Unsupervised
Speech Recognition

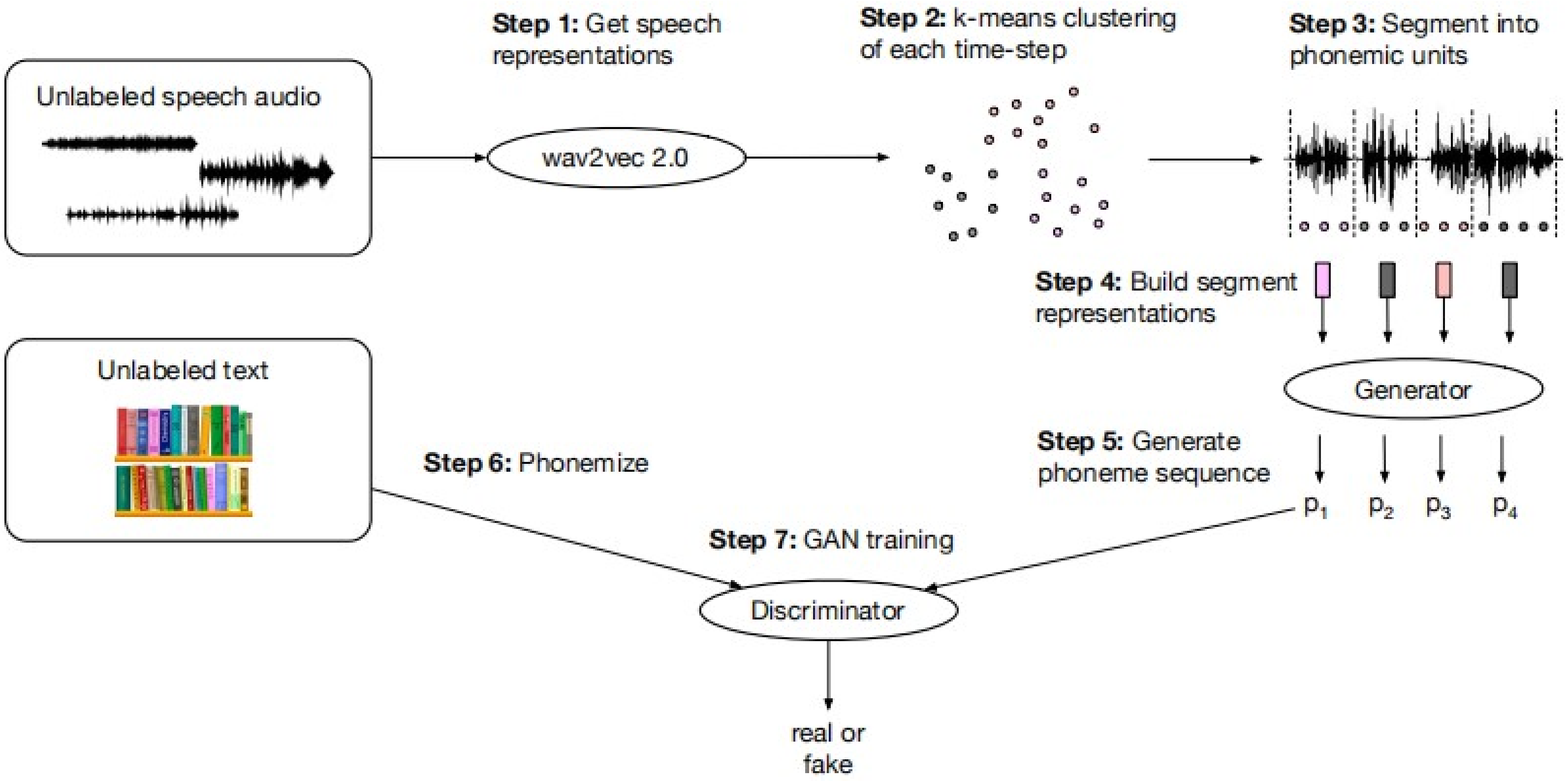
李思瑞 严子曦

https://arxiv.org/pdf/2105.11084.pdf

为什么要做无监督语音识别?

- 由于模型结构的改进和半监督学习、自监督学习的应用，语音识别的性能得到极大提升
- 但是这些技术都需要语音对应的文本标记
- 特别地，人类仅仅通过倾听周围的人，而没有明确的监督，就能学到很多关于说话的知识

