日本音響学会 2024年春季研究発表会 1-2-8

Emotion-controllable Speech Synthesis using Emotion Soft Label and Word-level Prominence

☆Xuan Luo, Shinnosuke Takamichi, Yuki Saito, Hiroshi Saruwatari

Graduate School of Information Science and Technology,

The University of Tokyo, Japan.

Research Topic

- Topic: Enable both **emotion** and word-level **prominence** control
  - Emotion and intention (expressed by prominence) consist of important paralinguistic information
  - Prominence is a similar concept to emotion strength but they are different

Emotion strength

- Metrics for emotion
- Only appears in words with emotion
- How strong I am feeling
- Example: Why is it so spicy? 

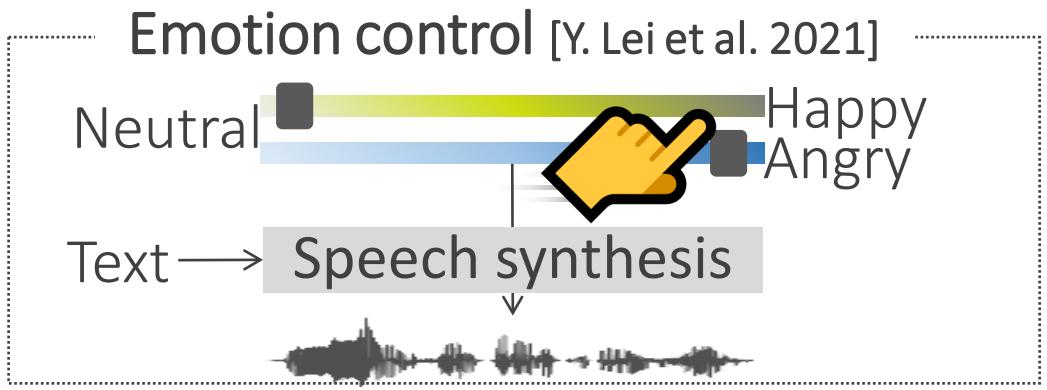
VS

Prominence

- Metrics for intention
- Possibly appears in every word
- How important these words are
- Example: We should focus more on products. 

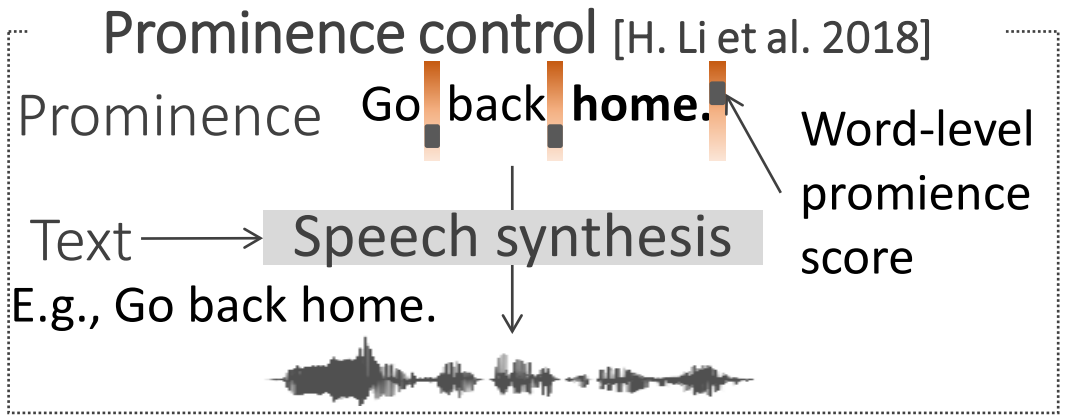
Related research

- Condition TTS model on emotion label and strength [Y. Lei et al. 2021]



Only emotion and emotion strength control

- Condition TTS model on word-level prominence [H. Li et al. 2018]



Only word-level prominence control

However, none of them can control **emotion** and **prominence at the same time**.

