HAL
open science

The Zero Resource Speech Challenge 2021: Spoken language modelling

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, Emmanuel Dupoux

To cite this version:

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, et al.. The Zero Resource Speech Challenge 2021: Spoken language modelling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, pp.1-1. 10.1109/TPAMI.2021.3083839 . hal-03329301v1

HAL Id: hal-03329301

https://inria.hal.science/hal-03329301v1

Submitted on 30 Aug 2021 (v1), last revised 11 Oct 2021 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Zero Resource Speech Challenge 2021: Spoken language modelling

Ewan Dunbar[1], *Mathieu Bernard*[2], *Nicolas Hamilakis*[2], *Tu Anh Nguyen*[2,3], *Maureen de Seyssel*[2], *Patricia Rozé*[2], *Morgane Rivière*[3], *Eugene Kharitonov*[3], *Emmanuel Dupoux*[2,3]

[1]University of Toronto, Canada
[2]Cognitive Machine Learning (ENS–CNRS–EHESS–INRIA–PSL Research University), France
[3]Facebook AI Research, France

ewan.dunbar@utoronto.ca, mathieu.a.bernard@inria.fr, nick.hamilakis562@gmail.com, nguyentuanh208@gmail.com, maureen.deseyssel@gmail.com, patricia.roze@ens.fr, mriviere@fb.com, kharitonov@fb.com, emmanuel.dupoux@gmail.com

arXiv:2104.14700v2 [cs.CL] 9 Aug 2021

Abstract

We present the Zero Resource Speech Challenge 2021, which asks participants to learn a language model directly from audio, without any text or labels. The challenge is based on the Libri-light dataset, which provides up to 60k hours of audio from English audio books without any associated text. We provide a pipeline baseline system consisting on an encoder based on contrastive predictive coding (CPC), a quantizer (k-means) and a standard language model (BERT or LSTM). The metrics evaluate the learned representations at the acoustic (ABX discrimination), lexical (spot-the-word), syntactic (acceptability judgment) and semantic levels (similarity judgment). We present an overview of the eight submitted systems from four groups and discuss the main results.

Index Terms: zero-resource; language modelling; low-resource; unsupervised speech; cognitive benchmarks

1. Introduction

Infants are able to learn their native language(s) through observation and interaction before they learn to read and write. They show that it is in principle possible to build a language model in the absence of textual resources, from sensory data only. Being able to reproduce this achievement through automatic means would open up speech and language technology to the majority of the world's languages which do not have enough textual resources to be served by current text-driven approaches.

The Zero Resource Speech Challenge Series aims at developing the building block necessary to construct textless AI applications. Previous iterations of the challenge have focused on the discovery of subword units (ZR15,17,19,20) and word units (ZR15,17,20). Here, we wish to push the envelope even further by aiming at learning a language model directly from audio without any annotation nor text.

Systems are only allowed to use the raw audio of the training set as input; they can use it to discover discrete units from it (pseudo-text) and then train a language model from it, or learn everything end-to-end without discrete units. Following the strategy of the previous challenges, evaluation is done through zero-shot metrics based on human psycholinguistics which do not require any training. Here we are probing four linguistic levels: acoustic, lexical, syntactic and semantic.

We provide baseline systems which are the concatenation of three unsupervised components: self-supervised contrastive representation learning (CPC) [1], clustering (k-means), language modeling (LSTM or BERT). The language models learn on the basis of the pseudo-text derived from clustering the learned representation. In order to take into account the computing resources of participants, we distinguish submissions by the amount of GPU budget used for training. Accordingly our baseline language models are sorted into a high and a low compute budget. As this benchmark series is about fostering new ideas, not getting the best numbers, we encouraged participants to submit systems in the low budget category.

2. Methods

2.1. Datasets.

Participants can use any training set provided they do not use textual labels besides the identity of the speaker. We encourage use of the LibriSpeech 960h English dataset [2], and for larger models, the clean-6k version of Libri-light [3], a huge collection of speech for unsupervised learning.

2.2. Evaluation metrics.

The evaluation metrics are described in detail in [4], and we only give here high level descriptions for lack of space.

