UNSUPERVISED ACCENT ADAPTATION THROUGH MASKED LANGUAGE MODEL CORRECTION OF DISCRETE SELF-SUPERVISED SPEECH UNITS

Jakob Poncelet, Hugo Van hamme

KU Leuven
Department Electrical Engineering ESAT-PSI, Leuven, Belgium

ABSTRACT

Self-supervised pre-trained speech models have strongly improved speech recognition, yet they are still sensitive to domain shifts and accented or atypical speech. Many of these models rely on quantisation or clustering to learn discrete acoustic units. We propose to correct the discovered discrete units for accented speech back to a standard pronunciation in an unsupervised manner. A masked language model is trained on discrete units from a standard accent and iteratively corrects an accented token sequence by masking unexpected cluster sequences and predicting their common variant. Small accent adapter blocks are inserted in the pre-trained model and fine-tuned by predicting the corrected clusters, which leads to an increased robustness of the pre-trained model towards a target accent, and this without supervision. We are able to improve a state-of-the-art HuBERT Large model on a downstream accented speech recognition task by altering the training regime with the proposed method.

Index Terms— self-supervised learning, pre-training, accent invariance, accent adaptation, speech recognition

1. INTRODUCTION

Recently, acoustic models pre-trained in a self-supervised fashion have made big waves in the field of speech processing. Hidden representations learned by HuBERT [1], wav2vec 2.0 [2], WavLM [3] and others [4, 5] can be used to achieve strong performance on a variety of speech processing tasks, without requiring copious amounts of task-specific labeled data. However, prior research has revealed that while these pre-trained acoustic models are able to discover informative speech units [6], they are sensitive to domain shifts between the (pre-)training and target domain [7], to speaker variations [8] and to contextual factors like co-articulation and allophony [9].

The learned pre-trained representations can be biased towards standardised canonical data characteristics of the pre-training corpora, and consequently perform worse on nonnative or accented speakers, as shown for HuBERT in [10]. Following [11], dialectal changes can to some extent be described as a collection of vocal shifts from a standard pronunciation paradigm. As the learning process of many self-supervised pre-trained acoustic models crucially involves quantisation or clustering of latent features in the prospect of discovering informative speech units, this shift will also be represented in the discretised features.

In this paper, we propose to improve accent-invariance by learning to remove the acoustic accent shift from the pre-trained model representations, without any supervision. We model the standard pronunciation statistics by training a masked language model [12] on discrete speech units from the source language, generated by clustering the pre-trained model's hidden features. The extracted pre-trained discrete clusters of accented speech are then corrected towards the standard accent by applying an iterative mask-and-decode algorithm to the language model. Finally, we fine-tune the pre-trained models with these corrected clusters as self-supervised targets and show improved invariance towards the target accent.

Our approach requires only unlabeled, non-parallel data from the source language and the target accent. In this work, we consider North American English to be the standard source accent, as it is the conventional language in most work on English speech recognition (e.g. LibriSpeech [13]). We evaluate the approach for several English accents from the CommonVoice dataset [14] on an automatic speech recognition (ASR) task. We perform continuous ASR with fine-tuned HuBERT model features and show improved Word Error Rates (WER) compared to the uncorrected models on accented speech.

2. PRELIMINARIES

2.1. BERT

BERT [12] is a language model based on Transformers [15] that learns strong bidirectional language representations. The main objective is masked language modeling: part of the input tokens are masked, and the model is optimised to fill in those masked tokens conditioning on both left and right context with attention layers. The masks can be applied randomly or in larger spans comprising several consecutive tokens, as in SpanBERT [16]. The output layer maps the feature representations back to the vocabulary of tokens. In the standard BERT model, masked language modeling is combined with a contrastive next sentence prediction objective. Finally, distilBERT [17] is a smaller, distilled version of BERT.

2.2. HuBERT

As audio and speech are continuous signals, a language model cannot be applied directly without a discretisation step. HuBERT [1], or Hidden Unit BERT, is a popular approach for self-supervised speech representation learning from raw audio. First, an offline clustering step is applied to discover discrete acoustic units in the audio. Applying the K-means algorithm to MFCC features for example gives meaningful units. Second, the model is trained to predict these aligned discrete units from the partly masked output of an acoustic feature encoder using a bidirectional BERT-like encoder. Hence, HuBERT intrinsically combines an acoustic and masked language model over continuous features of units. Finally, the model is improved iteratively by re-training with better discrete units, obtained

Research supported by Research Foundation Flanders (FWO) under grant S004923N of the SBO programme and by the Flemish Government under the "Flanders AI Research Program".