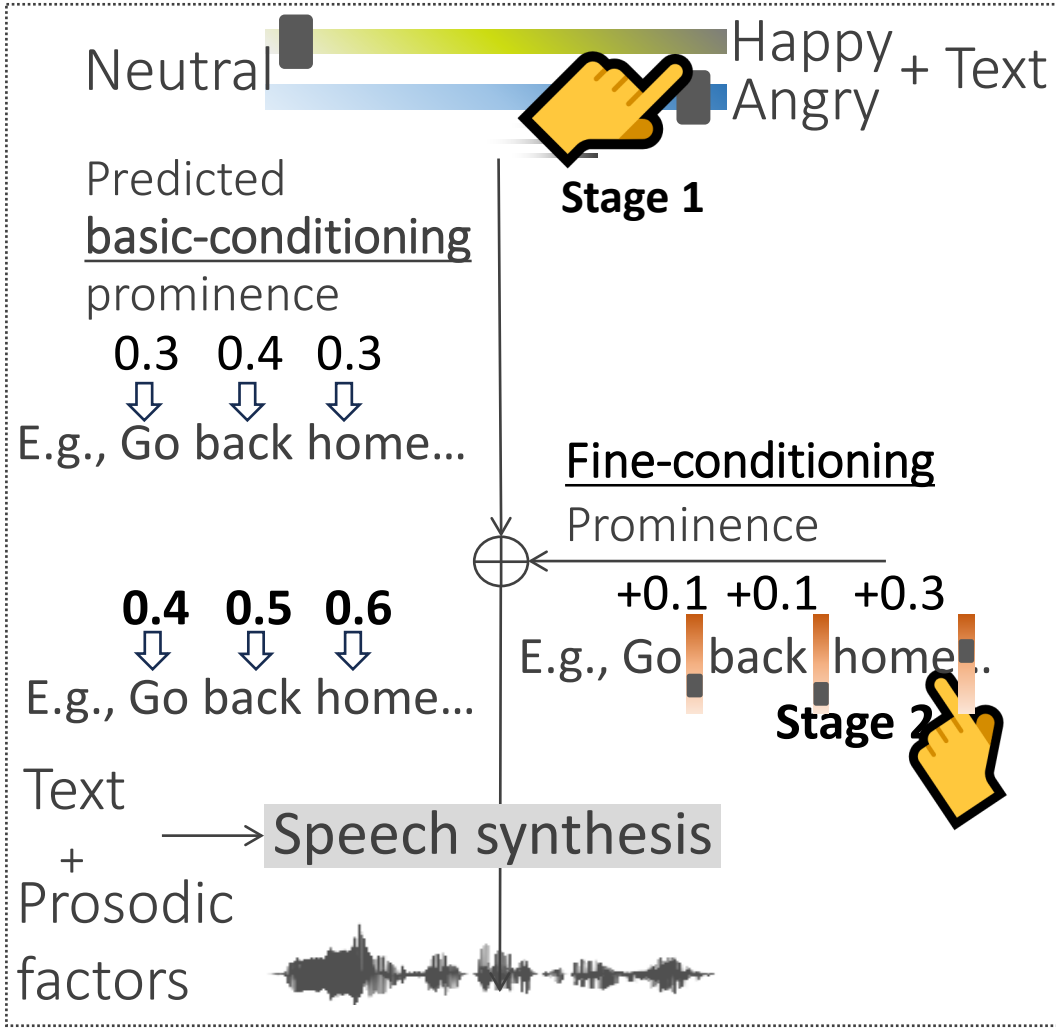
Proposed method and Result

- Proposed method
  - A two-stage emotion controllable TTS model that allows **emotion** and **word-level prominence** control using emotion soft labels and prominence.
- Result
  - **51%** emotion-distinguishable accuracy on 3 emotions
    - Fair emotion discrimination ability for synthesized speech
  - **0.95** linear controllability on prominence
    - Strongly linear controllability of word-level prominence
  - **3.9** MOS score on naturalness
    - Comparable to previous research

Concept of the proposed two-stage controlling TTS model

- Stage 1: Condition on emotion soft label
  - Emotion soft label ranges from 0 to 1
  - Predict **basic-conditioning prominence**

- Stage 2: Condition on fine-conditioning prominence
  - Fine-conditioning prominence ranges from 0 to 1
  - **Basic-conditioning prominence**(step1) + **fine-conditioning prosodic feature bias** (step 2) are summed up for conditioning TTS control



Our model : Enable both **emotion** and **word-level prominence** control

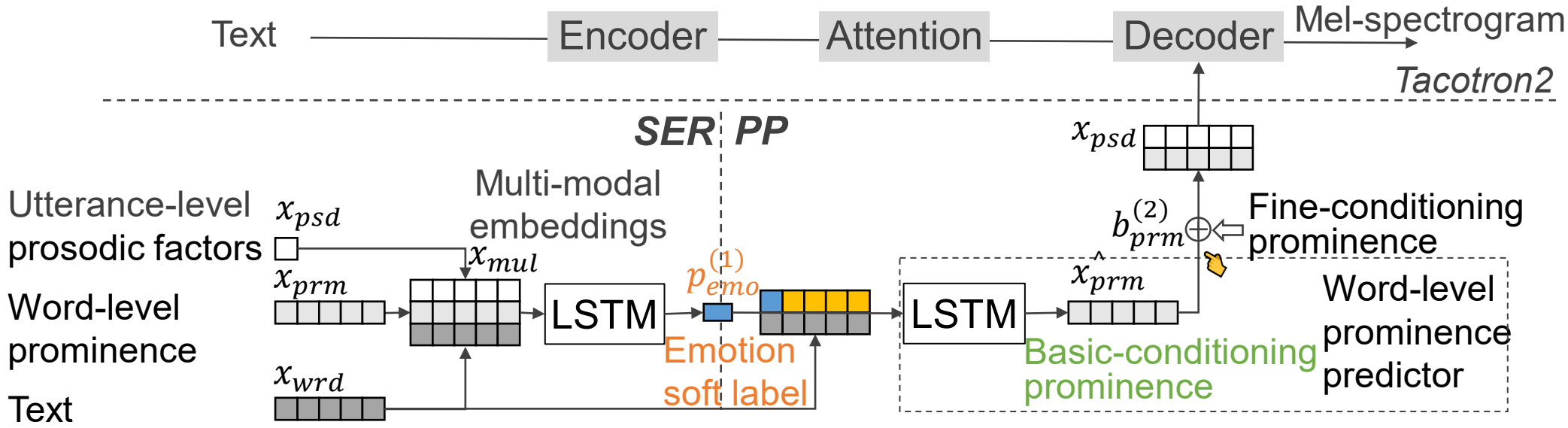
Architecture of proposed model (Training)