Acoustic: the Libri-light ABX metrics. The ABX metric consists in estimating, for two speech categories A and B (e.g., 'bit' and 'bet'), the probability that two exemplars x and a of the same category A are closer to one another than two exemplars x and b of different categories A and B. The score is aggregated across all pairs of triphones like 'bit' and 'bet', where the change occurs in the middle phoneme. This can be computed within-speaker (a, b and x are from the same speaker) or across-speaker (a and b are from the same speaker, and x from a different speaker). To compute this score, participants are required to provide an embedding for each input triphone and to specify a pseudo-distance between acoustic tokens. By default, we provide such a distance, which is the average along a Dynamic Time Warping path realigning a, b and x of a distance between embedding frames (angular distance). This metric is agnostic to the dimensionality of the embeddings, can work with discrete or continuous codes, and has been used to compare ASR speech features [5]. Here, we run it on the pre-existing Libri-light dev and test sets, which has been already used to evaluate several self-supervised models [3, 6].

Lexicon: the sWUGGY spot-the-word metrics. In this task, the models are presented with a pair of spoken tokens: an existing word and a matching nonword. Participants are to provide a number (probability or pseudo-probability) associated to

Table 1: ***Summary description of the four Zero Resouce Benchmark 2021 metrics***. *The metrics in light blue use a pseudo-distance d between embeddings (d_h being from human judgments), the metrics in light orange use a pseudo-probability p computed over the entire input sequence.*

Linguistic level	Metrics	Dataset	Task	Example
acoustic-phonetic	ABX	Libri-light	$d(a,x) < d(b,x)$? $a \in A, b \in B, x \neq a \in A$	within-speaker: $(\text{apa}_{s_1}, \text{aba}_{s_1}, \text{apa}_{s_1})$ across-speaker: $(\text{apa}_{s_1}, \text{aba}_{s_1}, \text{apa}_{s_2})$
lexicon	spot-the-word	sWUGGY	$p(a)>p(b)$?	(brick, *blick) (squalled, *squilled)
lexical semantics	similarity judgment	sSIMI	$d(a,b) \propto d_h(a,b)$?	(abduct, kidnap) : 8.63 (abduct, tap): 0.5
syntax	acceptability judgment	sBLIMP	$p(a) > p(b)$?	(dogs eat meat, *dogs eats meat) (the boy can't help himself, *the boy can't help herself)

each acoustic tokens, and models are evaluated on their average accuracy of word-nonword classification based on this probability (chance level: 0.5). The sWUGGY test and development sets consists of 20,000 and 5,000 pairs respectively, with the existing words being part of the LibriSpeech train vocabulary. We also prepared additional OOV-sWUGGY test and development sets consisting of 20,000 and 5,000 pairs respectively, with existing words which do not appear in the LibriSpeech training set. The nonwords are produced with WUGGY [7], which generates, for a given word, a list of candidate nonwords best matched in phonotactics and syllabic structure, which we additionally filtered for pronouncability using G2P, and for having on average the same unigram and bigram phoneme frequencies as words. Stimuli were produced with the Google Speech API.

Syntax: the sBLIMP acceptability metrics. This part of the benchmark is adapted from BLIMP [8], a set of linguistic minimal sentence pairs of matched grammatical and ungrammatical sentences. Similarly to sWUGGY, the task is to decide which of the two is grammatical based on the probability of the sentence. The test and dev sets contain 63,000 and 6,300 sentence pairs respectively, with no sentence pair overlap. Stimuli were filtered to contain LibriSpeech vocabulary and for natural prosodic contours, and synthesised as above.

Lexical Semantics: the sSIMI similarity metrics. Here, as in [9], the task is to compute the similarity of the representation of pairs of words and compare it to human similarity judgements. As for the ABX task, participants provide embeddings for input tokens as well as a distance to compute similarity. Here, we provide by default the cosine distance computed over pooled embeddings (with mean, max or min pooling). We used a set of 13 existing semantic similarity and relatedness tests: WordSim-353 [10], WordSim-353-SIM [11], mc-30 [12], rg-65 [13], Rare-Word (or rw) [14], simLex999 [15], simverb-3500 [16], verb-143 [17] , YP-130 [10] and the relatedness-based datasets include MEN [18], Wordsim-353-REL [11], mturk-287 [19], and mturk-771 [20]. All scores were normalised on a 0-10 scale, and pairs within a same dataset containing the same words in different order were averaged. Pairs containing a word absent from LibriSpeech train set [2] were discarded. We selected as development set the mturk-771 dataset and the other 12 datasets were used as test sets, making sure that no pair from the development set was present in any of the test sets.

