JSALT Team meeting

SSL-Prosody

Guan-Ting Lin, Chi-Luen Feng
National Taiwan University

2022/4/24

Outline

- Progress on SUPERB-prosody downstream tasks
  - Turn-taking
  - Pitch reconstruction
  - Sarcasm detection
  - Discussion
- Possible new downstream tasks
- Prosody probing
- Hierarchical self-supervised pretrained model

Progress on SUPERB-prosody downstream tasks

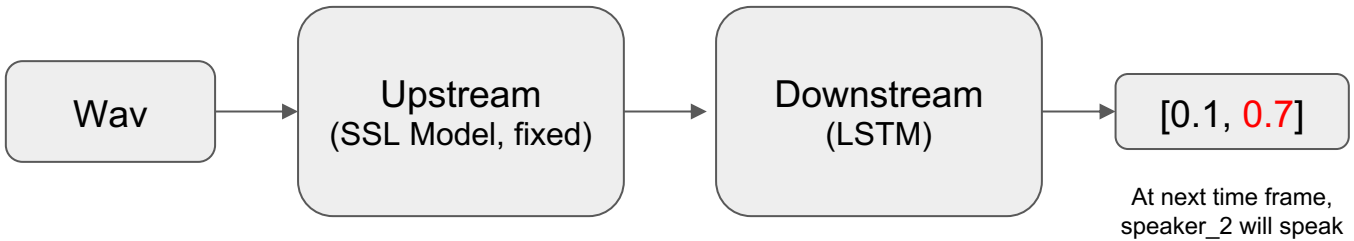
Turn-taking

Task setting

Task description: Predict who is speaking in the next time frame

Input: Wav files

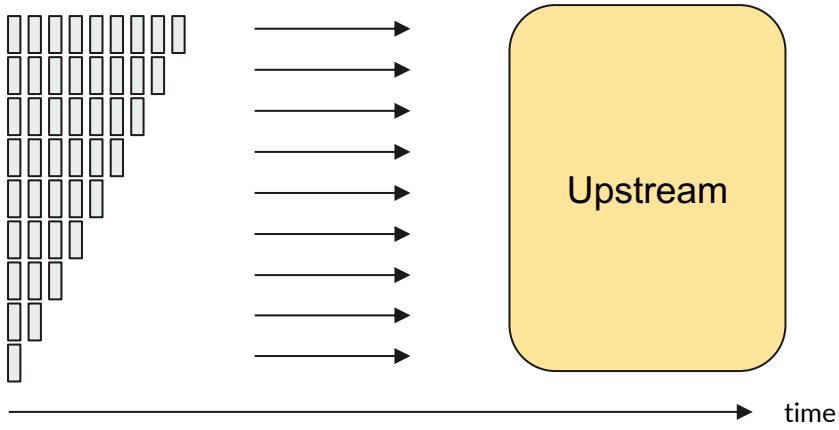
Output: Binary output (for two speakers setting)



Turn-taking

Progress and problem

1. **Finish Upstream/Downstream construction**
2. **Setting:**
   a. Dataset: Maptask
   b. Upstream: fbank/HuBERT
3. **Problem:**
   a. The training loss can't decrease
4. **Solution:**
   a. Use longer wavfile as Upstream input with mask mechanism