SER: Speech emotion recognizer

- Estimates emotion soft labels

PP: Prominence predictor

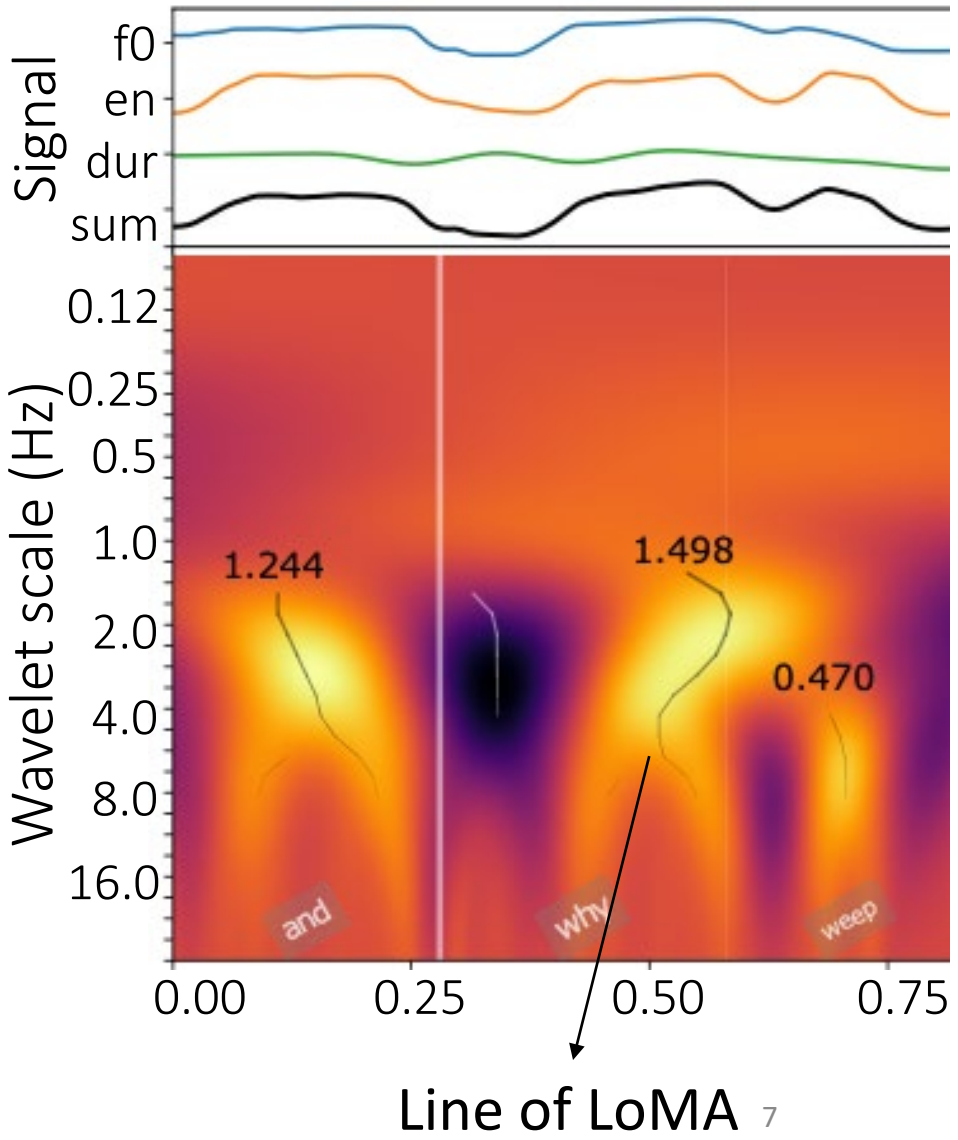
- Estimates basic-conditioning prominence



The proposed **SER** and **PP** models enabled both emotion and prominence control