We then created two subsets of audio files, one synthetic (using the Google API), one natural obtained by retrieving the audio extracts from LibriSpeech corresponding to each word, following the process presented in [9]. In this subset, each word can appear in multiple tokens, providing phonetic diversity; duplicated scores are averaged in the analysis step. The natural subset is smaller than its synthesised counterpart, as we had to discard pairs from the test and dev sets which were not present in the LibriSpeech test and dev sets respectively. The synthesized subset is composed of 9744 and 705 word pairs for the test and dev sets respectively, and the LibriSpeech subset is composed of 3753 and 309 pairs for the test and dev sets.

2.3. Toplines and baselines.

The Baseline models. We build baselines in three steps: acoustic modelling, clustering, and language modelling.

The acoustic model uses Contrastive Predictive Coding (CPC, [1]), where the representation of the audio is learned by predicting the future through an autoregressive model. The CPC model uses a convolutional encoder g_{enc} to map an input signal $\mathbf{x}$ as a sequence of embeddings $\mathbf{z} = (z_1, \ldots, z_T)$ at a given frame rate. At each time step t, a predictor network g_{pred} takes as input the available embeddings $z_1, \ldots, z_t$ and tries to predict the K next future embeddings $\{z_{t+k}\}_{1 \leq k \leq K}$ by minimizing a constrastive loss using $\mathcal{N}_t$ negative embedding samples. We used the PyTorch implementation of CPC[1] [6], which is a modified version of the CPC model with the following architecture: the encoder g_{enc} is a 5-layer 1D-convolutional network with kernel sizes of 10,8,4,4,4 and stride sizes of 5,4,2,2,2 respectively, resulting in a downsampling factor of 160, meaning that, for a 16KHz input, the embeddings have a rate of 100Hz.

The predictor g_{pred} is a multi-layer LSTM network, with the same hidden dimension as the encoder, followed by a 1-layer transformer. We report results from a 4-layer LSTM, trained on the clean-6k version of Libri-light (see [4] for additional results from a smaller model). Note that we are still able to train low- and higher-budget baselines on the output of this model, as we take the critical and potentially costly part of the pipeline to be language model training, and therefore calculate the GPU budget based on the language model training time only.

The clustering module uses k-means on the outputs of a given hidden layer of the autoregressive model. The clustering is done on the collection of all the output features at every time step of all the audio files in a given training set. After training the k-means clustering, each feature is assigned to a cluster, and each audio file can then be discretized to a sequence of discrete units corresponding to the index of the assigned cluster.

We trained the clustering module on the subset of LibriSpeech containing 100 hours of clean speech, using as input the second layer of the CPC model. Our baseline ABX scores are calculated on the framewise k-means units, for $k = 50$.

The language modeling module takes as input the dis-

[1] https://github.com/facebookresearch/CPC_audio

cretized version of the audio files and is trained with a predictive objective; we used two architectures: LSTM and BERT [21]. We trained a 3-layer LSTM model as a reference small-budget system (22M parameters). Following [22], we trained the BERT model with only the masked token prediction objective. We also followed [22] by masking a span of tokens in the input sequence instead of a single token (otherwise the prediction would be trivial to the model as discretized units tend to replicate). We masked M consecutive tokens for each span, where $M \sim \mathcal{N}(10, 10)$, with a total masking coverage of roughly half of the input tokens (spans may overlap). We trained the BERT model using a 12-layer transformer encoder and use this as a reference high-budget system (90M parameters). The implementation was done via fairseq ([23]). For further details of the models, see [4].

The Topline models. We trained a BERT model on force-aligned phoneme labels (one per frame) using the gold transcription of the LibriSpeech dataset. We also employed the span masking similarly to the baseline model. In addition to the forced-alignment BERT, we also included a BERT model trained on the gold phonetic transcription of the LibriSpeech dataset (no framewise repetitions), with the difference that we only mask one token instead of a span of tokens, since each token is the width of a phoneme rather than a frame. For an absolute topline comparison, we used the pretrained RoBERTa large model ([24]), which was trained on 50K subword units on a huge dataset of total 160GB, 3000 times bigger than the transcription of the LibriSpeech 960h dataset.

3. Submitted systems and results

Eight systems were submitted. All were low-budget. The systems are in principle less comparable than in previous years, because the choice of training sets was freer, but all systems made use of either the baseline representations or the same choices of corpus for re-training, with the exception of **HL**, which trained its units on LibriSpeech 100 hrs (clean), a smaller data set.