Sentence	WHERE IS ...																	
Input phones*	W	W	AH	AH	AH	AH	R	R	R	R	IH	IH	IH	Z	Z	Z	Z	Z
Input clusters	7	345	181	181	181	181	468	406	406	467	356	356	356	281	281	453	9	9
Masked inputs	7	345	[M]	[M]	[M]	[M]	468	406	406	467	[M]	[M]	[M]	281	281	453	9	9
Output clusters	7	345	109	109	264	264	468	406	406	467	356	356	356	281	281	453	9	9
Output phones*	W	W	W	W	EH	EH	R	R	R	R	IH	IH	IH	Z	Z	Z	Z	Z

Table 1: Example of the inputs and outputs of the MLM for part of an accented utterance in CommonVoice. Note that the input and output phone labels (*) are not used by the model, but generated afterwards as the most likely cluster to phone mapping. [M] denotes a mask.

from clustering the encoder features of a previous iteration model. Note that a HuBERT Large model (24 layers) is pre-trained using clusters from a 2nd iteration HuBERT Base model (12 layers) [1].

2.3. Adapters

Fine-tuning a large pre-trained model on a new dataset or task can be very computationally expensive, as the scale of pre-trained models is increasing rapidly. Instead of fine-tuning the entire model, nearly the same improvements can be accomplished by only fine-tuning small residual blocks (i.e. adapters) that are added to every layer [18] and freezing all pre-trained weights, drastically reducing the number of parameters to optimise. Furthermore, adapters alleviate catastrophic forgetting of learned knowledge during pre-training [19], as the pre-trained weights are not altered.

3. APPROACH

This work aims to improve the accent-invariance of pre-trained models using unlabeled speech from the source and target accent. A masked language model (MLM) is trained on discrete speech units from the source accent, of which generally lots of data is available, and afterwards the model is used to correct the discrete units for an accented utterance.

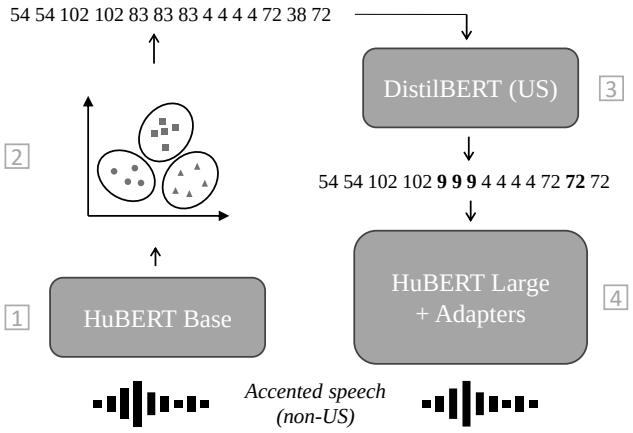


Fig. 1: Schematic of the proposed method. 1. Features are extracted for an accented utterance with the HuBERT Base model. 2. K-means quantisation to a discrete token sequence. 3. Clusters are corrected by the distilBERT model, trained on US data, using the mask and predict algorithm. 4. HuBERT Large model is adapted to the accent by optimising small adapter blocks during continual pre-training with the corrected clusters.

3.1. Standardised Unit Language Model

Offline clustering of the hidden features of a self-supervised pre-trained model leads to informative discrete speech units that correlate well with actual phonemes as in HuBERT [1], without having seen any labels. Previous work has shown that these discrete acoustic units are strong building blocks for modeling spoken language without supervision [20, 21]. To model the standard pronunciation in the source language, North American English, we train an MLM on the discretised utterances from that accent. To aid training, the MLM is initialised from a pre-trained distilBERT model[1]. The token embeddings are re-initialised as the vocabulary now only consists of 500 clusters id's. The MLM is optimised with the masked prediction objective, omitting the next sentence prediction. Following SpanBERT [16], the masks are applied in larger spans masking several consecutive frames, as one frame corresponds to only 20ms of audio.