Multi-modal features for SER

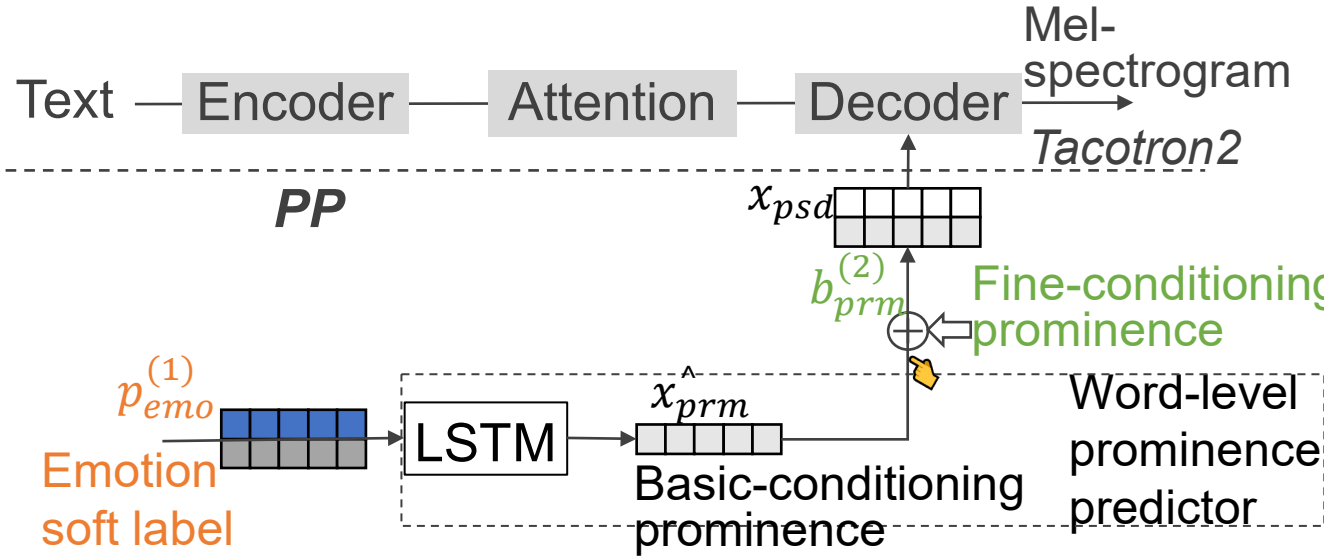
- Utterance-level prosodic factors
  - Mean, standard deviation, and range of pitch and energy [Sahu. 2019]
- Word-level prominence
  - Weighted sum of CWT (continuous wavelet transform) amplitudes which are on the lines of LoMA (maximum amplitude) at different scales [A. Suni et al. 2017]
- Text: fastText embedding [Bojanowski et al. 2017]



Controlling example (Inference)

- Emotion control
  - Don’t disappoint your public (Angry)
  - Don’t disappoint your public (Sad)

- Emotion and prominence control
  - Sad: They **forcefully** (+0.3) keep them at a black hotel.
  - Sad: They forcefully keep them at a black **hotel** (+0.3).

