PSEUDO-LABEL BASED SUPERVISED CONTRASTIVE LOSS FOR ROBUST SPEECH REPRESENTATIONS

Varun Krishna, Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science, Bangalore.
varunkrishna@iisc.ac.in, sriramg@iisc.ac.in

ABSTRACT

The self supervised learning (SSL) of speech, with discrete tokenization (pseudo-labels), while illustrating performance improvements in low-resource speech recognition, has faced challenges in achieving context invariant and noise robust representations. In this paper, we propose a self-supervised framework based on contrastive loss of the pseudo-labels, obtained from an offline k-means quantizer (tokenizer). We refer to the proposed setting as *pseudo-con*. The pseudo-con loss, within a batch of training, allows the model to cluster the instances of the same pseudo-label while separating the instances of a different pseudo-label. The proposed pseudo-con loss can also be combined with the cross entropy loss, commonly used in self-supervised learning schemes. We demonstrate the effectiveness of the pseudo-con loss applied for various SSL techniques, like hidden unit bidirectional encoder representations from transformers (HuBERT), best random quantizer (BEST-RQ) and hidden unit clustering (HUC). Our evaluations using the proposed pseudo-con framework achieves state of art results on various sub-tasks of ZeroSpeech 2021 challenge as well as on the context invariance benchmarks. Further, we show significant performance improvements over existing SSL approaches on the TIMIT phoneme recognition task as well as the Librispeech (100h) ASR experiments.

Index Terms— Self-supervised pre-training, Supervised Contrastive Loss, Context invariance, ZeroSpeech

1. INTRODUCTION

Self-supervised learning (SSL), an approach where a pretext task is defined to form pseudo-labels from raw data, has shown to be a promising pre-training framework for various domains like natural language processing (NLP) [1], computer vision [2] and audio processing [3]. In these settings, particularly on audio and text, the framework involves tokenizing (discretizing) the raw data in a pseudo-label space, masking portions of the data, and designing a transformer based model architecture to predict the sequence of labels that are masked in the input data. The SSL model acts as a unified pre-trained model that can be fine-tuned for a range of downstream tasks, including those which were not part of the design [4].

In speech processing, self-supervised pre-training in the early form used a single convolutional encoder to generate different acoustic representations using the raw audio as input, called the problem agnostic speech encoder (PASE) [5]. An autoregressive prediction of the speech representations was explored by Chung et al. [6]. The investigation of contrastive loss functions for SSL was proposed by [7], where the task was designed to predict the future frames of the raw audio in a constrastive fashion. The wav2vec series of models [8, 9], using a learned vector quantization module, enables the learning of representations with masked inputs and transformer based architectures. These models also convert the continuous audio signal into discrete tokens, a step called tokenization, that is shown to be beneficial for self-supervision based learning [10]. The subsequent works, like hidden unit bidirectional encoder representations from transformer (HuBERT) [3], using hidden layer representations and best random-quantizer (BEST-RQ) [11], using random quantizer on the input spectrogram, extend this idea using to improve the SSL modeling. The SSL approaches have been effective for a variety of speech tasks such as ASR, speaker identification, and spoken language modeling [8, 3, 12]. A overview of the various SSL models for speech is given in Mohamed et al. [13].

In recent years, there has been growing interest in analyzing and benchmarking SSL approaches [12, 14, 15]. In the “clean” environments, representations from pre-trained acoustic models appear to be equivalent to phonemes or phoneme states, as reported by Ma et al. [16]. The work by Hallap et al. [14] showed that representations learned by pre-trained acoustic models are sensitive to changes in the phonetic context. This may indicate that the representations learned by pre-trained acoustic models are more allophonic than phonemic. A separate study by Gat et al. [17] highlighted that SSL models are highly susceptible to noise and other distortions. Prior works [18, 19, 17, 20] leverage data augmentation techniques coupled with the idea of consistency regularization [21] or use speaker disentanglement techniques [22] to achieve robustness. While works reported in [22, 19, 17] mainly address the issue of robustness to noise and other perturbations, very little prior work has focused on the issue of context invariance of representations. This paper proposes our attempt in learning context invariant and robust speech representations.

To achieve context invariant representations of speech, it is necessary to cluster the representations belonging to the same phonetic class while disregarding speaker, noise and accent variations. The supervised contrastive loss, which has shown encouraging results in NLP and vision domains [23, 24, 25, 26, 27], on supervised labeling tasks, offers a potential choice for deriving context invariant representations of speech. The key difference in our proposal is the lack of labels in the SSL framework, thus enabling the investigation of discrete tokens (pseudo-labels) for the contrastive loss.

We propose the pseudo-con framework, where the first step is the discretization of the raw audio to generate the pseudo-labels. The subsequent step is the learning of the model that embeds the within pseudo-label representations in a clustered space. The pseudo-con loss, which is defined as the supervised contrastive loss applied on pseudo-labels, is used as the batch-level objective function, where the similarity between representations corresponding to the same pseudo-label are enhanced, while discriminatively separating the representations from distinctive pseudo-labels. The pseudo-con is applicable for predicting the cluster labels directly (for example, applications in SSL methods like hidden unit clustering (HUC) [28])

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

or in a masked language modeling (MLM) setting, (for example, those used in HuBERT [3] and BEST-RQ [11]).