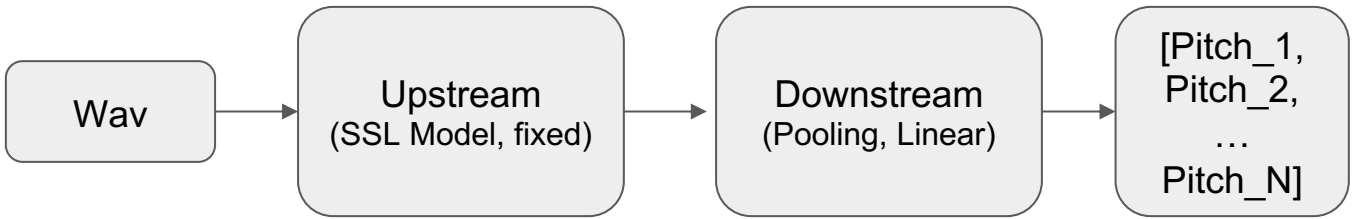
Pitch-reconstruction

Task setting

Task description: Reconstruct the pitch information for each frame of wav file

Input: Wav files

Output: Pitch for each frame



Pitch-reconstruction

Progress and result

1. **Finish Upstream/Downstream construction, test on different upstream model**
2. **Setting:**
    a. Dataset: LJSpeech, LibriTTS
    b. Upstream: fbank/mel/wav2vec2/HuBERT
    c. Label: “log-scale pitch” from PyWorld
    d. Metrics: Test loss
3. **Result:**
    a. Wav2vec2 get the best result
    b. Pitch information aggregate in first few layers

Pitch-reconstruction

Progress and result

Wav2vec2 get the best result:

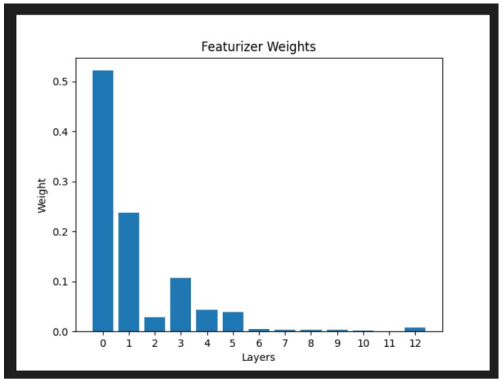
Wav2vec2 > HuBERT > Fbank > Mel

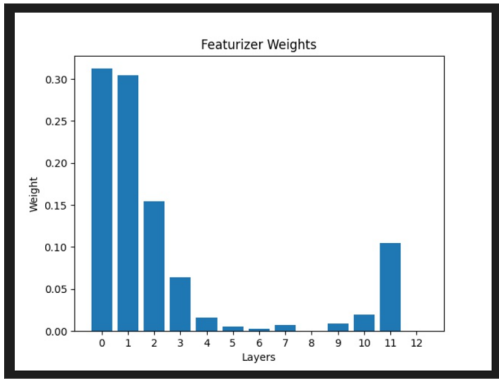
Test loss

	Fbank	Mel	HuBERT	Wav2vec2
LJSpeech	0.018	0.021	0.010	**0.009**
LibriTTS	0.125	bugged	0.021	**0.019**

Prosody information aggregate in first few layers:

Weights for each layer, first few layers have higher weight
(Left: HuBERT; Right: Wav2vec2)





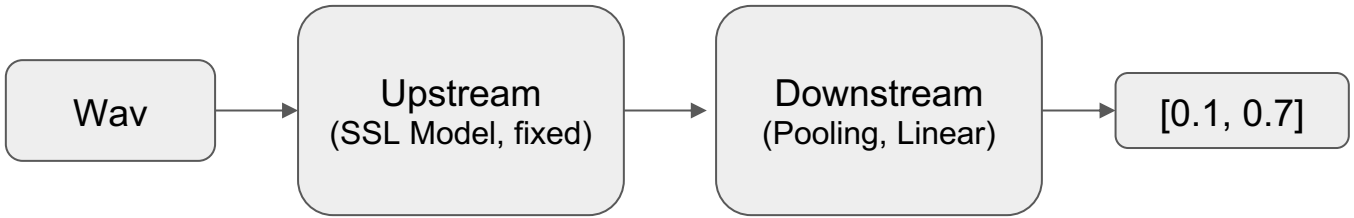
Sarcasm detection

Task setting

Task description: Predict whether a utterance contains sarcasm

Input: Wav files

Output: Binary classification



Sarcasm detection

Progress and result

1. **Finish Upstream/Downstream construction, test on different upstream model**
2. **Setting:**
    a. Dataset: MUStARD
    b. Upstream: fbank/mel/wav2vec2/HuBERT
    c. Downstream pooling method: Attentive pooling, Max pooling, Mean pooling
    d. Data augmentation: Shift/Gaussian noise/Stretch/Volume
3. **Result:**
    a. Get comparable result between MUStARD paper
    b. Use different data augmentation method to increase the performance