3.2. Accented Cluster Correction

Accents are characterised by a vocal shift from a standard pronunciation [11]. Vowel shifts are the most obvious examples, but not the only ones. For example, notice the difference between British and North American pronunciations for the words *dance* (/dɑːns/ vs /dæns/), *stupid* (/ˈstjuːpɪd/ vs /ˈstuːpɪd/) and *either* (/ˈaɪðəʳ/ vs /ˈiːðəʳ/)[2].

An acoustic model trained in a self-supervised way will identify these sounds as belonging to separate units, although they represent the same characters. To achieve accent-invariance, the masked language model trained on the standard pronunciation should correct this accent shift by changing the clusters that are affected back to the (statistically most-likely) standard pronunciation clusters, e.g. the *a* in the above example *dance*, taking context into account.

To correct a sequence of cluster id's corresponding to an accented utterance with the MLM, we apply an iterative mask-and-decode algorithm, similar to Mask-Predict [22]. As the MLM is not conditioned on the audio, conditional decoding algorithms like Mask-Predict [22], Mask-CTC [23] and BERT-CTC [24] cannot be trivially applied, and only a small fraction of the input sequence should be masked, or the model will not be able to recover the original sentence. Moreover, the masks are applied to groups of identical consecutive clusters, to correct entire speech units more easily.

The cluster rescoring algorithm works as follows. First, the input cluster sequence is forwarded through the MLM, and the frame-wise softmax output scores are used as confidences for the input tokens. Ideally, the aforementioned accent shifts are given a lower confidence by the LM, as it is trained to predict an alternative pronunciation for those sounds in a word. After scoring, consecutive identical

[1] https://huggingface.co/distilbert-base-uncased

[2] According to Cambridge Dictionary using IPA symbols.

Masking strategy	PER *dev*	PER *test*
Original clusters	85.69	85.76
Adaptive masking cluster groups	80.55	80.61
Adaptive masking phone groups*	76.09	76.14
Adaptive masking phone groups fill-all*	63.92	63.95

Table 2: Phone Error Rates of rescored British English utterances, converted from clusters to phones, w.r.t. a North-American English G2P transcription. For phone groups (*), masks are applied to groups of frames based on their phone class. *Fill-all* means that all masks are filled in every iteration, instead of a fraction.

clusters are grouped together. The score of every cluster group is determined by the maximum score of all clusters in the group, as in [23]. Then, the cluster groups with the lowest scores are masked, so that at least N_k frames are masked. Next, the masked inputs are digested by the MLM which gives it's predictions for the mask tokens. Finally, the M most confident predictions of these cluster groups are filled in. The masks are computed again for the corrected clusters, and the decoding continues. This procedure is carried out for K iterations, where the total number of masks adaptively shrinks every iteration k as $N_k = \lfloor N_{max} \cdot \frac{K-k+1}{K} \rfloor$ [23]. The number of masks M that are filled in every iteration is fixed as $M = \frac{N_{max}}{K}$. The maximum number of masks is determined by the length T of the utterance and the masking ratio P_{mask}, as $N_{max} = P_{mask} \cdot T$.

The number of iterations K and the masking percentage P_{mask} influence how much of the input sequence will be altered.

3.3. Accent-Invariant Model Adaptation

Previous work [10] has shown that pre-trained HuBERT models benefit from adaptation to a new accent by continually pre-training on data from the target accent. To this end, appending accent-specific adapters to the encoder layers and fine-tuning only the adapter parameters, reaches almost the performance of full encoder fine-tuning but with a strongly reduced number of parameters.

This work builds upon [10], continually pre-training HuBERT on corrected cluster sequences by fine-tuning accent-specific adapters, improving accent-invariance without supervision. We apply Houlsby bottleneck adapters [18] after every attention and feedforward block in the encoder. The adapters will learn to normalise the accent shifts that have been corrected by the MLM.

4. EXPERIMENTS

4.1. Experimental Setup

We use a subset of the CommonVoice [14] 7.0 release with 6 well represented English accents for pre-training. The training set consists of 827k utterances (611 hours) of English speech, including North American (350k), British (116k), Indian (71k), Australian (56k), Scottish (11k) and Irish (7k) accents. There are 1k utterances from each accent for validation and testing.

