during the k-means training steps for generating the discrete target tokens.

3.3.4. Longer-range Units

The discovered units by baseline HuBERT models are mainly at the phonetic or sub-phonetic level. With the growing interest in the textlessNLP research [7] and seq2seq pretraining, there were some proposals for longer-range units [25, 4] which capture the linguistic structure of the input beyond phonemes, with lengths reaching syllables or sub-word textual representations. Byte-pair Encoding [26] and Sentencepiece algorithm [30] were used to combine units into larger sequential clusters. An order of magnitude larger dictionary size would emerge with a much higher target length compression ratio (r), which is preferable for training autoregressive decoders or subsequent unit language models (ULM) [7]. The compression ratio (r) is the ratio between the total number of frames in the raw data (at 20ms rate) and the length of deduplicated and BPE-encoded sequence of units.

3.3.5. Brown Clustering

Brown Clustering (BC) [31] is a hierarchical clustering algorithm used extensively in NLP applications. The basic intuition of BC is that similar entities appear in a similar context, so given an initial clustering, BC iteratively merges similar clusters based on the neighboring cluster IDs they keep. We use BC to reduce the size of BPE-encoded vocabularies, e.g., Starting with 30k vocabulary down to a smaller one of 2k units. BC, while helping to reduce noise in the discovered units, enables the application of another BPE encoding on the dictionary of reduced size, which extends the granularity of the discrete units even more and provides a higher compression ratio.

4. EXPERIMENTAL SETUP

We use librispeech 960h for all pretraining experiments, and the 10h supervised subset of the Libri-light dataset [32] for fine-tuning. Since our focus is on analyzing the quality of learned representations, not reporting a single best number for a specific method, we have opted to only use greedy frame-level decoding without an external language model (LM) or lexicon constraints. Our clustering experiments are performed using the kmeans++ algorithm with 500 clusters unless stated otherwise. Clustering is done using the public HuBERT BASE model. We report the zero speech metric on the dev_clean subset and downstream ASR performance on dev_other and $test_other$. We do all model selection based on dev_other WER. Since we use the public HuBERT model for generating units with different strategies and using them for another round of pretraining, we trained a baseline model which uses the default k-means clustering for training. All these models represent third iterations of HuBERT training. We stick to the HuBERT BASE architecture for our encoders with the same training recipe for both pretraining and fine-tuning. We add a 6-layer decoder for encoder-decoder models and train the whole model for 100k updates for pretraining (with encoder initialization from public HuBERT) and 10k updates for fine-tuning. For BPE training, we use the BPE encoder from Huggingface tokenizer vocabulary size of 30,000.

5. RESULTS

5.1. Frame level units for encoder-only pretraining

Table 1 shows ASR and zero speech metrics for different types of units on encoder-only models. Aside from the MFCC features, which are expected to be a distant last, all other features yield comparable WER results except layer averaging. As found in other studies [33], the topmost layers of HuBERT are not the best feature representations. Averaging layers 6,7 and 8 led to slightly better results. Although they show good trends in the quality of learned representations, the zero speech metrics of discovered units are not good indicators of the encoder-only downstream performance. The best-performing system has the worst over-segmentation, R-value and F-score. This is an interesting finding pointing to the ability of the encoder transformer network to recover from noisy boundaries and over-segmentation as long as the units are packed with information (the case of averaged features). The units of different refinement proposals could lead to more costly mistakes than slightly bad segmentation and lousy cluster labels.

5.2. Optimizing Units for the encoder-decoder models

To evaluate the correlation between zero-speech metrics and encoder or encoder-decoder pretraining, we use the best-performing encoder-only unit strategy from table 1 and search for optimal zero-speech metrics over smoothing λ, number of km units and time averaging. We select the best units according to either R-value, V-measure or Compression Ratio. We then use the units to train encoder models or build longer-range, coarser units to train encoder-decoder models. For BPE, we build a 30k dictionary tokenizer. For Table 2, we picked the best DP smoothing λ and the number of k-means units based on their V-measures, R-values or Compression ratio (CR). Although improving zero speech metrics seems to hurt downstream ASR WER for encoder-only pretraining, they do not hurtthe quality of the final pretrained encoder-decoder model with the best dev number coming from the best CR units. Units with lower compression ratios achieve good final performances; however, this excludes low compression ratios achieved trivially by using a higher λ during the DP smoothing step, e.g., the 3rd and 4th rows in Table 2 artificially achieve low CR by over-smoothing initial units before BPE. Our results suggest a correlation between zero-speech metrics and compression ratio on one end and the performance of encoder-decoder pretraining. Since encoder-only relies on masked LM loss and encoder-decoder predicts the entire sequence of discrete targets, units optimized for encoder-only pretraining do not necessarily generalize to other pretraining losses.