Experiments show that models pre-trained with contrastive loss achieve better phonetic context invariance and robustness compared to prior works. The representations also show significant improvements in ASR and zero resource spoken language modeling tasks, like those defined as part of the ZeroSpeech 2021 challenge [12].

The key contributions from this paper are,

- Proposing the application of the supervised contrastive loss to pseudo-labeling task in a self-supervised setting. We call this loss as pseudo-con loss.
- Incorporating the pseudo-con loss on diverse settings of predicting the frame-level pseudo-label clusters in HUC [28] as well as in predicting the pseudo-labels for the masked audio regions in HuBERT [3] and BEST-RQ [11].
- Experimental validation on various downstream tasks like ZeroSpeech language modeling tasks [12], phoneme recognition and ASR tasks.
- Detailed analysis on the context invariance and noise robustness attributes of the proposed pseudo-con representations.

2. RELATED PRIOR WORK

The most common SSL frameworks can be broadly categorized based on the type of objective functions used in learning the models. The broad category of such approaches are,

1. Generative loss - audio wav2vec [29], PASE [5], autoregressive predictive coding (APC) [6], transformer encoder representations of audio (TERA) [30], non-autoregressive predictive coding (NPC) [31].
2. Contrastive loss - The wav2vec series of models, wav2vec-vq [32], wav2vec2.0 [8], wav2vec-BERT [9], and Speech SIMCLR [33].
3. Predictive loss - HuBERT [3], WavLM [34], and BEST-RQ [11].

For deriving representations that are robust to noise perturbations and speaker variations, the key directions pursued are a) with the use of the semantic content preserving audio augmentations [19, 17, 35] like pitch, reverberation and additive noise perturbation, and b) with a combination of augmentation and speaker information disentanglement [22]. We describe some of these approaches below.

2.1. Augmentation Invariance

The work proposed by Gat et al. [17] investigated the idea of consistency regularization [21, 36]. In this work, the given clean speech signal is forward passed through a model f, followed by a k-means quantizer, to obtain discrete sequence S_1. In parallel, the augmented signal is passed through same model f and a learnable multi-layer perceptron based quantizer e is used to obtain sequence S_2. Both sequences S_1 and S_2 are de-duplicated (for example a sequence $11, 11, 12, 12, 13$ is converted to $11, 12, 13$). The connectionist temporal classification (CTC) [37] loss between the sequences S_1 and S_2 is minimized to train the quantizer. The cascade of SSL model f and quantizer e outputs noise invariant representation.

2.2. CCC wav2vec 2.0

The clustering aided contrastive self-supervised representation learning (CCC-wav2vec) [19] introduces clustering based negative sampling module and an auxiliary cross-contrastive loss over the wav2vec 2.0 model [8]. Further it leverages data augmentation to achieve robustness.

2.3. Contentvec

The work termed contentvec [22] is based on masked prediction paradigm of HuBERT [3] and is made of 3 components - student speech representation network $f(.)$, predictor $p(.)$ and frozen teacher label generator $g(.)$. The model tries to learn robust representations using combination of speaker disentanglement techniques and contrastive loss minimization between the augmented speech samples. The functions $f(.)$ and $g(.)$ are pre-trained HuBERT models. The speaker information is disentangled from student network using SIMCLR [38] style of training. The student network minimizes the contrastive loss between the representation of masked copies X_1 and X_2, for the clean speech sample X.

2.4. HUC

Hidden unit clustering (HUC) [28] is a model trained using the cross-entropy loss on psuedo-labels. The HUC achieves speaker invariance and robustness to noise by processing the representations of CPC [7] model. An utterance level mean normalization of representations, before the k-means quantizer, generates pseudo labels that are speaker invariant. The supervision from these processed pseudo labels produces representations that are robust to speaker variations and noise perturbation.

3. PROPOSED PSUEDO-CON FRAMEWORK

We propose to use the supervised contrastive loss proposed by Khosla et al. [23] as an objective function in the models that use masked language model framework such as HuBERT and BEST-RQ or with models that directly predict the frame-level tokens, HUC [28].

3.1. Pseudo-Con loss

We use the pseudo labels obtained by quantizing the dense representation from the pre-trained self-supervised models into K classes, using an offline k-means step to provide supervision. These discretized units are shown to have good correlation with phoneme-like units [10] and are used as proxy for frame level ground truth phonetic transcription.

3.2. Architecture

We evaluated the pseudo-supervised contrastive loss on the HuBERT [3], BEST-RQ [11], and HUC [28] models without making any changes to the models' architectures. These models typically consist of a trainable convolutional feature extractor (with raw audio input) or mel-spectrogram feature extractor, followed by a transformer or LSTM feature encoder. The encoder output $\mathbf{H}$, is then fed to a linear layer with soft-max non-linearity.

