Towards End-to-end Unsupervised Speech Recognition

李思瑞

2022-07-15

Completely Unsupervised Speech Recognition By A Generative Adversarial Network Harmonized With Iteratively Refined Hidden Markov Models

Discriminator loss

$$\mathcal{L}_D = \frac{1}{K}\sum_{k=1}^{K} D(P^{gen(k)}) - \frac{1}{K}\sum_{k=1}^{K} D(P^{real(k)}) + \alpha\mathcal{L}_{gp}, \quad (1)$$

$$\mathcal{L}_{gp} = \frac{1}{K}\sum_{k=1}^{K}(||\nabla D(P^{inter(k)})|| - 1)^2, \quad (2)$$

Generator loss

$$\mathcal{L}_{intra} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i,j\in \mathbf{S}_k}(y_i - y_j)^2, \quad (3)$$

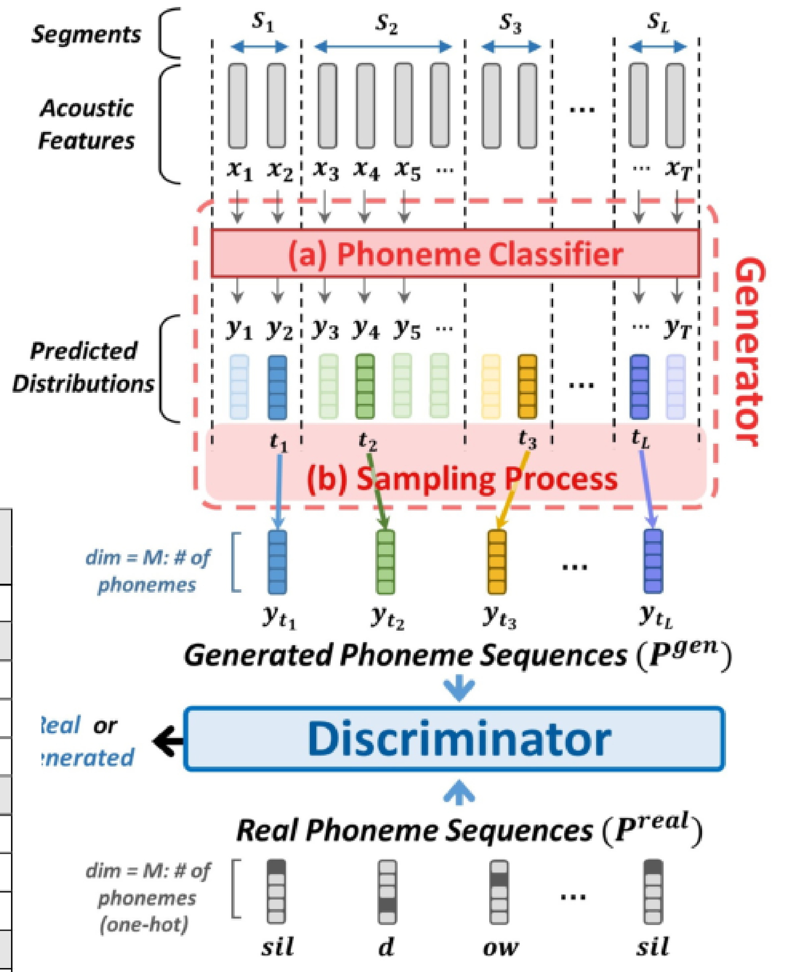
$$\mathcal{L}_G = -\frac{1}{K}\sum_{k=1}^{K} D(P^{gen(k)}) + \lambda\mathcal{L}_{intra}, \quad (4)$$

Algorithm 1: GAN/HMM Harmonization

Input: Real phoneme sequences P^{real}, Speech utterances, initial phoneme segmentation boundaries b
Output: Unsupervised ASR system
while *not converged* **do**
- Given b, in an unsupervised way train the GAN;
- Obtain transcriptions T of speech utterances using the generator within the GAN;
- Given T, train the HMMs;
- Obtain a new b by forced alignment with the HMMs.

Table 1: *Comparison of different methods*

Approaches	Matched (all 4000) FER	Matched (all 4000) PER	Nonmatched (3000/1000) FER	Nonmatched (3000/1000) PER
(I) Supervised (labeled)				
(a) RNN Transducer [22]	-	17.7	-	-
(b) standard HMMs	-	21.5	-	-
(c) Phoneme classifier	27.0	28.9	-	-
(II) Unsupervised (with oracle boundaries)				
(d) Mapping relationship GAN [21]	40.5	40.2	43.6	43.4
(e) Segmental empirical-ODM [22]	33.3	32.5	40.0	40.1
(f) Proposed: GAN	27.6	28.5	32.7	34.3
(III) Completely unsupervised (no label at all)				
(g) Segmental empirical-ODM [22]	-	36.5	-	41.6
Proposed, iteration 1, (h) GAN	48.3	48.6	50.3	50.0
Proposed, iteration 1, (i) GAN/HMM	-	30.7	-	39.5
Proposed, iteration 2, (j) GAN	41.0	41.0	44.3	44.3
Proposed, iteration 2, (k) GAN/HMM	-	27.0	-	35.5
Proposed, iteration 3, (l) GAN	39.7	38.4	45.0	44.2
Proposed, iteration 3, (m) GAN/HMM	-	26.1	-	33.1



