



Prosody - Progress

Guan-Ting Lin, Chi-Luen Feng

National Taiwan University

2022/6/12



Outline

- Overview
- Main downstream tasks
 - Turn taking
 - Sentiment analysis
 - Prosody reconstruction
 - Next frame prediction of prosody
 - Sarcasm detection
- Future work

Overview



Prosody - Quick recap

What we want to achieve:

- We want to create a **prosody-related track** in SUPERB, help others test SSL models for prosody information

What are our plan:

- Create downstream tasks which related to prosody information
- Turn taking, Prosody reconstruction, Sarcasm detection...



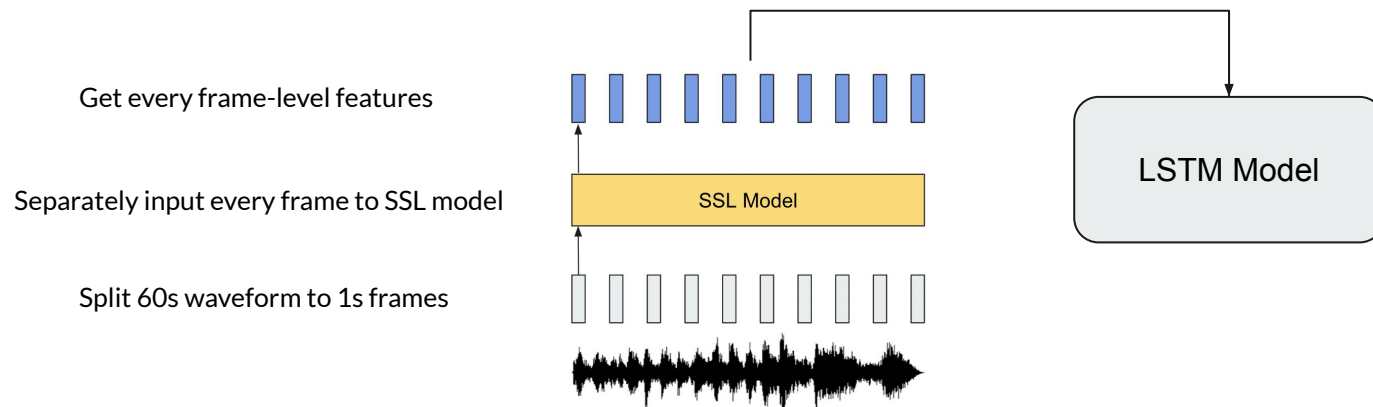
Prosody - overall progress

	Prepare	In progress	Have some results	Finish
Turn taking				
Sentiment analysis				
Prosody reconstruction				
Next frame prediction of prosody				
Sarcasm detection				
Persuasiveness prediction				