3.3. Loss function

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_T\}$ denote the windowed audio samples (or audio spectrogram), where the windowing is typically done at 20ms sampling. Let f_{enc} denote the encoder function realized by the SSL model, typically using a cascade of convolution and transformer layers. Let the output of the encoder be denoted as $f_{enc}(\boldsymbol{x}_t) = \boldsymbol{h}_t$. The encoded representation of the given audio utterance is then given by $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, ...\boldsymbol{h}_T\}$.

In all the settings considered in this work, the audio is also tokenized to a discrete label set $\mathcal{Z}$, with number of pseudo-labels (k-means clusters) denoted as K. Let $\{\boldsymbol{z}_1, \boldsymbol{z}_2, ...\boldsymbol{z}_T\}$ denote the one-hot encoded pseudo-label sequence for the given audio recording, and let $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ...\boldsymbol{y}_T\}$ denote the logits (linear layer with softmax) transformation of the encoded outputs $\{\boldsymbol{h}_1, ..., \boldsymbol{h}_T\}$.

In the case of masked inputs, a random binary mask $m_1, ..., m_T$ is applied on the input, where $m_t = 1$ indicates a masked window of the given audio. If $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{H}}$, denote the masked input, and encoder output for the masked input, and if $\{\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_T\}$ denote the logits output for the masked input, the standard learning objective used in masked language modeling (MLM) based SSL models (like BEST-RQ [11]) is given by,

$$\mathcal{L}_{\mathbb{CE}} = -\sum_{b=1}^{B}\sum_{t=1}^{T_b}\sum_{k=1}^{K} m_t^b z_{t,k}^b \log(\hat{y}_{t,k}^b), \tag{1}$$

where $z_{t,k}^b$, $\hat{y}_{t,k}^b$ denotes the k-th index of the one-hot vector $\boldsymbol{z}_t^b$ and the logits output $\hat{y}_t$ respectively, for the b-th utterance in the batch, containing B samples. Here, T_b denotes the length of the b-th utterance.

The pseudo-con loss for this setting is given by,

$$\mathcal{L}_{\mathbb{PC}} = \sum_{b=1}^{B}\sum_{t=1}^{T_b} \frac{-m_t^b}{|P(t,b)|} \sum_{p \in P(t,b)} \frac{exp(\hat{\boldsymbol{y}}_t^b \cdot \hat{\boldsymbol{y}}_p/\tau)}{\sum_{a \in A(t,b)} exp(\hat{\boldsymbol{y}}_t^b \cdot \hat{\boldsymbol{y}}_a/\tau)}, \tag{2}$$

where τ is the temperature parameter, $P(t,b)$ denotes the set of all positive instances (over all time windows $t = 1...T$ and all samples in the batch $b = 1...B$) which satisfy $\boldsymbol{z}_t^b = \boldsymbol{z}_p$, while $A(t,b)$ denotes the set of all negative instances which satisfy $\boldsymbol{z}_t^b \neq \boldsymbol{z}_a$, i.e.,

$$\begin{aligned} P(t,b) &= \{p : \boldsymbol{z}_t^b = \boldsymbol{z}_p \; \forall \, p = \{\{t\}_1^{T_b}\}_{b=1}^{B}\}\}, \\ A(t,b) &= \{a : \boldsymbol{z}_t^b \neq \boldsymbol{z}_a \; \forall \, a = \{\{t\}_1^{T_b}\}_{b=1}^{B}\}\}, \end{aligned} \tag{3}$$

$|P(t,b)|$ denotes the cardinality of the set $P(t,b)$, and $\cdot$ denotes the dot product.

A modified version of the cross-entropy loss, commonly used in BERT settings (like HuBERT [3]), $\mathcal{L}_{\mathbb{CE}-\mathbb{EMB}}$ is,

$$\mathcal{L}_{\mathbb{CE}-\mathbb{EMB}} = -\sum_{t,b} m_t^b log\left(\boldsymbol{p}_{emb}(\boldsymbol{z}_t^b|\hat{\boldsymbol{y}}_t^b)\right) \tag{4}$$

posterior vector $\boldsymbol{p}_{emb}(.)$ is defined as,

$$\boldsymbol{p}_{emb}(z_{t,k}|\hat{y}_{t,k}) = \frac{exp(sim(\hat{\boldsymbol{y}}_t, \boldsymbol{e}_k)/\tau)}{\sum_{k'=1}^{K} exp(sim(\hat{\boldsymbol{y}}_t, \boldsymbol{e}_{k'}))/\tau)}. \tag{5}$$

Here, $\boldsymbol{e}_k$ is the learnable code embedding corresponding to pseudo-label k, $sim(.,.)$ computes cosine similarity between the two vectors and τ is the temperature factor that is set to 0.1.

For SSL models without masking (like HUC [28]), the input $\boldsymbol{X}$ is fed without any masking, and the cross-entropy/pseudo-con loss is computed on all time windows.

The joint loss, combining the pseudo-con loss and the cross-entropy loss is then given by,

$$\mathcal{L}_{\mathbb{TOT}} = \alpha\mathcal{L}_{\mathbb{PC}} + (1-\alpha)\mathcal{L}_{\mathbb{CE}-\mathbb{EMB}} \tag{6}$$

α is the weighting parameter, $\alpha \in [0,1]$. For the implementation with HuBERT/HUC/BEST-RQ model, the total loss (Equation (6)) uses $\mathcal{L}_{\mathbb{CE}}$ instead of the $\mathcal{L}_{\mathbb{CE}-\mathbb{EMB}}$. The setting $\alpha = 0$ reverts the pre-training similar to that of vanilla models, while $\alpha = 1$ ignores the cross-entropy loss term and uses the pseudo-con loss alone.