The baseline HuBERT Large model consists of 7 convolutional feature extraction layers and 24 Transformer layers, with a total of 316M parameters. It has been pre-trained on the 60k hour LibriVox corpus with target clusters extracted from the 9th layer of the 2nd iteration HuBERT Base model (12 Transformer layers), which was pre-trained on 960 hours of LibriSpeech. The K-means quantiser that generated the clusters has 500 centroids learned on LibriSpeech train-clean-100h. Both models and the quantiser are officially released checkpoints in fairseq[3].

For continually pre-training the HuBERT Large model, we insert Houlsby adapters at every Transformer layer, after the attention block and after the feed-forward block. The adapters consist of a feedforward down-projection layer, a ReLU, a feedforward up-projection layer and a layer normalisation block. The bottleneck dimension is set to 1024, which performed best in [10]. The target clusters are extracted with the same HuBERT Base 2nd iteration model and the respective quantiser. The learning rate is set to 1.5e-3 and the model is updated for 30k steps with a linear learning rate schedule with 5k warm-up steps. We use a batch size of 32 with a maximum of 300k tokens. For a dataset of 100h, this setup corresponds to 30 full epochs. The adapter-HuBERT model is fine-tuned on only one accent per experiment.

The distilBERT masked language model consists of 6 Transformer layers. It is trained on masked token sequences where spans of 10 tokens are masked and on average 20% of the entire utterance is masked ($P_{mask} = 0.2$). Following BERT, masked tokens have a 20% chance to be replaced with either the original or a random token, to reduce the train-test mismatch. The MLM is trained for 18 epochs with a batch size of 32, or about 200k steps. The learning rate follows a linear decay starting from 5e-5. The MLM is trained on the North American English split of the dataset.

The ASR model consists of a 2-layer BiLSTM (as in the SUPERB challenge [25]) with 512 hidden units, corresponding to around 10M parameters. It is optimised with a CTC loss [26] to predict characters. The input to the ASR model is a learnable weighted sum of the pre-trained model's Transformer layer hidden features. SpecAugment is applied to the inputs and the model is trained for 100k steps with a learning rate of 1e-4. The ASR model is trained on the North American English split of the dataset, and evaluated on all English accents to show unsupervised accent adaptation.

4.2. Phone Modeling

In this first experiment, we evaluate the cluster corrections of the MLM towards the North-American accent. As there are only 500 different clusters, the discrete units are better interpretable. Based on force-aligned phoneme annotations[4] for LibriSpeech-100h, we determine the most likely mapping from the cluster id to the corresponding phoneme (40 classes). Then, we use a North-American G2P[5] on the transcripts of the accented dev and test set to get a reference phoneme transcription for these utterances. We rescore the original clusters with the MLM and the proposed mask-decode algorithm to get corrected clusters. Next, we map both the original clusters and the rescored cluster sequences to their phoneme equivalent based on the learned cluster-phoneme mapping. Table 1 shows an example of input and output clusters and their mapping to phonemes. Finally, we evaluate the Phoneme Error Rate (PER) between the reference G2P phonemes and the original/corrected phonemes. Results are shown for the British English utterances in the dataset in Table 2, applying the masking algorithm either to cluster groups or to phone groups, where the phone group is derived from the cluster id based on the learned mapping.

[3] https://github.com/facebookresearch/fairseq/tree/main/examples/hubert

[4] https://zenodo.org/record/2619474

[5] https://pypi.org/project/g2p-en/

Model / Masking strategy	ASR w/ LS960				ASR w/ CV-US			
	dev-clean	*test-clean*	*dev-eng*	*test-eng*	*dev-clean*	*test-clean*	*dev-eng*	*test-eng*
HuBERT Large [1]	**5.77**	**6.01**	24.39	19.82	**6.84**	**7.22**	16.79	14.26
HuBERT + Adapters [10]	6.14	6.50	22.11	17.92	7.18	7.64	15.87	12.45
Adaptive masking cluster groups	6.18	6.72	21.52	18.00	7.49	7.85	15.41	**11.71**
Adaptive masking only vowels if $k > 3$	6.17	6.58	**21.01**	**17.47**	7.34	7.63	**15.06**	11.89

Table 3: ASR experiments with WER results. The ASR model is either trained on LibriSpeech (LS960) or on the North-American English split of CommonVoice (CV-US), and evaluated on LibriSpeech dev-clean/test-clean or on the British English (ENG) part of the CommonVoice dev/test set. Row 3 indicates a model fine-tuned with the outlined algorithm (K=10). Row 4 first applies the masking algorithm to make basic corrections in the first 3 iterations, and focuses on vowel shifts by only masking vowels in the following iterations (based on the cluster id).

