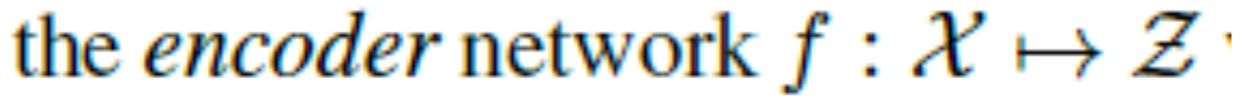
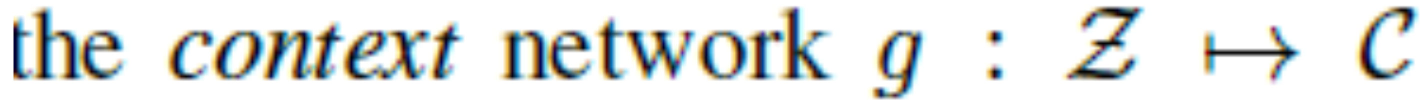
WAV2VEC: Unsupervised Pre-Training for Speech Recoginiton

Wenwei Dong

WAV2VEC: a convolutional neural network that takes raw audio as input and computes a general representation that can be input to a speech recognition system.



the *encoder* network $f : \mathcal{X} \mapsto \mathcal{Z}$



the *context* network $g : \mathcal{Z} \mapsto \mathcal{C}$

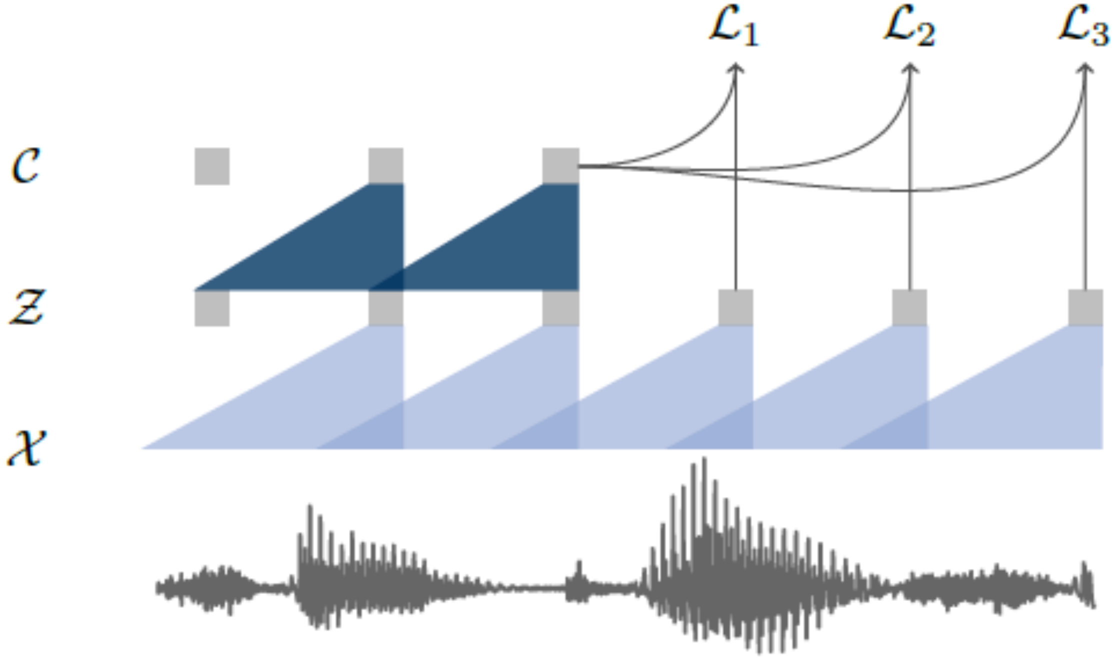


Figure 1: Illustration of pre-training from audio data $\mathcal{X}$ which is encoded with two convolutional neural networks that are stacked on top of each other. The model is optimized to solve a next time step prediction task.

$$\mathcal{L}_k = -\sum_i \left(\log \sigma(\mathbf{z}_{i+k}^\top h_k(\mathbf{c}_i)) + \lambda \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p_n} [\log \sigma(-\tilde{\mathbf{z}}^\top h_k(\mathbf{c}_i))] \right)$$

ASR dataset

- TIMIT: standard train dev and tst
- Train set: si284, dev set:nov93dev , testset: nov92