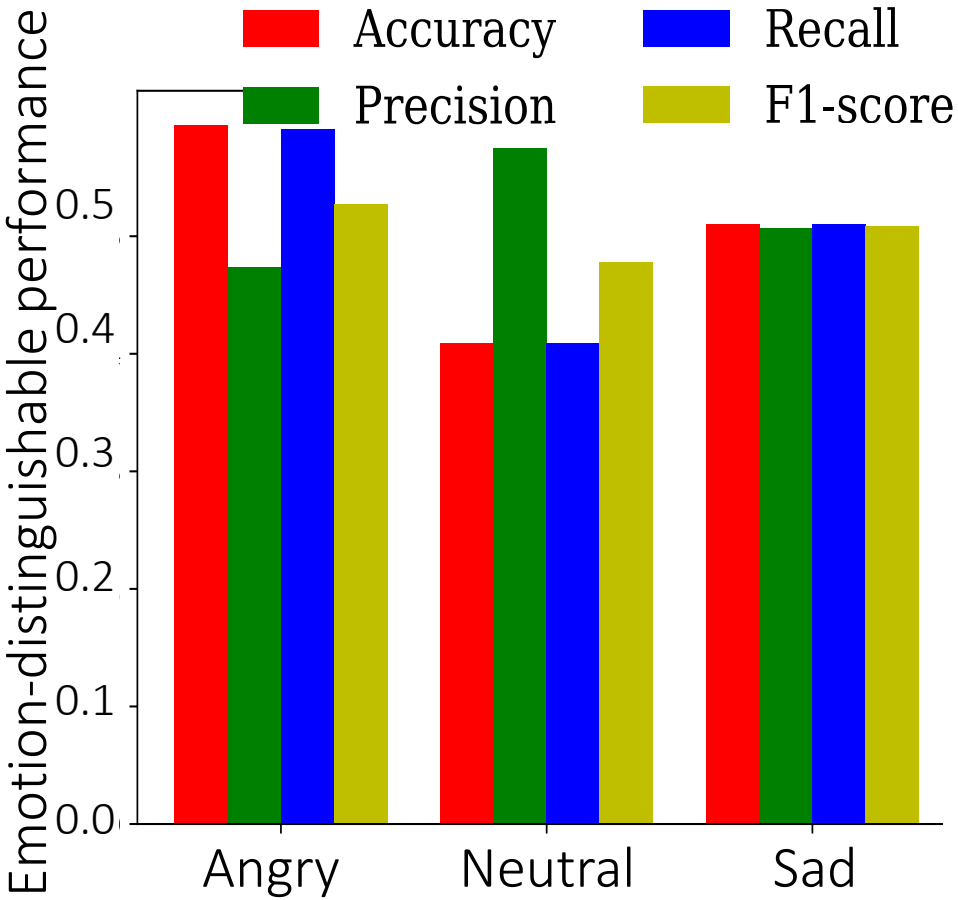


Experiment setup

- Data
  - IEMOCAP [Busso+08]: Used for pre-training the SER and PP models (12 hours)
  - Blizzard2013 [King+14]: Used for training proposed TTS model (75 hours)
- Emotion labels
  - Emotion: Angry/Sad/Neutral
  - Emotion labels in Blizzard2013 are predicted by SER pre-trained on IEMOCAP
- Backbone TTS model: Tacotron2 [Shen et al. 2018]
- Vocoder: Parallel WaveGAN [Yamamoto et al. 2020]

Evaluation1: Controllability of emotion soft labels

- Evaluation purpose
  - Whether the emotion can be controlled.
- Experiment set
  - 10 utterances for each of angry, neutral, and sad
  - 10 sets of ang/neu/sad with the same sentence
  - Each of 50 evaluators was required to listen 10 sets and select an utterance with a given emotion
- Result
  - Accuracy, precision, recall, and F1-score were **51%, 52%, 50%, and 51%** on average of 3 emotions.
  - Specifically, the accuracy of angry speech was **60%**