Model	US		ENG		AUS		IND		IRE		SCO	
	dev	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*
HuBERT Large [1]	17.29	18.41	16.79	14.26	10.34	13.63	24.96	25.84	14.02	10.56	15.09	16.12
HuBERT + Adapters [10]	15.09	**15.82**	15.87	12.45	**8.93**	11.97	17.87	18.04	**12.20**	10.54	11.67	12.93
Ours MLM K=5	15.09	16.04	15.48	**11.14**	9.11	11.72	18.18	18.90	12.24	10.63	**10.99**	12.53
Ours MLM K=10	**15.05**	16.05	**15.41**	11.71	8.98	**11.66**	**17.44**	**17.97**	12.31	**10.45**	11.08	**12.35**

Table 4: ASR experiments with WER results for multiple English accents. The proposed models are adapted to the target accent in every column separately. The ASR model is trained on the North-American English (US) split of CommonVoice, and evaluated on the target accent. The accents are abbreviated as US (North-American), ENG (British), AUS (Australian), IND (Indian) and SCO (Scottish).

4.3. Accented Speech Recognition

The main goal of this research is improving the invariance of a pre-trained model towards an accent without having labeled data for that accent. We train an ASR model on North-American English data and evaluate it on the other accents, using the features of the adapted HuBERT models for every accent separately. For the British English accent, Table 3 displays the results for different masking schemes, for an ASR model trained on the US part of CommonVoice (accent mismatch) or on LibriSpeech (accent and dataset mismatch). For one of the results, the MLM focuses on replacing vowel shifts, if there is some labeled source data available (as in 4.2). Table 4 shows the WER results for the experiments on all accents with an ASR model trained on CommonVoice, where the outlined unsupervised mask-and-predict algorithm is applied to cluster groups for K iterations.

5. DISCUSSION

Phone modeling As the PER w.r.t. the US G2P transcription drops, the clusters are closer to the North American pronunciation. Empirically, we see the MLM corrects some vowels and deviant pronunciations, and reduces cluster artifacts by smoothing similar clusters into one group. However, changing too much can be harmful, and there is no real ground truth for cluster correctness. There is a trade-off between smoothing and removing insertions on the one hand, and keeping temporal detail for ASR purposes on the other hand.

Speech recognition The experiments show that for most accents there is an improvement over regular fine-tuning of HuBERT with uncorrected clusters of [10], indicating an improved accent invariance by rescoring the clusters. The idea of modeling the discrete units with a language model on canonical data and correct atypical speech is conceptually simple and applicable to many models. If there is labeled data from the source accent available, the clusters can be mapped to phones and accent-specific characteristics can be targeted (e.g. vowel shifts between British and American English).

6. RELATED WORK

Several works have built on the idea of combining discretised self-supervised learning with a language model [20]. In vq-wav2vec [27, 28], the learned quantiser of a pre-trained wav2vec model is used to generate discrete audio codes, which are then modelled by a BERT model. Afterwards, the continuous BERT features are used for downstream tasks like ASR. There are several differences with our work. First of all, they quantise into 12.5k unique codes, which makes the BERT training more difficult (128 GPU's) and the clusters less interpretable. Furthermore, we use a novel mask-and-decode algorithm so that the rescored clusters can be used to adapt a HuBERT model or for discrete lightweight ASR [29]. Finally, we make adaptations to specifically target the case of accented speech.

Most research on accent robustness has focused on adaptation in multi-accent speech recognition models [30, 31, 32]. Other recent work [33] has investigated codebook regularisation to phonemes in wav2vec 2.0 to improve accent invariance during pre-training without supervision. Finally, [34] uses discriminators to disentangle accent-specific information during auto-encoder pre-training.