5.3. Long-range units using Brown Clustering

Table 3 shows the results of both encoder-only models and encoder-decoder models using baseline units, best short units from Table 1 with and without Brown Clustering (BC) applied. It confirms our earlier observation that longer-range units serve the encoder-decoder pretraining better than the encoder-only one. We see an almost opposite trend for both systems concerning unit granularity.

Table 1. Results showing zero speech metrics and WER using different unit creation approaches.

Units	P ↑	R ↑	F ↑	OS ↓	R-value	H ↑	C ↑	V ↑	Dev/Test WER ↓
MFCC	17.1	98.2	29.1	476.0	-306.9	24.1	5.9	9.5	20.8/21.3
Baseline (3rd iteration HuBERT)	34.7	96.4	51.1	177.7	-53.0	68.4	38.0	48.8	16.4/16.7
Instance Normalization	32.4	98.3	48.8	205.5	-75.7	69.5	38.3	49.3	15.9/16.1
Layers Avg 7-9	29.2	99.5	45.1	240.7	-105.7	60.1	33.2	42.8	**15.6**/16.0
Layers Avg 10-12	32.6	98.9	49.0	203.7	-74.4	69.2	38.1	15.7	15.7/16.0
Time Avg	48.7	92.2	63.7	89.4	20.7	68.8	38.3	49.2	16.0/16.3
BC 1k $\rightarrow$ 800	51.6	86.9	64.8	68.3	36.4	73.9	37.9	50.1	15.8/16.2
BC 1k $\rightarrow$ 500	51.8	86.6	64.8	67.3	37.3	69.5	38.2	49.3	15.9/16.3
DP Smoothing	51.6	86.9	64.7	68.4	36.3	75.0	37.7	50.1	16.2/16.2
DP Smoothing $\lambda = 0.5$ + Time Avg	62.1	44.6	51.9	-28.1	59.3	70.4	39.4	50.5	16.4/16.1
DP Smoothing $\lambda = 0.5$ + Layers Avg 7-9	47.4	94.0	63.0	98.3	13.9	72.4	39.7	51.3	16.0/16.2
Time Avg + + Layers Avg 7-9	32.3	98.9	48.7	206.1	-76.3	69.3	38.2	49.3	16.1/16.3

Table 2. Results of units optimized for zero speech metrics and encoder-decoder pretraining.

Units	P↑	R↑	F↑	OS ↓	R-value	H↑	C↑	V ↑	CR ↓	Enc Dev/Test ↓	Enc-Dec Dev/Test ↓
Avg 7-9	29.2	99.5	45.1	240.7	-105.7	60.1	33.2	42.8	18.9	15.6/16.0	15.3/15.8
+ DP λ = 3	47.4	94.0	63.0	98.3	**13.9**	72.4	39.7	51.3	14.4	16.8/16.6	15.2/15.8
+ DP λ = 7	58.2	83.1	**68.4**	42.9	55.8	73.6	40.6	52.3	11.9	16.6/16.6	15.5/16.0
+ DP λ = 13	62.9	65.1	64.0	3.5	68.9	74.4	41.1	**53.0**	9.7	17.1/17.1	15.6/16.1
100 km baseline	39.6	96.4	56.1	143.6	-23.9	54.2	40.8	46.5	12.7	16.6/16.9	15.6/16.0
100 KM+ λ = 1.0	56.0	87.1	68.2	55.4	47.4	60.2	44.8	51.4	**9.3**	16.8/17.1	15.2/15.8

Table 3. Results of Applying BPE and/or Brown Clustering to frame level units to obtain long range units

	Dev/Test Enc	Dev/Test Enc-Dec
Baseline Hubert units	16.4/16.7	18.1/19.4
Best short units	15.6/16.0	17.9/19.0
baseline + BPE	17.7/18.2	15.9/16.7
Best short units + BPE	17.3/18.0	15.3/15.9
+ BC 2k + BPE	17.2/18.0	15.1/15.6
Best Short units + smoothed + BPE	17.4/18.1	15.2/15.8
+ BC 2k + BPE	17.7/18.3	15.3/15.8

