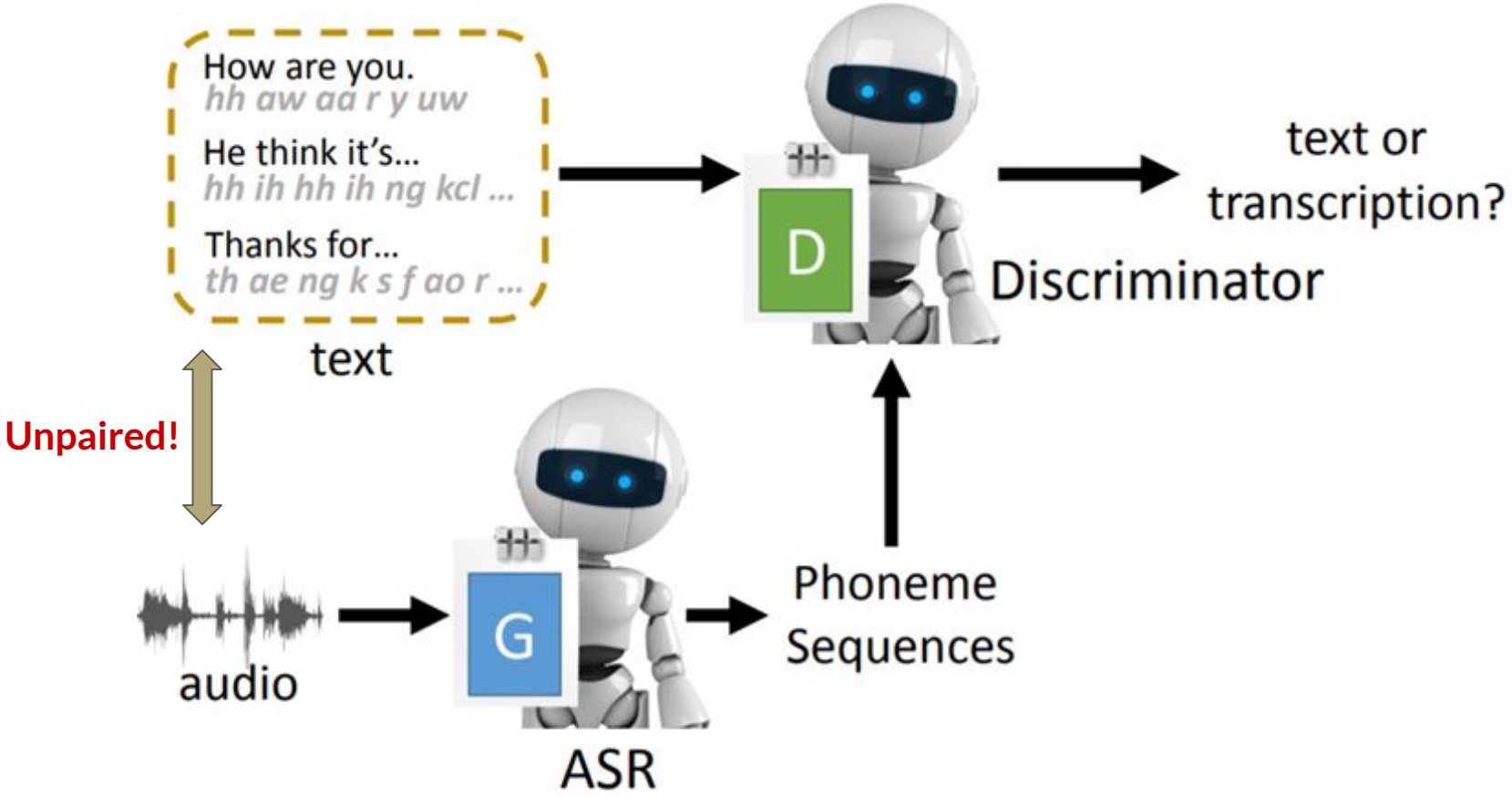
Unsupervised ASR

伏宇寬
r11942083@ntu.edu.tw

Outline

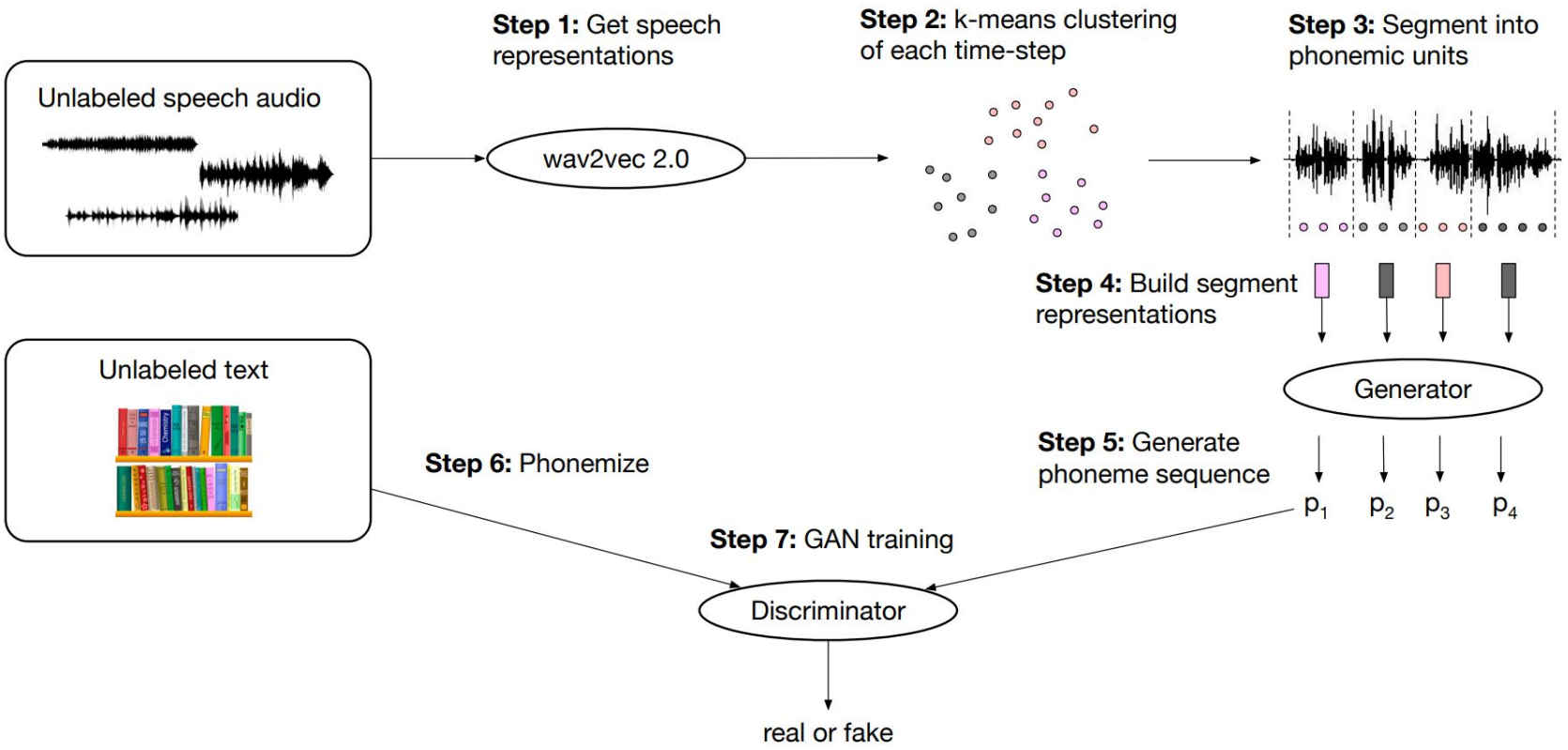
- Unsupervised ASR
- Wav2vec-U framework
  - SSL representations
  - Data preprocessing
  - GAN training
