Pretrained Models for Prosody: Track Plan

Guan-Ting Lin, Chi-Luen Feng, Nigel Ward, Diego Aguirre, Hung-Yi Lee

Outline

- Original concept
- Changes in direction

Ward, 5 minutes

- New tasks for SUPERB

Lin and Feng, 5 minutes

- Other thoughts
- Discussion

all, 7 minutes

Original Concept

- Current pretrained models probably ignore all interesting prosody
  - Yet they outperform MFCCs, which convey prosodic information
- Evaluation sets for pretrained models are deficient in dialog-specific and pragmatics-related functions
  - Yet this is changing: SLUE (Shon 2021), CALC (Weston, 2021)

Aims

1. Augment SUPERB with prosody-intensive tasks

   (pre-workshop: Guan-Ting Lin, Chi-Luen Feng, Nigel Ward)

2. Characterize adequacy of existing pretrained models for these tasks

   (at the workshop: quantitative analysis + failure analysis? by who?)

3. Side Activities on prosody and dialog

Prosody track plan

Guan-Ting Lin, Chi-Luen Feng

Outline

- Introduction
- Three main tasks (Finish before JSALT)
  - Turn taking
  - Pitch reconstruction
  - Sarcasm detection
- Potential Future work (During JSALT)
- Timeline overview

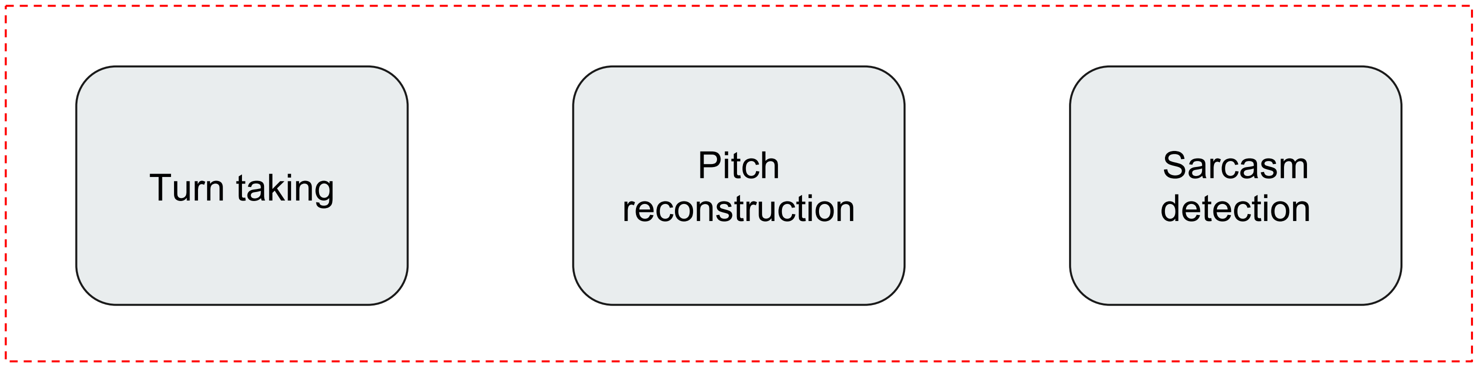
Introduction

SUPERB prosody track preparation

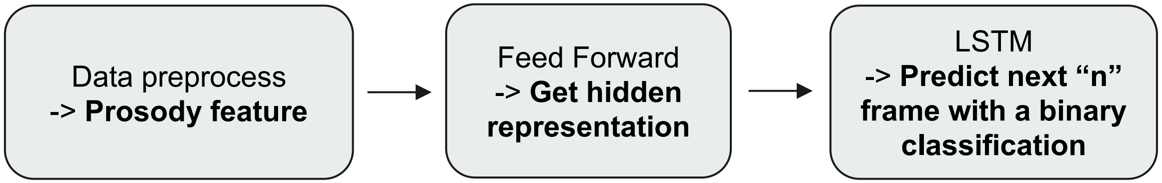
SUPERB toolkit structure and limitation

1. There are no exist toolkit to measure the prosody aspect of model
2. The goal is to construct a “**Prosody track**” in SUPERB toolkit

GOAL



Turn taking


Data preprocess
-> **Prosody feature**
Feed Forward
-> **Get hidden representation**
LSTM
-> **Predict next "n" frame with a binary classification**

Turn taking:

1. **Given a dialogue data(for two persons), try to predict who will speak at next time frame**
2. Example of input/output
   a. Input: Conversation between two person
   b. Output: At time t, speaker 1 will speak, speaker 2 will be quiet
3. Expectation:
   a. Extract useful information(ex: prosody feature) to increase the performance

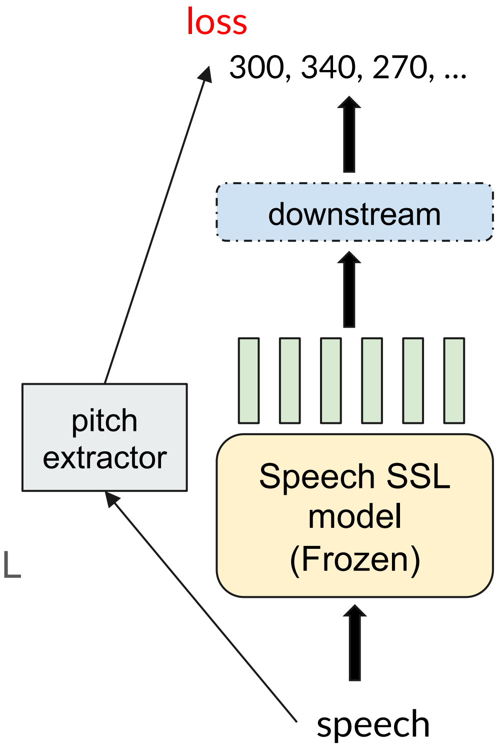
Pitch reconstruction

Probing task:

- **Given:** Raw waveform + Frozen speech ssl model + light downstream model
- **Goal:** Reconstruct quantized/ continuous pitch

Expectation:

- The reconstruction is nearly perfect -> give us confident that SSL models are rich in prosody!
- To find which **layer** of SSL model encodes most pitch related information



SUPERB prosody track - MUStARD dataset

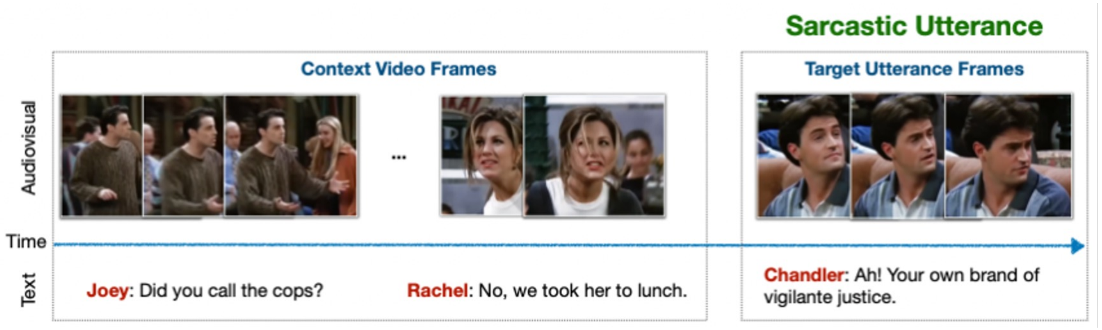
sarcasm detection

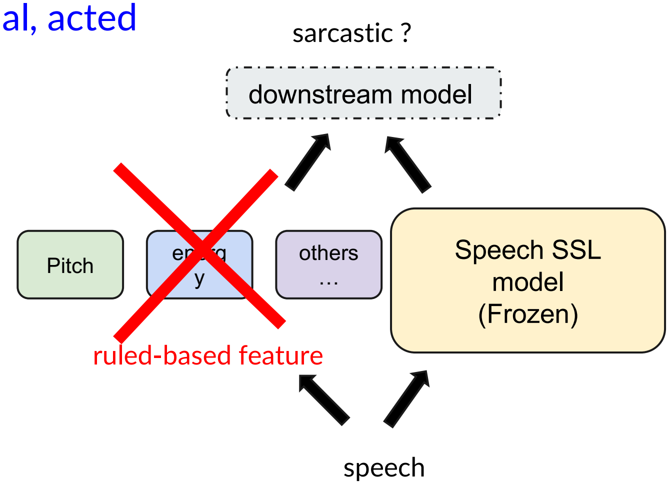
American TV shows, conversational, acted

1. **Problem definition:**

   Binary classify whether the target utterance is sarcastic or not

1. **Format:**
   a. Input: target utterance (with Context conversation)
   b. Output: Sarcastic / non-sarcastic


Sarcastic Utterance
Context Video Frames
Target Utterance Frames
Audiovisual
...
Time
Text
Joey: Did you call the cops?
Rachel: No, we took her to lunch.
Chandler: Ah! Your own brand of vigilante justice.



reference: Towards Multimodal Sarcasm Detection

Potential future work

- Predicting Action from Speech (Video game)
- Prediction of response prosody
- Dissatisfication detection in phone conversation
- Prosody-aware SSL model
- ...

Timeline Overview

schedule

Feb - Mar	Apr - May	Jun	July	Aug
Paper reading & Data processing	Build downstream model for baseline tasks & run experiment	Finish SUPERB prosody track & start future work	**Workshop** dive into future work	**Workshop** Paper writing

Aims

1. Augment SUPERB with prosody-intensive tasks

   (pre-workshop: Guan-Ting Lin, Chi-Luen Feng, Nigel Ward)

2. Characterize adequacy of existing pretrained models for these tasks

   (at the workshop: quantitative analysis + failure analysis? by who?)

3. Side activities on prosody and dialog

Possible Side Activities

1. A concise pretrained model for pragmatics-related prosody
2. A dialog-aware pretrained speech model
3. Predicting actions from speech

1. A concise pretrained model for pragmatics-related prosody

- A very low-dimensional representation of prosody
- Using a hand-crafted model structure

- Baseline already created (Ward & Avila, submitted)
- Need datasets for eval, experiments with more models
- Timeline: post-workshop(?)

2. A dialog-aware pretrained speech model

Use dialog data in pretraining

- assuming interlocutor’s orient to the important aspects of speech, this should discover them faster

- potentially supporting good pretraining on less data

- pretraining task may be masked prediction of the interlocutor’s track

- Timeline: post-workshop(?)

3. Prediction domain actions from speech

- Predict in-game actions from both participants' speech assuming

- Potentially as another SUPERB task
- Timeline: post-workshop(?)

Current Unknowns

What exactly will happen at the workshop?

Who will do it?

What publications are we planning?