Pre-training

- WSJ 81hours
- Librispeech 80 hours clean data
- Librispeech 960 hours

	nov93dev		nov92	
	LER	WER	LER	WER
Deep Speech 2 (12K h labeled speech; Amodei et al., 2016)	-	4.42	-	3.1
Trainable frontend (Zeghidour et al., 2018a)	-	6.8	-	3.5
Lattice-free MMI (Hadian et al., 2018)	-	$5.66^{\dagger}$	-	$2.8^{\dagger}$
Supervised transfer-learning (Ghahremani et al., 2017)	-	$4.99^{\dagger}$	-	$2.53^{\dagger}$
4-gram LM				
Baseline	3.34	8.42	2.39	5.83
wav2vec (Libri 80h)	3.71	9.11	2.17	5.55
wav2vec (Libri 960h)	2.81	7.43	1.84	4.77
wav2vec (Libri + WSJ 1041h)	2.91	7.59	1.67	4.61
Word ConvLM (Zeghidour et al., 2018b)				
Baseline	2.57	6.27	1.51	3.60
wav2vec Libri (960h)	2.22	5.39	1.25	2.87
Char ConvLM (Likhomanenko et al., 2019)				
Baseline	2.77	6.67	1.53	3.46
wav2vec Libri (960h)	2.14	5.31	1.15	2.78

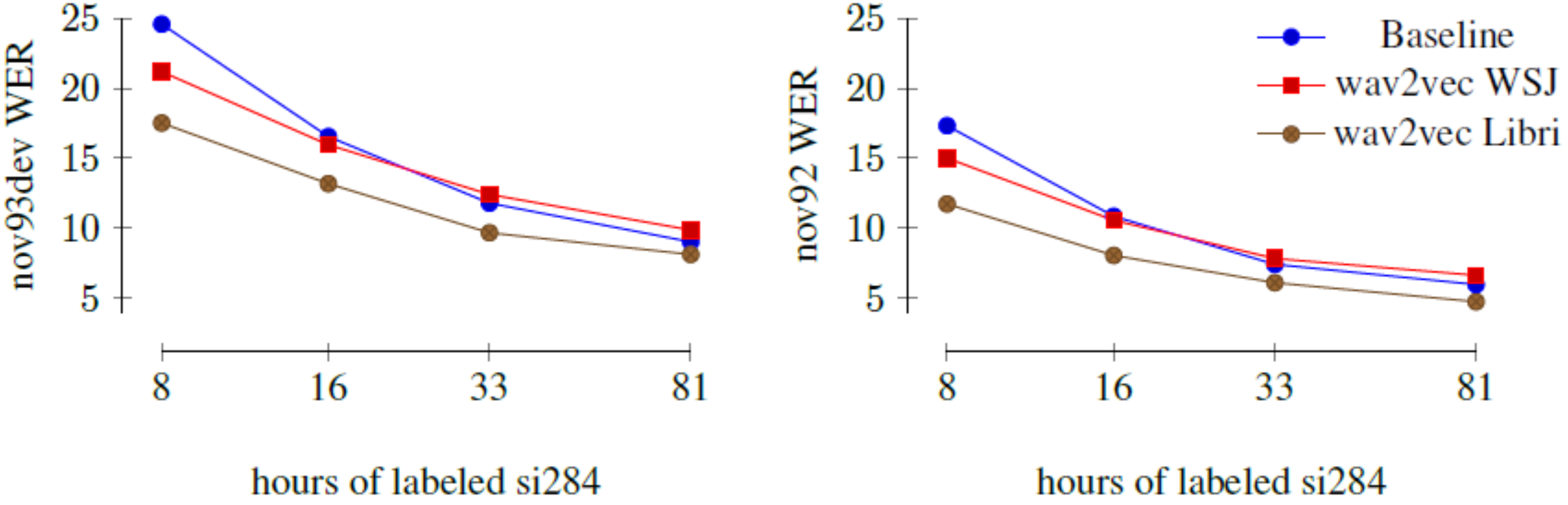
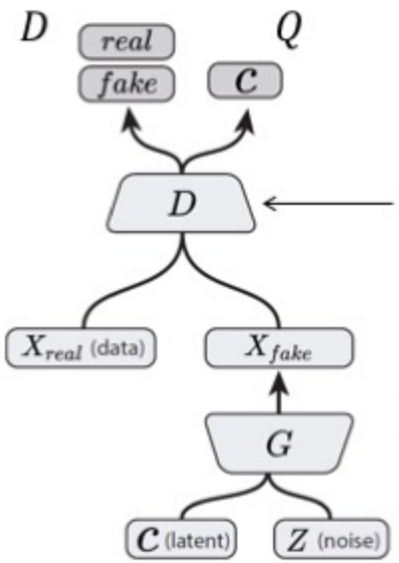

nov93dev WER
25
20
15
10
5
8
16
33
81
hours of labeled si284
nov92 WER
25
20
15
10
5
8
16
33
81
hours of labeled si284
Baseline
wav2vec WSJ
wav2vec Libri