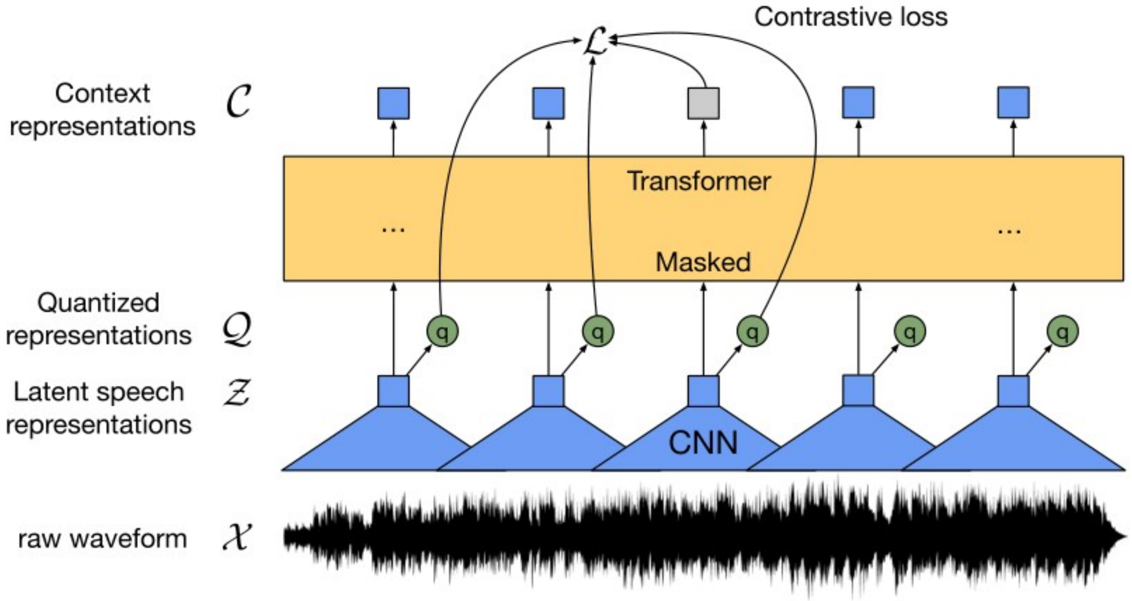
wav2vec-Unsupervised



Unlabeled speech audio
Step 1: Get speech representations
wav2vec 2.0
Step 2: k-means clustering of each time-step
Step 3: Segment into phonemic units
Step 4: Build segment representations
Generator
Step 5: Generate phoneme sequence
p_1 p_2 p_3 p_4
Unlabeled text
Step 6: Phonemize
Step 7: GAN training
Discriminator
real or fake

Speech and Text Representations

Self-supervised Learning of Speech Audio Representations


Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
...
Masked
Quantized representations $\mathcal{Q}$
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

Choosing Audio Representations

- Removing Silences：rVAD, an unsupervised voice activity detection (VAD)
- Speech Audio Representations：选择wav2vec 2.0的输出作为表示

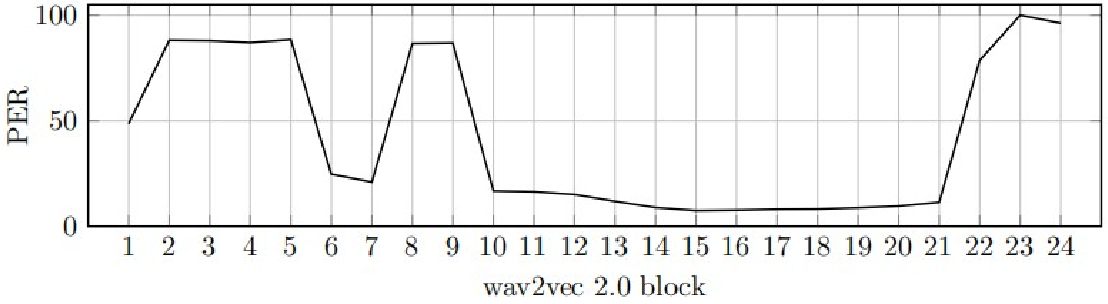


Figure 2: Supervised phoneme classification using representations from different wav2vec 2.0 blocks on dev-other of English Librispeech. Low and high blocks do not provide good features, while as blocks 14-19 do. Block 15 performs best.

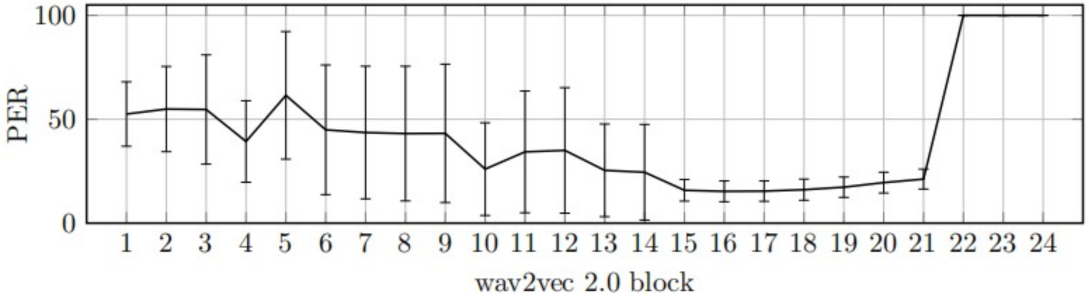
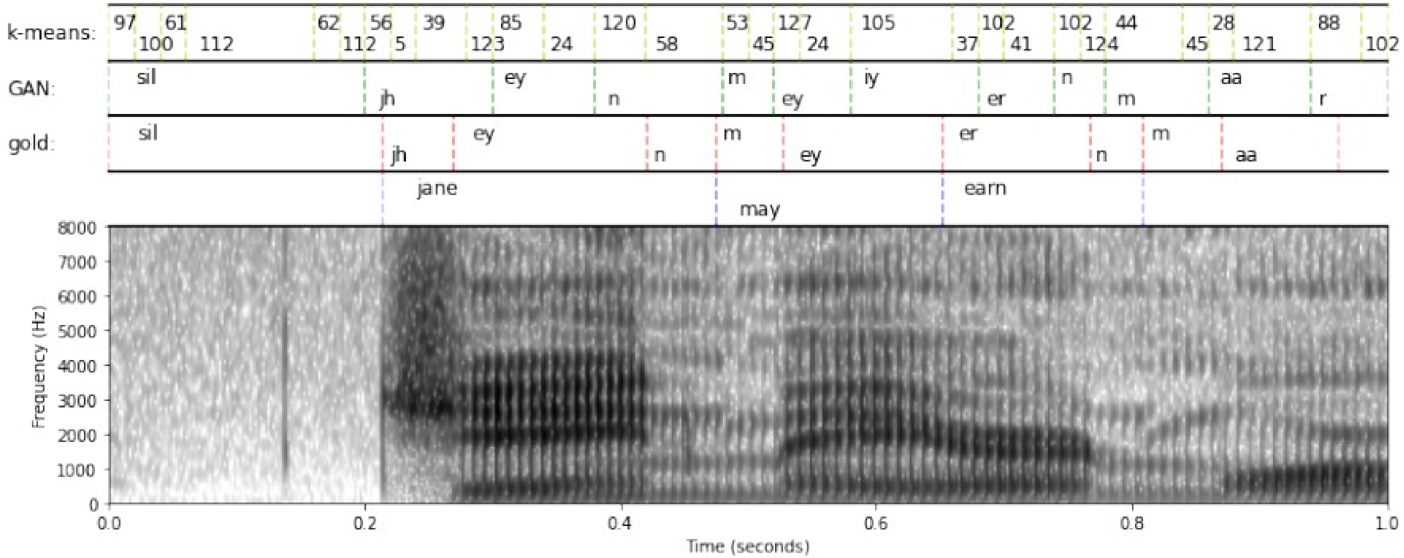


Figure 3: Supervised phoneme classification on eight languages of the MLS dataset in terms of mean PER and standard deviation for different wav2vec 2.0 blocks to represent the raw audio (cf. Figure 2). We consider English, German, Spanish, French, Italian, Dutch, Polish and Portuguese.