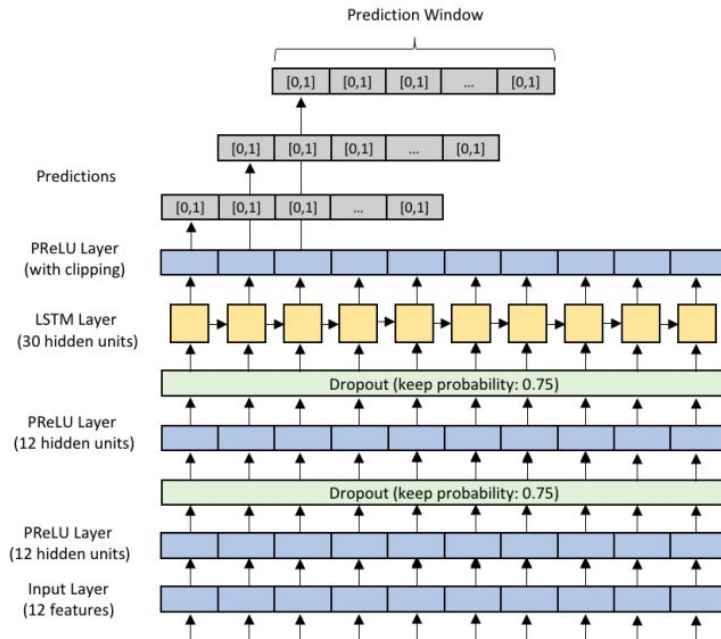
Turn taking

Turn Taking - Setup

- Split waveform to frames
- Get representation
- Put to downstream model

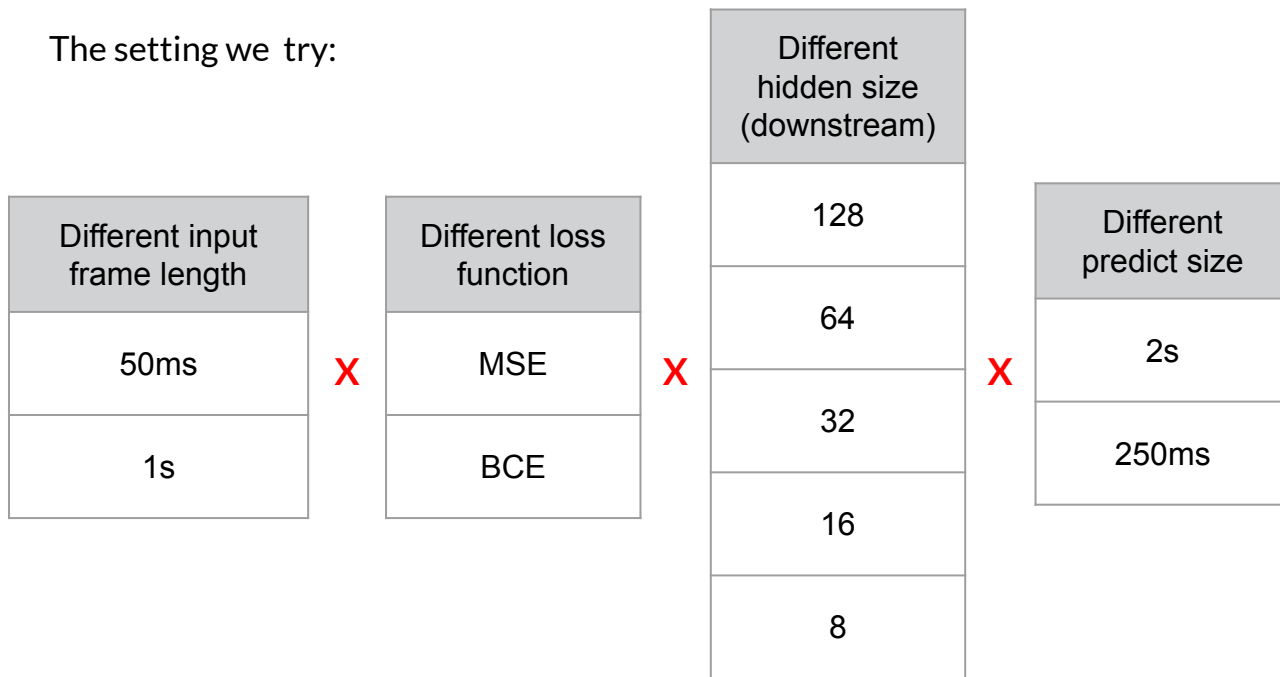


Turn Taking - Downstream setup



Turn Taking - What we have done

The setting we try:



Train accuracy:
above **90%**

Test accuracy:
around **70%**



Turn Taking - Try to reproduce the original work.

- [Turn-Taking Predictions across Languages and Genres Using an LSTM Recurrent Neural Network \(SLT 2018\)](#)
- We use the tools, trying to reproduce the work
 - The toolkit used Matlab, and there are some technical issues when we used this toolkit
- Our teammates rewrite the code last week and will test these codes.



Turn taking - Conclusion

Conclusion

1. Overfit for the dataset, can't get comparable results
2. Test previous work:
 - a. Face some problem when trying to reproduce original outcome

Future work

1. First we will try to reproduce the result in the paper
2. We will set a deadline, if we can't finish this downstream task, we will move on to other tasks

Sentiment analysis



Sentiment analysis

Task setting

- Given a wav file, we will try to predict the sentiment label of this utterance
- Classification task, 2 and 7 labels classification

Dataset:

- CMU-MOSEI
- Each utterance is annotated for sentiment on a $[-3,3]$



Sentiment analysis - Downstream setup

Label setting:

- **Setting 1:** 2 labels (negative / non-negative) $[-3, 0]$ $[0, 3]$ (Follow setting prior to 2019)
- **Setting 2:** 2 labels (negative / positive) $[-3, 0]$ $(0, 3]$ (Follow setting after 2019)
- **Setting 3:** 7 labels $[-3, 3]$

Downstream model setting

- Simple linear layer with mean pooling
- Macro F1 score

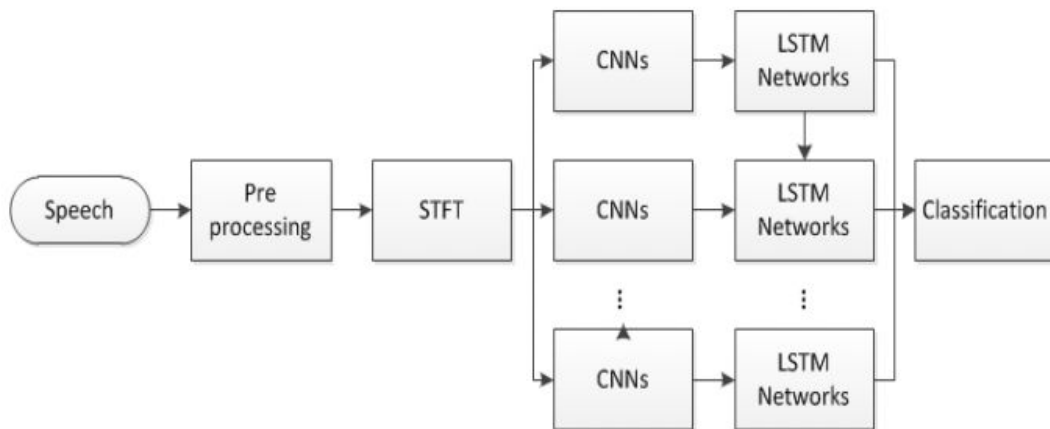
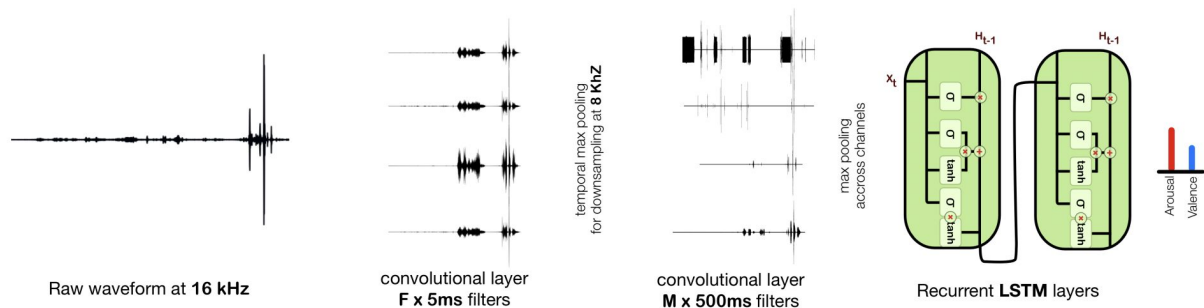


Sentiment analysis - Previous SOTAs

Previous work (Our Comparison):

- Because SOTA1 and SOTA2 is prior to 2019, so we include Mockingjay which is after 2019

SOTA1	SOTA2	Mockingjay
Upstream: COVAREP software Downstream: AdieuNet	Upstream: MFCC Downstream: SER-LSTM	Just Mockingjay model






Sentiment analysis - 2 labels conclusion

In two setting, the SSL model always get better results

<i>Prior 2019</i>	Test acc
SOTA1	74.2
SOTA2	74.2
fbank	71.60
wav2vec 2.0 base	75.74
HuBERT base	77.45

<i>After 2019</i>	Test acc
Mockingjay large	71.05
fbank	64.39
wav2vec 2.0 base	73.66
HuBERT base	75.01