ıre 1: *Overview of the proposed approach. The generator in- les (a) phoneme classifier transforming the acoustic features predicted phoneme distributions, and (b) a phoneme distribu- sampled from each segment. The discriminator is trained to inguish between the generated and real phoneme sequences. HMMs are not shown.*

Unsupervised Speech Recognition

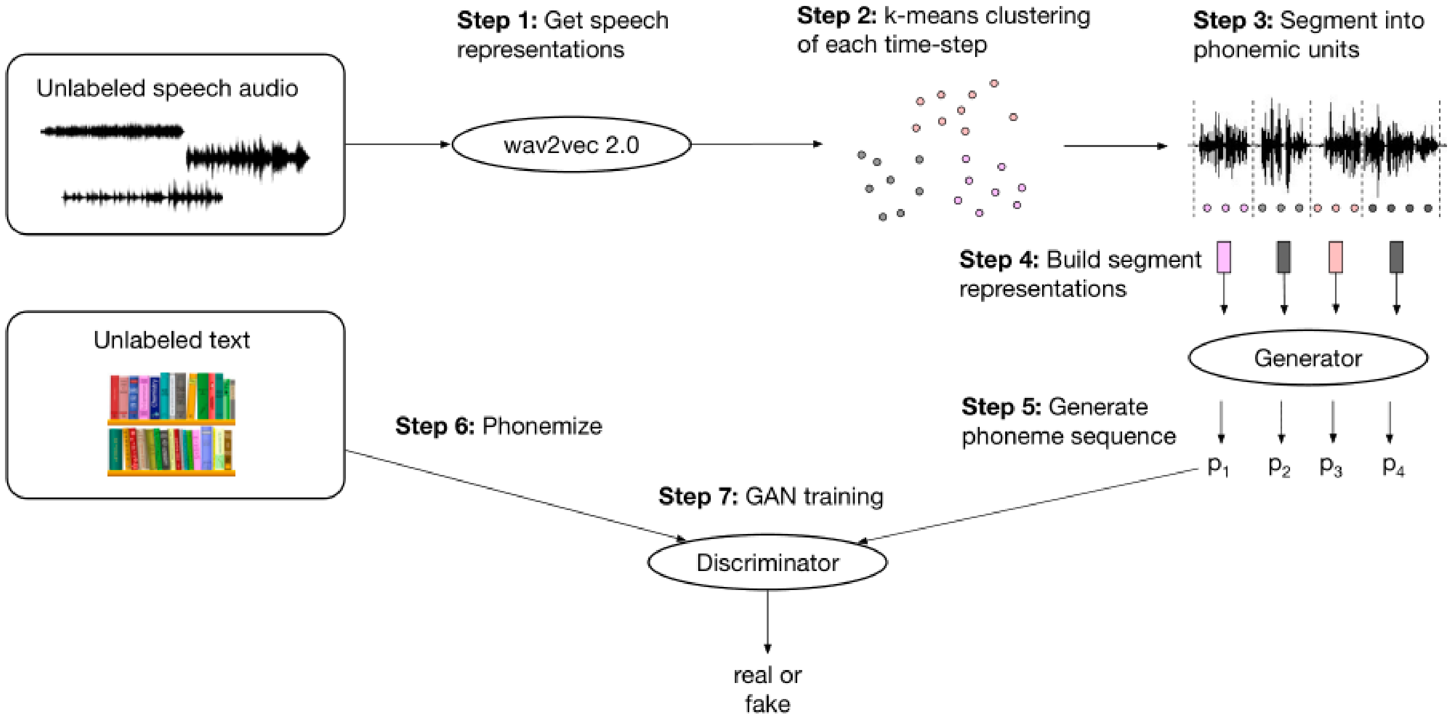


Figure 1: Illustration of wav2vec Unsupervised: we learn self-supervised representations with wav2vec 2.0 on unlabeled speech audio (Step 1), then identify clusters in the representations with k-means (Step 2) to segment the audio data (Step 3). Next, we build segment representations by mean pooling the wav2vec 2.0 representations, performing PCA and a second mean pooling step between adjacent segments (Step 4). This is input to the generator which outputs a phoneme sequence (Step 5) that is fed to the discriminator, similar to phonemized unlabeled text (Step 6) to perform adversarial training (Step 7).