4. EXPERIMENTS

4.1. Pre-training data

In our experimental comparison, all the models are pre-trained on the same dataset of Librispeech [39] 960 hours. This dataset consists of English read speech (audio books) from 1000 speakers. The speech utterances, sampled at 16kHz, are of duration 3-7 seconds.

4.2. SSL Implementation

We pre-train the HuBERT Base [3], BEST-RQ [11] and HUC [28] models by optimizing the loss function given by Equation (6). Since HuBERT model uses computationally expensive iterative clustering and pre-training, we use pseudo labels obtained by quantizing the 12-th layer output of HuBERT-base [3] model with 200 clusters using k-means algorithm for pre-training. We call the HuBERT model trained with pseudo-con loss as HuBERT-pseudo-con In this setting, pseudo-con objective with $\alpha = 0$ corresponds to an iterative continuation of HuBERT Base pre-training with random initialization of the model architecture and with the cross entropy based MLM loss.

4.3. ZeroSpeech 2021 Task

ZeroSpeech 2021 challenge [12] defines metrics to measure phonetic, lexical, syntactic and semantic properties of the representations.

Phonetic (ABX): For a triplet of tri-phone words, **A**,**B** and **X**, where **A** and **B** differ in the center phoneme, while **A** and **X** are two different instances of the same word, the ABX metric [41] computes the fraction of instances when **A** and **X** are more distant than **A** and **B**. The angular distance averaged along dynamic time warped (DTW) path is used to compute the distance between word utterances.

Lexical: The sWUGGY "spot-the-word" [42] measures the ability of the model to identify a legitimate word. Given a set of word/non-word pairs, sWUGGY is computed as the fraction of the pairs where the likelihood of the legitimate word is higher than non-word.

Syntactic: The sBLIMP metric [43] measures the ability of model to identify grammatically correct sentences. It is computed as the fraction of instances where the likelihood of grammatically correct sentence is higher than an incorrect one.

Semantic: The sSIMI metric is used to assess the lexical semantics and is computed as the Spearman's rank correlation coefficient ρ between the semantic similarity scores computed for representations from pair of words given by the model and the human scores in the dataset.

A lower value of ABX error is prefered, while higher values are prefered for the other three metrics. The encoder representations from the self-supervised learning (SSL) model are used to compute the ABX metric. For the rest of the metrics, the challenge [12] puts forth a cascade of models - SSL feature extractor, followed by a k-means quantizer and a BERT language model. The k-means quantizer trained on the representations from feature extractor generates

Table 1: Results for various models on ZeroSpeech 2021 challenge dataset.

Model	ABX↓				sWUGGY↑	sBLIMP↑	sSIMI↑	
	Clean Within	Clean Across	Other Within	Other Across			librispeech	synthetic
BERT-small language model [12]								
Chorowski et al. [40]	2.95	3.60	4.50	6.99	74.40	52.97	4.60	-7.75
HUC [28]	2.92	3.50	4.47	6.95	74.97	55.01	7.94	**5.47**
HuBERT Base [3]	3.40	4.16	4.47	6.97	70.11	54.54	3.31	2.17
wav2vec 2.0 Base [8]	4.51	5.38	5.47	7.64	66.73	52.44	6.01	-0.87
CCC-wav2vec [19]	4.65	5.52	5.75	8.17	65.01	53.26	2.44	0.70
ContentVec [22]	2.98	3.51	3.70	5.16	73.47	54.86	5.56	1.62
HuBERT-pseudo-con ($\alpha = 0$)	3.22	3.85	4.25	6.11	71.81	54.05	4.44	1.53
HuBERT-pseudo-con ($\alpha = 0.5$)	**2.42**	**2.84**	**3.14**	**4.87**	**77.30**	**62.94**	**9.42**	2.06
HuBERT-pseudo-con ($\alpha = 1$)	4.45	5.52	6.17	8.82	60.95	52.32	2.00	-2.90
BERT-Big Language model [12]								
Nguyen et al. [10]	3.26	3.81	4.00	5.91	**83.29**	61.93	**9.73**	2.48

Table 2: ABX (%) error for HUC and BEST-RQ on ZeroSpeech 2021 challenge dataset with different values of α

Model	ABX↓			
	clean within	clean across	other within	other across
HUC ($\alpha = 0$)	2.92	3.50	4.47	6.95
HUC ($\alpha = 0.5$)	**2.71**	**3.16**	**4.00**	**6.23**
HUC ($\alpha = 1$)	4.10	4.95	5.75	8.36
BEST-RQ ($\alpha = 0$)	4.63	5.22	5.48	8.17
BEST-RQ ($\alpha = 0.5$)	**3.87**	**4.96**	**5.10**	**7.05**
BEST-RQ ($\alpha = 1$)	5.71	7.90	8.62	10.63

a second level of discretization of the utterance. During evaluation, likelihood scores computed using the BERT language model for the discretized sequences are used for language modeling tasks. Our experiments use k-means quantizer with $K = 200$ to discretize the utterance and BERT-small [12] language model is trained on 960 hours of LibriSpeech data for the language modeling tasks.