7. CONCLUSION

This work improves the accent-invariance of a state-of-the-art HuBERT Large model by correcting the discrete acoustic units towards standard pronunciation, with a simple adaptation of its training scheme. A masked language model is trained on the discrete units of a common accent, and corrects the accented token sequences with an iterative mask-and-decode algorithm. Fine-tuning the pre-trained model with adapters on these corrected clusters improves speech recognition for the target accent, without any labels.

In future work, we plan to explore combining multiple adapters to enable multi-accent training and be resilient against L2 speakers, and experiment with bigger language models. It is also interesting to combine speech and MLM encoders in an end-to-end fashion.

8. REFERENCES

[1] 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*, vol. 29, pp. 3451–3460, 2021.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449–12460.

[3] Sanyuan Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *Proc. ICML*, 2022, pp. 3915–3924.

[5] Yu-An Chung et al., "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. ASRU*. 2021, pp. 244–250, IEEE.

[6] Danni Ma, Neville Ryant, and Mark Liberman, "Probing acoustic representations for phonetic properties," in *Proc. ICASSP*. 2021, pp. 311–315, IEEE.

[7] Wei-Ning Hsu et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech*, 2021, pp. 721–725.

[8] Kaizhi Qian et al., "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *Proc. ICML*, 2022, pp. 18003–18017.

[9] Mark Hallap, Emmanuel Dupoux, and Ewan Dunbar, "Evaluating context-invariance in unsupervised speech representations," in *Proc. Interspeech*, 2023, pp. 2973–2977.

[10] Anshu Bhatia et al., "Don't stop self-supervision: Accent adaptation of speech representations via residual adapters," in *Proc. Interspeech*, 2023, pp. 3362–3366.

[11] John C. Wells, *Accents of English*, vol. 2, Cambridge University Press, 1982.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*. 2015, pp. 5206–5210, IEEE.

[14] Rosana Ardila et al., "Common Voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020, pp. 4218–4222.

[15] Ashish Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017.

[16] Mandar Joshi et al., "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *Proc. NeurIPS-EMC²*, 2019.

[18] Neil Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. ICML*, 2019.

[19] Steven Vander Eeckt and Hugo Van Hamme, "Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition," in *Proc. ICASSP*. 2023, IEEE.

[20] Kushal Lakhotia et al., "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[21] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, "Unsupervised speech recognition," in *Proc. NeurIPS*, 2021, pp. 27826–27839.

[22] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," in *Proc. EMNLP-IJCNLP*, 2019, pp. 6112–6121.

[23] Yosuke Higuchi, Shinji Watanabe, Nanxin Chen, Tetsuji Ogawa, and Tetsunori Kobayashi, "Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict," in *Proc. Interspeech*, 2020, pp. 3655–3659.

[24] Yosuke Higuchi et al., "BERT meets CTC: New formulation of end-to-end speech recognition with pre-trained masked language model," in *Proc. EMNLP*, 2022, pp. 5486–5503.

[25] Shu wen Yang et al., "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[26] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, p. 369–376.

[27] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. ICLR*, 2020.

[28] Alexei Baevski and Abdelrahman Mohamed, "Effectiveness of self-supervised pre-training for ASR," in *Proc. ICASSP*. 2020, pp. 7694–7698, IEEE.

[29] Xuankai Chang, Brian Yan, Yuya Fujita, Takashi Maekaku, and Shinji Watanabe, "Exploration of efficient end-to-end ASR using discretized input from self-supervised learning," in *Proc. Interspeech*, 2023.

[30] Xian Shi et al., "The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," in *Proc. ICASSP*. 2021, pp. 6918–6922, IEEE.

[31] Yanmin Qian, Xun Gong, and Houjun Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[32] Song Li et al., "End-to-end multi-accent speech recognition with unsupervised accent modelling," in *Proc. ICASSP*. 2021, pp. 6418–6422, IEEE.

[33] Chendong Zhao et al., "Learning invariant representation and risk minimized for unsupervised accent domain adaptation," in *Proc. SLT*. 2023, pp. 509–516, IEEE.

[34] Yi-Chen Chen et al., "AIPNet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition," in *Proc. ICASSP*. 2020, pp. 6979–6983, IEEE.