Sentiment analysis - 7 labels conclusion

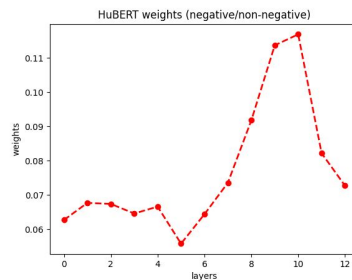
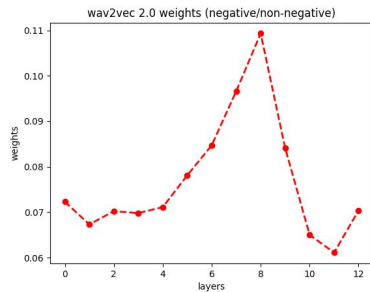
In this setting, previous work get better results, but SSL model is better than fbank

<i>Prior 2019</i>	Test acc
SOTA1	42.1
SOTA2	42.4
fbank	32.75
wav2vec 2.0 base	37.71
HuBERT base	39.49

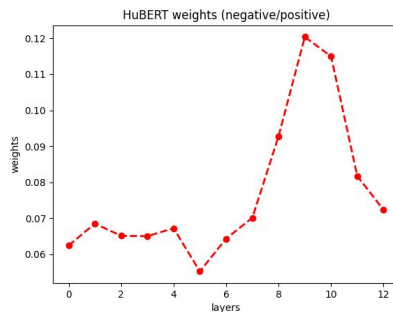
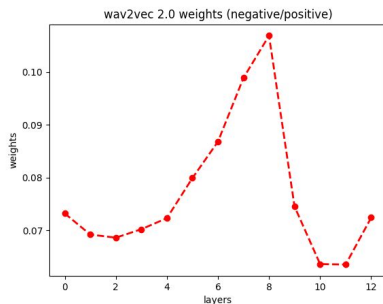
Because we finish this task before one weeks, we will try different learning rate or other setting to get better result

Sentiment analysis - layerwise analysis

2 labels
(prior 2019)



2 labels
(after 2019)



x-axis: weights
y-axis: layer number

left: wav2vec 2
right: hubert

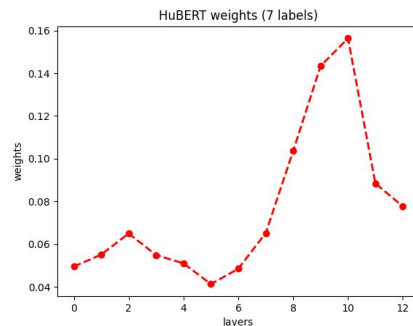
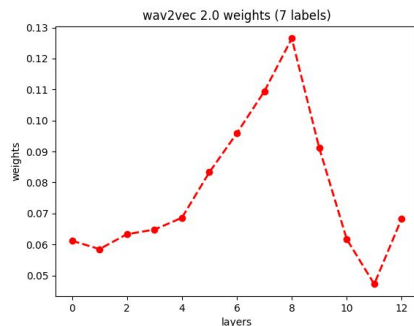
Last few layers
have larger
weight

Sentiment analysis - layerwise analysis

x-axis: weights
y-axis: layer number

left: wav2vec 2
right: hubert

7 labels



Last few layers
have larger
weight



Conclusion

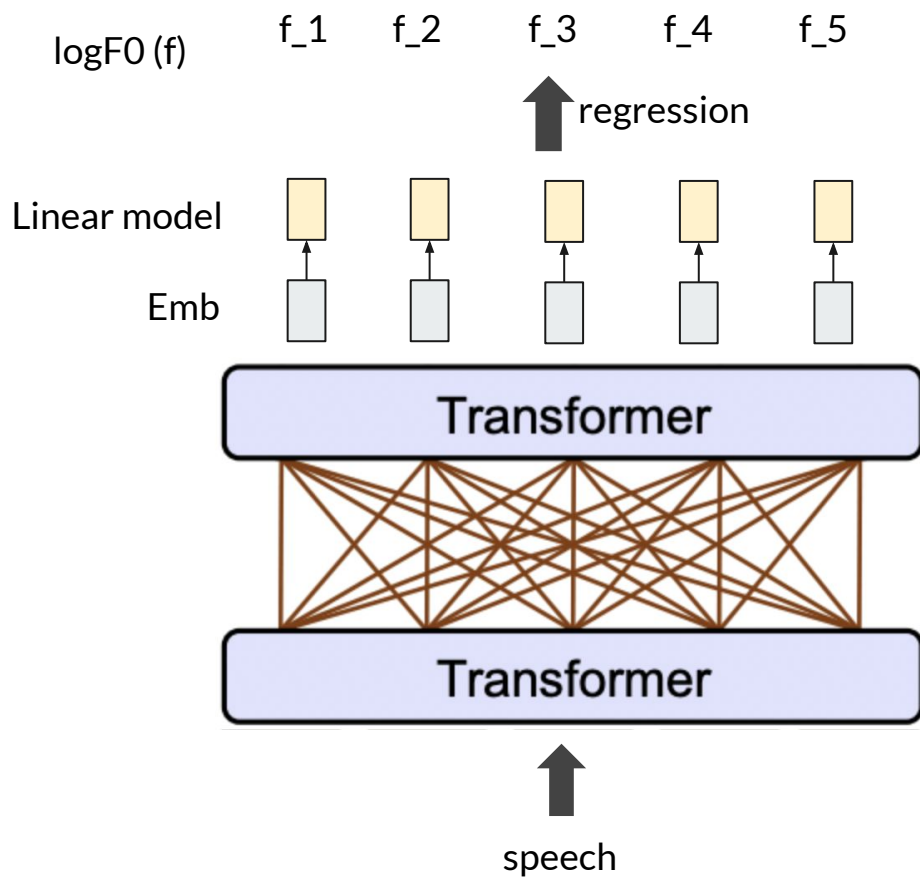
1. Get comparable results on 2 labels setting
2. Still need to do more experiments on 7 labels setting
3. After see the weight graph, have to test whether the first few layers information (prosody) help with the sentiment analysis
 - a. **Only use last few layers v.s. Use first and last few layers together**