4.3.1. Results

The results are reported in Table 1. The first set of rows, report the results for various baseline comparisons, while the subsequent set of rows report the results for the proposed pseudo-con applied on the HuBERT framework. As seen in these results, the HuBERT-pseudo-con improves over all the baseline systems compared on the ABX tasks, syntactic, lexical and Librispeech semantic tasks. On ABX and lexical tasks, the results reported in this Table also constitute the best published results[1]. The choice of $\alpha = 0.5$, with equal weighting of the cross-entropy and the pseudo-con loss yields the best performance. We also report the ABX results for the HUC [28] and BEST-RQ models [11] in Table 2. The pseudo-con loss is seen to improve both the BEST-RQ and HUC based representations as well. Comparing Table 1 and Table 2, we observe that the best ABX result is achieved by the HuBERT-pseudo-con model, with $\alpha = 0.5$.

Table 3: Word error rate (WER) on test and dev set of LibriSpeech ($100h$ setup) of various models. No language model was used for CTC decoding. PER column reports the phoneme error rate of various models on the test split of TIMIT dataset

Model	ASR (WER↓)				PER↓
	dev clean	dev other	test clean	test other	
HuBERT Base [3]	6.4	14.0	6.6	13.7	13.8
wav2vec 2.0 Base [8]	6.2	14.2	6.2	13.5	14.0
CCC-wav2vec [19]	6.1	13.4	6.4	13.1	13.1
ContentVec [22]	6.1	13.7	6.3	13.0	12.8
HuBERT-pseudo-con ($\alpha = 0$)	6.2	13.9	6.3	13.1	13.1
HuBERT-pseudo-con ($\alpha = .5$)	**5.2**	**12.8**	**5.3**	**12.6**	**11.8**
HuBERT-pseudo-con ($\alpha = 1$)	7.5	17.1	7.8	16.7	19.5
HUC ($\alpha = 0$)	11.6	19.8	10.9	20.0	20.6
HUC ($\alpha = .5$)	**9.1**	**18.3**	**9.7**	**18.4**	**19.7**
HUC ($\alpha = 1$)	16.7	23.0	13.8	23.1	27.8
BEST-RQ ($\alpha = 0$)	6.4	13.9	6.7	13.9	15.4
BEST-RQ ($\alpha = .5$)	**5.9**	**13.0**	**6.0**	**12.8**	**14.6**
BEST-RQ ($\alpha = 1$)	10.1	19.4	9.1	19.3	22.1

4.4. TIMIT Phoneme Recognition Task

The TIMIT [44] corpus consists of 5 hours of English read speech sampled at 16kHz. The phonetic transcription for each utterance is manually generated. The dataset contains recordings from 630 speakers belonging to 8 different dialects of American English. Training split consists of 3 hours of data with 3696 utterances. The evaluation set consists of 1344 utterances. For the phoneme recognition task we fine tune the SSL models on 3 hours of training data from the TIMIT dataset. The model is optimized after affixing a linear layer using CTC loss [37] on the encoder representations. During training, the convolutional feature extractors are kept frozen. A learning rate of $1\mathrm{e} - 5$ is used and the batch consists of 4 utterances. The phoneme error rate (PER) is reported on the test set of TIMIT [44] and this is shown in Table 3. As seen in Table, among the various baselines compared, the contentVec [22] gives the best

[1] `https://zerospeech.com/challenge_archive/2021/06_results/`

Table 4: Modified Levenshtein distance (mean and standard error) measured on LibriSpeech test-split averaged over 5 runs. The system HuBERT-pseudo-con is reported with value of α in braces.

Model	**Transformation**		
	Pitch↓	**Noise**↓	**Revb.**↓
HuBERT [3]	233.1±0.5	173.6±2.6	183.7±0.2
wav2vec 2.0 [8]	378.3±0.9	316.8±1.9	353.9±0.6
CCC-wav2vec [19]	176.4±0.1	147.6±1.2	139.7±0.1
ContentVec [22]	106.5±0.4	111.9±2.1	114.8±0.2
HuBERT-pseudo-con (0)	245.6±0.6	170.0±2.3	186.9±0.2
HuBERT-pseudo-con (0.5)	**94.1**±**0.1**	**89.7**±**0.8**	**90.5**±**0.1**
HuBERT-pseudo-con (1)	116.2±0.1	108.6±0.0	109.3±0.3
HUC ($\alpha = 0$)	182.1±0.4	145.8±0.7	148.6±0.4
HUC ($\alpha = 0.5$)	**87.3**±**0.1**	**81.6**±**0.3**	**84.8**±**0.2**
HUC ($\alpha = 1$)	104.0±0.6	99.8±0.5	100.4±0.1
BEST-RQ ($\alpha = 0$)	236.4±0.8	203.1±0.4	219.7±0.2
BERT-RQ ($\alpha = 0.5$)	**114.8**±**0.1**	**118.0**±**0.1**	**119.3**±**0.2**
BEST-RQ ($\alpha = 1$)	193.8±0.0	118.0±0.3	119.3±0.1

results. The proposed HuBERT-pseudo-con provides a significant improvement in PER (average relative improvement of 8% over the best baseline system). The other SSL approaches also illustrate gains using the pseudo-con, while the overall WER/PER of the HuBERT-pseudo-con is seen to be the best.