System **BN** [25] begins with the baseline CPC representations and applies speaker normalization before re-running k-means. An LSTM language model architecture is used.

The two systems of [26] both use continuous representations for the acoustic evaluation. **JC1** uses the baseline CPC representations for the acoustic tests, while **JC2** improves on the baseline representations using speaker embeddings, and by pulling the vectors in the direction of a cluster mean. For the other evaluation measures, these submissions use different approaches for each measure. While this does not directly assess language modelling, it helps understand the limits of the acoustic representations for finding and encoding different types of linguistic information. Beginning from a (k-means) discretized version of the respective acoustic CPC representations, for sWUGGY, they treat the training corpus as a dictionary and extract a distance to the best match; for sBLIMP, they train an LSTM; and, for sSIMI, they perform segmentation to discover word types. They then train word embeddings using word2vec (**JC1**) or fast-text (**JC2**) and use this embedding space to calculate similarity.

System **HL** [27] trains representations using Mockingjay, an approach based on bidirectional self-attention [28]. These are used as pseudo-labels in a teacher-student training scheme.

Finally, systems **TM1–TM4** [29] train a first round of CPC representations (using different variants of the baseline CPC models), followed by clustering. These cluster labels are then used as a classification objective for a second trained network. The language models are based on a smaller BERT recipe (28M parameters) we provided [4].

Acoustic-phonetic. Results are presented in Table 2. Since the ABX scores do not rely on the language models, the BERT and LSTM baselines are the same. Most systems improve on the baseline, with the exception of **HL** (which is the only representation not based on CPC). Multiple systems are tied for first place, but the bigger picture is that, since the baseline is already excellent, the English triphone ABX measure is approaching what one might consider "solved."

Note that the ABX task has not changed fundamentally since the first ZeroSpeech challenge, and, importantly, it continues to evaluate the discriminability of allophones only. As the test items are always in the same phonetic context, the test does not measure whether phonemes have a context-invariant representation. For this reason, low ABX scores may not be a sufficient basis for learning language models. We use our other measures to assess this.

Lexical. From Table 2, the systems tend to improve quite a bit on the low-budget LSTM baseline (with **BN** taking the lead). There remains a big gap between the gold transcription, which supports near-perfect performance on identifying real word-forms, and any of the unsupervised representations proposed, giving room for improvement in future challenges.

Syntactic. The sBLIMP scores show smaller improvements over the LSTM baseline than sWUGGY. Unlike for sWUGGY, the toplines show that access to the gold transcription alone is not sufficient: while it helps, good performance is only achieved by RoBERTa, which notably trains on much more data than our other toplines. We discuss the implications below. Thus, this appears to be a harder metric to do well on.

Lexical semantics. sSIMI is even harder . Even the best-performing reference model, the RoBERTa topline, shows fairly weak correlations with human judgements. And, as with syntax, the gold transcription alone is not sufficient even to reach this weak correlation. More is needed, as evidenced by the large drop in performance for the phone topline compared to RoBERTa. Here, however, it would appear that an additional factor matters: the temporal resolution of the input representations. The force-aligned topline, which takes frames as input, shows much poorer performance—in some cases worse than the random baseline (note also, for the LibriSpeech test set, the random baseline is actually the best). In spite of the difficulty of the task, the word-discovery approach of **JC1** appears promising.

4. Discussion

In the span of six years and five challenges, astonishing progress has been made on the triphone ABX task. Systems starting from CPC modelling (all but one here) achieve extremely good ABX scores. Previous work has shown that a low ABX score appears to be a very good predictor of good performance on TTS without T [30, 31, 32]. Nevertheless, previous results of the TTS without T task indicate that finding low-bitrate representations—with coarse-grained resolution both temporally and spatially—is a substantial constraint. As hinted at above, future challenges may also introduce more difficult ABX tasks, which require strictly phoneme-level invariance.

The 2021 challenge takes a different route. Rather than adding extrinsic constraints to push representations to be more abstract and text-like, we asked whether unsupervised speech representations could solve one of the *problems* classically solved by text: language modelling. Results are promising.