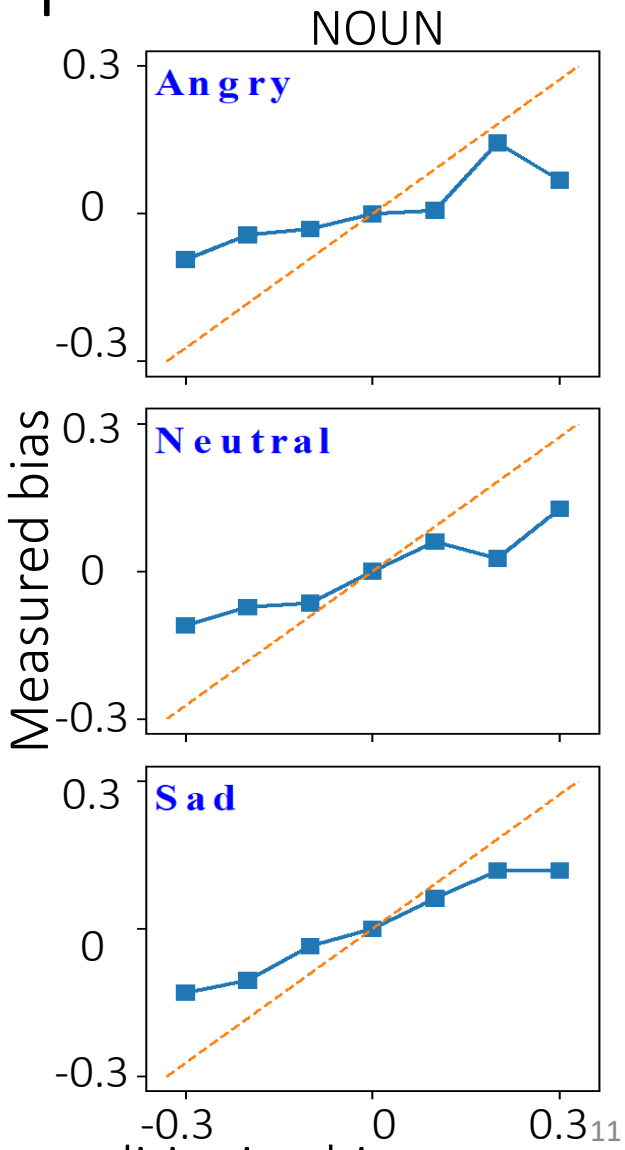


-> Fair emotion discrimination ability for synthesized speech

Evaluation2: Controllability of word-level prominence

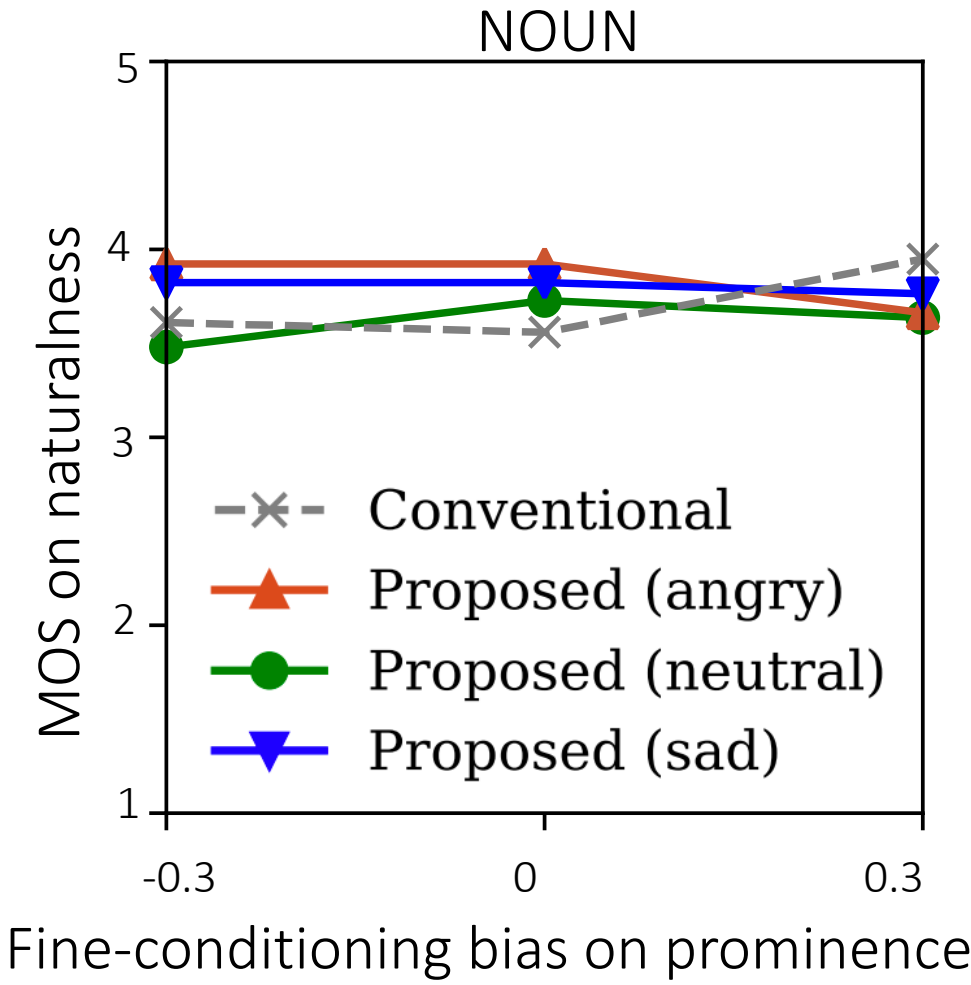
- Evaluation purpose
  - Whether the prominence can be linearly controlled
- Experiment set
  - Experiment with content words
  - Set 7 fine-conditioning prominence biases ranged from −0.3 to 0.3 by 0.1 step
  - 50 sentences and total 2,100 synthetic speech
- Result
  - Average PCC (Pearson Correlation Coefficient) scores are **0.93** (angry), **0.97** (neutral), **0.96** (sad)
  - Strong linear controllability on the prominence of the NOUN, VERB, ADJ, and ADV words.

-> Strongly linear controllability of word-level prominence



Evaluation3: Subjective test of word-level prominence controlling

- Experimental purpose
  - Whether the quality of synthetic speech is good
- Experiment set
  - 10 speech audio
  - 50 listeners to evaluate MOS on naturalness
- Result
  - Shows equal performance (MOS = 3.9) synthetic speech quality equal to the conventional method [H. Li et al. 2018]