Table 4. Results of baseline and best short and long units on encoder-only and encoder-decoder models with comparable capacities

Units	12 Layers (Enc)	12 Layers(Enc) +6 layers(Dec)	18 Layers (Enc)
baseline	16.4/16.7	18.1/19.4	15.4/15.9
Best Frame-level Units	15.6/16.0	17.9/19.0	**14.3**/14.2
Best Long-range Units	17.2/18.0	**15.1**/15.6	16.3/16.7

5.4. How good is encoder-decoder pretraining?

Given the results in Table 3, encoder-decoder pretraining shows better downstream WER than encoder-only pretraining. These gains could come from better language modeling in the decoder or merely due to their increased modeling capacity (6 more transformer layers). Table 4 compares the downstream ASR performance when models of similar capacity are pretrained. There is no significant performance difference between encoder-only and encoder-decoder pretraining, given that we use the best discrete target units for each system. The best systems achieve WER of about 15% (both systems do not use a lexicon nor an external LM during decoding). We believe that the gains of encoder-decoder pretraining demonstrated in [12] can be attributed to their use of a dual loss during fine tuning and inference, which is known to improve the ASR results regardless of the pretraining strategy.

6. CONCLUSION AND FUTURE WORK

This paper focused on analyzing the performance of different target discrete acoustic units with increasing granularity for encoder-only and encoder-decoder pretraining. We systematically built and examined many proposals for acoustic unit smoothing and aggregation to cover an extended range of input audio. Encoder-only and encoder-decoder pretraining benefit from different unit granularity. Encoder-decoder pretraining was found to benefit more from long-range units with large dictionaries. Our results show the comparable performance of encoder-only and encoder-decoder pretraining when adjusting for the model capacity and target units, contradicting previous work on encoder-decoder pretraining and long-range acoustic units. Our future work includes a more in-depth study of the long-range units for TextlessNLP and generative speech applications that could benefit from its high target compression abilities.

7. REFERENCES

[1] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *NeurIPS*, 2017.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[3] 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*, 2021.

[4] Felix Wu, Kwangyoun Kim, Shinji Watanabe, Kyu Han, Ryan McDonald, Kilian Q. Weinberger, and Yoav Artzi, "Wav2seq: Pre-training speech-to-text encoder-decoder models using pseudo languages," 2022.

[5] Abdelrahman et al. Mohamed, "Self-supervised speech representation learning: A review," *arXiv*, 2022.

[6] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[7] Kushal et. al. Lakhotia, "On generative spoken language modeling from raw audio," *TACL*, 2021.

[8] 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," *arXiv*, 2022.

[9] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv*, 2021.

[10] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu, "Textless speech-to-speech translation on real data," in *NAACL*, 2022.

[11] Herman Kamper, "Word segmentation on discovered phone units with dynamic programming and self-supervised scoring," *arXiv*, 2022.

[12] Junyi Ao, Ziqiang Zhang, Long Zhou, Shujie Liu, Haizhou Li, Tom Ko, Lirong Dai, Jinyu Li, Yao Qian, and Furu Wei, "Pre-training transformer decoder for end-to-end asr model with unpaired speech data," *arXiv preprint arXiv:2203.17113*, 2022.

[13] Tu Anh Nguyen, Maureen de Seyssel, Robin Algayres, Patricia Roze, Ewan Dunbar, and Emmanuel Dupoux, "Are word boundaries useful for unsupervised language learning?," *arXiv*, 2022.

[14] Andrew Rosenberg and Julia Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *EMNLP-CoNLL*, 2007.

[15] Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar, "An improved speech segmentation quality measure: the r-value," in *INTERSPEECH*, 2009.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv*, 2019.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, 2018.

[22] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[23] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[24] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu, "Self-supervised learning with random-projection quantizer for speech recognition," *arXiv*, 2022.

[25] Shuo Ren, Shujie Liu, Yu Wu, Long Zhou, and Furu Wei, "Speech pre-training with acoustic piece," *arXiv*, 2022.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *ACL*, 2016.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv*, 2019.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[29] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," 2021.

[30] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv*, 2018.

[31] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer, "Class-based n-gram models of natural language," *Computational linguistics*, 1992.

[32] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020.

[33] Puyuan Peng and David Harwath, "Word discovery in visually grounded, self-supervised speech models," 2022.