Prosody reconstruction



Introduction

- Goal:
 - Whether SSL model extract more accessible prosody feature?
 - Reconstruct rule-based prosodic feature by frozen SSL model with the lightweight downstream model





Task setup

- **Pitch reconstruction**
 - Use LibriTTS dataset
 - Use pYAPPT to extract pitch groundtruth.
 - Use LogMSE loss, ignore zero pitch frames.
- **Energy reconstruction**
 - Use LibriTTS dataset
 - Use LogMSE loss



Experimental setup

- Model
 - Upstream: fbank, wav2vec, wav2vec2, hubert
 - Downstream: linear layer (model dim \rightarrow 1)
- Special model ***fbank-fair***: Increase the number of downstream parameters of fbank for fair comparison (even give some advantage) to SSL models.
- Provide additional 3 baselines
 - Predict a constant (average).
 - Predict a constant (average) for each speaker.
 - Predict a constant (average) for each utterance.



Implementation Details

- LibriTTS data split: train-clean-100, dev-clean, test-clean
- Filter out >15s audios for larger batch size (32).
- For pitch reconstruction, discard frames with undefined pitch.
- Downstream model is single linear layer (upstream dim \rightarrow 1).



Upstream Comparison

Test Loss - LogMSE

	fbank	fbank-fair	hubert	wav2vec	wav2vec2	All Avg.	Spk Avg.	Utter Avg.
Pitch	0.089	0.051	0.018	0.035	0.022	0.124	0.057	0.050
Energy	0.517	0.429	0.243	0.318	0.240	2.567	2.478	2.401

Conclusion: SSL models > fbank and naive baselines



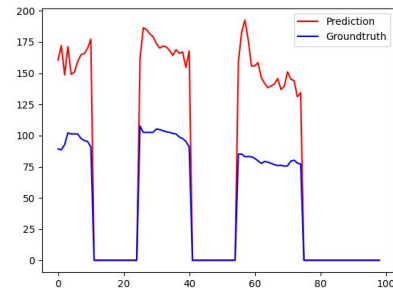
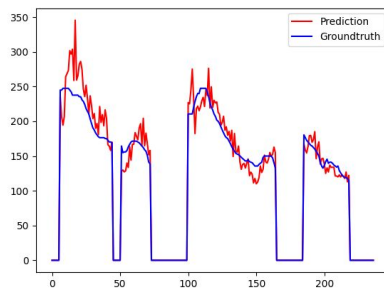
Visualization

- Prediction results
 - Show pitch / energy reconstruction without zero frames.
 - From top row to bottom row: fbank, hubert, wav2vec2.
- Weight analysis
 - Considering features' norms of different layers, averaged from 1024 samples.

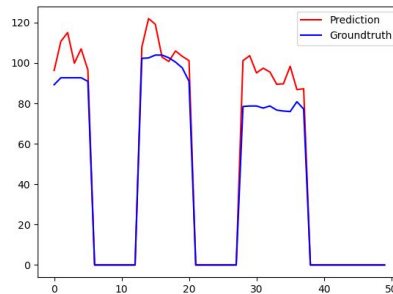
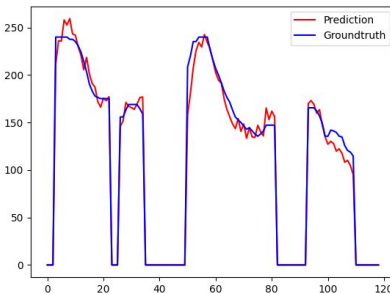
fbank



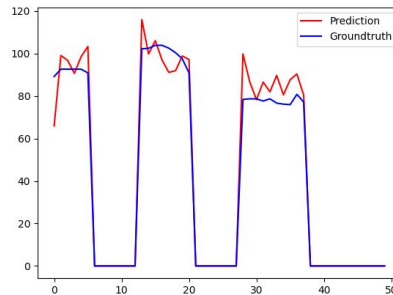
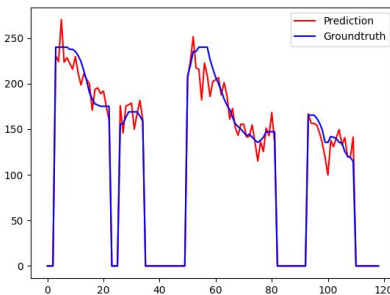
Pitch reconstruction



hubert



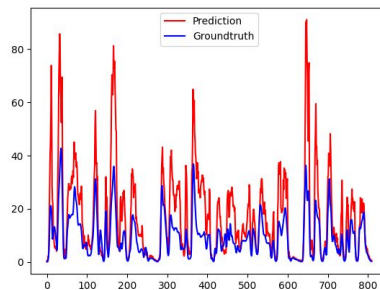
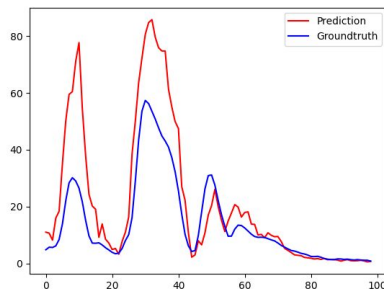
wav2vec 2.0



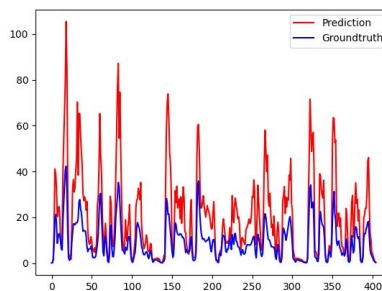
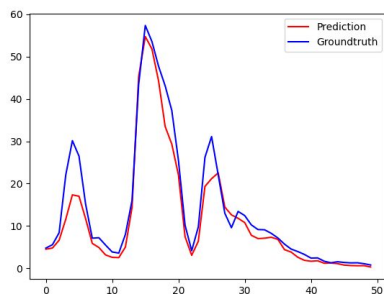
Energy reconstruction