- Homework
  - problem 1
  - problem 2
  - problem 3
- Reference

Unsupervised Speech Recognition

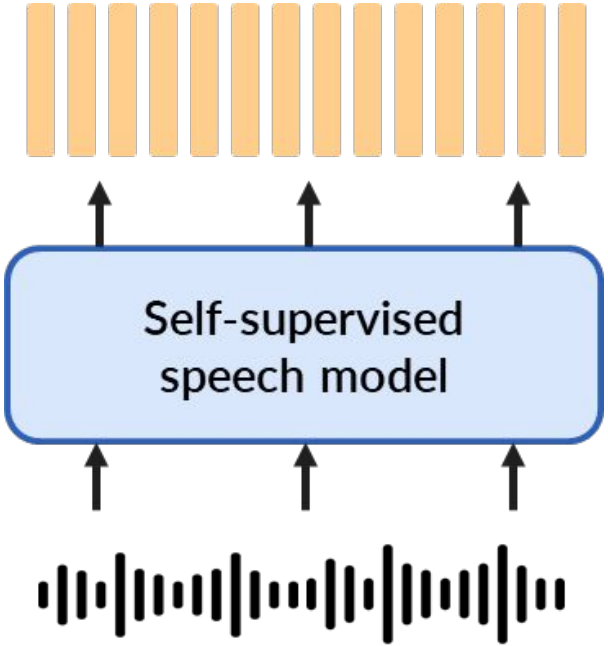


Reference: [DLHLP 2020] Text Style Transfer and Unsupervised Summarization/Translation/Speech Recognition