Table 2: Results for phoneme recognition on TIMIT in terms of PER. All our models use the CNN-8L-PReLU-do0.7 architecture Ravanelli et al. (2018).

	dev	test
CNN + TD-filterbanks Zeghidour et al. (2018a)	15.6	18.0
Li-GRU + MFCC Ravanelli et al. (2018)	–	16.7 ± 0.26
Li-GRU + FBANK Ravanelli et al. (2018)	–	15.8 ± 0.10
Li-GRU + fMLLR Ravanelli et al. (2018)	–	14.9 ± 0.27
Baseline	16.9 ± 0.15	17.6 ± 0.11
wav2vec (Libri 80h)	15.5 ± 0.03	17.6 ± 0.12
wav2vec (Libri)	13.6 ± 0.20	15.6 ± 0.23
wav2vec (Libri + WSJ)	**12.9** ± **0.18**	**14.7** ± **0.42**

infoGan



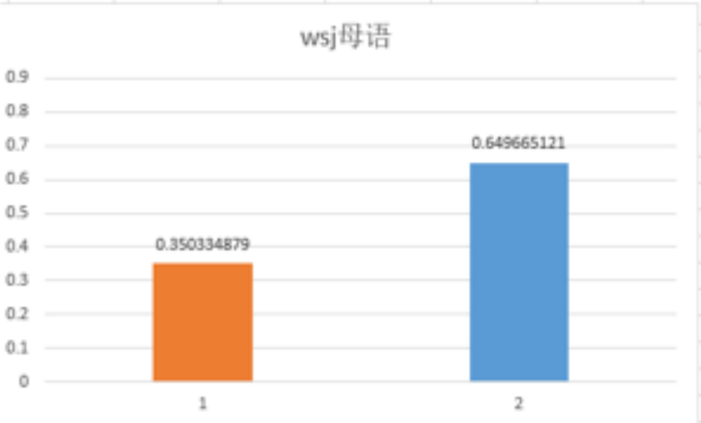
Convolution layers of the discriminator

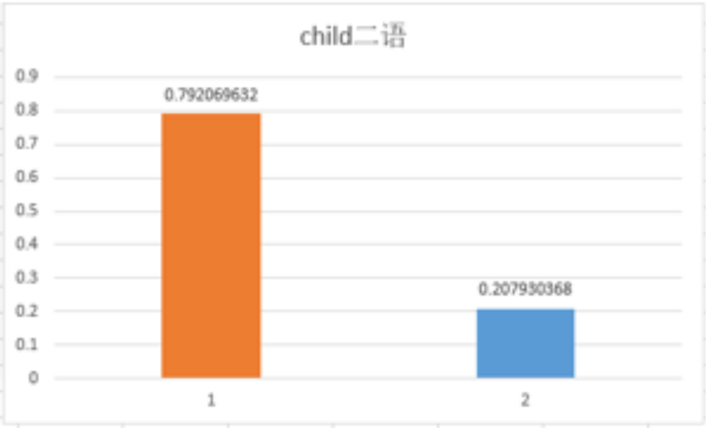
Given DCGANs,
InfoGAN comes for negligible additional costs!

Image from Odena, *et al.*, arXiv:1610.09585.