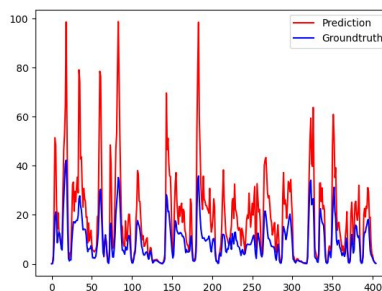
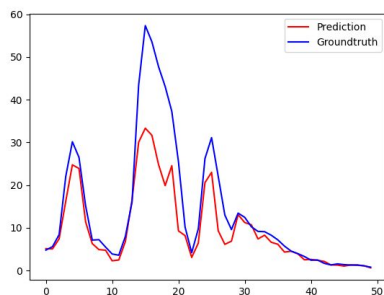
fbank



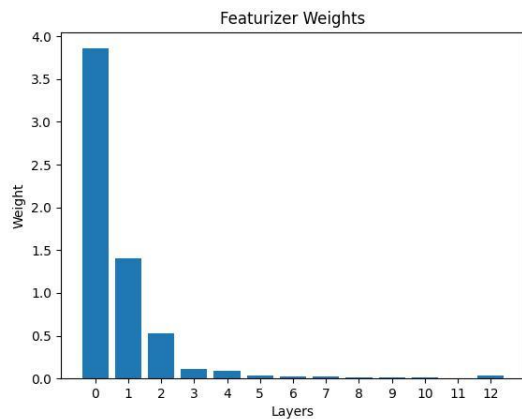
hubert



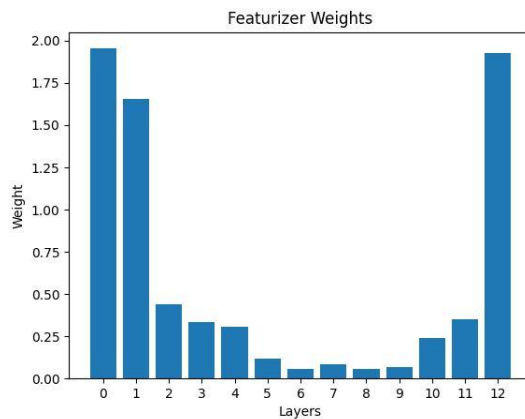
wav2vec 2.0



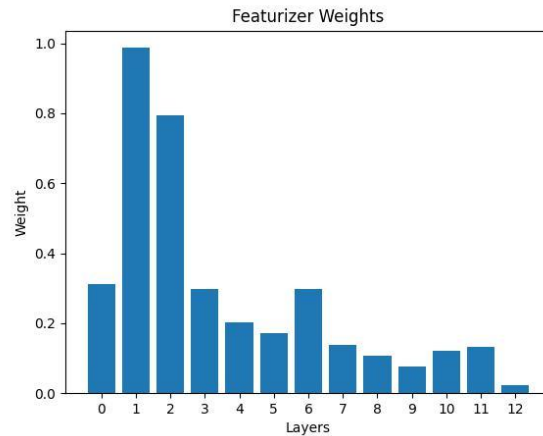
Weight visualize (pitch)



hubert

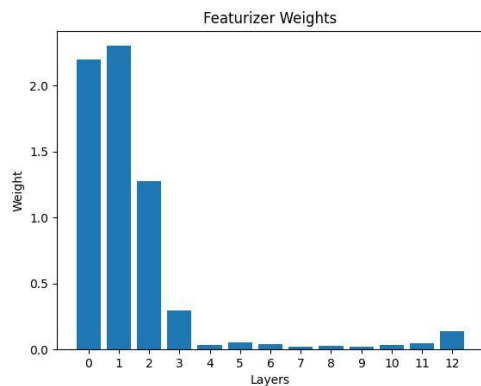


wav2vec 2.0

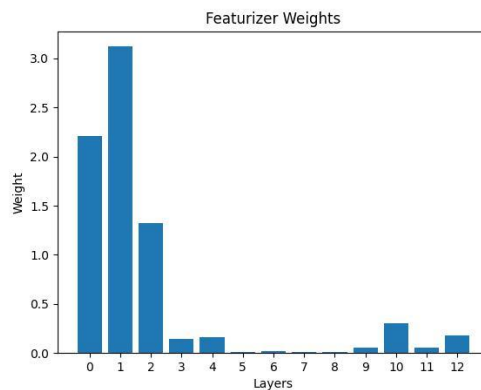


wav2vec

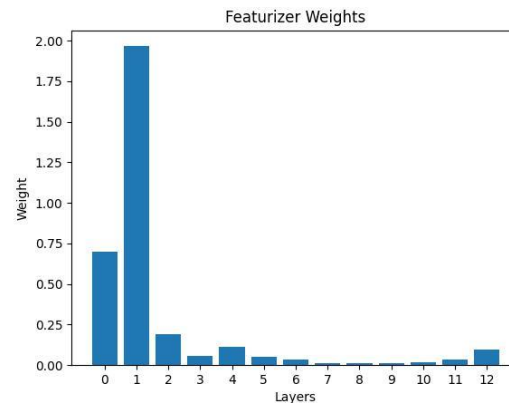
Weight visualize (energy)



hubert



wav2vec 2.0



wav2vec



Conclusion & future work

- The SSL models provide more accessible representation to reconstruct prosody-related features with a simple linear layer
- SSL models tend to store prosodic information in the first few layers