Segmenting the Audio Signal

- Identifying Speech Audio Segments：K-means
- Building Segment Representations：PCA、mean-pool

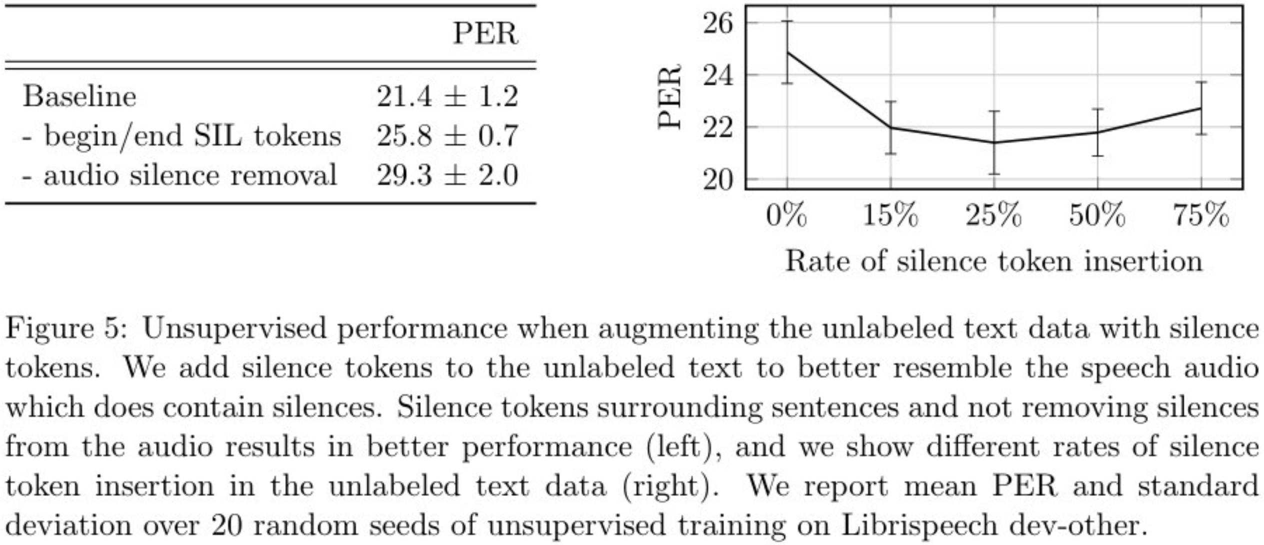


Method	Precision	Recall	F1
DAVEnet + peak detection (Harwath and Glass, 2019)	.893	.712	.792
CPC + peak detection (Kreuk et al., 2020)	.839	.836	.837
k-means on wav2vec 2.0 features	.935	.379	.539
wav2vec-U Viterbi prediction	.598	.662	.629

Table 1: Quantitative evaluation of segment boundaries with respect to human labeled segment boundaries. We report precision, recall and f-measure using a 20ms tolerance.

Pre-processing the Text Data

- Phonemization： G2P
- Silence token insertion



	PER
Baseline	21.4 ± 1.2
- begin/end SIL tokens	25.8 ± 0.7
- audio silence removal	29.3 ± 2.0

Figure 5: Unsupervised performance when augmenting the unlabeled text data with silence tokens. We add silence tokens to the unlabeled text to better resemble the speech audio which does contain silences. Silence tokens surrounding sentences and not removing silences from the audio results in better performance (left), and we show different rates of silence token insertion in the unlabeled text data (right). We report mean PER and standard deviation over 20 random seeds of unsupervised training on Librispeech dev-other.

Unsupervised Learning

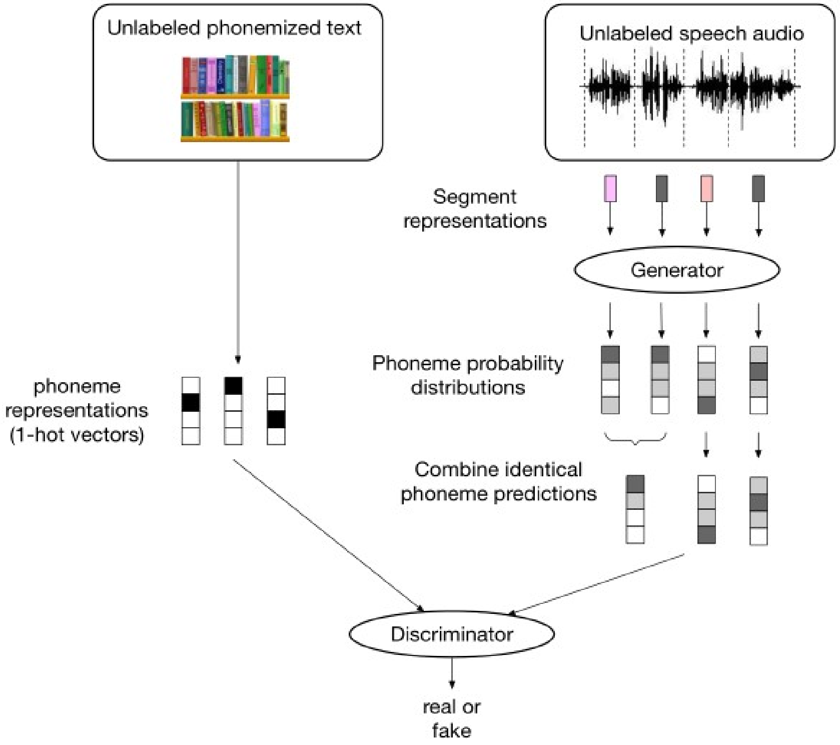

Unlabeled phonemized text
Unlabeled speech audio
Segment representations
Generator
Phoneme probability distributions
phoneme representations (1-hot vectors)
Combine identical phoneme predictions
Discriminator
real or fake

Figure 6: Illustration of how generator outputs and real phonemized text are turned into inputs to the discriminator. Real text is represented as a sequence of 1-hot vectors and generator outputs for different segment representations are collapsed if consecutive segments result in the same argmax prediction.