Sarcasm detection

Progress and result

Attentive Pooling & Max Pooling & Mean Pooling

Pooling method	F1-Score
Attentive	0.67
Max	0.63
Mean	0.73

**Best F1 score on MUStARD paper(Only audio): 62.7

Sarcasm detection

Progress and result

Data augmentation result:

1. Setting: Mean pooling
2. No major difference between baseline

Method	F1-score
Baseline(No aug)	0.7333
Gaussian	**0.7416**
Stretch	0.7333
Shift	0.7333
Volume	0.7333
All method	0.7356

All downstream tasks

Next step

1. **Try different upstream models**
2. **Fix bugs in turn-taking downstream task**
3. **Finish the downstream tasks in Prosody for SUPERB Leaderboard**

Discussion

Pitch reconstruction:

1. How to create a baseline for the pitch reconstruction (We can only know the relationship between different upstream model)
2. Fairness of different upstream model, we directly transform the output of upstream model to “1” dimension.
    a. Ex: fbank: 80 -> 1 ; HuBERT-L: 1,024 -> 1

Possible new downstream tasks

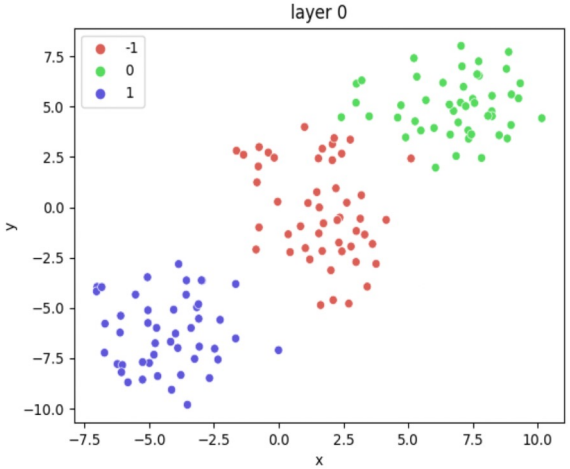
Prosody-related tasks

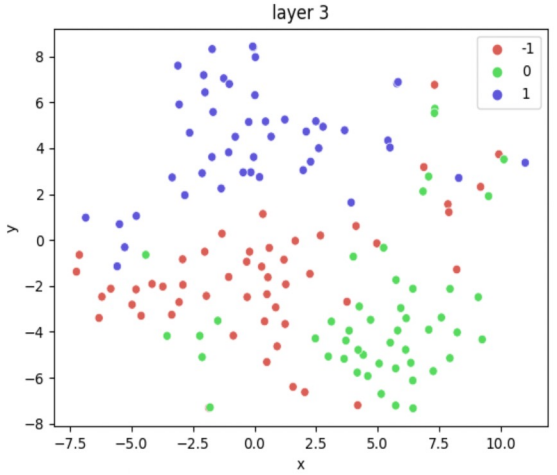
- Voice sentiment: CMU-MOSEI
- Depression Diagnosis: DAIC-WOZ
- More … (any suggestion?)