TODO

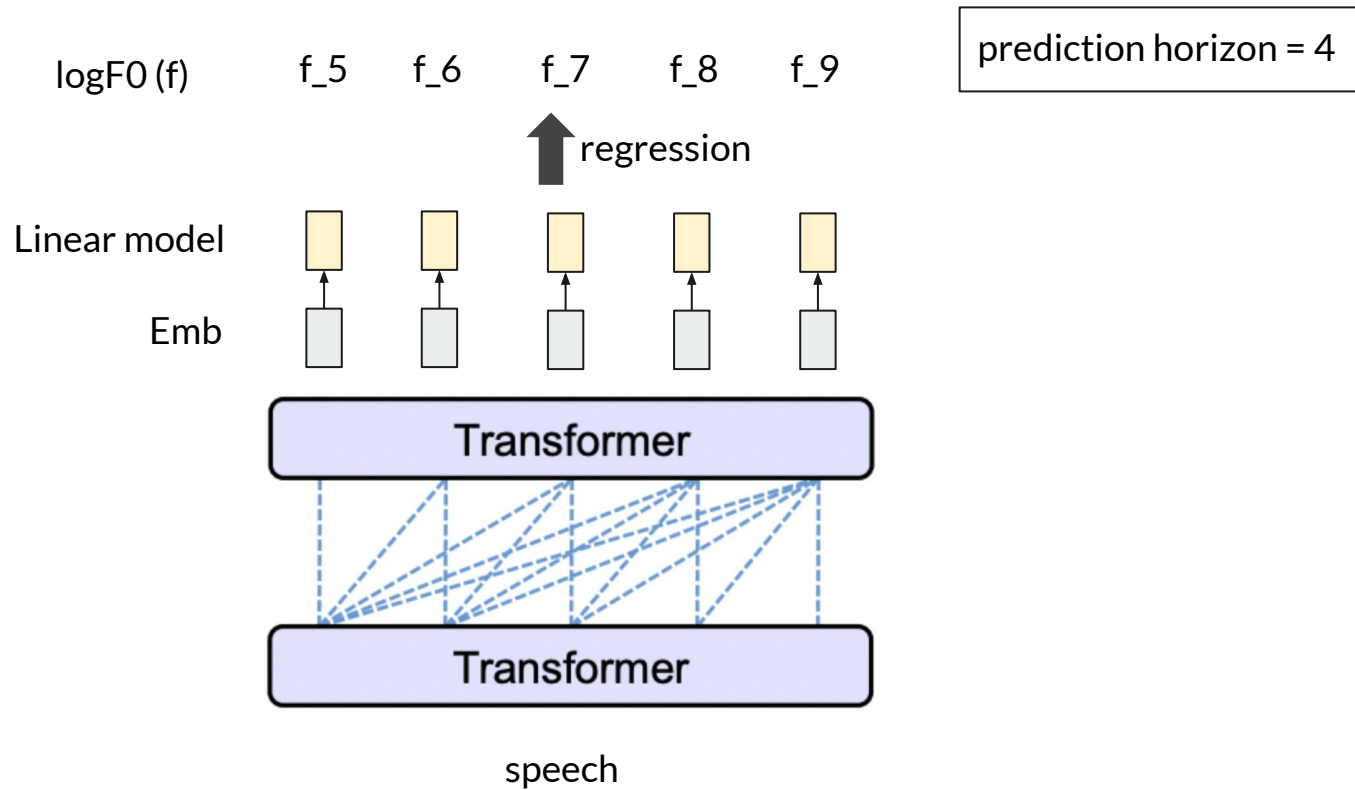
- Use other popular pitch extractors
- Run on wide range of SSL models

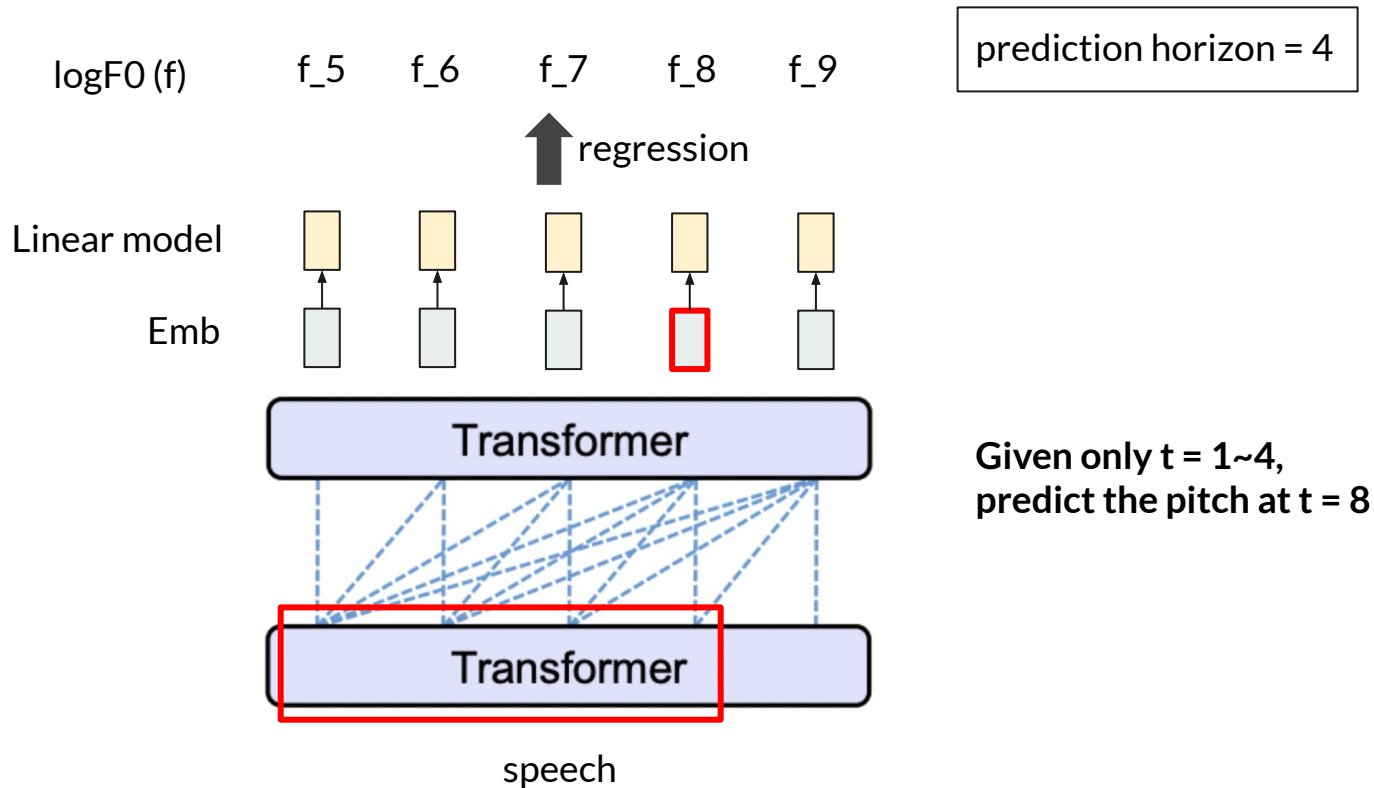
Next frame prediction of prosody



Next Frame Prediction

- Beside reconstruction of prosody information, can we **predict prosody information in the future time frames**?
- Similar setup as prosody reconstruction task, but a few modifications:
 - The prediction target is prosody feature in the future time frames.
 - Most SSL models contain self-attention layer: Adopt the **causal attention mask** to avoid leaking the future information.