Objective

$$\min_{\mathcal{G}} \max_{\mathcal{C}} \mathop{\mathbb{E}}_{P^r \sim \mathcal{P}^r} [\log \mathcal{C}(P^r)] - \mathop{\mathbb{E}}_{S \sim \mathcal{S}} [\log (1 - \mathcal{C}(\mathcal{G}(S)))] - \lambda \mathcal{L}_{gp} + \gamma \mathcal{L}_{sp} + \eta \mathcal{L}_{pd}$$

- Gradient penalty

$$\mathcal{L}_{gp} = \mathop{\mathbb{E}}_{\tilde{P} \sim \tilde{\mathcal{P}}} \left[\left(\|\nabla \mathcal{C}(\tilde{P})\| - 1 \right)^2 \right]$$

- Segment smoothness penalty

$$\mathcal{L}_{sp} = \sum_{(p_t, p_{t+1}) \in \mathcal{G}(S)} \|p_t - p_{t+1}\|^2$$

- Phoneme diversity loss

$$\mathcal{L}_{pd} = \frac{1}{|B|} \sum_{S \in B} -H_{\mathcal{G}}(\mathcal{G}(S))$$

Unsupervised Cross-Validation Metric

- 我们使用这个度量是为了能够提前停止训练，并且选择训练的超参数
- language model entropy：表示转录的流畅度
- vocabulary usage：模型通过维特比解码输出的音素词汇的比例

$$\hat{\mathcal{P}} = \arg\min\nolimits_{\mathcal{P}} H(\mathcal{P}) - \log U(\mathcal{P}) \ \text{ where } U(\mathcal{P}) \in [0, 1]$$

$$H(\mathcal{P}) < H(\hat{\mathcal{P}}) + \log\left(1.2 \times \frac{U(\mathcal{P})}{U(\hat{\mathcal{P}})}\right)$$

$$\arg\min_{\mathcal{P}} H_{LM}(\mathcal{P}) = \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_{t=1}^{M} p_{LM}(p_t^j) \log p_{LM}(p_t^j), M = |P^j|, P^j = [p_1^j, \ldots, p_M^j]$$

How effective is this metric?

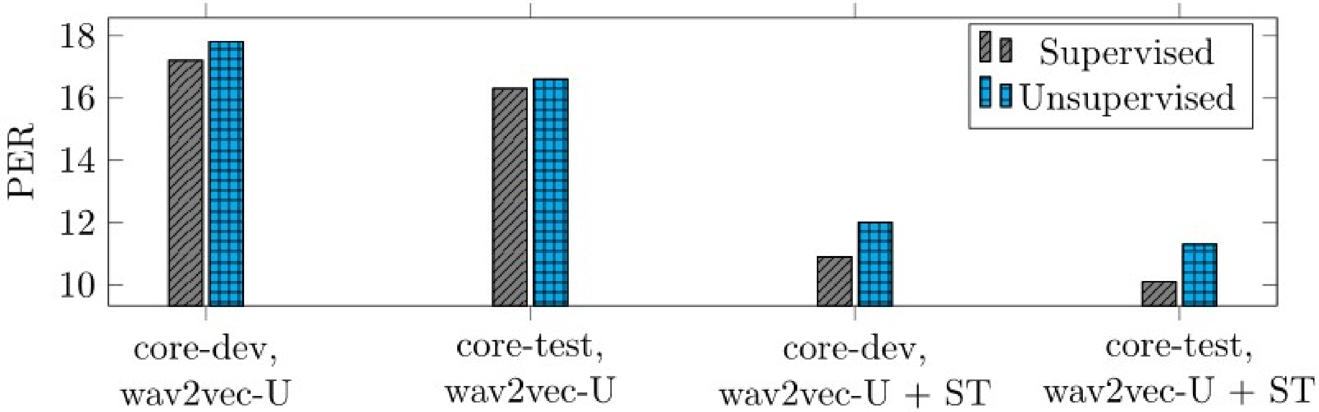


Figure 7: Effectiveness of the unsupervised cross-validation metric for model development compared to using a labeled development set (Supervised). We report PER on TIMIT core-dev/test (§ 5.1) for the GAN (wav2vec-U) and with self-training (wav2vec-U + ST).

Decoding

- 使用Pykaldi构建WFST进行解码
- 在解码过程中，我们将声学尺度作为参数提供给kaldi解码器，并在空白令牌发射中添加标量v(GaN模型为silence，其他模型为blank)。我们通过最小化衡量输出流畅性和模型输出忠诚度的量来调整这两个参数的最佳权重。

$$\sum_{j=1}^{N_s} H_{LM}(\bar{P}_j) \times \max\left(ED(\bar{P}_j, P_j), \mu\right)$$