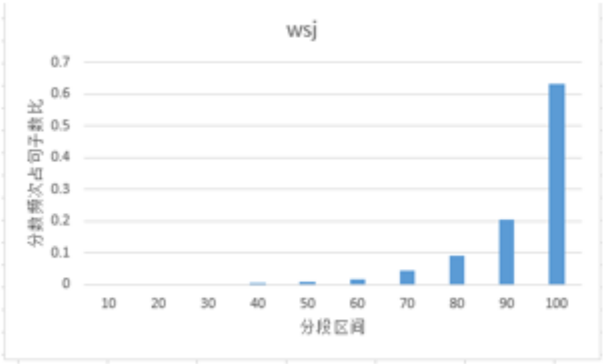
C=2
feature length =1s

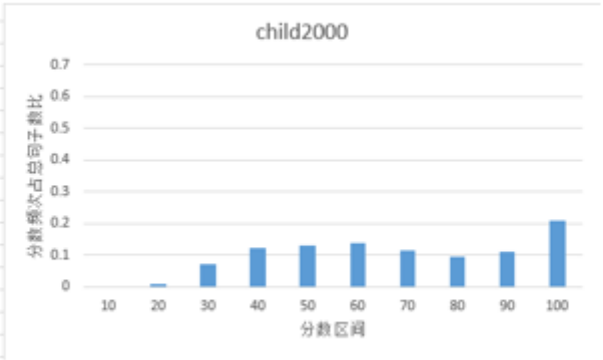
epochs=100





C=10
Feature length = 1s





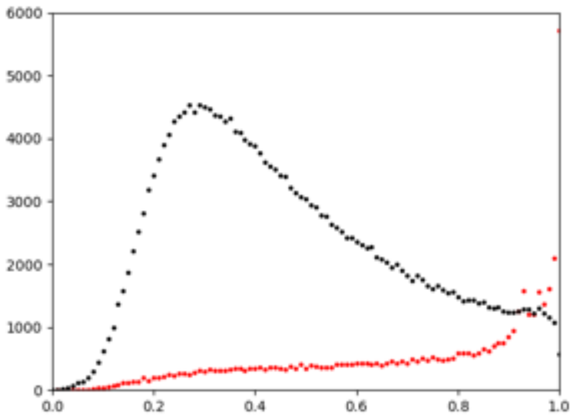
Test set :
1:1000 sentences of Chinese children speak English

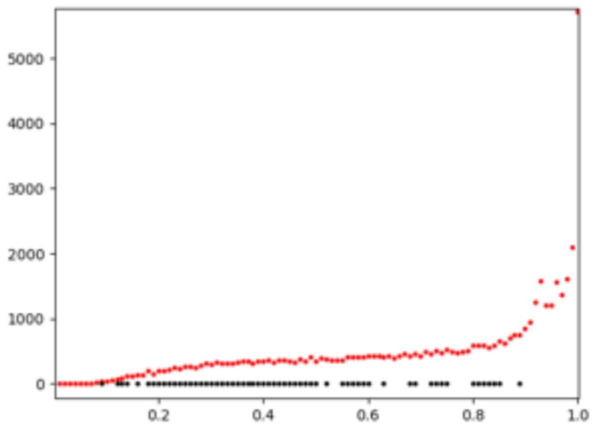
model	corrcoef
GAN	0.0314
infoGan+logistics regression	0.1285

2:995 sentence Japanese speak English

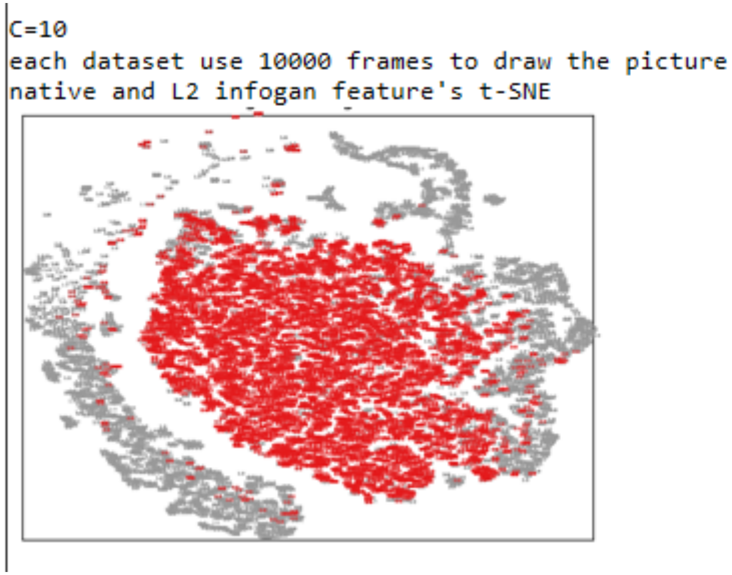
Human-human	0.5397
infoGan+logistics regression	0.02

C=10
feature length =1 frame





C=10
feature length =1 frame



```
C=10
each dataset use 10000 frames to draw the picture
native and L2 infogan feature's t-SNE
```

model	corrcoef
Human-human	0.5397
infoGan+logistic regression(C=10 mean)	0.013
infoGan(C=2 mean)	0.237
infoGan(C=2 mode)	0.190