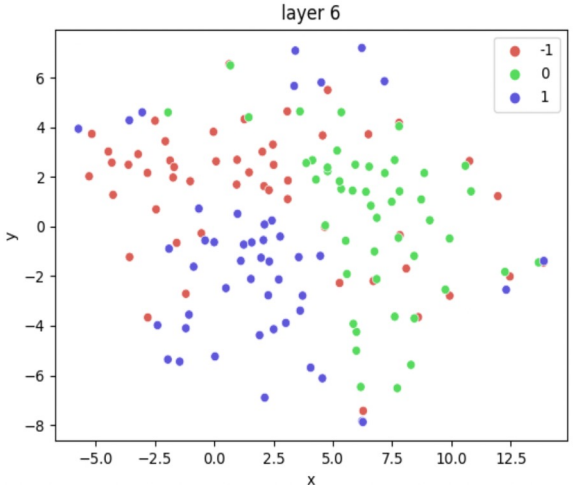
Prosody/voice sentiment probing

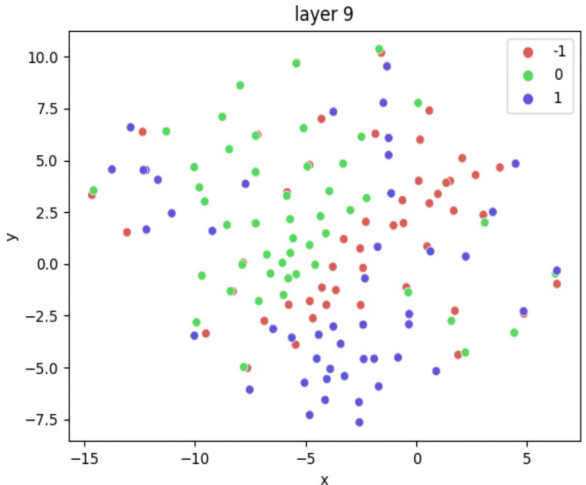
- Given the same speaker, speak same text sentence by different voice sentiments / emotions.
- Text sentences are neutral.
- Datasets
  - EmoV_DB (En)
  - Korean emotion speech (Kr)
- Speech SSL model: Hubert-base