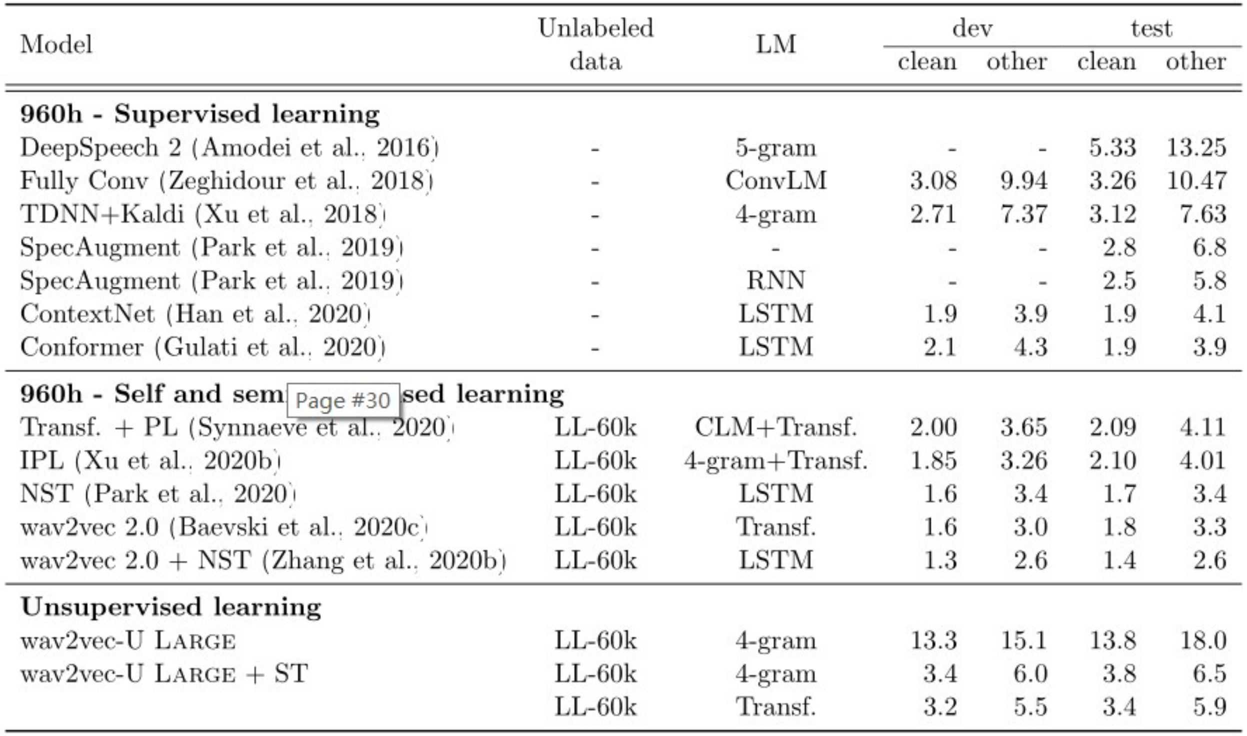
Self-Training

- 对于自训练，我们执行两次迭代：
- 首先，我们使用无监督 GAN 模型对训练数据进行伪标记，并在伪标记上训练 HMM。
- 其次，我们使用 HMM 重新标记训练数据，然后使用具有 CTC 损失的 HMM 伪标签对原始 wav2vec 2.0 模型进行微调。
- HMM 模型使用音素作为输出，而 wav2vec 2.0 模型使用字母。两者都使用 WFST 解码器解码成字。

Results

- Comparison to Supervised Speech Recognition on Librispeech
- Comparison to Prior Unsupervised Work
- Performance on non-English languages
- Application to Low-resource Languages

Librispeech



Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
960h - Supervised learning						
DeepSpeech 2 (Amodei et al., 2016)	-	5-gram	-	-	5.33	13.25
Fully Conv (Zeghidour et al., 2018)	-	ConvLM	3.08	9.94	3.26	10.47
TDNN+Kaldi (Xu et al., 2018)	-	4-gram	2.71	7.37	3.12	7.63
SpecAugment (Park et al., 2019)	-	-	-	-	2.8	6.8
SpecAugment (Park et al., 2019)	-	RNN	-	-	2.5	5.8
ContextNet (Han et al., 2020)	-	LSTM	1.9	3.9	1.9	4.1
Conformer (Gulati et al., 2020)	-	LSTM	2.1	4.3	1.9	3.9
960h - Self and sem Page #30 **sed learning**						
Transf. + PL (Synnaeve et al., 2020)	LL-60k	CLM+Transf.	2.00	3.65	2.09	4.11
IPL (Xu et al., 2020b)	LL-60k	4-gram+Transf.	1.85	3.26	2.10	4.01
NST (Park et al., 2020)	LL-60k	LSTM	1.6	3.4	1.7	3.4
wav2vec 2.0 (Baevski et al., 2020c)	LL-60k	Transf.	1.6	3.0	1.8	3.3
wav2vec 2.0 + NST (Zhang et al., 2020b)	LL-60k	LSTM	1.3	2.6	1.4	2.6
Unsupervised learning						
wav2vec-U LARGE	LL-60k	4-gram	13.3	15.1	13.8	18.0
wav2vec-U LARGE + ST	LL-60k	4-gram	3.4	6.0	3.8	6.5
	LL-60k	Transf.	3.2	5.5	3.4	5.9

Table 2: WER on the Librispeech dev/test sets when using 960 hours of unlabeled audio data from Librispeech (LS-960) or 53.2k hours from Libri-Light (LL-60k) using representations from wav2vec 2.0 LARGE. Librispeech provides clean dev/test sets which are less challenging than the other sets. We report results for GAN training only (wav2vec-U) and with subsequent self-training (wav2vec-U + ST).