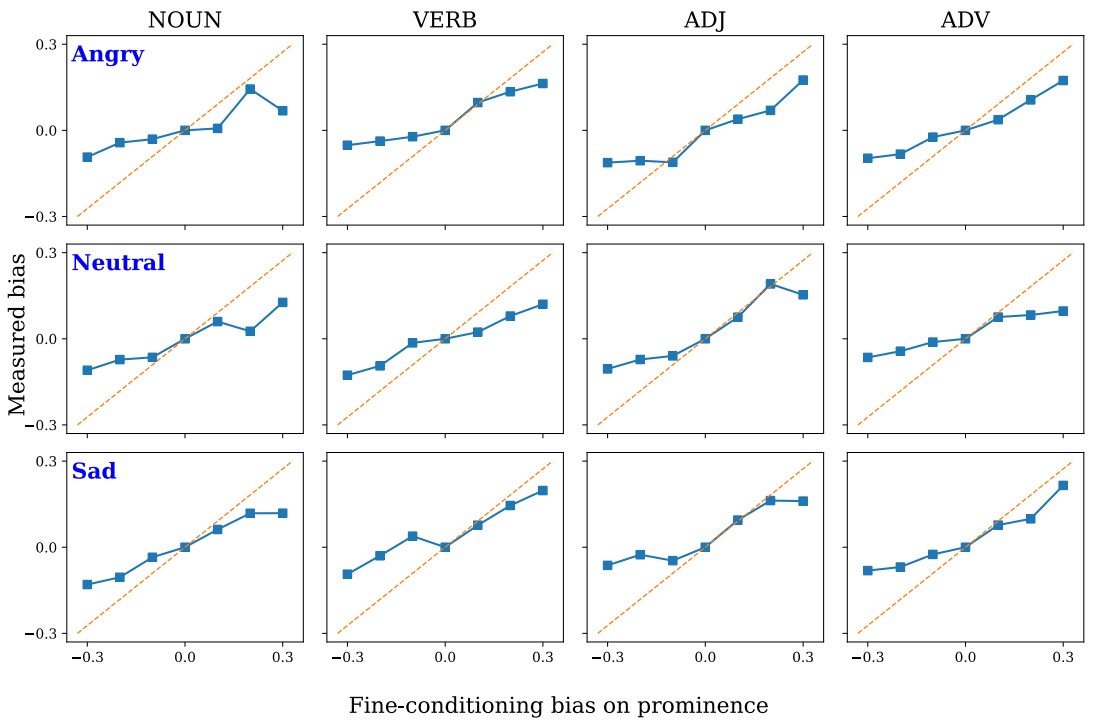
Summary

- Purpose
  - Enable emotion and word-level prominence control
- Model
  - A two-stage emotion controllable TTS model that allows **emotion** and **word-level prominence** control using emotion soft labels and prominence.
- Result
  - **51%** emotion-distinguishable accuracy
  - **0.95** linear controllability on prominence
  - **3.9** MOS score on naturalness
- Future work
  - Better emotion-distinguishable speech
  - Towards phoneme-level prominence control

Appendix

Evaluation2: Controllability of word-level prominence

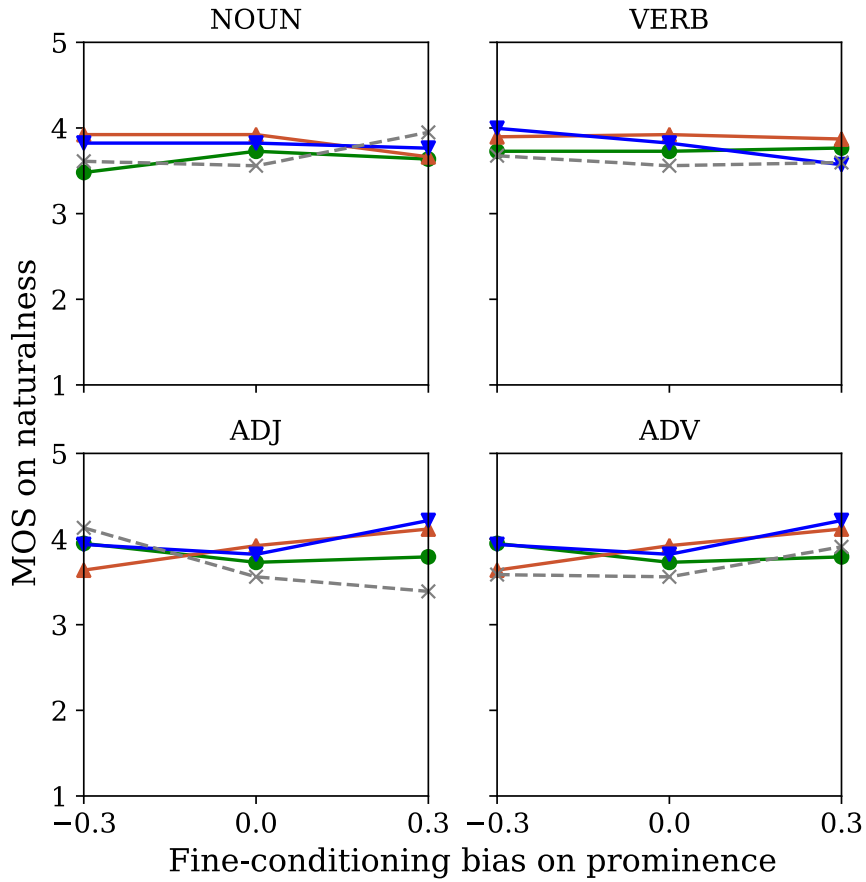
- Evaluation purpose
  - Whether the prominence can be linearly controlled
- Experiment set
  - Experiment on content words
  - Set 7 fine-conditioning prominence biases ranged from −0.3 to 0.3 by 0.1 step
  - 50 sentences and totally 2,100 synthetic speech
- Result
  - Average PCC (Pearson Correlation Coefficient) scores are **0.93** (angry), **0.97** (neutral), **0.96** (sad)
  - Strong linear controllability on the prominence of the NOUN, VERB, ADJ and ADV words.