Most prominently, progress has been made in the span

Table 2: ***Leaderboard.*** *Bolded results have the best score in the column among the submitted systems for the given task.*

			ABX-with.		ABX-across		sWUGGY	sBLIMP	sSIMI	
System	Budget	Set	clean	other	clean	other			synth.	Libri.
Random Baseline	0	dev	0.49	0.5	0.5	0.5	0.5	0.49	-1.48	6.79
		test	0.5	0.49	0.5	0.5	0.5	0.5	0.17	6.44
Bert Baseline	1536	dev	0.03	0.05	0.04	0.08	0.68	0.56	6.25	4.35
		test	0.03	0.05	0.04	0.08	0.68	0.56	5.17	2.48
LSTM Baseline	60	dev	(idem)	(idem)	(idem)	(idem)	0.61	0.52	4.42	7.07
		test	(idem)	(idem)	(idem)	(idem)	0.61	0.53	7.35	2.38
BN	60	dev	0.05	0.09	0.07	0.13	**0.64**	**0.54**	4.29	7.69
		test	0.05	0.09	0.07	0.13	**0.65**	**0.54**	**9.23**	-1.14
JC1	60	dev	**0.03**	0.05	**0.04**	0.08	0.63	0.52	**5.90**	**10.20**
		test	**0.03**	0.05	**0.04**	0.08	0.64	0.53	2.42	**9.02**
JC2	60	dev	**0.03**	**0.04**	**0.04**	**0.07**	**0.64**	0.53	-7.75	4.60
		test	**0.03**	**0.04**	**0.04**	**0.07**	0.64	0.53	5.15	-0.85
AL	60	dev	0.17	0.20	0.25	0.30	0.51	0.52	3.16	1.79
		test	0.17	0.20	0.24	0.31	0.52	0.52	7.30	-4.33
TM1	60	dev	**0.03**	0.05	**0.04**	0.08	0.61	**0.54**	-0.81	5.45
		test	**0.03**	0.05	**0.04**	0.09	0.61	**0.54**	7.00	-1.47
TM2	60	dev	**0.03**	0.05	**0.04**	0.08	0.58	**0.54**	-1.65	4.81
		test	**0.03**	0.05	**0.04**	0.08	0.59	**0.54**	2.89	-1.67
TM3	60	dev	0.04	0.06	0.05	0.10	0.62	0.53	-0.17	7.07
		test	0.04	0.06	0.05	0.11	0.62	0.53	5.93	0.56
TM4	60	dev	0.04	0.06	0.05	0.10	0.60	0.53	-2.10	8.89
		test	0.04	0.06	0.05	0.10	0.60	0.53	6.74	2.03
Top: Frame labels	1536	dev	0	0	0	0	0.92	0.64	7.92	4.54
		test	0	0	0	0	0.92	0.63	8.52	2.41
Top: Phone labels	1536	dev	-	-	-	-	0.98	0.67	9.86	16.11
		test					0.98	0.67	12.23	20.16
Top: RoBERTa	24576	dev	-	-	-	-	0.97	0.82	32.28	28.96
		test	-	-	-	-	0.96	0.82	33.16	27.82

of one iteration of this challenge on the sWUGGY metric. This is great news for LM applications that are heavily driven by the lexicon. Thus, the approaches taken here should already be helpful for ASR decoding (note also that some recent self-supervised pretraining approaches to ASR have shown good results without explicit LMs: [33]). There is a big gap between discovered units and text—so, still major room for improvement—but the speed of progress is very encouraging.

Above the word level, things are harder. As for sWUGGY, sBLIMP performance shows promising improvements over the baselines. Still, the gaps with text-based approaches are more pronounced for both sBLIMP and sSIMI. Speech-based language modelling needs to make more progress before being applicable to tasks like translation and dialogue, for which good language modelling depends on making syntactic and semantic predictions. For syntax, even the topline, character-based models are middling, and the much-improved performance of the pre-trained RoBERTa model comes at a cost: the training set used by that model is the equivalent of many human lifetimes worth of speech. Training on this much speech is not a plausible cognitive model, nor useful for low-resource tasks. For semantics, even our best text-based model is far from the target.

One possible missing piece in both tasks is words. Most systems start from representations with the temporal granularity of phonemes or less. RoBERTa may also be limited by its use of word-pieces as input, as even simple word embeddings obtain better correlations than those seen here on metrics related to sSIMI [34]. Indeed, **JC1** demonstrates that explicit term discovery is useful for the sSIMI task. Future challenges may do well to introduce a distinction between fine-grained ("character"-based) and coarse-grained ("word"-based) metrics.