4.5. ASR Task

All the pre-trained models are fine-tuned on LibriSpeech 100h split using CTC loss. During fine-tuning the transformer encoder layers and the randomly initialized softmax layer are made trainable. The CTC target vocabulary consists of 26 English alphabets, a space/silence token, an apostrophe and a special CTC blank symbol. The ASR fine-tuning is done using *Torch-audio ASR* pipeline[2]. The default hyper-parameter settings of the toolbox are used for our experiments. CTC decoding was done without language model. The word error rates (WER) are reported on development and test split of LibriSpeech [39] in Table 3. Similar to the phoneme recognition experiments, the proposed HuBERT-pseudo-con ($\alpha = 0.5$) improves over all the baseline systems on this task. The improvement over the HuBERT model is substantial (average relative improvements of 11.8% in WER).

4.6. Robustness Measure

We follow the approach proposed by Gat et al. [17] to measure the robustness of the pre-trained models to noise and other semantically invariant perturbations. Given a speech sample $x \in \mathbb{R}^{\mathbb{T}}$, non-semantic perturbations $g : R^T \mapsto R^T$, such as pitch, reverberation or additive noise is applied to obtain x'. Then, x and x' are fed to the pre-trained model $f : R^T \mapsto R^{T'}$. The encoder representations $f_{enc}(x)$ and $f_{enc}(x')$ are quantized using k-means quantizer $E : R^{T'} \mapsto \{1....K\}^{T'}$, which was trained on representations from “clean” data. The modified Levenshtein distance [45] $UED_{\mathbb{D}}$ between the deduplicated discretized sequences corresponding to x and x' are used to compute the robustness measure.

Table 5: Variance ratio (VR) measure, obtained as the ratio of the inter-class dispersion to the intra-class dispersion, measured on phonetic alignments from LibriSpeech 100h data.

Model	**VR**↑
HuBERT Base [3]	95448
wav2vec 2.0 Base [8]	61979
CCC-wav2vec [19]	101539
ContentVec [22]	108002
HuBERT-pseudo-con ($\alpha = 0$)	94234
HuBERT-pseudo-con ($\alpha = 0.5$)	**157029**
HuBERT-pseudo-con ($\alpha = 1$)	133678
HUC ($\alpha = 0$)	102563
HUC ($\alpha = 0.5$)	**181458**
HUC ($\alpha = 1$)	110842
BEST-RQ ($\alpha = 0$)	87195
BEST-RQ ($\alpha = 0.5$)	**105774**
BEST-RQ ($\alpha = 1$)	99361

Table 6: Performance of various models on sub-tasks defined as part of the NOSS benchmark [15].

Models	**Task**		
	Spkr ID	**Lang. ID**	**Emotion Rec.**
HuBERT Base [3]	80.96	99.51	**81.31**
wav2vec 2.0 Base [8]	79.18	96.56	78.67
CCC-wav2vec [19]	76.13	95.20	78.85
ContentVec [22]	39.79	98.30	64.16
HuBERT-pseudo-con ($\alpha = 0$)	**81.19**	99.50	**80.88**
HuBERT-pseudo-con ($\alpha = 0.5$)	51.00	**99.70**	72.41
HuBERT-pseudo-con ($\alpha = 1$)	47.91	79.87	66.20
HUC ($\alpha = 0$)	20.11	82.13	**68.26**
HUC ($\alpha = 0.5$)	17.92	**86.44**	65.39
HUC ($\alpha = 1$)	**22.30**	78.82	58.58
BEST-RQ ($\alpha = 0$)	**70.66**	91.97	**76.01**
BEST-RQ ($\alpha = 0.5$)	61.94	**94.71**	72.47
BEST-RQ ($\alpha = 1$)	58.10	88.49	69.51

$$UED_{\mathbb{D}} = \sum_{\mathbf{x} \in \mathbb{D}} \frac{1}{T'} LEV((E \circ f)(x), (E \circ f \circ g)(x)) \qquad (7)$$

Here, $\mathbb{D}$ is evaluation data and LEV is Levenshtein distance. For a given model, representations for LibriSpeech 100h subset is used to train k-means quantizer. The value of K is set to 200. The transformations applied are pitch perturbation, uniformly sampled between scales of -300 to 300, reverberation by uniformly sampling room responses with scale between 0 to 100, and additive noise, sampled from MUSAN dataset [46] with SNR between 5 to 15 dB. The transformations are implemented using Wav-Augment [18] toolbox. The results are reported in Table 4. As seen in these results, the proposed HuBERT-pseudo-con and the HUC-pseudo-con achieve the best UED distance, indicating that the representations trained with the pseudo-con loss are robust to audio transformations. While both $\alpha = 0.5$ and $\alpha = 1$ improve the UED metric, the best results are observed for equal contribution of pseudo-con loss and the cross-entropy loss ($\alpha = 0.5$).