Next Frame Prediction

fbank

Prediction horizon	Pitch	Energy
2	0.094	0.518
4	0.094	0.558
6	0.099	0.835

hubert

Prediction horizon	Pitch	Energy
2	0.020	0.305
4	0.022	0.329
6	0.026	0.402



Next Frame Prediction

- To validate that shifting does provide different target distribution, compare to the loss using **baseline predictions (with no shift)**.
- **Frame shift = 6 (120 ms)** is chosen.

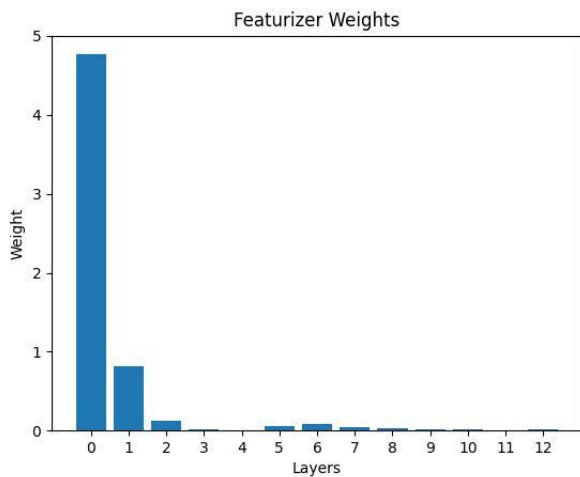
hubert (pred/baseline pred)

Prediction horizon	Pitch	Energy
2	0.020/0.029	0.305/1.309
4	0.022/0.045	0.329/2.519
6	0.026/0.056	0.402/3.124

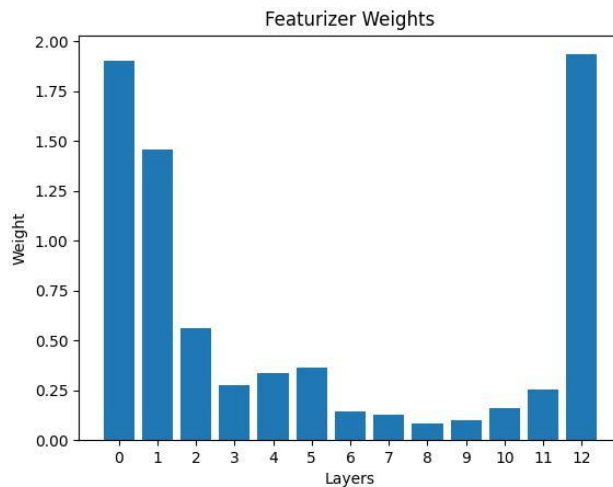
Prediction horizon = 6 (120 ms)

Upstream	Pitch	Energy
fbank	0.099	0.835
hubert	0.026	0.402
wav2vec2	0.034	0.429

Weight visualize (pitch)

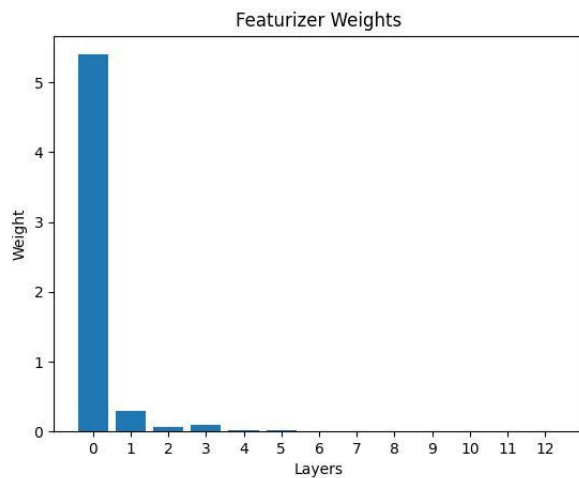


Hubert

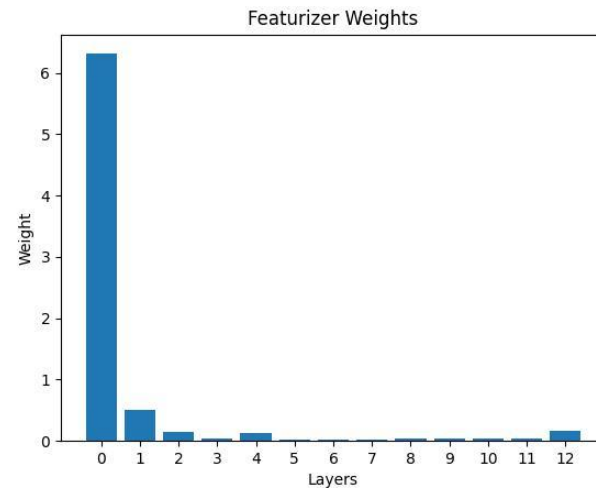


wav2vec 2.0

Weight visualize (energy)



Hubert



wav2vec 2.0

Sarcasm detection

Sarcasm in Speech

- Previously, **hand-crafted** prosodic feature or spectrogram
- Prosodic cues
 - such as slower speaking rate and greater intensity

Utterances

1)



Chandler : Yes and we are very excited about it.

2)



SA_man: You got off to a really good start with the group.

