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

Shreya G. Upadhyay, Luz Martinez-Lucas, Bo-Hao Su, Wei-Cheng Lin, Woan-Shiuan Chien, Ya-Tse Wu, William Katz, Carlos Busso, Chi-Chun Lee

THE UNIVERSITY OF TEXAS AT DALLAS
UT Dallas MSP Multimodal Signal Processing Laboratory
國立清華大學 NATIONAL TSING HUA UNIVERSITY
biic

- June 8, 2023 -

Overview

- **Aim**: To develop a new approach for cross-lingual *speech emotion recognition* (SER) by integrating phonetic constraints as an anchor
- **Propose**: A twofold approach
  - First analyzes emotion-specific phonetic commonalities (vowels) across languages
  - Leverages these common vowels as an anchoring mechanism to facilitate cross-lingual SER

Outline

- Introduction
- Literature Survey
- Emotion-specific Commonality
  - Phonetic Analysis
  - Emotion-Specific SER Analysis
- Anchor-based Cross-lingual SER
- Conclusion and Future Work

Introduction

- *Speech Emotion Recognition* (SER) systems diverse application needs generalization across different domains
- Common formulation:
  - Mitigate mismatches of Source <--> Target domains
    - Transfer learning, semi-supervised learning, few-shot learning etc.
  - Optimizing to decrease a distance metric of Source <--> Target features
    - Variations on Generative Adversarial Network (GAN)
- Models are useful but come purely from a computational angle
- In case of cross-lingual scenario, what about knowledge of the languages?

Language Agnostics?

Literature Says

- Emotion perception and the acoustic feature space depend on the language
- Discriminative emotional information can be observed at the phonetic-level
- Some of these emotional patterns at phone-level generalize to other languages
- Simple phoneme-class dependent emotion classifiers and fine-tuned deep models (e.g., Wav2Vec2) can effectively improve emotion recognition rates

Research Entails

- Two investigations:
  - 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

- Two large-scale in-the-wild natural speech emotion corpora considered:
  - **MSP-Podcast** (*American English*) : Intonation language
  - **BIIC-Podcast** (*Taiwanese Mandarin*) : Tonal language

Emotion-Specific Commonality

- Commonalities over the set of ``common ground'' vowels

  [i, ə, ɑ, ɛ, ɔ, u]

- Considered emotional classes

  [Happiness, Anger, Sadness, and Neutrality]

- Two Analyses:
  - 1) Phonetic Analysis
  - 2) Emotion-Specific SER Analysis

Phonetic Analysis : Vowel space plot

- Common vowels span and their positions are consistent with the expected from the literature
- Visible vowel commonality over corpora
  - Vowels /i/ and /ə/ cover similarity regions in their respective languages