[2] https://github.com/pytorch/audio/tree/main/examples/hubert

Table 7: ABX context-independence evaluation [14]. Lower scores are better. Here, "W/in-ctx" and "W/out-ctx" denotes within context and without context respectively, for clean and other data splits of the context-invariance benchmark [14].

Model	Clean				Other			
	within-speaker		across-speaker		within-speaker		across-speaker	
	W/in-ctx↓	W/out-ctx↓	W/in-ctx↓	W/out-ctx↓	W/in-ctx↓	W/out-ctx↓	W/in-ctx↓	W/out-ctx↓
HuBERT Base [3]	2.05	7.60	2.71	8.01	4.38	10.44	6.57	11.26
wav2vec 2.0 Base [8]	2.33	10.33	2.94	10.87	3.95	12.23	6.23	13.53
CCC-wav2vec Base [19]	2.59	10.27	3.42	10.64	4.32	12.15	6.76	13.34
ContentVec [22]	1.24	6.03	1.70	6.12	2.97	7.49	**4.22**	7.89
Nguyen et al. [10]	1.56	7.26	2.13	8.04	3.08	8.64	4.78	10.09
HuBERT-Iter ($\alpha = 0$)	1.86	7.08	2.22	7.21	3.62	9.18	5.49	9.73
HuBERT-Iter ($\alpha = 0.5$)	**1.21**	**5.66**	**1.64**	**5.83**	**2.82**	**7.11**	**4.35**	**7.55**
HuBERT-Iter ($\alpha = 1$)	3.11	8.62	3.90	8.87	4.87	12.95	6.62	14.02
HUC ($\alpha = 0$)	1.92	7.09	2.44	7.06	3.72	10.06	6.12	10.94
HUC ($\alpha = 0.5$)	**1.86**	**6.98**	**2.22**	**6.87**	**3.62**	**9.18**	**5.49**	**9.73**
HUC ($\alpha = 1$)	2.54	8.24	3.21	8.44	4.73	10.52	6.95	11.21
BEST-RQ ($\alpha = 0$)	2.46	7.27	3.29	7.42	4.95	9.37	6.85	9.84
BEST-RQ ($\alpha = 0.5$)	**2.20**	**7.16**	**2.86**	**7.01**	**4.31**	**9.11**	**6.54**	**9.45**
BEST-RQ ($\alpha = 1$)	3.02	9.88	3.98	10.48	5.04	12.43	8.16	14.02

4.7. Cluster Compactness Measure

Representations that are more compact within a phonetic class tend to be more robust to perturbations in the input raw audio. To measure the cluster compactness, we use the variance ratio measure [47]. Given a set of speech utterances and corresponding phonetic alignments (using ASR force alignment), we compute mean representations for each phoneme class. Then, the following are computed, i) the sum of distances of representations from the mean representations of their corresponding phoneme, called intra-class dispersion and, ii) the sum of pairwise distance between mean representations of different classes, called the inter-class dispersion. The variance ratio is the ratio of inter-class dispersion to the intra-class dispersion. A higher value of VR implies better compactness. Table 5 reports the VR measure on LibriSpeech 100h split for various models. Similar to the robustness measures, we observe that the HuBERT model and the HUC model with the proposed pseudo-con objective provides the best VR measures for this evaluation.

4.8. Non-semantic Speech Tasks

The non-semantic evaluations are part of the NOSS benchmark defined by Shor et al. [15]. We investigate the effectiveness of the pseudo-con loss for emotion recognition, speaker and language identification tasks (Table 6). For all the tasks, we derive the utterance level means of the representations and the SSL model is frozen. A linear SVM is trained on the pooled representations for classification tasks. The tasks that were explored here are our own implementations of the sub-tasks in the non-semantic speech (NOSS) benchmark [15].

The VoxCeleb-1 [48] dataset, which contains 1251 speakers, was used for speaker identification tasks. The accuracy score on the test split of the dataset was used as the evaluation metric. The VoxForge [49] dataset, which contains 6 languages, was used for language identification tasks. The CREMA-D [50] dataset was used for emotion recognition tasks. The 5-fold cross-validation score was used to evaluate all the systems for emotion recognition and language identification tasks.

The results reported in the Table 6 show that inclusion of pseudo-con loss only favours language identification tasks, while it deteriorates the performance in other tasks. The results indicate that pseudo-con loss may lead to phonetically rich representations, while compromising on speaker and emotion encoding.

4.9. Context Invariance

We use the **ABX-LS** proposed by Hallep et al. [14] to measure the phonetic context invariance of representations. ABX-LS extracts phonemes in isolation, rather than triphone tokens. In the within-context condition, the phonemes that precede and follow the target phoneme are the same for all three stimuli (A, B, and X). In the without-context condition, there are no such constraints. Computing the ABX metric on two of these conditions measures the invariance to changes in context. The results are reported in Table 7.