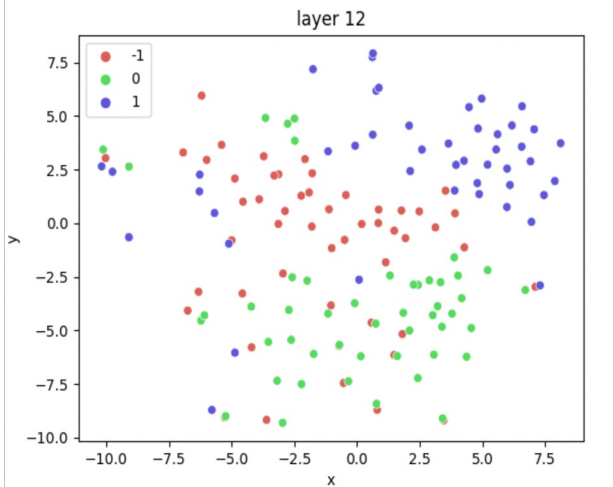
Speaker A


layer 0
-1
0
1
x
y

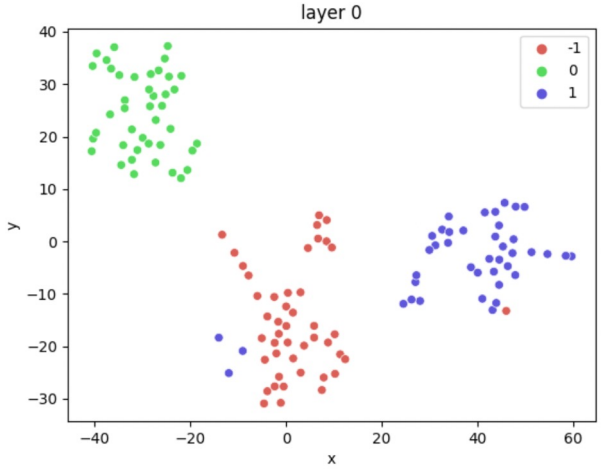

layer 3
-1
0
1
x
y

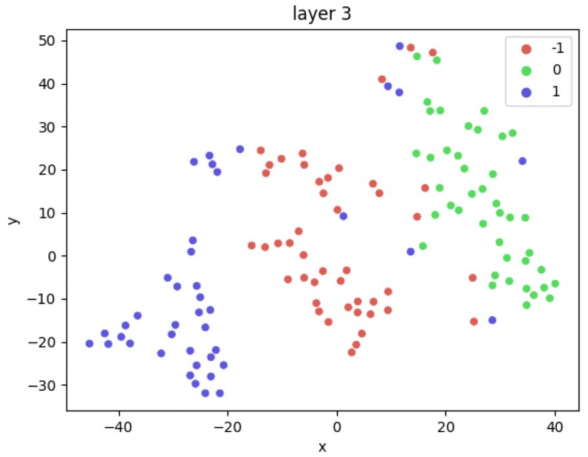

layer 6
-1
0
1
x
y

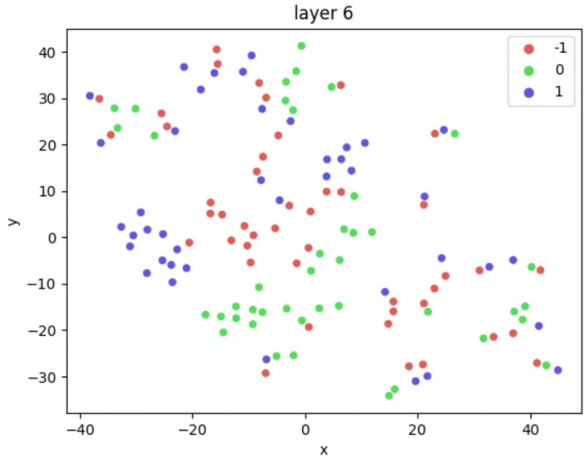

layer 9
-1
0
1
x
y

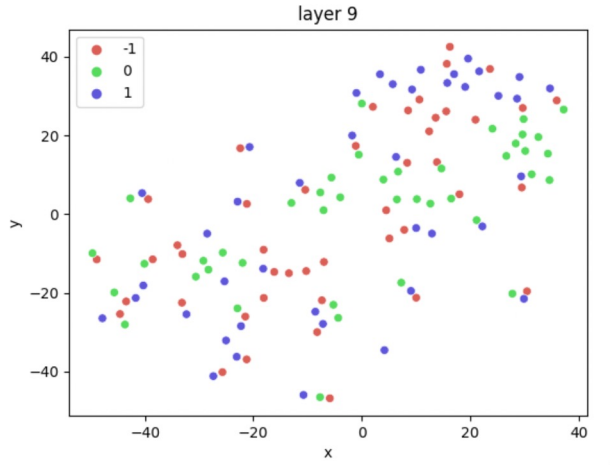

layer 12
-1
0
1
x
y

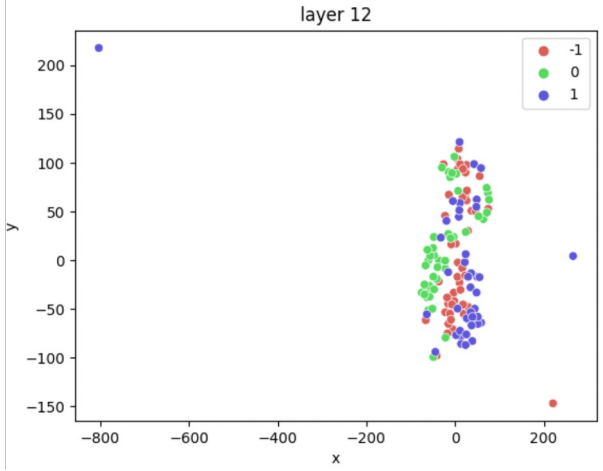
Speaker B