TIMIT

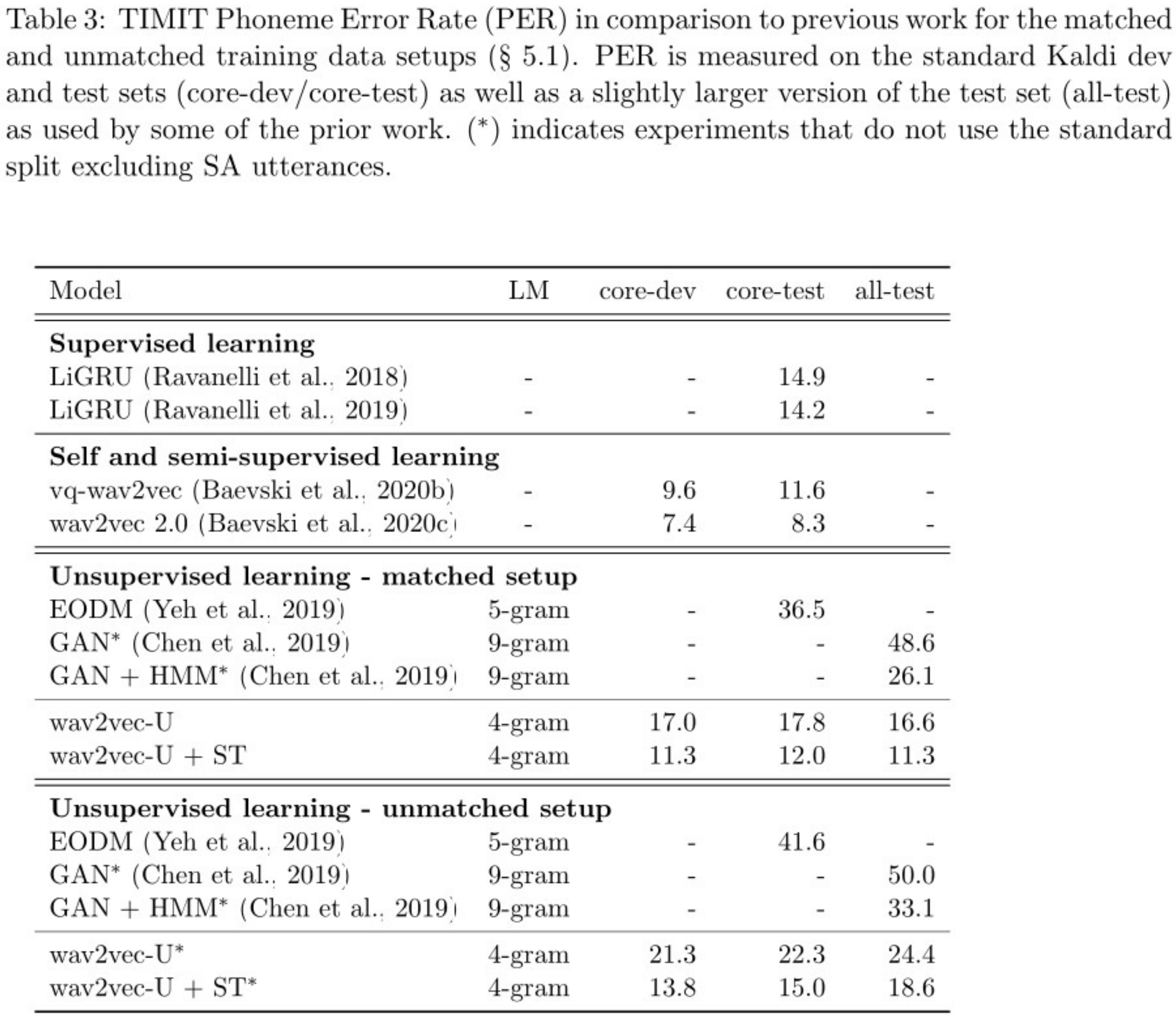
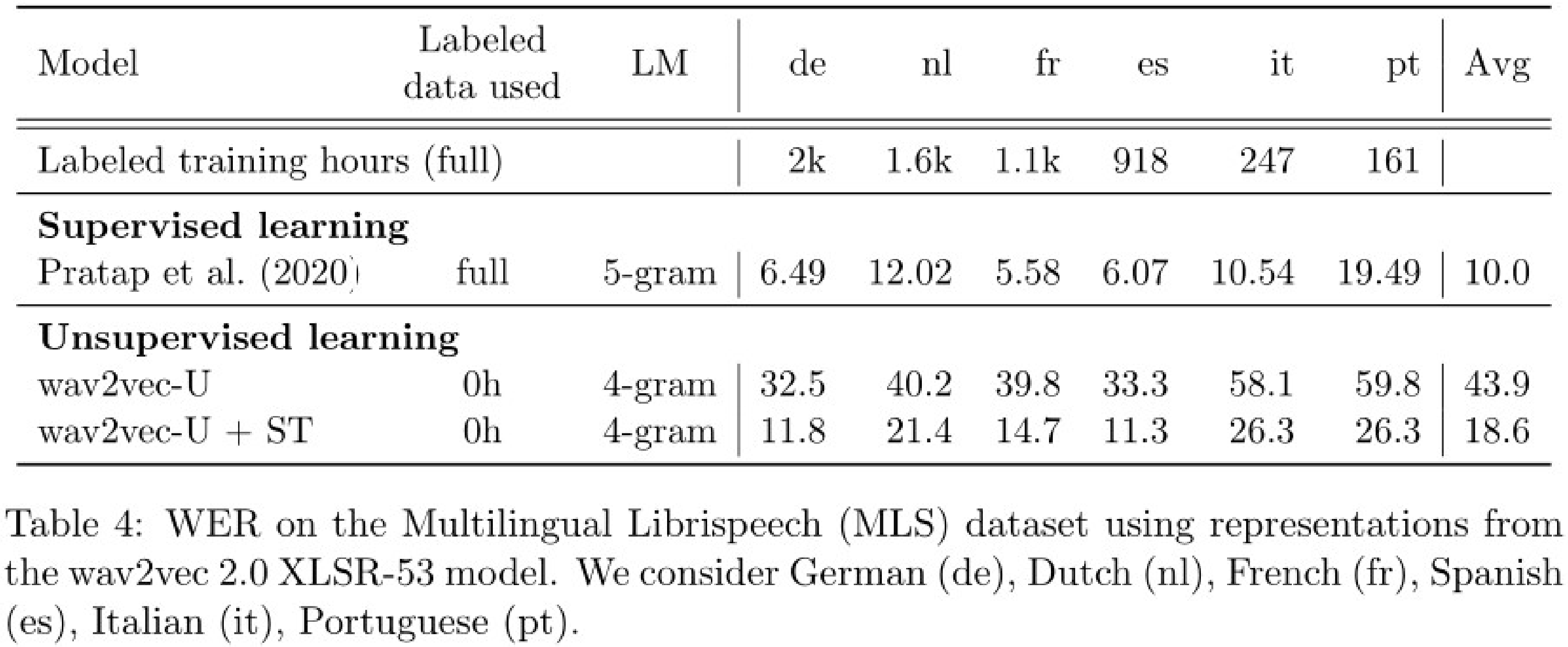


Table 3: TIMIT Phoneme Error Rate (PER) in comparison to previous work for the matched and unmatched training data setups (§ 5.1). PER is measured on the standard Kaldi dev and test sets (core-dev/core-test) as well as a slightly larger version of the test set (all-test) as used by some of the prior work. (*) indicates experiments that do not use the standard split excluding SA utterances.