Remarks

- **Text** and **Video**: positive indication.
- **Audio** : stressed word

Figure 6: Vocal stress in sarcasm.

SUPERB prosody track - MUsTARD dataset

sarcasm detection

American TV shows, conversational

1. Problem definition:

Binary classify whether the target utterance is sarcastic or not (given the context audio)

2. Format:

- Input: target utterance (with Context conversation)
- Output: Sarcastic / non-sarcastic



Reference:

[Towards Multimodal Sarcasm Detection](#)

The simple baseline from MUsTARD

Algorithm	Modality	Precision	Recall	F-Score
Majority	-	32.8	57.3	41.7
Random	-	51.1	50.2	50.4
SVM	T	60.9	59.6	59.8
	A	65.1	62.6	62.7
	V	54.9	53.4	53.6
	T+A	64.7	62.9	63.1
	T+V	62.2	61.5	61.7
	A+V	64.1	61.8	61.9
	T+A+V	64.3	62.6	62.8
$\Delta_{multi-unimodal}$		$\downarrow 0.4\%$	$\uparrow 0.3\%$	$\uparrow 0.4\%$
Error rate reduction		$\downarrow 1.1\%$	$\uparrow 0.8\%$	$\uparrow 1.1\%$

Audio clue is important for sarcasm detection !

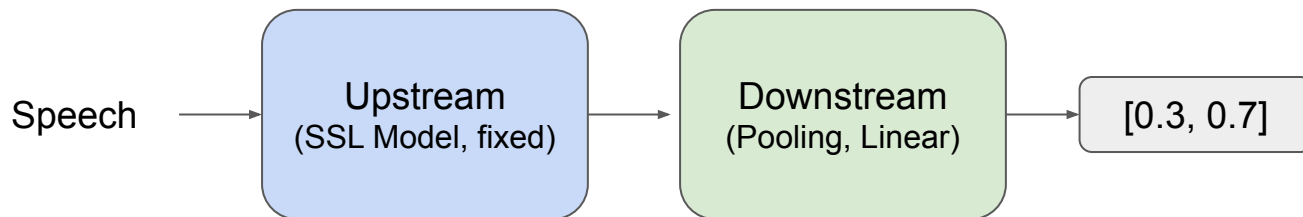
Table 3: Multimodal sarcasm classification. Evaluated using an speaker-independent setup. Note: T=text, A=audio, V=video.

Task setting

Task description: Binary classify whether a utterance contains sarcasm

Upstream: Hubert, fbank

Downstream: mean-pooling + linear model



Experiment Setup

Speaker Dependent

- Five-fold cross validation in a stratified manner

Speaker Independent

- Training set: Big Bang Theory, The Golden Girls and Sarcasmaholics Anonymous
- Testing set: Friends

Best Results: Speaker Dependent

Upstream	Layer	Downstream	Aug	F1	Precision	Recall
Fbank		Linear	No	0.6493	0.6506	0.6493
Hubert	3	Linear	No	0.6868	0.6984	0.6884
Hubert	3	SVM	No	0.6903	0.6953	0.6899
Baseline		SVM		0.646	0.659	0.646

Best Results: Speaker Independent

Upstream	Layer	Downstream	Aug	F1	Precision	Recall
Fbank		Linear	No	0.6943	0.6942	0.6966
Hubert	1	Linear	No	0.7092	0.7134	0.7079
Baseline		SVM		0.627	0.651	0.626

Speaker Independent

- Upstream: Hubert
- Downstream: Linear
- Learning rate: 3e-5
- Batch size: 32
- Step: 3000
- No augmentation

Layer	0	1	2	3	4	5	6
F1	0.6519	0.7092	0.7056	0.6234	0.6281	0.6499	0.6899
Precision	0.7309	0.7134	0.7311	0.6670	0.6421	0.6701	0.6983
Recall	0.6601	0.7079	0.7051	0.6264	0.6264	0.6489	0.6882

Layer	7	8	9	10	11	12	W.S.
F1	0.6758	0.6241	0.6279	0.6479	0.6273	0.6209	0.6647
Precision	0.6817	0.6478	0.6445	0.6589	0.6485	0.6471	0.6735
Recall	0.6742	0.6236	0.6264	0.6460	0.6264	0.6208	0.6629

Speaker Dependent

- Upstream: Hubert
- Downstream: Linear
- Learning rate: 3e-5
- Batch size: 32
- Step: 1000
- No augmentation

Layer	0	1	2	3	4	5	6
F1	0.6590	0.6743	0.6687	0.6868	0.6794	0.6745	0.6791
Precision	0.6815	0.7012	0.6821	0.6984	0.6906	0.6819	0.6869
Recall	0.6638	0.6797	0.6710	0.6884	0.6812	0.6754	0.6797

Layer	7	8	9	10	11	12	W.S.
F1	0.6768	0.6770	0.6708	0.6743	0.6707	0.6621	0.6752
Precision	0.6806	0.6804	0.6769	0.6766	0.6785	0.6669	0.6874
Recall	0.6768	0.6768	0.6710	0.6739	0.6710	0.6623	0.6768

Conclusion & future work

- Frozen SSL model with linear model outperforms previous baseline
- Obtain better performance while using the representation from the first few layer
- Weighted sum method does not yield better performance

TODO

- Evaluate on wide range of SSL models
- Hyper-parameter tuning

Future plan & Timeline



Future work

1. Finish testing on wide range of upstream models
 - Use the upstream setting as SUPERB paper
2. Persuasiveness prediction task
3. Analyze the performance of different upstream models according to their characteristics.
4. Submit paper to SLT
5. [Sam project]: Better leverage the content and prosody information in SSL model for the tasks that require both information:
 - Speech sentiment analysis
 - Speech emotion recognition
 - Paralinguistic transferability of different languages



Timeline

June	July	Aug
<ul style="list-style-type: none">- Persuasiveness prediction task preparation- SLT paper draft- Run on more SSL models	<ul style="list-style-type: none">- Analysis- Start Sam project- Submit paper to SLT (7/21)- Finalize the downstream tasks on superb-prosody repo	<ul style="list-style-type: none">- Finish Sam project- Propose future research directions