layer 0
-1
0
1
x
y


layer 3
-1
0
1
x
y

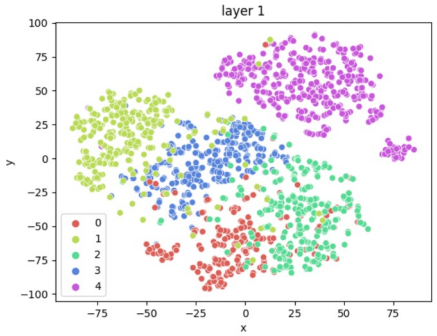

layer 6
-1
0
1
x
y

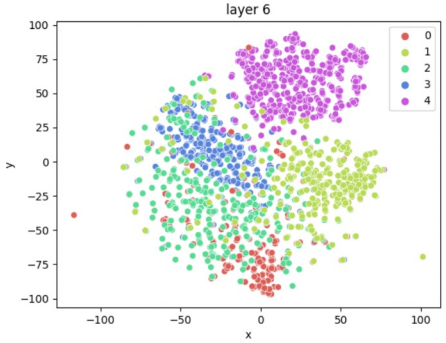

layer 9
-1
0
1
x
y

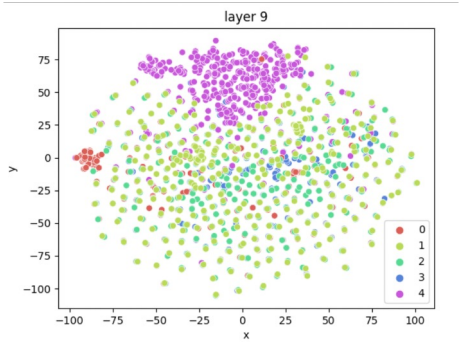

layer 12
-1
0
1
x
y

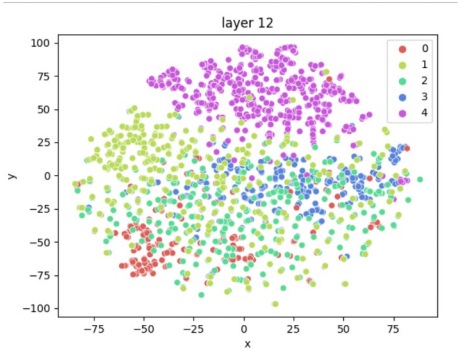
EmoV_DB

speaker bea

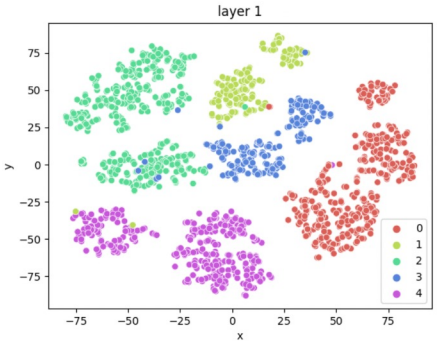


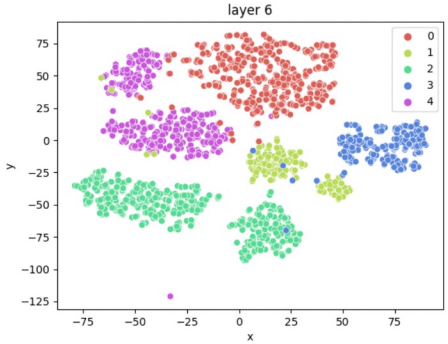