$$\min_{\mathcal{G}} \max_{\mathcal{C}} \mathop{\mathbb{E}}_{P^r \sim \mathcal{P}^r} [\log \mathcal{C}(P^r)] - \mathop{\mathbb{E}}_{S \sim \mathcal{S}} [\log (1 - \mathcal{C}(\mathcal{G}(S)))] - \lambda \mathcal{L}_{gp} + \gamma \mathcal{L}_{sp} + \eta \mathcal{L}_{pd} \qquad (6)$$

- wav2vec-U is a framework which enables building speech recognition models without labeled data. It embeds and segments the speech audio with self-supervised representations from wav2vec 2.0, learns a mapping to phonemes with adversarial learning, and cross-validates hyper-parameter choices as well as early stopping with an unsupervised metric.

Towards End-to-end Unsupervised Speech Recognition

- However, existing methods still heavily rely on hand-crafted pre-processing. We introduce wav2vec-U 2.0 which does away with all audio-side pre-processing and improves accuracy through better architecture.
- we introduce an auxiliary self-supervised objective that ties model predictions back to the input.

$$\mathcal{L}_{ss} = -\sum_t \log P_{\mathcal{G}}(z_t|X)$$

$$\min_{\mathcal{G}} \max_{\mathcal{C}} \mathop{\mathbb{E}}_{Y_u}[\log \mathcal{C}(Y_u)] - \mathop{\mathbb{E}}_{X}[\log(1 - \mathcal{C}(\mathcal{G}(X)))] - \lambda\mathcal{L}_{gp} + \gamma\mathcal{L}_{sp} + \eta\mathcal{L}_{pd} + \delta\mathcal{L}_{ss}$$

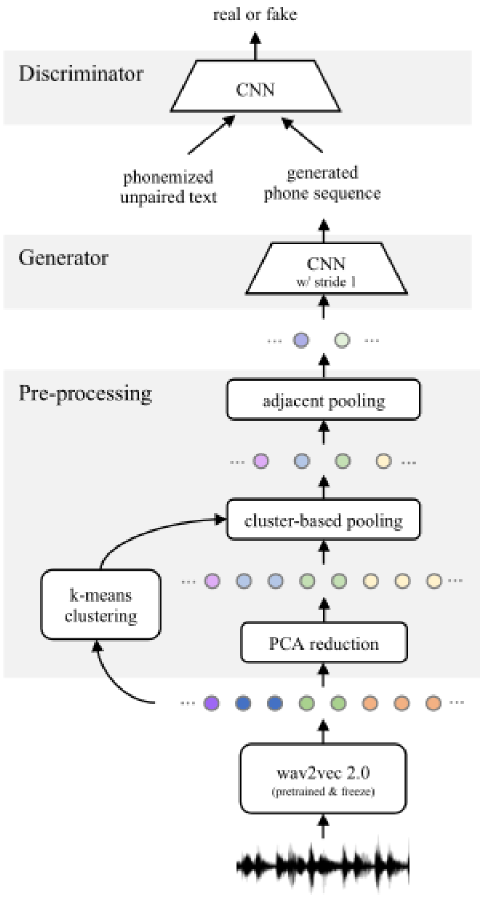


Figure 1: Wav2vec-U [6]. The input wav2vec2.0 feature is pre-processed before feeding into the generator as described in Section 2.2. The generator is optimized through adversarial training against the discriminator as described in Section 3.1.