5. Summary of contributions

We have presented the results of eight systems submitted to the Zero Resource Speech Challenge 2021. This serves as a definitive jumping-off point for language modelling from speech, both because of the standard evaluation provided, and because of the novel reference systems submitted. The systems demonstrated such excellent phonetic discriminability scores that the bar must be pushed higher in future challenges in order to be able to demonstrate progress. However, higher-order tasks remain in their infancy, with even the easiest tasks proving difficult for current unsupervised representations.

6. Acknowledgements

The work for MS, PR and for EDupoux and TAN in their EHESS role was supported by the Agence Nationale de la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and grants from CIFAR (Learning in Minds and Brains) and Facebook AI Research (Research Grant). The work for EDunbar was supported by a Google Faculty Research Award and by the Agence Nationale de la Recherche (ANR-17-CE28-0009 GEOMPHON, ANR-18-IDEX-0001 U de Paris, ANR-10-LABX-0083 EFL).

7. References

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[3] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, and et al., "Libri-light: A benchmark for asr with limited or no supervision," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020. [Online]. Available: http://dx.doi.org/10.1109/ICASSP40776.2020.9052942

[4] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *arXiv preprint arXiv:2011.11588*, 2020.

[5] T. Schatz, "Abx-discriminability measures and applications," Ph.D. dissertation, Paris 6, 2016.

[6] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," 2020.

[7] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior research methods*, vol. 42, no. 3, pp. 627–633, 2010.

[8] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman, "Blimp: A benchmark of linguistic minimal pairs for english," *arXiv preprint arXiv:1912.00582*, 2019.

[9] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *arXiv preprint arXiv:1803.08976*, 2018.

[10] D. Yang and D. M. Powers, *Verb similarity on the taxonomy of WordNet*. Masaryk University, 2006.

[11] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," 2009.

[12] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.

[13] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[14] M.-T. Luong, R. Socher, and C. D. Manning, "Better word representations with recursive neural networks for morphology," in *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 2013, pp. 104–113.

[15] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[16] D. Gerz, I. Vulić, F. Hill, R. Reichart, and A. Korhonen, "Simverb-3500: A large-scale evaluation set of verb similarity," *arXiv preprint arXiv:1608.00869*, 2016.

[17] S. Baker, R. Reichart, and A. Korhonen, "An unsupervised model for instance level subcategorization acquisition," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 278–289.

[18] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional semantics in technicolor," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012, pp. 136–145.

[19] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: computing word relatedness using temporal semantic analysis," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 337–346.

[20] G. Halawi, G. Dror, E. Gabrilovich, and Y. Koren, "Large-scale learning of word relatedness with constraints," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1406–1414.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[22] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS

[23] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://www.aclweb.org/anthology/N19-4009

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[25] B. van Niekerk, L. Nortje, M. Baas, and H. Kamper, "Analyzing speaker information in self-supervised models to improve zero-resource speech processing," *Submitted to Interspeech, 2021*.

[26] J. Chorowski, G. Ciesielski, J. Dzikowski, A. Łancucki, R. Marxer, M. Opala, P. Pusz, P. Rychlikowski, and M. Stypułkowski, "Submission to interspeech, 2021," *Submitted to Interspeech, 2021*.

[27] Liu, "Submission to interspeech, 2021," *Submitted to Interspeech, 2021*.

[28] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[29] T. Maekaku, Y. Fujita, X. Chang, L.-W. Chen, S. Watanabe, and A. Rudnicky, "Submission to interspeech, 2021," *Submitted to Interspeech, 2021*.

[30] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, and E. Dupoux, "The zero resource speech challenge 2019: Tts without t," 2019.

[31] E. Dunbar, J. Karadayi, M. Bernard, X.-N. Cao, R. Algayres, L. Ondel, L. Besacier, S. Sakriani, and E. Dupoux, "The zero resource speech challenge 2020: Discovering discrete subword and word units," in *INTERSPEECH, perception;bootstrapping/modeling;clustering/bootphon*, 2020.

[32] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "Generative spoken language modeling from raw audio," *arXiv preprint arXiv:2102.01192*, 2021.

[33] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[34] R. Dror, G. Baumer, M. Bogomolov, and R. Reichart, "Replicability analysis for natural language processing: Testing significance with multiple datasets," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 471–486, 2017.