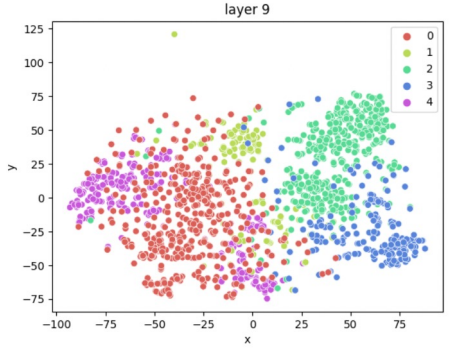


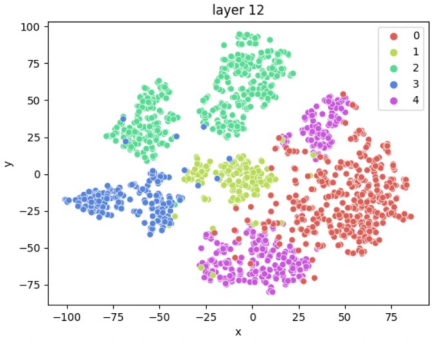


speaker jenie



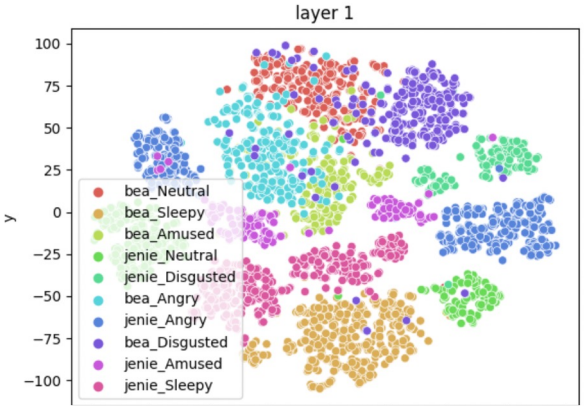


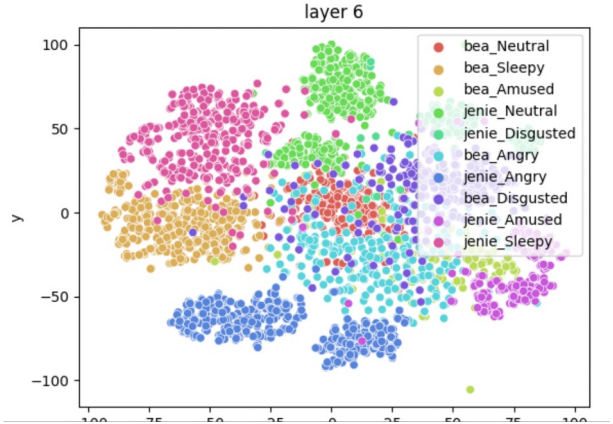


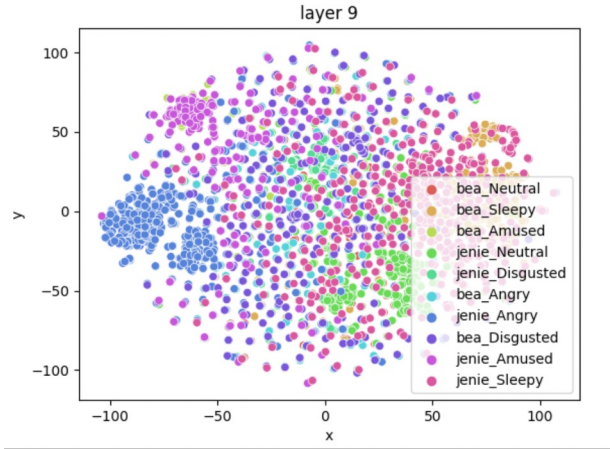


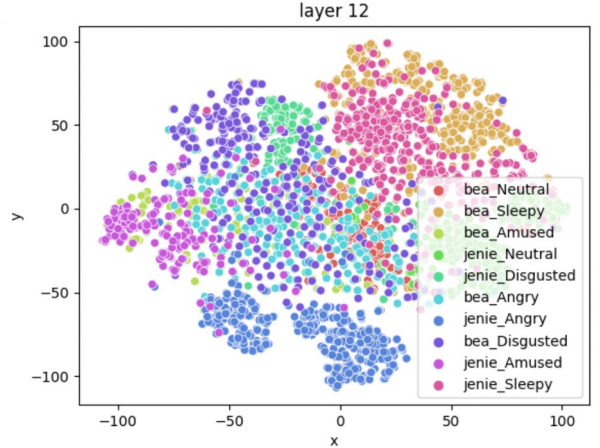
speaker bea, jenie

same voice sentiment in a cluster, but not in the same cluster for different speakers