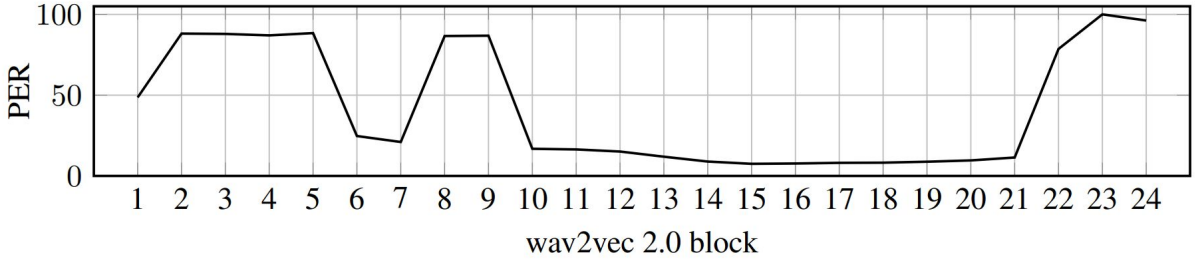
Wav2vec-U


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
Unlabeled text
Step 5: Generate phoneme sequence
p1 p2 p3 p4
Step 6: Phonemize
Step 7: GAN training
Discriminator
real or fake

SSL representation



Extract from
15th transformer layer
of w2v2-large



Speech pre-processing

1. Silence removal: robust Voice Activity Detection (rVAD) [ref]

2. Extract representation: 15th layer of w2v2-large

3. Identifying speech audio segments: K-means clustering

4. Reduce dimension: PCA > mean-pooling (merge by cluster ids & adjacent segments)

Text pre-processing

1. Phonemization: Use Espeak phonemizer

Ex: get stronger. our peoples and our economies

g ɛ t s t ɹ ɔ ŋ g ɚ aʊ ɚ p iː p əl z æ n d aʊ ɚ ɪ k ɑː n ə m i z

2. Insert SIL tokens: start & end + between words (25% insertion rate)

Ex:

<SIL> s ə p oʊ z <SIL> ɔː l θ ɹ iː ʌ v <SIL> ð ə f ɑː l oʊ ɪ ŋ k ə n d ɪ ʃ ə n z <SIL> h oʊ l d <SIL>

GAN training objective

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

WGAN-GP [ref]

- 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:

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

Self-training

- HMM
- finetuning by pseudo-label

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

Homework

You only need to do:

~~1. Preprocessing speech & text~~

2. Gan training

~~3. Self-training~~

Instructions: See README

Datasets

Speech:

Voxpopuli (unlabeled)

Text:

Voxpopuli (asr transcriptions)

[link of voxpopuli]

Other things you can do

You might try:

- different text corpus
    - lr rate
    - loss weight
    - model depth
- different w2v2 versions

Q&A

有任何問題都可以直接在 facebook 社團貼文底下留言!

Reading list

1. **“Unsupervised speech recognition”**, Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and MichaelAuli
2. **“Completely unsupervised phoneme recognition by adversari-ally learning mapping relationships from audio embeddings”**, Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, and Lin shan Lee
3. **“Completely unsupervised speech recogni-tion by a generative adversarial network harmonized with iter-atively refined hidden markov models”**, Kuan-Yu Chen, Che-Ping Tsai, Da-Rong Liu, Hung-Yi Lee,and Lin shan Lee
4. **“Improved Training of Wasserstein GANs”,** Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville
5. **“Unsupervised automatic speech recognition: A review”**, Hanan Aldarmaki, Asad Ullah, and Nazar Zak

Reference

ASR:

intro.: https://www.youtube.com/watch?v=AIKu43goh-8&list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG&index=2

LAS: https://www.youtube.com/watch?v=BdUeBa6NbXA&list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG&index=3

CTC & RNN-T:
https://www.youtube.com/watch?v=CGuLuBaLIeI&list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG&index=4

LM: https://www.youtube.com/watch?v=dymfkWtVUdo&list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG&index=8

UASR:

https://www.youtube.com/watch?v=WROBoprE0js&list=PLJV_el3uVTsO07RpBYFsXg-bN5Lu0nhdG&index=25

感謝曾亮軒助教授權提供投影片與處理好的資料