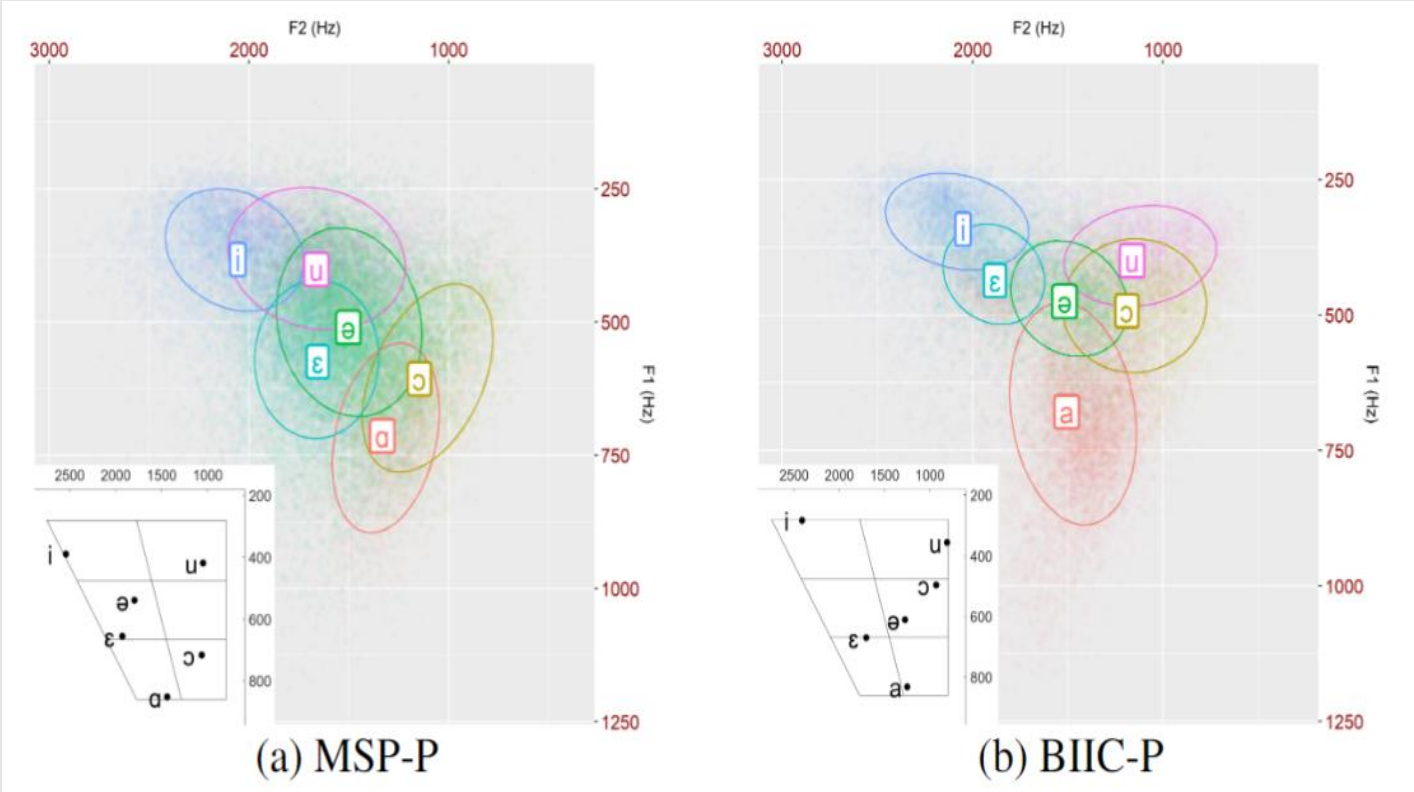

Figure 1. The vowel space using the first two formants (F1 and F2).

Phonetic Analysis: Vowel triangle plot

- First, data are normalized using the *Nearey* normalization to remove speaker differences due to individual vocal tract disparities and gender
- Example, for *Neutral* speech, closest distances across languages for vowels are /i/ and /ə/
- These vowels are potential candidates for serving as anchors in our transfer learning strategy

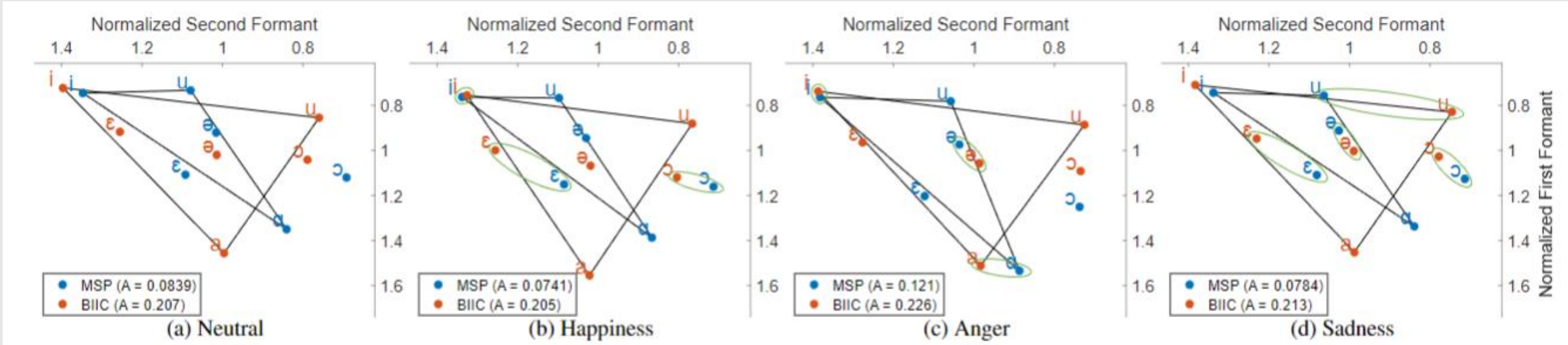

Figure 2. A plot of the average F1 and F2 values with respect to four emotional classes

