2023「中技社科技獎學金」
2023 CTCI Foundation Science and Technology Scholarship
境外生研究獎學金
Research Scholarship for Overseas Students

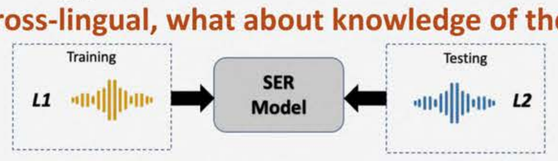
Phonetic Anchor-Based Transfer Learning to Facilitate Unsupervised Cross-Lingual Speech Emotion Recognition

Shreya Upadhyay, 5th grade PhD
Advisor: Prof. Chi-Chun Lee
Department of Electrical Engineering, National Tsing Hua University

Research Entails

Aim: To develop a novel approach for cross-lingual Speech Emotion Recognition (SER) using phonetic anchors

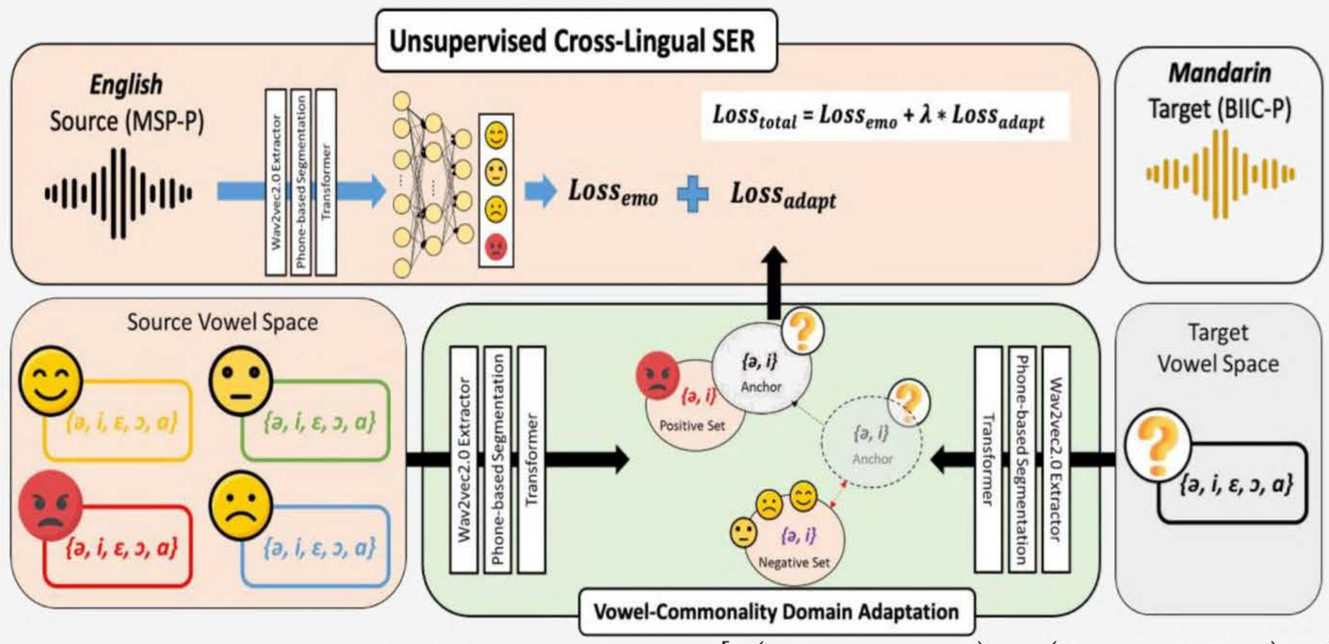
In the case of a cross-lingual, what about knowledge of the languages?



Is it really a Language Agnostics case?

- Proposes: A two-fold approach
  - **Analyze the emotion-specific commonality at the phonetic level across languages:** To find some vowels present emotion-specific commonality
  - **Devise an anchoring mechanism:** To leverage the phonetic commonalities across languages
- Uses two large-scale in-the-wild natural speech emotion corpora:
  - **MSP-Podcast (American English) :** Intonation language
  - **BIIC-Podcast (Taiwanese Mandarin) :** Tonal language

Anchor-based Cross-lingual SER



$$Loss_{emo} = \mathbb{E}_{X_S,y_S}[\|CE(T(X_S), y_S)\|] \qquad Loss_{adapt} = \sum_i^N \left[d\left(f\left(X_i^{t_{ph}}\right), f\left(X_i^{S_{p_{ph}}}\right)\right) - d\left(f\left(X_i^{t_{ph}}\right), f\left(X_i^{S_{n_{ph}}}\right)\right) + \alpha \right]$$

$$Loss_{total} = Loss_{emo} + \lambda * Loss_{adapt}$$

Emotion-Specific Commonality

- Commonalities over the set of "common ground" vowels

[i, ə, ɑ, ɛ, ɔ, u]

- Considered emotional classes

[Happiness, Anger, Sadness, and Neutral]

Phonetic Analysis :

1) Vowel space plot:

- Visible vowel commonality over corpora

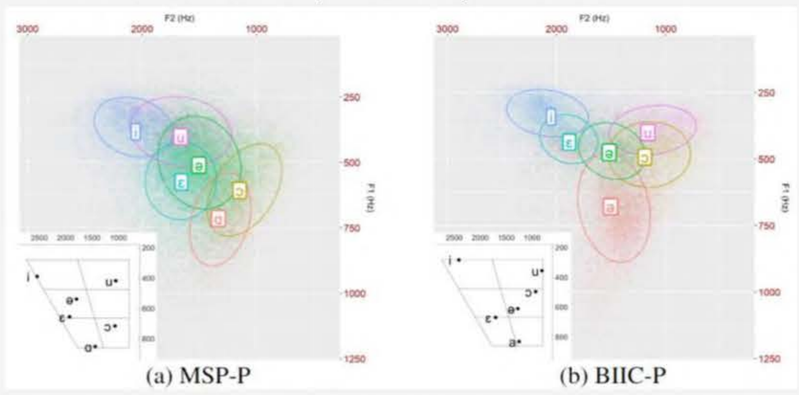


Figure 1. The vowel space using the first two formants (F1 and F2)

2) Vowel triangle plot:

- Green circled vowels are potential candidates for serving as anchors in our transfer learning strategy

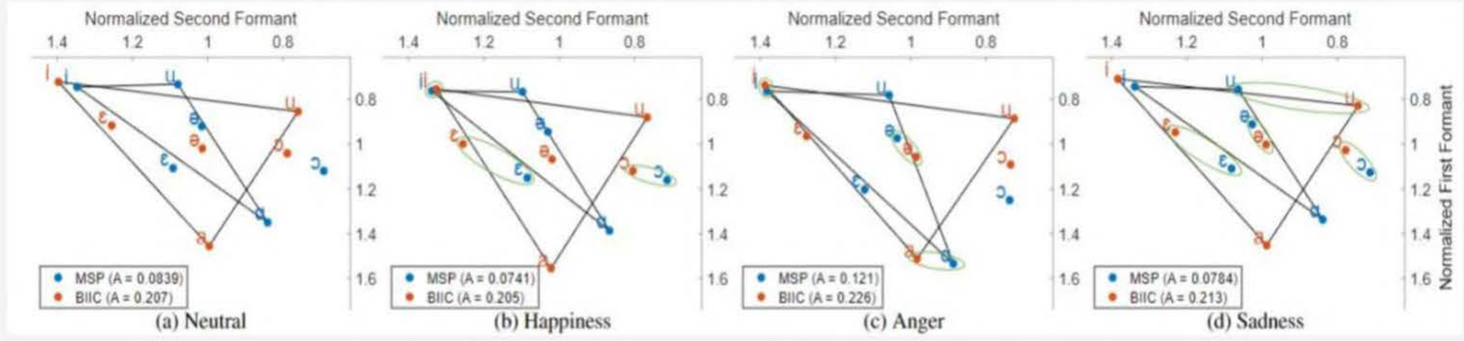
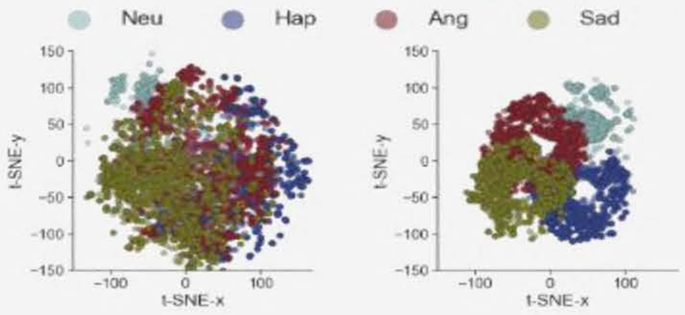


Figure 2. A plot of the average F1 and F2 values with respect to four emotional classes



- Feature analysis t-SNE plot of baseline (CL) and proposed (GA-CL)
- **Our proposed GA-CL has better emotion class feature separation**

Models	4-category	Neutral	Happiness	Anger	Sadness
CL	51.75	65.61	62.77	64.47	58.53
FM-CL	56.92	70.40	67.32	69.83	65.59
GA-CL	**58.64**	**72.83**	**69.69**	**70.15**	**68.17**
BA-CL	**55.33**	70.23	**68.74**	**67.83**	63.91
WA-CL	55.21	**70.43**	61.45	66.26	**64.62**

- Compare UAR results for Group-Vowel Anchored (GA-CL), Feature-Matching (FM-CL), and ablations (BA-CL, WA-CL) in the table
- **Our proposed GA-CL outperforms other models**

Insights

- Emotion-specific commonality analysis indicated that some vowels are more similar between corpora after emotion modulations
- Our learning approach used these vowels as phonetic constraints to control the variability between two languages enhancing the learning for unsupervised cross-lingual SER

Future Research:

Innovative Phonetic Anchoring

- Merge with state-of-the-art (SOTA) domain adaptation approaches for enhanced generalization
- Incorporate common ground consonants, including fricatives, affricates, and approximants

Acknowledgements

NSTC 國家科學及技術委員會 National Science and Technology Council
NSF