The proposed HuBERT-pseudo-con with $\alpha = 0.5$ achieves the best performances. In line with the other semantic tasks, the improvement in context-invariance is seen with equal weighting of cross-entropy and pseudo-con loss ($\alpha = 0.5$).

5. SUMMARY

This paper presents our work on self-supervised framework using supervised contrastive loss. We propose pseudo-con loss that leverages the pseudo labels to minimize the supervised contrastive loss. The improved within-cluster merging with discriminative separation across clusters, a property derived with the pseudo-label based contrastive loss, allows the proposed model to learn representations that are linguistically grounded. Experiments show that pseudo-con loss is a simple and effective auxiliary module that can be easily integrated into SSL models which use masked language model/hidden unit clustering frameworks.

The downstream evaluations show that inclusion of pseudo-con loss improves the model performances for several semantic tasks (like ASR, phoneme recognition and ZeroSpeech tasks), while achieving robustness and context invariance properties (measured through different settings). It is also noteworthy that both ContentVec [22] (best baseline for all tasks except for non-semantic speech tasks) and the proposed HuBERT-pseudo-con are built on HuBERT base [3] model. However, with the addition of pseudo-con loss, the proposed framework is seen to outperform ContentVec [22] on all the semantic task evaluations.

6. REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[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*, vol. 29, pp. 3451–3460, 2021.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[5] Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[6] Yu-An Chung and James Glass, "Generative pre-training for speech with autoregressive predictive coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3497–3501.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

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

[10] Tu Anh Nguyen, Benoit Sagot, and Emmanuel Dupoux, "Are discrete units necessary for spoken language modeling?," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1415–1423, 2022.

[11] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3915–3924.

[12] Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *arXiv preprint arXiv:2011.11588*, 2020.

[13] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[14] Mark Hallap, Emmanuel Dupoux, and Ewan Dunbar, "Evaluating context-invariance in unsupervised speech representations," *arXiv preprint arXiv:2210.15775*, 2022.

[15] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv, "Towards learning a universal non-semantic representation of speech," *arXiv preprint arXiv:2002.12764*, 2020.

[16] Danni Ma, Neville Ryant, and Mark Liberman, "Probing acoustic representations for phonetic properties," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 311–315.

[17] Itai Gat, Felix Kreuk, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi, "On the robustness of self-supervised representations for spoken language modeling," *arXiv preprint arXiv:2209.15483*, 2022.

[18] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," *arXiv preprint arXiv:2007.00991*, 2020.

[19] Vasista Sai Lodagala, Sreyan Ghosh, and S Umesh, "Ccc-wav2vec 2.0: Clustering aided cross contrastive self-supervised learning of speech representations," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1–8.

[20] Changfeng Gao, Gaofeng Cheng, and Pengyuan Zhang, "Multi-variant consistency based self-supervised learning for robust automatic speech recognition," *arXiv preprint arXiv:2112.12522*, 2021.

[21] Philip Bachman, Ouais Alsharif, and Doina Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, 2014.

[22] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18003–18017.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 18661–18673, Curran Associates, Inc.

[24] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[25] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu, "Contrastive learning for label efficient semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10623–10633.

[26] Li Yi, Sheng Liu, Qi She, A Ian McLeod, and Boyu Wang, "On learning contrastive representations for learning with noisy labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16682–16691.

[27] Varsha Suresh and Desmond C Ong, "Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification," *arXiv preprint arXiv:2109.05427*, 2021.

[28] Tarun Sai Varun Krishna and Sriram Ganapathy, "Representation learning with hidden unit clustering for low resource speech applications," *arXiv preprint arXiv:2307.07325*, 2023.

[29] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[30] Andy T Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[31] Alexander H Liu, Yu-An Chung, and James Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *arXiv preprint arXiv:2011.00406*, 2020.

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

[33] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li, "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," *arXiv preprint arXiv:2010.13991*, 2020.

[34] 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*, vol. 16, no. 6, pp. 1505–1518, 2022.

[35] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 215–222.

[36] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[37] Alex Graves and Alex Graves, "Connectionist temporal classification," *Supervised sequence labelling with recurrent neural networks*, pp. 61–93, 2012.

[38] 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.

[39] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[40] Jan Chorowski, Grzegorz Ciesielski, Jarosław Dzikowski, Adrian Lańcucki, Ricard Marxer, Mateusz Opala, Piotr Pusz, Paweł Rychlikowski, and Michał Stypułkowski, "Information retrieval for zerospeech 2021: The submission by university of wroclaw," *arXiv preprint arXiv:2106.11603*, 2021.

[41] Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1–5.

[42] Gaël Le Godais, Tal Linzen, and Emmanuel Dupoux, "Comparing character-level neural language models using a lexical decision task," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 125–130.

[43] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman, "Blimp: The benchmark of linguistic minimal pairs for english," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 377–392, 2020.

[44] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[45] Vladimir I Levenshtein et al., "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*. Soviet Union, 1966, vol. 10, pp. 707–710.

[46] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[47] Tadeusz Caliński and Jerzy Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[48] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[49] Voxforge.org, "Free speech... recognition (linux, windows and mac) - voxforge.org," `http://www.voxforge.org/`, accessed 06/25/2014.

[50] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.