Emotion-Specific SER Analysis

- Within-Corpus Vowel Discriminability Analysis: Matched conditions

Neutral, SER models for /*i*/ and /*ə*/ lead to better UAR for both corpora

Happiness, the SER models for /*i*/, /*ə*/, and /*ɑ*/

Anger, the SER models for /*ɑ*/, and /*ɛ*/

Sadness, the SER models for /*ɛ*/, /*ɔ*/, and /*u*/

- Cross-Lingual Vowel Discriminability Analysis: Mismatched conditions

Models with MSP-P corpus do not work well in recognizing emotions for the BIIC-P samples

Sadness, the SER model for /*ɔ*/ shows low performance, even in the matched condition /*ɔ*/ have relatively good performance for both languages

		Neutral		Happiness		Anger		Sadness	
		UAR	*Exp*	*UAR*	*Exp*	*UAR*	*Exp*	*UAR*	*Exp*
/*i*/	M → M	**77.78**	G	**76.47**	G B	73.36		65.30	
	B → B	**77.62**		**75.04**		72.53		67.87	
	M → B	60.80		60.28		60.19		59.96	
/*ɛ*/	M → M	69.45		73.90		**75.78**	G	**67.34**	G
	B → B	75.66		68.24		**75.22**		**70.19**	
	M → B	58.34		60.10		55.53		51.76	
/*ə*/	M → M	**76.34**	G B	**75.78**	G	73.65		64.35	W
	B → B	**77.15**		**75.50**		72.52		65.19	
	M → B	61.55		63.23		63.89		50.40	
/*ɑ*/	M → M	69.36	W	**75.61**	G	**76.56**	G B	67.45	
	B → B	75.31		**74.31**		**75.14**		68.34	
	M → B	61.93		61.41		61.45		53.02	
/*ɔ*/	M → M	74.38		72.53	W	70.89		**68.76**	G B
	B → B	76.19		70.99		74.62		**70.82**	
	M → B	58.93		57.82		59.20		58.48	
/*u*/	M → M	76.45		77.01		70.35	W	**66.89**	G
	B → B	73.36		71.23		72.29		**69.28**	
	M → B	51.04		52.69		53.02		52.24	

Anchor-based Cross-lingual SER: Architecture

$$L_{ec} = \mathbb{E}_{X_S, y_S}[\|CE(T(X_S), y_S)\|]$$

$$L_{ad} = \sum_i^N \left[d\left(f\left(X_i^{t_{ph}}\right), f\left(X_i^{S_{p_{ph}}}\right)\right) - d\left(f\left(X_i^{t_{ph}}\right), f\left(X_i^{S_{n_{ph}}}\right)\right) + \alpha \right]$$