Model	LM	core-dev	core-test	all-test
Supervised learning				
LiGRU (Ravanelli et al., 2018)	-	-	14.9	-
LiGRU (Ravanelli et al., 2019)	-	-	14.2	-
Self and semi-supervised learning				
vq-wav2vec (Baevski et al., 2020b)	-	9.6	11.6	-
wav2vec 2.0 (Baevski et al., 2020c)	-	7.4	8.3	-
Unsupervised learning - matched setup				
EODM (Yeh et al., 2019)	5-gram	-	36.5	-
GAN* (Chen et al., 2019)	9-gram	-	-	48.6
GAN + HMM* (Chen et al., 2019)	9-gram	-	-	26.1
wav2vec-U	4-gram	17.0	17.8	16.6
wav2vec-U + ST	4-gram	11.3	12.0	11.3
Unsupervised learning - unmatched setup				
EODM (Yeh et al., 2019)	5-gram	-	41.6	-
GAN* (Chen et al., 2019)	9-gram	-	-	50.0
GAN + HMM* (Chen et al., 2019)	9-gram	-	-	33.1
wav2vec-U*	4-gram	21.3	22.3	24.4
wav2vec-U + ST*	4-gram	13.8	15.0	18.6

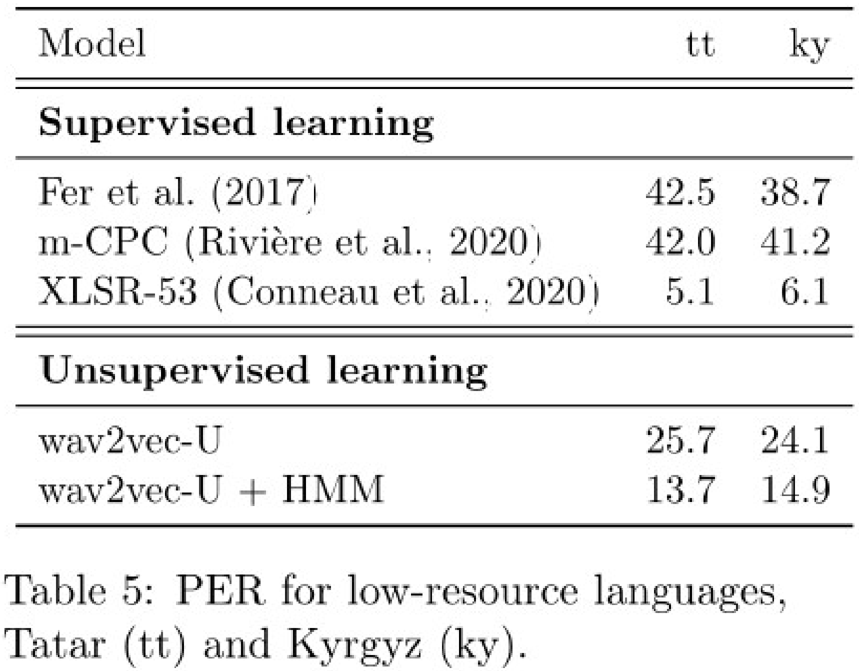
non-English languages



Model	Labeled data used	LM	de	nl	fr	es	it	pt	Avg
Labeled training hours (full)			2k	1.6k	1.1k	918	247	161	
Supervised learning									
Pratap et al. (2020)	full	5-gram	6.49	12.02	5.58	6.07	10.54	19.49	10.0
Unsupervised learning									
wav2vec-U	0h	4-gram	32.5	40.2	39.8	33.3	58.1	59.8	43.9
wav2vec-U + ST	0h	4-gram	11.8	21.4	14.7	11.3	26.3	26.3	18.6

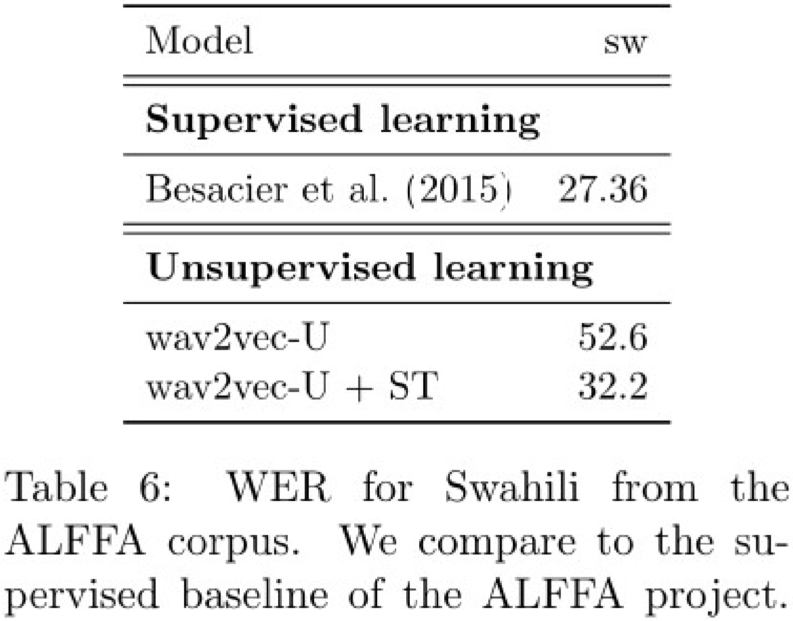
Table 4: WER on the Multilingual Librispeech (MLS) dataset using representations from the wav2vec 2.0 XLSR-53 model. We consider German (de), Dutch (nl), French (fr), Spanish (es), Italian (it), Portuguese (pt).