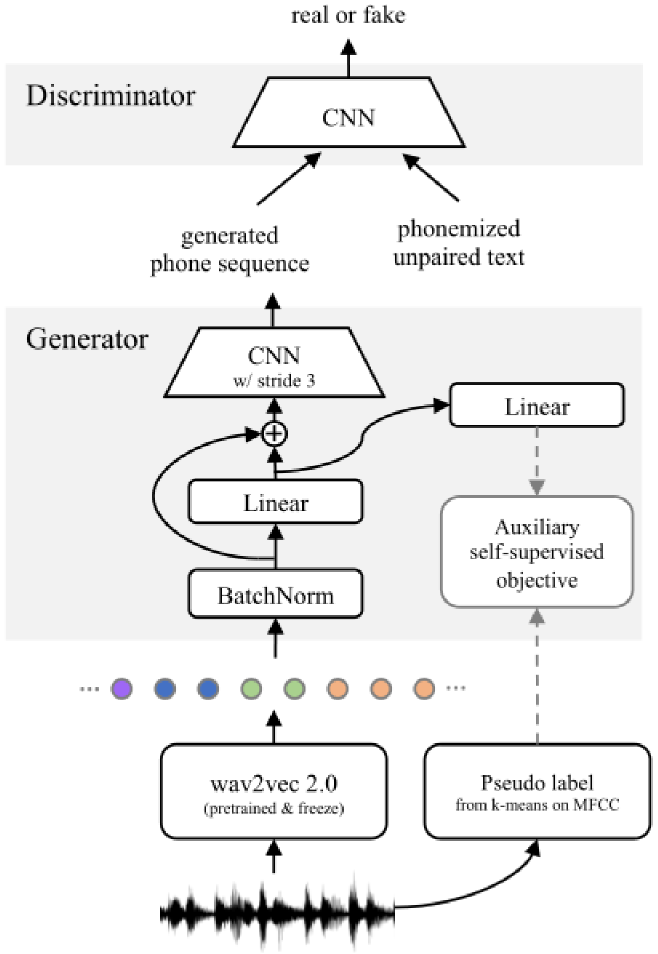


Figure 2: Proposed wav2vec-U 2.0. The generator takes raw wav2vec 2.0 feature as input without pre-processing step as described in Section 3.2. In addition to adversarial training, an auxiliary self-supervised objective is introduced with pseudo label derived from the raw waveform as described in Section 3.3.

Table 1: Interpolation from wav2vec-U (Fig. 1) to wav2vec-U 2.0 (Fig. 2). Phone Error Rate (PER) computed with greedy decoding on LibriSpeech `dev-other` set averaged over 8 runs. *Freq.* refers to the frequency of sequence, i.e. number of tokens per second.

	Pre-processing			Generator configuration				Result	
	Adjacent pooling	Cluster pooling	PCA reduction	Batch norm.	Linear proj.	Auxiliary loss	Stride	Freq. (Hz)	Average PER
wav2vec-U	✓	✓	✓	-	-	-	1	14	18.8 ± 0.9
step (i)	-	✓	✓	-	-	-	1	28	> 100
step (ii)	-	✓	✓	-	-	-	2	14	18.5 ± 0.6
step (iii)	-	-	✓	-	-	-	2	25	> 100
step (iv)	-	-	✓	-	-	-	3	16	19.0 ± 0.9
step (v)	-	-	-	-	-	-	3	16	> 100
step (vi)	-	-	-	✓	-	-	3	16	16.4 ± 0.7
step (vii)	-	-	-	✓	✓	-	3	16	15.9 ± 1.1
wav2vec-U 2.0	-	-	-	✓	✓	✓	3	16	$\mathbf{13.6 \pm 0.9}$
input wav2vec 2.0 feature								50	-
ground truth phone sequence								~10	-

Table 4: Word Error Rate (WER) on the Multilingual Librispeech (MLS) for German (de), Dutch (nl), French (fr), Spanish (es), Italian (it) and Portuguese (pt).

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. [22]	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 2.0	0h	4-gram	23.5	35.1	35.7	25.8	46.9	48.5	35.9
Unsupervised learning + self-training									
wav2vec-U	0h	4-gram	11.8	21.4	14.7	11.3	26.3	26.3	18.6
wav2vec-U 2.0	0h	4-gram	11.5	17.6	12.8	10.9	18.6	20.6	15.3

Table 3: Word Error Rate (WER) on LibriSpeech with different language models (LM) on the standard LibriSpeech dev/test sets.

Model	Unlabeled speech (hours)	LM	dev		test	
			clean	other	clean	other
Supervised learning w/ 960 hours of speech						
DeepSpeech 2 [34]	-	5-gram	-	-	5.33	13.25
Fully Conv [35]	-	ConvLM	3.08	9.94	3.26	10.47
TDNN+Kaldi [36]	-	4-gram	2.71	7.37	3.12	7.63
SpecAugment [18]	-	RNN	-	-	2.5	5.8
ContextNet [2]	-	LSTM	1.9	3.9	1.9	4.1
Conformer [1]	-	LSTM	2.1	4.3	1.9	3.9
Semi-supervised learning w/ 960 hours of speech						
Transf. + PL [26]	54k	CLM+Transf.	2.00	3.65	2.09	4.11
IPL [37]	54k	4-gram+Transf.	1.85	3.26	2.10	4.01
NST [38]	54k	LSTM	1.6	3.4	1.7	3.4
wav2vec 2.0 [15]	54k	Transf.	1.6	3.0	1.8	3.3
wav2vec 2.0 + NST [39]	54k	LSTM	1.3	2.6	1.4	2.6
Unsupervised learning						
wav2vec-U	54k	4-gram	13.3	15.1	13.8	18.0
wav2vec-U 2.0	54k	4-gram	9.8	13.1	9.9	13.9
Unsupervised learning + Self-Training						
wav2vec-U	54k	4-gram	3.4	6.0	3.8	6.5
wav2vec-U 2.0	54k	4-gram	3.5	6.0	3.7	6.3

- An end-to-end approach for unsupervised ASR is key to increasing applicability to low-resource languages. In this work, we move towards this goal by removing the need for human-engineered pre-processing and by improving accuracy.