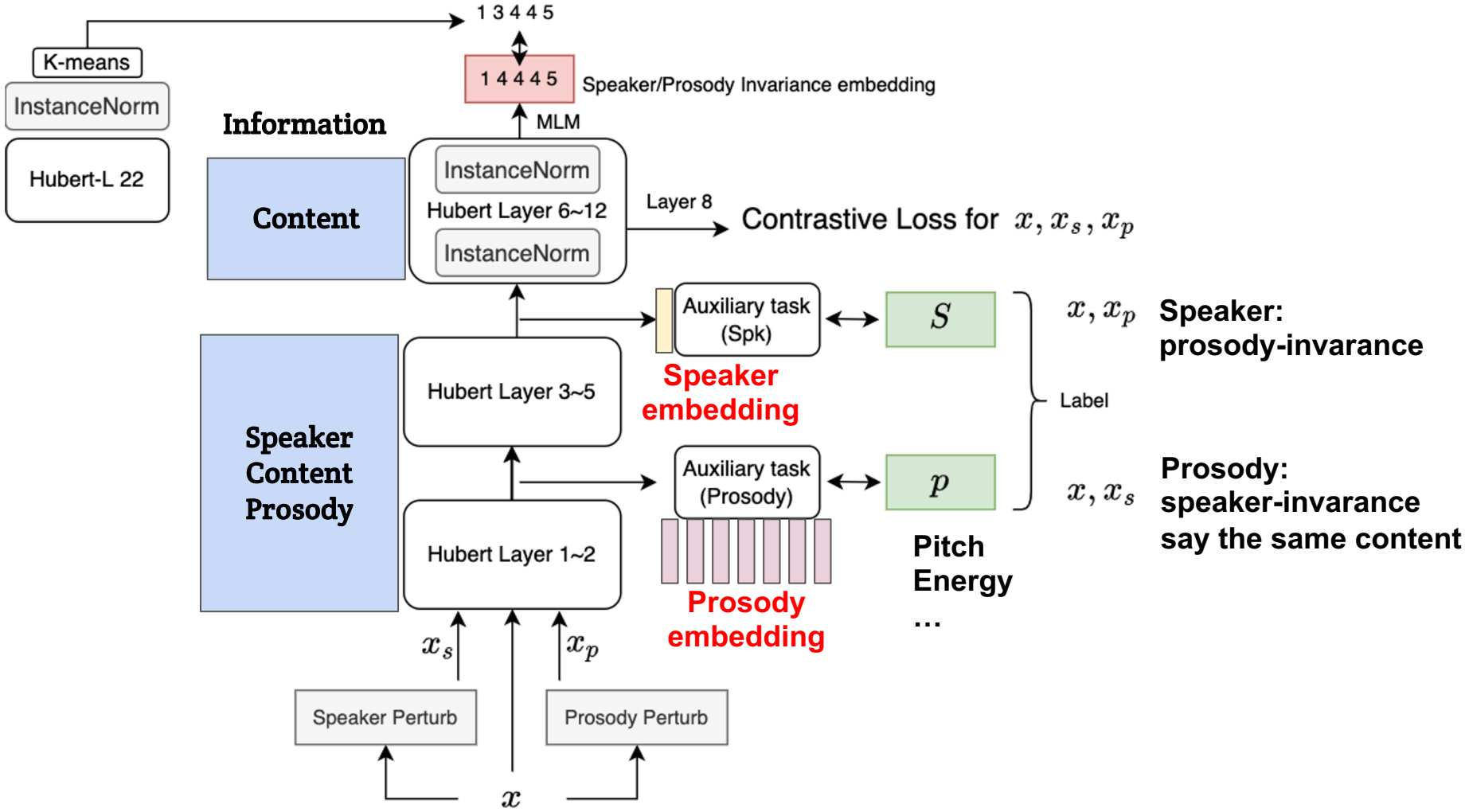




Hierarchical self-supervised pretrained model

Motivation

- Now we roughly know that speech self-supervised stored different types of information in different layers (**weakly disentanglement**)
  - Content: middle~top
  - speaker: bottom~middle
  - Prosody/paralinguistic: bottom
- Can we disentangle different information in a hierarchical manner by more guidance (**strongly disentanglement**)?
- Multi-task learning: by decomposing speech to hierarchical features, model can learn better and store all important information instead of focusing on one aspect (i.e. content).


1 3 4 4 5
K-means
InstanceNorm
Hubert-L 22
1 4 4 4 5
Speaker/Prosody Invariance embedding
MLM
Information
Content
InstanceNorm
Hubert Layer 6~12
InstanceNorm
Layer 8
Contrastive Loss for x, x_s, x_p
Auxiliary task (Spk)
S
x, x_p
Speaker: prosody-invarance
Speaker embedding
Speaker Content Prosody
Hubert Layer 3~5
Label
Auxiliary task (Prosody)
p
x, x_s
Prosody: speaker-invarance say the same content
Pitch
Energy
…
Hubert Layer 1~2
Prosody embedding
x_s
x_p
Speaker Perturb
Prosody Perturb
x