-> Strongly linear controllability of word-level prominence

Evaluation3: Subjective test of word-level prominence controlling

- Experimental purpose
  - Whether the quality of synthetic speech is good
- Experiment set
  - 10 speech audio
  - 50 listeners to evaluate MOS on naturalness
- Result
  - Shows equal performance (MOS = 3.9) synthetic speech quality equal to the conventional method [H. Li et al. 2018]



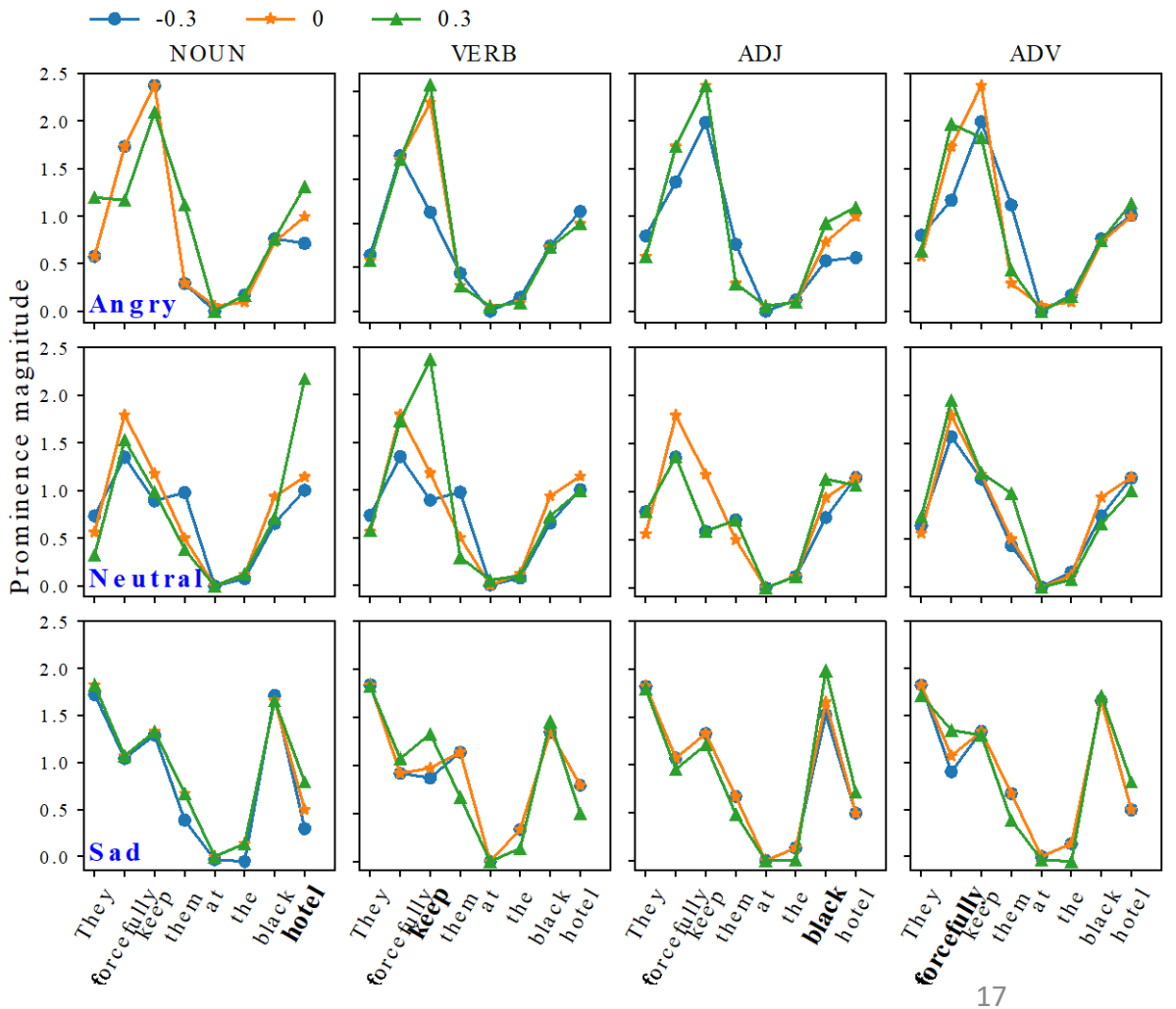
-0.3 0 0.3

Evaluation3: Prominence contours when conditioning on word-level prominence

One example!
X, Y description

- Prominence contours of fine-conditioned words **increased** when the conditioning bias **increased**

- Prominence contours of fine-conditioned words **decreased** when the conditioning bias **decreased**



Prominence

$$\boldsymbol{x}_{\text{prm}} = W_s(a_0, t_{i_0,0}) + \ldots + \log(j+1)a^{-j/2}W_s(a_0a^j, t_{i_j,j}), \quad (1)$$

where $\boldsymbol{x}_{\text{prm}}$ is word-level prominence, a_0 denotes the finest scale in CWT, a defines the spacing between chosen scales, j denotes sale, $t_{i_j,j}$ is a time point where the local maxima occurred in the a_0a^j scale. $W_s(a_0a^j, t_{i_j,j})$ denotes the CWT amplitude in $t_{i_j,j}$ time point at a_0a^j level scale.