$$L_{ec} = L_{ec} + L_{ad}$$

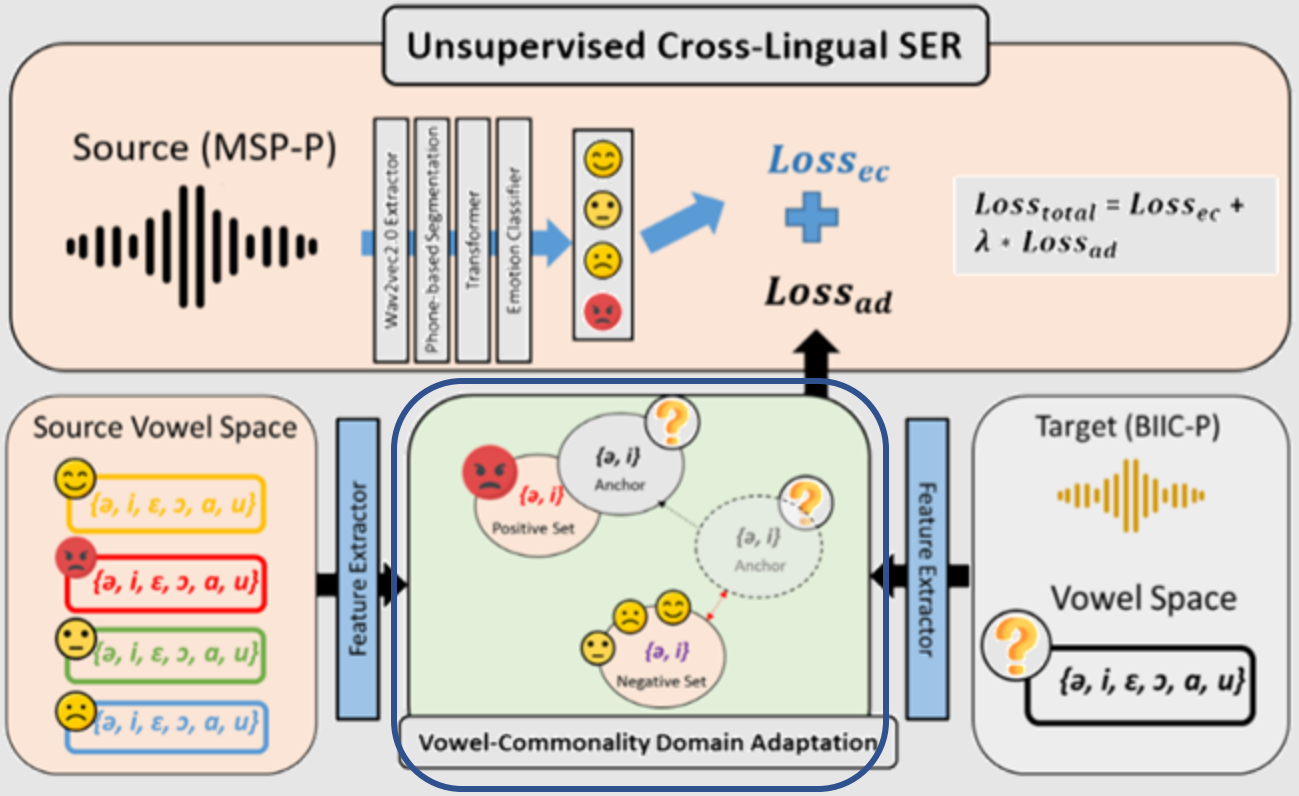

Figure 3. Proposed cross-lingual SER architecture

Anchor-based Cross-lingual SER: Performance Table

- Group-vowel-anchored (GA-CL) for unsupervised cross-lingual SER outperforms (in absolute UAR gains)
  - GA-CL ↔ CL : 6.89%
  - GA-CL ↔ FM-CL : 2.72 %
- Single vowel as *Anchor*
  - Best-vowel-anchored (BA-CL)
  - Worst-vowel-anchored (WA-CL)
  - BA-CL ↔ WA-CL : significant gain in *Happiness* and *Anger*

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**

Table 2. Cross-lingual SER performance (in UAR) with proposed group-vowel-anchored (GA-CL), feature-matching (FM-CL), and some ablation results with best-vowel-anchored (BA-CL) and worst-vowel-anchored (WA-CL)

Conclusion

- Proposed a phonetic anchoring mechanism for unsupervised cross-lingual SER
  - Based on initial evidence of emotion-specific commonality of vowels
- Emotion-specific commonality analysis indicated that some vowels are more similar between corpora after emotion modulations
- The contrastive learning approach used these vowels as phonetic constraints to control the variability between two languages
  - Enhancing the learning for unsupervised cross-lingual SER
- The proposed model GA-CL (58.64%) of UAR outperforms the FM-CL (56.92%) and CL (51.75%) baselines models

Future Work

- Merge this novel phonetic knowledge-driven anchoring mechanism with recent SOTA approaches on domain adaptation for better generalization
- Include common ground consonants (particularly fricatives, affricates, and approximants) to improve cross-lingual SER performances

Thank You !

UT Dallas MSP
Multimodal Signal Processing Laboratory

biic
National Tsing Hua University

NSF

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