Low-resource Languages



Model	tt	ky
Supervised learning		
Fer et al. (2017)	42.5	38.7
m-CPC (Rivière et al., 2020)	42.0	41.2
XLSR-53 (Conneau et al., 2020)	5.1	6.1
Unsupervised learning		
wav2vec-U	25.7	24.1
wav2vec-U + HMM	13.7	14.9

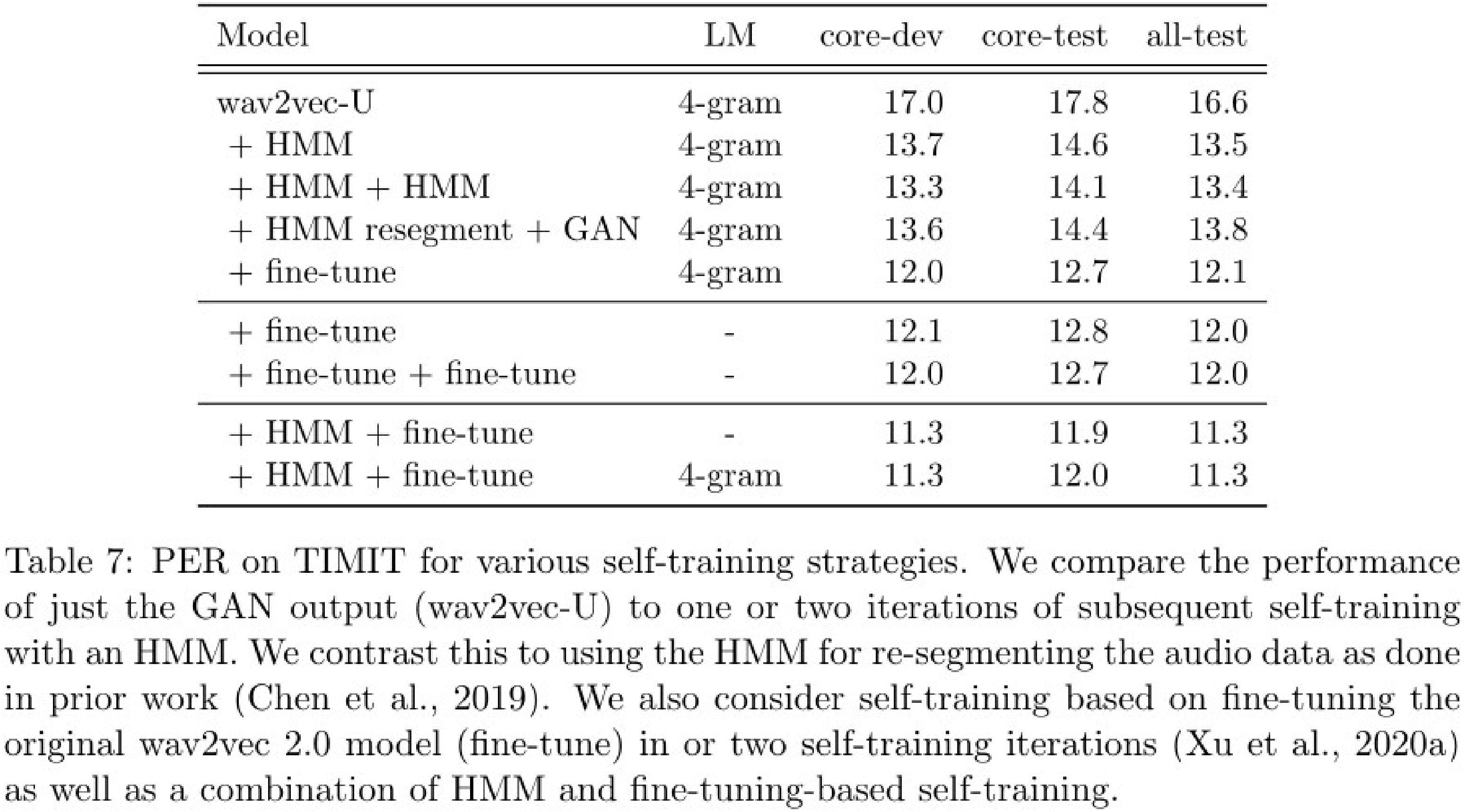
Table 5: PER for low-resource languages, Tatar (tt) and Kyrgyz (ky).



Model	sw
Supervised learning	
Besacier et al. (2015)	27.36
Unsupervised learning	
wav2vec-U	52.6
wav2vec-U + ST	32.2

Table 6: WER for Swahili from the ALFFA corpus. We compare to the supervised baseline of the ALFFA project.

Self-training Strategies



Model	LM	core-dev	core-test	all-test
wav2vec-U	4-gram	17.0	17.8	16.6
+ HMM	4-gram	13.7	14.6	13.5
+ HMM + HMM	4-gram	13.3	14.1	13.4
+ HMM resegment + GAN	4-gram	13.6	14.4	13.8
+ fine-tune	4-gram	12.0	12.7	12.1
+ fine-tune	-	12.1	12.8	12.0
+ fine-tune + fine-tune	-	12.0	12.7	12.0
+ HMM + fine-tune	-	11.3	11.9	11.3
+ HMM + fine-tune	4-gram	11.3	12.0	11.3

Table 7: PER on TIMIT for various self-training strategies. We compare the performance of just the GAN output (wav2vec-U) to one or two iterations of subsequent self-training with an HMM. We contrast this to using the HMM for re-segmenting the audio data as done in prior work (Chen et al., 2019). We also consider self-training based on fine-tuning the original wav2vec 2.0 model (fine-tune) in or two self-training iterations (Xu et al., 2020a) as well as a combination of HMM and fine-tuning-based self-training.

Conclusion

- wav2vec-U是一个无监督语音识别建模的框架
- 在Librispeech 上，与最好的有标签的模型性能十分接近
- 在TIMIT上，与先前的无监督的识别相比，将PER从26.1降低到了11.3
- 在除英语之外的数据集上，也展现出该系统的可行性
- 仅仅使用未标记的音频和未标记的文本构建语音识别模型，减少了在世界上更多语言上实现语音识别技术的工作量